STATISTICAL GUIDELINES: NEW DEVELOPMENTS IN STATISTICAL METHODS AND PSYCHOMETRIC TOOLS

EDITED BY: Pietro Cipresso and Jason C. Immekus
PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88974-761-0
DOI 10.3389/978-2-88974-761-0

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

STATISTICAL GUIDELINES: NEW DEVELOPMENTS IN STATISTICAL METHODS AND PSYCHOMETRIC TOOLS

Topic Editors:
Pietro Cipresso, University of Turin, Italy
Jason C. Immekus, University of Louisville, United States

Citation: Cipresso, P., Immekus, J. C., eds. (2022). Statistical Guidelines: New Developments in Statistical Methods and Psychometric Tools. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-761-0

Table of Contents

ORIGINAL RESEARCH
published: 25 October 2019
doi: 10.3389/fpsyg.2019.02387

Exploring the Correlation Between Multiple Latent Variables and Covariates in Hierarchical Data Based on the Multilevel Multidimensional IRT Model

Jiwei Zhang[1]*, *Jing Lu*[2], *Feng Chen*[3] and *Jian Tao*[2]

[1] School of Mathematics and Statistics, Yunnan University, Kunming, China, [2] School of Mathematics and Statistics, Northeast Normal University, Changchun, China, [3] Department of East Asian Studies, The University of Arizona, Tucson, AZ, United States

In many large-scale tests, it is very common that students are nested within classes or schools and that the test designers try to measure their multidimensional latent traits (e.g., logical reasoning ability and computational ability in the mathematics test). It is particularly important to explore the influences of covariates on multiple abilities for development and improvement of educational quality monitoring mechanism. In this study, motivated by a real dataset of a large-scale English achievement test, we will address how to construct an appropriate multilevel structural models to fit the data in many of multilevel models, and what are the effects of gender and socioeconomic-status differences on English multidimensional abilities at the individual level, and how does the teachers' satisfaction and school climate affect students' English abilities at the school level. A full Gibbs sampling algorithm within the Markov chain Monte Carlo (MCMC) framework is used for model estimation. Moreover, a unique form of the deviance information criterion (DIC) is used as a model comparison index. In order to verify the accuracy of the algorithm estimation, two simulations are considered in this paper. Simulation studies show that the Gibbs sampling algorithm works well in estimating all model parameters across a broad spectrum of scenarios, which can be used to guide the real data analysis. A brief discussion and suggestions for further research are shown in the concluding remarks.

Keywords: education assessment, teacher satisfactions, multidimensional item response theory, multilevel model, Bayesian estimation

OPEN ACCESS

Edited by:
Jason C. Immekus,
University of Louisville, United States

Reviewed by:
Peida Zhan,
Zhejiang Normal University, China
Yong Luo,
Educational Testing Service,
United States

****Correspondence:***
Jiwei Zhang
zhangjw713@nenu.edu.cn

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *25 April 2019*
Accepted: *07 October 2019*
Published: *25 October 2019*

Citation:
Zhang J, Lu J, Chen F and Tao J (2019) Exploring the Correlation Between Multiple Latent Variables and Covariates in Hierarchical Data Based on the Multilevel Multidimensional IRT Model. Front. Psychol. 10:2387. doi: 10.3389/fpsyg.2019.02387

1. INTRODUCTION

With the increasing interest in multidimensional latent traits and the advancement in estimation techniques, multidimensional item response theory (IRT) has been developed vigorously which made the model estimation become easy to implement and effective. Single-level multidimensional IRT (MIRT) models were proposed decades ago, as it have the primary features of modeling the correlations among multiple latent traits and categorical response variables (Mulaik, 1972; Reckase, 1972, 2009; Sympson, 1978; Whitely, 1980a,b; Way et al., 1988; Ackerman, 1989; Muraki and Carlson, 1993; Kelderman and Rijkes, 1994; Embretson and Reise, 2000; Béguin and Glas, 2001; Yao and Schwarz, 2006). The MIRT models later incorporated covariates to elucidate the connection between multiple latent traits and predictors (Adams et al., 1997; van der Linden, 2008; De Jong and Steenkamp, 2010; Klein Entink, 2009; Klein Entink et al., 2009; Höhler et al., 2010; Lu, 2012; Muthén and Asparouhov, 2013).

It has become frequent practice to regard IRT model calibration's latent ability as a dependent variable in resulting regression analysis in relation to educational and psychological measurement. Measurement error within latent ability estimates is ignored in this two-stage treatment resulting in statistical inferences that may be biased. Specially, measurement error can reduce the statistical power of impact studies and deteriorate the researchers' ability to ascertain relationships among different variables affecting student outcomes (Lu et al., 2005). One error that can reduce the statistical capabilities of impact studies and make it difficult for researchers to identify relationships between variables related to student outcomes is the measurement error.

Taking a multilevel perspective on item response modeling can avoid issues that arise when analysts use latent regression (using latent variables as outcomes in regression analysis) (Adams et al., 1997). The student population distribution is commonly handled as a between-student model with the IRT model being placed at the lowest level as a within-subject model within the structure of multilevel or hierarchical models. Using a multilevel IRT model gives analysts the ability to estimate item and ability parameters along with structural multilevel model parameters at the same time (e.g., Adams et al., 1997; Kamata, 2001; Hox, 2002; Goldstein, 2003; Pastor, 2003). This results in measurement error associated with estimated abilities being accounted for when estimating the multilevel parameters (Adams et al., 1997).

Although the multilevel IRT models have been deeply studied in the last 20 years, there are significant differences between our multilevel IRT models and the existing literatures in the problem to be solved and the viewpoint of modeling. Next, we discuss the differences from many aspects. Multidimensional IRT models that have a hierarchical structure relationship between specific ability and general ability were developed in 2007 by Sheng and Wikle. Specifically, general ability has a linear relationship with specific ability, or all specific abilities linearly combine within a general ability. However, the hierarchical structure in our study refers to the nested data structure, for example, the students are nested in classes while classes are nested in schools, rather than the hierarchical relationships between specific ability and general ability. The modeling method similar to Sheng and Wikle (2007) also includes Huang and Wang (2014) and Huang et al. (2013). Note that in Huang and Wang (2014), not only the hierarchical abilities models are discussed, but also the multilevel data are modeled. Muthén and Asparouhov (2013) proposed the multilevel multidimensional IRT models to investigate elementary student aggressive-disruptive behavior in school classrooms and the model parameters were estimated in Mplus (Muthén and Muthén, 1998) using Bayes. Although Muthén and Asparouhov (2013) and our current study also focus on the multilevel multidimensional IRT modeling, there are great differences in the model construction. In the multilevel modeling, they suggested that the ability (factor) of each dimension has between-and within-cluster variations. However, the sources of the between—and within—cluster variations are not taken into account. More specifically, whether these two types of variation are affected by the between cluster covariates and within individual background variables have not been further analyzed. Similarly, in the works of both Höhler et al. (2010) and Lu (2012) demonstrated the same modeling method. In our study, the between—and within—cluster variations are further explained by considering the effects of individual and school covariates on multiple dimensional latent abilities. For example, we can consider whether the gender difference between male and female has an important influence on the vocabulary cognitive ability and reading comprehension ability. Moreover, Chalmers (2015) proposed an extended mixed-effects IRT models to analyze PISA data. By using a Metropolis-Hastings Robbins-Monro (MH-RM) stochastic imputation algorithm (cf. Cai, 2010a,b,c, 2013), it evaluates fixed and random coefficients. Rather than directly explaining the multiple dimensional abilities, the individual background (level-1) and school (level-2) covariates are used to model the fixed effects.

In order to illustrate the interactions between unidimensional ability and individual—and school—level covariates where the ability parameters possess a hierarchical nesting structure, Fox and Glas (2001) and Kamata (2001) proposed multilevel IRT models. In this current research, we broaden Fox and Glas (2001) and Kamata (2001)'s models by swapping their unidimensional IRT model with a multidimensional normal ogive model because we want to assess students' four types of abilities from a large-scale English achievement test. We particularly pay attention to investigating the connection between multiple latent traits and covariates. Taking the proposed multilevel multidimensional IRT models as the basis, the following issues will be addressed. (1) According to the model selection results, which model is the best to fit the data and how can judge the individual-level regression coefficients be judged as fixed effect or random effect? (2) How will students from different ends of the socioeconomic-status (SES) score in English performance as tested in four types of latent abilities, based on the level-2 gender (GD), level-3 teacher satisfaction (ST) and school climate (CT) [The details of the Likert questionnaires for measuring teacher satisfaction and school climate, please refer to (Shalabi, 2002)]. (3) What relationship exists between males and females' performances in different latent abilities by controlling for SES, ST and CT. (4) What effects, if any, are seen with different teachers' or schools' effects (covariates)? (5) Is it possible to use a measurement tool to determine whether items' factor patterns correlate to the subscales of the test battery? In particular, will the four subtests of the test battery be discernable according to the discrimination parameters on the four dimensions?

The rest of the article is organized as follows. Section 2 presents the detailed development of the proposed multilevel multidimensional IRT models and procedure for hierarchical data. Section 3 provides a Bayesian estimation method to meet computational challenges for the proposed models. Meanwhile, Bayesian model assessment criteria is discussed in section 3. In section 4, simulation studies are conducted to examine the performances of parameter recovery using the Gibbs sampling algorithm. In addition, a real data analysis of the education quality assessment is given in section 5. We conclude this article with a brief discussions and suggestions for further research in section 6.

2. MULTILEVEL MULTIDIMENSIONAL IRT MODEL

The model contains three levels. At the first level, a multidimensional normal ogive IRT model is defined to model the relationship between items, persons, and responses. At the second level, personal parameters are predicted by personal-level covariates, such as an individual's social economic status (SES). At the third level, persons are nested within schools, and school-level covariates are included such as school climate and teacher satisfaction.

- The measurement model at level 1 (multidimensional two parameter normal ogive model; Samejima, 1974; McDonald, 1999; Bock and Schilling, 2003)

$$p_{ijk} = P\left(Y_{ijk} = 1 \left| \boldsymbol{\theta}_{ij},\ \boldsymbol{\xi}_k \right.\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\eta_{ijk}} e^{-\frac{t^2}{2}} dt. \tag{2.1}$$

In terms of notation, let $j = 1, \ldots, J$ indicate J schools (or groups), and within school j, there are $i = 1, \ldots, n_j$ individuals. The total number of individuals is $n = n_1 + n_2 + \ldots + n_J$. $k = 1, \ldots, K$ indicate the items. In Equation (2.1), Y_{ijk} denotes the response of the ith individual in the jth group answering the kth item. The corresponding correct response probability can be expressed as p_{ijk}, and $\boldsymbol{\theta}_{ij}$ denotes a Q-dimensional vectors of ability parameters for the ith individual in the jth group, i.e., $\boldsymbol{\theta}_{ij} = \left(\theta_{ij1}, \theta_{ij2}, \ldots, \theta_{ijQ}\right)'$, and $\boldsymbol{\xi}_k = \left(a_{k1}, a_{k2}, \ldots, a_{kQ}, b_k\right)'$ denotes the vector of item parameters, in which $\boldsymbol{a}_k = \left(a_{k1}, a_{k2}, \ldots, a_{kQ}\right)'$ is a vector of discrimination or slope parameters, and b_k is the difficulty or intercept parameter. Let $\eta_{ijk} = \sum_{q=1}^{Q} a_{kq}\theta_{ijq} - b_k$. The latent abilities of different dimensions can be explained by individual-level background covariates. Note that the multidimensional IRT model used in this paper actually belongs to the within-items multidimensional IRT model. That is, each item measures multiple dimensional abilities, and each test item has loadings on all these abilities. Unlike the between-items multidimensional IRT model, each item has a unity loading on one dimensional ability and zero loadings on other dimensional abilities. For a further explanation of the model used in this paper, please see **Table 1** in the following simulation study 1.

- Multilevel structural model at level 2 (individual level) can be represented by

$$\theta_{ijq} = \beta_{0jq} + x_{1ij}\beta_{1jq} + x_{2ij}\beta_{2jq} + \ldots + x_{hij}\beta_{hjq} + e_{ijq}, \tag{2.2}$$

In Equation (2.2), the level-2 individual covariates are denoted as $\boldsymbol{X}_{ij} = \left(x_{1ij}, x_{2ij}, \ldots, x_{hij}\right)$, where h is the number of individual background covariates. $\boldsymbol{X}_{ij}$ can contain both continuous and discrete variables (e.g., socio-economic status, gender). The residual term, $\boldsymbol{e}_{ij} = \left(e_{ij1}, e_{ij2}, \ldots, e_{ijQ}\right)'$ is assumed to follow a multivariate normal distribution $N\left(\mathbf{0},\ \boldsymbol{\Sigma}_e\right)$. Here, $\boldsymbol{\Sigma}_e$ is a Q-by-Q variance-covariance matrix. The individuals' abilities are considered to be the latent outcome variables of the multilevel regression model. Differences in abilities among individuals within the same school are modeled given student-level characteristics. Therefore, the explanatory information $\boldsymbol{X}_{ij}$ at the individual level explains variability in the latent abilities within school.

- Level 3 (school level) model in this current study can be expressed as follows:

$$\beta_{hjq} = \gamma_{h0q} + w_{1j}\gamma_{h1q} + w_{2j}\gamma_{h2q} + \ldots + w_{sj}\gamma_{hsq} + u_{hjq}, \tag{2.3}$$

In Equation (2.3), the level-3 school covariates are represented by $\boldsymbol{w}_j = \left(w_{j1}, w_{j2}, \ldots, w_{js}\right)'$, where s is the number of school covariates at level 3. Each level-2 random regression coefficient parameter is β_{hjq}, which can be interpreted by school level covariates. The level-3 residual $\left(u_{0jq}, u_{1jq}, \ldots u_{hjq}\right)'$ is multivariate normally distributed with mean $\mathbf{0}$ and $\left(h+1\right)$-by-$\left(h+1\right)$ covariance matrix $\boldsymbol{T}_q$, $q = 1, \ldots, Q$. The variation across schools is modeled given background information at the school level. To control the model complexity, we assume that the level-3 residual covariance between different dimensions is 0; that is

$$\begin{aligned} Cov\left(u_{hjq_1}, u_{hjq_2}\right) = 0,\ q_1, q_2 \in 1, 2, \ldots, Q, \text{ and } q_1 \neq q_2, \\ j = 1, 2, \ldots, J,\ h = 1, 2, \ldots \end{aligned} \tag{2.4}$$

Different from Equation (2.2) in this paper, Huang and Wang (2014) proposed a high-order structure model to construct ability parameters with hierarchical strucutre. More specifically, all specific abilities linearly combine within a general ability. Assuming that there are two order of ability, including $\theta_{iqv}^{(1)}$ and $\theta_{iv}^{(2)}$, their relationship is described by the following model

$$\theta_{iqv}^{(1)} = \beta_{0qv} + \beta_{1qv}\theta_{iv}^{(2)} + \varepsilon_{iqv}^{(1)}, \tag{2.5}$$

where $\theta_{iqv}^{(1)}$ and $\theta_{iv}^{(2)}$ denote first-order ability and second-order ability for the ith student sampled from school v, the subscript q denotes the dimension of the first-order ability. β_{0qv}, β_{1qv}, and $\varepsilon_{iqv}^{(1)}$ are the intercept, slope, and residual for the qth first-order ability in the vth school, respectively. $\varepsilon_{iqv}^{(1)}$ is the within-school residual and is typically assumed to be homogeneous across schools and normally distributed with a mean of zero and a variance of σ_ε^2 and independent of the other $\boldsymbol{\varepsilon}$ and $\boldsymbol{\theta}$. However, in this current study, we only focus on the specific abilities of four dimensions without the general ability, which is the different between Huang and Wang (2014) and us in the construction of the ability structure model.

Moreover, in Huang and Wang (2014)'s paper, the multilevel data structure is investigated by introducing the individual level predictions directly into the above-mentioned higher-order ability model (Equation 2.5). The specific model is as follows:

$$\theta_{iqv}^{(1)} = \beta_{0qv} + \beta_{1qv}\theta_{iv}^{(2)} + \sum_{h=2}^{H}\beta_{hqv}G_{hiv} + \varepsilon_{iqv}^{(1)}, \tag{2.6}$$

where G_{hiv} is the hth individual level predictor for the ith student in the vth school and β_{hqv} is its corresponding regression weight for the qth ability and school v. At the school level, the random coefficients $\boldsymbol{\beta}$ can be modeled as

$$\begin{aligned}\beta_{0qv} &= \gamma_{00q} + u_{0qv},\\ \beta_{1qv} &= \gamma_{10q} + u_{1qv},\\ \beta_{hqv} &= \gamma_{h0q} + u_{hqv},\end{aligned} \tag{2.7}$$

where $h = 2, \ldots, H$, and the residuals $\boldsymbol{u}_v{}' = (\mu_{0qv}, \mu_{1qv}, \ldots, \mu_{Hqv})$ are assumed to follow a multivariate normal distribution with a mean vector of zero and a covariance matrix of Σ_u. Further, school level predictors (e.g., school type, school size) can be added to the random intercept model. That is,

$$\beta_{0qv} = \gamma_{00q} + \sum_{k=1}^{K} \gamma_{kq} W_{kv} + u_{0qv}, \tag{2.8}$$

where W_{kv} is the kth school level predictor and γ_{kv} is its corresponding regression weight for the qth ability.

However, in this current study, the multiple dimensional abilities are directly built into the random regression models through the individual level predictors (Equation 2.2). It is not the same as Huang and Wang (2014, p. 498, Equation 4) that constructs hierarchical structure ability and multilevel data in one model. In addition, when constructing the school level models in our paper, school level predictive variables, such as teacher satisfaction, school climate, are used to model the random intercept and random slopes (Equation 2.3). Considering if different predictors are added to the school level model, multiple versions of the school level models are generated. Therefore, we can use the Bayesian model assessment to select the best-fitting model. However, Huang and Wang (2014) only model the random intercept by predictive variables at school level, without considering the impact of predictive variables on other random coefficients (page 498, Equation 8).

3. BAYESIAN PARAMETER ESTIMATION AND MODEL SELECTION

3.1. Identifying Restrictions

In this current study, the multilevel multidimensional IRT models are identified based on discrimination and difficulty parameters (Fraser, 1988; Béguin and Glas, 2001; Skrondal and Rabe-Hesketh, 2004). The most convenient method is to set Q item parameters b_k equal to 0 if $k = q$, and impose the restrictions $a_{kq} = 1$, where $k = 1, 2, \ldots Q$, and $q = 1, \ldots, Q$. If $k \neq q$, $a_{kq} = 0$. If $k > q$, b_k and a_{kq} will be free parameters to estimate. The basic idea is to identify the model by anchoring several item discrimination parameters to an arbitrary constant, typically $a_{kq} = 1$. Meanwhile, the location identification constrains is required by restricting the difficulty parameters for given items, typically, $b_k = 0$. Based on the fixed anchoring values of item parameters, other parameters are estimated on the same scale. The estimated difficulty or discrimination values of item parameters are interpreted based on their relative positions to the corresponding anchoring values (Béguin and Glas, 2001, p. 545). Additionally, in order to have a clear understanding of the process of restricting the identifiability, we illustrate the identifiability of the two-dimensional models. For details, please refer to item 1 and item 2 in **Tables 1**, **2** for the restrictions of discrimination and difficult parameters.

3.2. Gibbs Sampling Within the MCMC Framework

In the framework of frequentist, two commonly used estimation methods are used to estimate the complex IRT models. One is the marginal maximum likelihood estimation (MMLE; Bock and Aitkin, 1981), and the other is the weighted least squares means and variance adjusted (WLSMV; Muthén et al., 1997). However, the main disadvantage of the marginal maximum likelihood method is that it inevitably needs to approximate the tedious multidimensional integral by using numerical or Monte Carlo integration, which will increase large the computational burden. Another disadvantage of the MMLE are that it is difficulty to incorporate uncertainty (standard errors) into parameter estimates (Patz and Junker, 1999a), and the comparison method of the MMLE is simplistic, except the RMSEA (Root Mean Square Error of Approximation) which is often used, other comparison methods are seldom used. In addition, there are some disadvantages in WLSMV compared with Bayesian method used in this paper. Firstly, Bayesian method outperforms WLSMV solely in case of strongly informative accurate priors for categorical data. Even if the weakly informative inaccurate priors are used when the sample size is moderate and not too small, the performance of Bayesian method does not deteriorate (Holtmann et al., 2016). Secondly, compared with WLSMV, Bayesian method does not rely on asymptotic arguments and can give more reliable results for small samples (Song and Lee, 2012). Thirdly, Bayesian method allows the possibility to analyze models that are computationally heavy or impossible to estimate with WLSMV (Asparouhov and Muthén, 2012). For example, the computational burden of the WLSMV becomes intensive especially when a large number of items is considered. Fourth, Bayesian method has a better convergence rate compared with WLSMV. Fifth, Bayesian method can be used to evaluate the plausibility of the model or its general assumptions by using posterior predictive checks (PPC; Gelman et al., 1996). For the above-mentioned reasons, Bayesian method is chosen for estimating the following multilevel multidimensional IRT models.

In fact, Bayesian methods have been widely applied to estimate parameters in complex multilevel IRT models (e.g., Albert, 1992; Bradlow et al., 1999; Patz and Junker, 1999a,b; Béguin and Glas, 2001; Rupp et al., 2004). Within the framework of Bayesian, a series of BUGS softwares can be used to estimate these multilevel IRT models, including OpenBUGS (Spiegelhalter et al., 2003) and JAGS (Plummer, 2003). However, in this paper, we implement the Gibbs sampling by introducing the augmented variables rather than by constructing an envelope of the log of the target density as in a series of BUGS softwares. The auxiliary or latent variable approach has several important advantages. First, the approach is very flexible and can handle almost all sorts of discrete responses. Typically, the likelihood of the observed response data has a complex structure but the likelihood of the augmented (latent) data has a known distribution with convenient mathematical

TABLE 1 | Estimation of simulated item parameter estimation using Gibbs sampling algorithm in simulation study 1.

	a_{k1}			a_{k2}			b_k		
Item	**True**	**EAP**	**HPDI**	**True**	**EAP**	**HPDI**	**True**	**EAP**	**HPDI**
1	1^*	1^*	–	0^*	0^*	–	0^*	0^*	–
2	0^*	0^*	–	1^*	1^*	–	0^*	0^*	–
3	0.914	0.877	[0.711, 1.044]	0.686	0.672	[0.551, 0.795]	−1.182	−1.154	[−1.327, −1.005]
4	1.102	1.127	[0.915, 1.355]	1.468	1.485	[1.250, 1.717]	0.441	0.426	[0.203, 0.629]
5	2.055	2.046	[1.674, 2.466]	1.428	1.453	[1.214, 1.678]	−1.197	−1.367	[−1.683, −1.101]
6	2.291	2.361	[1.876, 2.835]	1.146	1.159	[0.877, 1.406]	−2.536	−2.524	[−3.068, −2.187]
7	2.131	2.185	[1.834, 2.576]	0.758	0.760	[0.595, 0.930]	1.782	1.759	[1.448, 2.081]
8	1.027	1.009	[0.806, 1.214]	1.720	1.736	[1.491, 2.009]	0.152	0.159	[−0.229, 0.225]
9	0.569	0.564	[0.403, 0.713]	1.119	1.152	[0.973, 1.324]	0.964	0.927	[0.735, 1.093]
10	0.578	0.550	[0.342, 0.761]	2.129	2.094	[1.776, 2.471]	1.462	1.485	[1.215, 1.745]
11	0.795	0.797	[0.615, 0.980]	1.445	1.466	[1.261, 1.691]	0.619	0.600	[0.376, 0.787]
12	2.279	2.389	[1.191, 2.867]	1.148	1.132	[0.875, 1.412]	−2.020	−2.028	[−2.388, −1.696]
13	0.714	0.616	[0.391, 0.864]	2.225	2.210	[1.867, 2.532]	0.602	0.577	[0.293, 0.826]
14	2.200	2.216	[1.797, 2.651]	1.465	1.471	[1.217, 1.721]	0.127	0.091	[−0.219, 0.381]
15	1.565	1.589	[1.349, 1.847]	0.728	0.711	[0.558, 0.867]	−0.587	−0.605	[−0.817, −0.419]
16	2.419	2.439	[2.076, 2.866]	2.408	2.380	[2.015, 2.796]	−0.218	−0.225	[−0.635, 0.094]
17	1.561	1.595	[1.342, 1.869]	1.398	1.388	[1.182, 1.621]	0.830	0.789	[0.533, 1.022]
18	2.457	2.470	[1.981, 2.900]	2.111	2.152	[1.792, 2.547]	1.558	1.560	[1.182, 1.926]
19	0.714	0.686	[0.545, 0.843]	0.918	0.883	[0.743, 1.030]	1.504	1.487	[1.320, 1.670]
20	2.447	2.482	[2.023, 2.942]	1.704	1.754	[1.490, 2.018]	0.126	0.110	[−0.221, 0.421]
21	1.588	1.562	[1.217, 1.905]	2.170	2.177	[1.825, 2.534]	−0.760	−0.789	[−1.123, −0.521]
22	1.724	1.721	[1.456, 2.037]	1.590	1.571	[1.320, 1.800]	0.769	0.671	[0.397, 0.912]
23	2.273	2.244	[1.909, 2.616]	0.948	0.917	[0.738, 1.119]	0.265	0.105	[−0.156, 0.343]
24	1.228	1.198	[0.902, 1.505]	2.782	2.755	[2.353, 3.128]	−1.398	−1.429	[−1.834, −1.115]
25	0.687	0.674	[0.456, 0.923]	2.261	2.275	[1.925, 2.651]	1.802	1.778	[1.429, 2.111]
26	1.665	1.666	[1.427, 1.928]	0.572	0.568	[0.443, 0.709]	0.033	0.021	[−0.172, 0.208]
27	2.383	2.400	[1.904, 2.823]	1.871	2.021	[1.626, 2.359]	1.307	1.285	[0.915, 1.620]
28	1.778	1.772	[1.443, 2.111]	2.326	2.305	[1.957, 2.641]	−0.871	−0.875	[−1.193, −0.581]
29	1.522	1.541	[1.175, 1.975]	2.909	2.934	[2.460, 3.505]	0.241	0.232	[−0.175, 0.588]
30	1.173	1.178	[1.940, 1.434]	1.703	1.710	[1.458, 1.977]	0.397	0.363	[0.104, 0.577]

**indicates the constraints for model identification. True denotes the true value of parameter. EAP denotes the expected a priori estimation. HPDI denotes the 95% highest posterior density intervals.*

TABLE 2 | Parameter estimates of the fixed effect, Level-2 variance-covariance and Level-3 variance-covariance in simulation 1.

Fixed effect	True	EAP	HPDI	Fixed effect	True	EAP	HPDI
γ_{001}	1.000	0.982	[0.928, 1.225]	γ_{002}	−0.350	−0.377	[−0.659, −0.115]
γ_{011}	0.300	0.326	[0.129, 0.510]	γ_{012}	0.300	0.281	[−0.046, 0.524]
γ_{101}	0.500	0.521	[0.244, 0.807]	γ_{102}	0.500	0.522	[0.296, 0.824]
γ_{111}	0.350	0.325	[0.134, 0.501]	γ_{112}	−1.000	−0.986	[−1.234, −0.736]

Level-2 random effect	True	EAP	HPDI
$\sigma_{e_1}^2$	0.300	0.323	[0.269, 0.387]
$\sigma_{e_1e_2}$	0.075	0.093	[0.053, 0.136]
$\sigma_{e_2e_1}$	0.075	0.093	[0.053, 0.136]
$\sigma_{e_2}^2$	0.500	0.529	[0.438, 0.648]

Level-3 T_1	True	EAP	HPDI	Level-3 T_2	True	EAP	HPDI
τ_{001}	0.100	0.115	[0.016, 0.380]	τ_{002}	0.100	0.073	[−0.058, 0.369]
τ_{011}	0	0.013	[−0.229, 0.140]	τ_{012}	0	0.017	[−0.143, 0.192]
τ_{101}	0	0.013	[−0.229, 0.140]	τ_{102}	0	0.017	[−0.143, 0.192]
τ_{111}	0.100	0.074	[−0.068, 0.436]	τ_{112}	0.100	0.119	[−0.093, 0.298]

properties. Second, conjugate priors, where the posterior has the same algebraic form as the prior, can be more easily defined for the likelihood of the latent response data, which has a known distributional form, than for the likelihood of the observed data. Third, the augmented variable approach facilitates easy formulation of a Gibbs sampling algorithm based on data augmentation. It will turn out that by augmenting with a latent continuous variable, conditional distributions can be defined based on augmented data, from which samples are easily drawn. Fourth, the conditional posterior given augmented data has a known distributional form such that conditional probability statements can be directly evaluated for making posterior inferences. The likelihood of the augmented response data is much more easily evaluated than the likelihood of the observed data and can be used to compare models. In summary, in this study, we adopt the Gibbs sampling algorithm (Geman and Geman, 1984) with data augmentation (Tanner and Wong, 1987) to estimate multilevel multidimensional IRT models. In particular, let $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ denote the vectors of all person and item parameters. Define an augmented variable Z_{ijk} that is normally distributed with mean $\eta_{ijk} = \sum_{q=1}^{Q} a_{kq}\theta_{ijq} - b_k$ and variance 1.

The joint posterior distribution of the parameters given the data is as follows:

$$p\left(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\beta}, \Sigma_e, \boldsymbol{\gamma}, \boldsymbol{T} \,|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{W}\right) \propto \prod_{i=1}^{n_j}\prod_{j=1}^{J}\prod_{k=1}^{K}\prod_{q=1}^{Q} p\left(Z_{ijk}\left|\theta_{ijq}, \boldsymbol{\xi}_k, Y_{ijk}\right.\right) p\left(\theta_{ijq}\left|\boldsymbol{\beta}_{jq}, \sigma_q^2, \boldsymbol{X}_j\right.\right) \times p\left(\boldsymbol{\beta}_{jq}\left|\boldsymbol{\gamma}_q, \boldsymbol{T}_q, \boldsymbol{W}_j\right.\right) p\left(\boldsymbol{\gamma}_q\left|\boldsymbol{T}_q\right.\right) p\left(\boldsymbol{\xi}_k\right) p\left(\Sigma_e\right) p\left(\boldsymbol{T}_q\right). \quad (3.1)$$

where σ_q^2 is the conditional variance given the other ability dimensions. It can be obtained from $\boldsymbol{\Sigma}_e$. The details of the Gibbs sampling are shown as follows

Step 1: Sampling $\boldsymbol{Z}$ given the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, where the random variable Z_{ijk} is independent

$$Z_{ijk}\,|\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{Y} \sim \begin{cases} N\left(\sum_{q=1}^{Q} a_{kq}\theta_{ijq} - b_k, 1\right) \text{ truncated at the left by 0 if } Y_{ijk} = 1, \\ N\left(\sum_{q=1}^{Q} a_{kq}\theta_{ijq} - b_k, 1\right) \text{ truncated at the right by 0 if } Y_{ijk} = 0. \end{cases} \quad (3.2)$$

Step 2: Sampling $\boldsymbol{\theta}_{ij}$ according to Gibbs sampling characteristics. A divide-and-conqueror strategy is used to draw each sampling element of $\boldsymbol{\theta}_{ij} = \left(\theta_{ij1}, \boldsymbol{\theta}_{ij(-1)}\right)'$, where $\boldsymbol{\theta}_{ij(-1)} = \left(\theta_{ij2}, \cdots, \theta_{ijQ}\right)$. Let $\boldsymbol{\beta}_j = \left(\boldsymbol{\beta}_{j1}, \cdots, \boldsymbol{\beta}_{jQ}\right)'$, $\boldsymbol{\mu} = \left(\boldsymbol{X}_{ij}\boldsymbol{\beta}_{j1}, \boldsymbol{\mu}_1^{(2)}\right)'$, where $\boldsymbol{\mu}_1^{(2)} = \left(\boldsymbol{X}_{ij}\boldsymbol{\beta}_{j2}, \cdots, \boldsymbol{X}_{ij}\boldsymbol{\beta}_{jQ}\right)$ and $\Sigma_e = \begin{pmatrix} \sigma_{e_1}^2 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. The conditional prior distribution of θ_{ij1} can be written as

$$p\left(\theta_{ij1}\left|\boldsymbol{\theta}_{ij(-1)}, \boldsymbol{\beta}_j, \Sigma_e\right.\right) \sim N\left(\mu_{ij}^1, \sigma_1^2\right),$$

$$\mu_{ij}^1 = \boldsymbol{X}_{ij}\boldsymbol{\beta}_{j1} + \Sigma_{12}\Sigma_{22}^{-1}\left(\boldsymbol{\theta}_{ij(-1)} - \boldsymbol{\mu}_1^{(2)}\right), \; \sigma_1^2 = \sigma_{e_1}^2 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Therefore, the full conditional posterior density of θ_{ij1} (Lindley and Smith, 1972; Box and Tiao, 1973) is given by

$$\theta_{ij1}\left|\boldsymbol{Z}_{ij}, \boldsymbol{\theta}_{ij(-1)}, \boldsymbol{\xi}, \boldsymbol{\beta}_{j1}, \sigma_1^2\right. \sim N\left(\left(\nu + \sigma_1^2\right)^{-1}\left(\widetilde{\theta}_{ij1}\sigma_1^2 + \mu_{ij}^1\nu\right), \left(\nu + \sigma_1^2\right)^{-1}\left(\nu\sigma_1^2\right)\right). \quad (3.3)$$

where

$$\widetilde{\theta}_{ij1} = \left(\sum_{k=1}^{K} a_{k1}^2\right)^{-1} \left[\sum_{k=1}^{K} a_{k1}\left(Z_{ijk} + b_k - a_{k2}\theta_{ij2} - \cdots - a_{kQ}\theta_{ijQ}\right)\right],$$

$v = \left(\sum_{k=1}^{K} a_{k1}^2\right)^{-1}$. For $q = 2, \ldots, Q$, θ_{ijq} can be drawn in the same manner.

Step 3: Sampling $\boldsymbol{\xi}_k$, $\boldsymbol{\xi}_k = \left(a_{k1}, \cdots, a_{kQ}, b_k\right)'$, Given $\boldsymbol{\theta}$, $\boldsymbol{Z}_k = \left(Z_{11k}, \cdots, Z_{n_1 1k}, \cdots, \ Z_{n_J Jk}\right)'$, Here n $\left(n = n_1 + n_2 + \cdots + n_J\right)$ represents the total number of individuals in different groups. The residual can be written as $\boldsymbol{\varepsilon}_k = \left(\varepsilon_{11k}, \cdots, \varepsilon_{n_1 1k}, \cdots, \varepsilon_{n_J Jk}\right)'$ and each element is distributed as $N\left(0, 1\right)$. Therefore, we have

$$\boldsymbol{Z}_k = \left[\boldsymbol{\theta} \ -1\right] \boldsymbol{\xi}_k + \boldsymbol{\varepsilon}_k.$$

Let $\boldsymbol{H} = \left[\boldsymbol{\theta} \ -1\right]$, the likelihood function of $\boldsymbol{\xi}_k$ is normally distributed with mean $\widetilde{\boldsymbol{\xi}}_k = \left(\boldsymbol{H}'\boldsymbol{H}\right)^{-1} \boldsymbol{H}'\boldsymbol{Z}_k$ and $\boldsymbol{H}_0 = \left(\boldsymbol{H}'\boldsymbol{H}\right)^{-1}$. Suppose that the priors of the discrimination and difficult parameters are $\boldsymbol{a}_k \sim N\left(\boldsymbol{\mu}_a, \Sigma_a\right) \mathrm{I}\left(\boldsymbol{a}_k \left| a_{kq} > 0, q = 1, \ldots, Q\right.\right)$ and $b_k \sim N\left(\mu_b, \sigma_b^2\right)$, respective, Here $\boldsymbol{\mu}_a = \left(\mu_{a1}, \ldots, \mu_{aQ}\right)'$ and $\Sigma_a = diag\left(\sigma_{a1}^2, \ldots, \sigma_{aQ}^2\right)$. The prior of item parameter $\boldsymbol{\xi}_k$ is a multivariate normal distribution with mean $\boldsymbol{\mu}_{\xi_0} = \left(\mu_{a1}, \ldots, \mu_{aQ}, \mu_b\right)'$ and $\Sigma_{\xi_0} = diag\left(\sigma_{a1}^2, \ldots, \sigma_{aQ}^2, \sigma_b^2\right)$. Therefore, the full conditional posterior distribution of the item parameters is given by

$$\boldsymbol{\xi}_k \left| \boldsymbol{\theta}, \ \boldsymbol{Z}_k, \ \boldsymbol{Y} \right. \sim N\left(\left(\boldsymbol{H}_0^{-1} + \Sigma_{\xi_0}^{-1}\right)^{-1} \left(\boldsymbol{H}'\boldsymbol{Z}_k + \Sigma_{\xi_0}^{-1}\boldsymbol{\mu}_{\xi_0}\right), \right.$$
$$\left. \left(\boldsymbol{H}_0^{-1} + \Sigma_{\xi_0}^{-1}\right)^{-1}\right) \mathrm{I}\left(\boldsymbol{a}_k \left| a_{kq} > 0, q = 1, \ldots, Q\right.\right). \tag{3.4}$$

Step 4: Sampling $\boldsymbol{\beta}_j = \left(\boldsymbol{\beta}_{j1}, \ldots, \boldsymbol{\beta}_{jQ}\right)'$, given $\boldsymbol{\theta}$, σ_q^2, $\boldsymbol{\gamma}$ and $\boldsymbol{T}$. Dawn an element of vector $\boldsymbol{\beta}_j$, $\boldsymbol{\beta}_{j1} = \left(\beta_{0j1}, \ldots, \beta_{hj1}\right)'$. Let $\boldsymbol{\theta}_{j1} = \left(\theta_{1j1}, \ldots, \theta_{n_j j1}\right)'$, and $\boldsymbol{X}_j = \left(\boldsymbol{X}_{1j}, \ldots, \boldsymbol{X}_{n_j j}\right)'$, with $\boldsymbol{X}_{ij}$ as defined in the part of model introduction. The level-2 residual $\boldsymbol{e}_{j1}$ can be defined as $\boldsymbol{e}_{j1} = \left(e_{1j1}, \ldots, e_{n_j j1}\right)'$. Therefore, we have

$$\boldsymbol{\theta}_{j1} = \boldsymbol{X}_j \boldsymbol{\beta}_{j1} + \boldsymbol{e}_{j1}.$$

The level-2 likelihood function of $\boldsymbol{\beta}_{j1}$ is normally distributed with mean $\widetilde{\boldsymbol{\beta}}_{j1} = \left(\boldsymbol{X}_j'\boldsymbol{X}_j\right)^{-1} \boldsymbol{X}_j'\boldsymbol{\theta}_{j1}$ and variance $\boldsymbol{\Sigma}_{j1} = \sigma_1^2 \left(\boldsymbol{X}_j'\boldsymbol{X}_j\right)^{-1}$. Furthermore, $\boldsymbol{w}_j$ is the direct product of $\boldsymbol{w}_{js} = \left(1, w_{j1}, \ldots, w_{js}\right)$ and a $\left(h+1\right)$ identity matrix, that is, $\boldsymbol{w}_j = \boldsymbol{I}_{(h+1)} \otimes \boldsymbol{w}_{js}$. The random regression coefficient $\boldsymbol{\beta}_{j1}$ is induced by a normal prior at level 3 with mean $\boldsymbol{w}_j\boldsymbol{\gamma}_1$ and covariance $\boldsymbol{T}_1$, where $\boldsymbol{\gamma}_1 = \left(\gamma_{001}, \gamma_{011} \ldots, \gamma_{0s1}, \ldots, \gamma_{h01}, \gamma_{h11}, \ldots, \gamma_{hs1}\right)'$. The level-3 residual $\boldsymbol{u}_{j1}$ can be defined as $\boldsymbol{u}_{j1} = \left(u_{0j1}, \ldots, u_{hj1}\right)'$. Therefore, we have

$$\boldsymbol{\beta}_{j1} = \boldsymbol{w}_j\boldsymbol{\gamma}_1 + \boldsymbol{u}_{j1}.$$

Thus, the fully conditional posterior distribution of $\boldsymbol{\beta}_{j1}$ is given by

$$\boldsymbol{\beta}_{j1} \left| \boldsymbol{\theta}_{j1}, \sigma_1^2, \ \boldsymbol{\gamma}_1, \ \boldsymbol{T}_1 \right. \sim N\left(\left(\boldsymbol{\Sigma}_{j1}^{-1} + \boldsymbol{T}_1^{-1}\right)^{-1}\right.$$
$$\left.\left(\boldsymbol{\Sigma}_{j1}^{-1}\widetilde{\boldsymbol{\beta}}_{j1} + \boldsymbol{T}_1^{-1}\boldsymbol{w}_j\boldsymbol{\gamma}_1\right), \ \left(\boldsymbol{\Sigma}_{j1}^{-1} + \boldsymbol{T}_1^{-1}\right)^{-1}\right), \tag{3.5}$$

and $\boldsymbol{\beta}_{jq}$, $q = 2, \ldots, Q$, is drawn in the same manner.

Step 5: Sampling $\boldsymbol{\gamma}$, $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_Q\right)$. An element of vector $\boldsymbol{\gamma}$ is drawn, and the matrix $\boldsymbol{\gamma}_1$ is the matrix of regression coefficients corresponding to the regression of $\boldsymbol{\beta}_{j1}$ on $\boldsymbol{w}_j$. An improper noninformative prior density for $\boldsymbol{\gamma}_1$ is used. Similar prior is used as shown in Fox and Glas (2001). Therefore, the full conditional posterior distribution of $\boldsymbol{\gamma}_1$ is given by

$$\boldsymbol{\gamma}_1 \left| \boldsymbol{\beta}_{j1}, \ \boldsymbol{T}_1 \right. \sim N\left(\left(\sum_{j=1}^{J} \boldsymbol{w}_j'\boldsymbol{T}_1^{-1}\boldsymbol{w}_j\right)^{-1} \sum_{j=1}^{J} \boldsymbol{w}_j'\boldsymbol{T}_1^{-1}\boldsymbol{\beta}_{j1}, \ \left(\sum_{j=1}^{J} \boldsymbol{w}_j'\boldsymbol{T}_1^{-1}\boldsymbol{w}_j\right)^{-1}\right), \tag{3.6}$$

and $\boldsymbol{\gamma}_q$ is drawn in the same manner for $q = 2, \cdots, Q$.

Step 6: Sampling the residual variance-covariance structure Σ_e. A prior for Σ_e is an Inverse-Wishart$\left(\nu_0, \Sigma_0^{-1}\right)$ distribution. The full conditional posterior distribution of Σ_e is given by

$$\Sigma_e \left| \boldsymbol{\theta}, \ \boldsymbol{\beta} \right. \sim \text{Inverse-Wishart}\left(\nu_0 + N, \left(S + \Sigma_0\right)^{-1}\right) \tag{3.7}$$

where $S = \sum_{j=1}^{J}\sum_{i=1}^{n_j} \left(\boldsymbol{\theta}_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\beta}_j\right)\left(\boldsymbol{\theta}_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\beta}_j\right)'$,where $N = J \times n_j$.

Step 7: Sampling the level-3 variance-covariance structure $\boldsymbol{T} = diag\left(\boldsymbol{T}_1, \cdots, \boldsymbol{T}_Q\right)$. $\boldsymbol{T}_1$ is drawn first. A prior for $\boldsymbol{T}_1$ is an Inverse-Wishart$\left(\nu_1, \Sigma_1^{-1}\right)$ distribution. The full conditional posterior distribution of $\boldsymbol{T}_1$ is given by

$$\boldsymbol{T}_1 \left| \boldsymbol{\beta}_{j1}, \ \boldsymbol{\gamma}_1 \right. \sim \text{Inverse-Wishart}\left(\nu_1 + J, \left(S_1 + \Sigma_1\right)^{-1}\right) \tag{3.8}$$

where $S_1 = \sum_{j=1}^{J} \left(\boldsymbol{\beta}_{j1} - \boldsymbol{w}_j\boldsymbol{\gamma}_1\right)\left(\boldsymbol{\beta}_{j1} - \boldsymbol{w}_j\boldsymbol{\gamma}_1\right)'$,and $\boldsymbol{T}_q$ is drawn in the same manner for $q = 2, \cdots, Q$.

3.3. Model Selection

The deviance information criterion (DIC) was introduced by Spiegelhalter et al. (2002) as a model selection criterion for the Bayesian hierarchical models. Similar to many other criteria (such as the Bayesian information criterion or BIC; BIC is not intended to predict out-of-sample model performance but rather is designed for other purposes, we do not consider it further here (Gelman et al., 2014), it trades a measure of model adequacy against a measure of complexity. Specifically, the DIC is defined as the sum of a deviance measure and a penalty term for the effective number of parameters based on a measure of model complexity. The model with a larger DIC has a better fit to the data. In the framework of a multilevel IRT models, the performances of DICs based on five versions of deviances have been investigated in Zhang et al. (2019). The DIC used in this current study belongs to the top-level marginalized DIC in their paper. The reason for using the top-level marginalized DIC in our paper is that our main purpose is to investigate the influences of fixed effects ($\boldsymbol{\gamma}$) on the multiple dimensional abilities. Therefore, the deviance is defined at the highest level fixed effects ($\boldsymbol{\gamma}$), where the random effects of intermediate processes, such as the second-level random individual ability effects $\boldsymbol{\theta}$ or the third-level random coefficient effects $\boldsymbol{\beta}$, will not be considered in the defined deviance. Next, the calculation formula of the top-level marginalized DIC is given.

Let $\boldsymbol{\Omega}_1 = (\boldsymbol{\xi}, \boldsymbol{\Sigma}_e, \boldsymbol{T})$ ($\boldsymbol{\Omega}_1$ do not include the intermediate process random parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$). According to the augmented data likelihood $p(\boldsymbol{Z}|\boldsymbol{\Omega}_1)$, we can obtain the following deviance

$$D(\boldsymbol{\gamma}) = -2\log p(\boldsymbol{Z}|\boldsymbol{\Omega}_1).$$

Then the top-level marginalized DIC is defined as

$$\begin{aligned}\text{DIC} &= \int [\text{DIC}|\boldsymbol{Z}, \boldsymbol{\Omega}_1] \cdot p(\boldsymbol{Z}, \boldsymbol{\Omega}_1|\boldsymbol{Y})\, d\boldsymbol{Z}d\boldsymbol{\Omega}_1 \\ &= \int \left[D(\overline{\boldsymbol{\gamma}}|\boldsymbol{Z}, \boldsymbol{\Omega}_1) + 2p_D(\boldsymbol{Z}, \boldsymbol{\Omega}_1)\right] \cdot p(\boldsymbol{Z}, \boldsymbol{\Omega}_1|\boldsymbol{Y})\, d\boldsymbol{Z}d\boldsymbol{\Omega}_1 \\ &= E_{\boldsymbol{Z}, \boldsymbol{\Omega}_1}\left[D(\overline{\boldsymbol{\gamma}}) + 2p_D(\boldsymbol{Z}, \boldsymbol{\Omega}_1)|\boldsymbol{Y}\right] \qquad (3.9)\end{aligned}$$

In Equation (3.9), the conditional DIC is a function of $\boldsymbol{Z}$ and $\boldsymbol{\Omega}_1$, which can be written as $[\text{DIC}|\boldsymbol{Z}, \boldsymbol{\Omega}_1]$. $D(\overline{\boldsymbol{\gamma}})$ denotes the deviance of the posterior estimation mean given augmented data $\boldsymbol{Z}$ and $\boldsymbol{\Omega}_1$. $p_D(\boldsymbol{Z}, \boldsymbol{\Omega}_1)$ is the effective number of parameters given the augmented data $\boldsymbol{Z}$ and $\boldsymbol{\Omega}_1$, which can be expressed as $p_D(\boldsymbol{Z}, \boldsymbol{\Omega}_1) = \overline{D(\boldsymbol{\gamma})} - D(\overline{\boldsymbol{\gamma}})$.

An important advantage of DIC is that it can be easily calculated from the generated samples. It can be obtained by MCMC sampling augmentation auxiliary variable $\boldsymbol{Z}$ and structural parameters $\boldsymbol{\Omega}_1$ from the joint posterior distribution $p(\boldsymbol{Z}, \boldsymbol{\Omega}_1|\boldsymbol{Y})$.

4. SIMULATION

4.1. Simulation 1

A simulation study is conducted to evaluate the performance of the proposed Gibbs sampler MCMC method for recovering the parameters of the multilevel IRT models. For illustration purposes, we only consider one explanatory variable on both levels, and the number of dimensions is fixed at 2 $(q = 2)$. The true structural multilevel model is simplified as

The individual-level model:

$$\theta_{ijq} = \beta_{0jq} + x_{ij}\beta_{1jq} + e_{ijq}, \qquad (4.1)$$

where

$$\boldsymbol{e} = \begin{pmatrix} e_{ij1} \\ e_{ij2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{e_1} & \sigma_{e_1e_2} \\ \sigma_{e_2e_1} & \sigma^2_{e_2} \end{pmatrix}\right). \qquad (4.2)$$

The school-level model:

$$\begin{aligned}\beta_{0jq} &= \gamma_{00q} + \gamma_{01q}w_j + u_{0jq}, \qquad (4.3) \\ \beta_{1jq} &= \gamma_{10q} + \gamma_{11q}w_j + u_{1jq},\end{aligned}$$

where

$$\begin{pmatrix} u_{0jq} \\ u_{1jq} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{T}\right), \boldsymbol{T} = \begin{pmatrix} \tau_{00q} & \tau_{01q} \\ \tau_{10q} & \tau_{11q} \end{pmatrix}. \qquad (4.4)$$

We use the multidimensional two-parameter normal ogive model to generate the responses. The test length is set to 30. In the multidimensional item response theory book, Reckase (2009, p. 93) points out that the each element of discrimination parameter vectors, a_{kq}, can take on any values except the usual monotonicity constraint that requires the values of the elements of $\boldsymbol{a}_k$ be positive, where $\boldsymbol{a}_k = (a_{k1}, a_{k2})'$. Therefore, we adopt the truncated normal distribution with mean 1.5 and variance 1 to generate the true value of the each element of discrimination parameter vectors $\boldsymbol{a}_k$. That is, $a_{kq} \sim N(1.5, 1)\, I\left(a_{kq} > 0\right)$, $q = 1, 2$, $k = 1, \ldots, 30$. For the difficulty parameter, the selection of the true values is the same as that of the traditional unidimensional IRT models. Here we assume that the difficult parameters are generated from the standard normal distribution. That is, $b_k \sim N(0, 1)$, $k = 1, \ldots, 30$. The ability parameters of 2,000 students from population $N\left(\boldsymbol{X}_{ij}\boldsymbol{\beta}_j, \boldsymbol{\Sigma}_e\right)$ are divided into $J = 10$ groups, with n_j (200) students in each group. The fixed effect $\boldsymbol{\gamma}$ is chosen as an arbitrary value between -1 and 1. For simplicity, we suppose that at level 3, each of the dimensional covariances τ_{01q} and τ_{10q} is equal to 0 for $q = 1, 2$, which means that the level-3 residuals between random coefficients $\boldsymbol{\beta}_q = \left(\beta_{0jq}, \beta_{01jq}\right)$ are independent of each other. The level-3 variances τ_{00q} and τ_{11q} are, respectively, set equal to 0.100, for $q = 1, 2$ such that they have very low stochastic volatility in the vicinity of the level-3 mean. The level-2 residual variance-covariance (VC) are set to 0.300, 0.500, and 0.075. The explanatory variables $\boldsymbol{X}$ and $\boldsymbol{W}$ are drawn from $N(0.25, 1)$ and $N(0.5, 1)$, respectively.

The posterior distribution in the Bayesian framework can be obtained by connecting with the likelihood function (sample information) and prior distribution (prior information). In general, the two kinds of information have important influence on the posterior distribution. In large scale educational assessment, the number of examinees is often very large, for example, in our real data study, the number of examinees and items, respectively, reach 2000 and 124. Therefore, the likelihood information plays a dominant role, and the selection of different priors (informative or non-informative) has no significant influence on the posterior inferences. As a result, the non-informative priors are often used in many educational measurement studies, e.g., van der Linden (2007) and Wang et al. (2018). In this paper, the prior specification will be uninformative enough for the data to dominate the prior, so that the influence of the prior on the results will be minimal. Next, we give the prior distributions of parameters involved in the simulation 1. The priors of the discrimination parameters and difficulty parameters are set as the non-informative priors

$\boldsymbol{a}_k \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix}\right) \mathrm{I}\left(\boldsymbol{a}_k \,|a_{k1} > 0, a_{k2} > 0\right)$ and $N(0, 100)$. The fixed effect $\boldsymbol{\gamma}$ follows a uniform distribution $U(-2, 2)$. The prior to the VC matrix of the level-2 ability dimensions is a 2-by-2 identity matrix. As used in many educational and psychological research studies (see Fox and Glas, 2001; Kim, 2001; Sheng, 2010), the priors to the VC matrices of the level-3, $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$, are set to the non-informative priors based on Fox and Glas (2001)'s paper (see Fox and Glas, 2001), where $p\left(\boldsymbol{T}_q\right) \propto 1, q = 1, 2$.

The convergence of Gibbs sampler is checked by monitoring the trace plots of the parameters for consecutive sequences of 20,000 iterations. The trace plots of two items randomly selected, fixed-effect parameters, level-2 residual variance-covariance component parameters and level-3 residual variance-covariance component parameters are shown in **Supplementary Material**. The trace plots show that all parameter estimates stabilize after 5,000 iterations and then converge quickly. Thus, we set the first 5,000 iterations as the burn-in period. In addition, the Brook-Gelman ratio diagnostic Brooks and Gelman (1998) ($\widehat{R}$; as updated Gelman-Rubin statistic) plots are used to monitor the convergence and stability. Four chains started at overdispersed starting values are run for monitoring the convergence. Our Brook-Gelman ratios are close to 1.2. The true values, the expected a priori (EAP) estimation and the 95% highest posterior density intervals (HPDIs) for item parameters are shown in **Table 1**. **Table 2** presents the true values and the estimated values of fixed effects $\boldsymbol{\gamma}$, level-2 covariance components, and level-3 variance components $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$.

The accuracy of the parameter estimates is measured by two evaluation indexes, namely, Bias and root mean squared error (RMSE). The recovery results are based on 100 times MCMC repeated iterations. That is, 100 replicas are generated. The results of the accuracy of the parameter estimates are displayed in **Tables 3**, **4**. From **Tables 3**, **4**, we see that Gibbs sampling algorithm provides accurate estimates of the item parameters and multilevel structure parameters in the sense of having small Bias and RMSE values.

4.2. Simulation 2

The purpose of this simulation study is to verify whether the Gibbs sampling algorithm can guarantee the accuracy of parameters estimation when the dimensions of latent ability increase so that it can be used to guide real data analysis later. The simulation design is as follows.

The number of dimensions is fixed at 4. The multidimensional normal ogive IRT model is used to generate responses. Two factors and their varied conditions are considered: (a) number of individuals, N = 1,000, 2,000, or 3,000; (b) number of items, K = 40, 100, or 200, and for per subtest number of itmes, 10, 25, or 50. Fully crossing the different levels of these two factors yield 9 conditions. Individuals (N = 1,000, 2,000, 3,000) are equally distributed to 10 schools (J = 10). True values of item parameters and priors of all of parameters are generated by the same in simulation 1. The true values of the fixed effects are, respectively, $1.000\left(\gamma_{00q}\right)$, $0.300\left(\gamma_{01q}\right)$, $0.500\left(\gamma_{10q}\right)$ and $0.350\left(\gamma_{11q}\right)$, $q = 1, 2, 3, 4$, and the level-2 variance are $0.300\left(\sigma^2_{e_1}\right)$, $0.500\left(\sigma^2_{e_2}\right)$, $0.750\left(\sigma^2_{e_3}\right)$, and $1.000\left(\sigma^2_{e_4}\right)$, and the covariance are set to 0.075. The level-3 variance are 0.1 $\left(\tau_{00q}, \tau_{11q}\right)$, and the covariance are 0 $\left(\tau_{01q}, \tau_{10q}\right)$. The multilevel structural models (Equations 2.2 and 2.3) in simulation study 1 are used, but the dimensions are fixed at 4.

The accuracy of the parameter estimates is measured by two evaluation indexes, namely, Bias and RMSE. The recovery results are based on the MCMC iterations repeated 100 times. The detail results of the accuracy of the parameter estimates under nine conditions are display in **Table 5**. The Biases are −0.089∼0.094 for the fixed effect parameters, −0.063∼0.117 for the level-2 variance-covariance component parameters, −0.069∼0.105 for the level-3 variance-covariance component parameters. The RMSEs are 0.152∼0.311 for the fixed effect parameters, 0.147∼0.438 for the level-2 variance-covariance component parameters, 0.132∼0.382 for the level-3 variance-covariance component parameters. Furthermore, the Bias and RMSE have a smaller trend with the increase in the number of individuals and items; in other words, increasing the number of individuals and items helps to improve the estimation accuracy of the structural parameters. In summary, the Gibbs sampling algorithm is effective for various numbers of individuals and items, and it can be used to guide practices.

5. REAL DATA ANALYSIS—EXAMINING THE CORRELATION BETWEEN DIFFERENT ABILITY DIMENSIONS AND COVARIATES

To illustrate the applicability of the multidimensional two-parameter normal ogive model in operational large-scale assessments, we consider a data set about students' English achievement test for junior middle schools conducted by NENU Branch, Collaborative Innovation Center of Assessment toward Basic Education Quality at Beijing Normal University. The analysis of the test data will help us to gain a better understanding of the practical situation of students' English academic latent traits and to explore the factors that affect their English academic latent traits. The results of this analysis will be potentially very valuable for development and improvement of educational quality monitoring mechanism in China.

5.1. Data Description

The data contain a two-stage cluster sample of 2,029 students in grade 7. These students are from 16 schools, with 121–134 students in each school. In the first stage, the sampling population is classified according to district, and schools are selected at random. In the second stage, students in grade 7 are selected at random from each school. The English test is a test battery consisting of four subscales: vocabulary (40 items), grammar (24 items), comprehensive reading (40 items), and table computing (20 items). All 124 multiple-choice items are scored using a dichotomous format. The Cronbach's alpha coefficients for vocabulary, grammar, reading comprehension and table computing items are 0.942, 0.875, 0.843, and 0.816, respectively. Level-2 and level-3 background covariates of individuals, teacher

TABLE 3 | Evaluating the accuracy of item parameter estimation.

	a_{k1}			a_{k2}			b_k		
Item	**True**	**Bias**	**RMSE**	**True**	**Bias**	**RMSE**	**True**	**Bias**	**RMSE**
1	1*	0	0	0*	0	0	0*	0	0
2	0*	0	0	1*	0	0	0*	0	0
3	0.914	−0.037	0.114	0.686	−0.014	0.090	−1.182	0.028	0.144
4	1.102	0.025	0.098	1.468	0.017	0.125	0.441	−0.015	0.093
5	2.055	−0.010	0.073	1.428	0.025	0.047	−1.197	−0.170	0.137
6	2.291	0.070	0.153	1.146	0.013	0.084	−2.536	0.012	0.126
7	2.131	0.054	0.119	0.758	0.002	0.035	1.782	−0.023	0.149
8	1.027	−0.018	0.159	1.720	0.016	0.140	0.152	0.007	0.094
9	0.569	−0.005	0.136	1.119	0.033	0.102	0.964	−0.037	0.072
10	0.578	−0.019	0.180	2.129	−0.035	0.185	1.462	0.023	0.103
11	0.795	0.002	0.088	1.445	0.021	0.137	0.619	−0.019	0.081
12	2.279	0.110	0.153	1.148	−0.016	0.098	−2.020	−0.008	0.053
13	0.714	−0.098	0.142	2.225	−0.015	0.053	0.602	−0.025	0.091
14	2.200	0.016	0.093	1.465	0.006	0.039	0.127	0.036	0.127
15	1.565	0.024	0.120	0.728	−0.017	0.092	−0.587	−0.018	0.116
16	2.419	0.020	0.162	2.408	−0.028	0.164	−0.218	−0.007	0.092
17	1.561	0.034	0.105	1.398	−0.010	0.072	0.830	−0.041	0.115
18	2.457	0.013	0.091	2.111	0.041	0.109	1.558	0.002	0.150
19	0.714	−0.028	0.155	0.918	−0.035	0.156	1.504	−0.017	0.197
20	2.447	0.035	0.198	1.704	0.050	0.143	0.126	−0.016	0.156
21	1.588	−0.026	0.185	2.170	0.007	0.124	−0.760	0.029	0.256
22	1.724	−0.003	0.147	1.590	−0.019	0.128	0.769	−0.098	0.153
23	2.273	−0.029	0.084	0.948	−0.031	0.060	0.265	−0.160	0.179
24	1.228	−0.030	0.189	2.782	−0.027	0.194	−1.398	−0.031	0.132
25	0.687	−0.013	0.075	2.261	0.014	0.107	1.802	0.024	0.193
26	1.665	0.001	0.120	0.572	−0.004	0.068	0.033	−0.012	0.090
27	2.383	0.017	0.148	1.871	0.015	0.095	1.307	0.022	0.158
28	1.778	−0.008	0.113	2.326	−0.021	0.140	−0.871	−0.004	0.083
29	1.522	0.019	0.096	2.909	0.025	0.163	0.241	0.009	0.127
30	1.173	0.005	0.181	1.703	0.007	0.098	0.397	−0.034	0.221

**indicates the constraints for model identification. RMSE denotes the root mean squared error.*

satisfaction, and school climate (teachers and schools constitute level 3) are measured. At the individual level, gender (0=male, 1=female) and socioeconomic statuses are measured; the latter is measured by the average of two indicators: the father's and mother's education, which are five-point Likert items; scores range from 0 to 8. At the teacher and school levels, teacher satisfaction is measured by 20 five-point Likert items, and school environment from the principal's perspective is measured by 23 five-point Likert items.

5.1.1. Prior Distributions

Based on the setting of priors in the simulation 1, we give the prior distributions of parameters involved in following the real data analysis. The priors of the difficulty parameters and discrimination parameters are set from $b_k \sim N(0,1)$ and $\boldsymbol{a}_k = (a_{k1}, a_{k2}, a_{k3}, a_{k4})' \sim N(\mathbf{0}, 100\mathbf{I}_{4\times4})\,\mathrm{I}(\boldsymbol{a}_k\,|a_{k1} > 0, a_{k2} > 0, a_{k3} > 0, a_{k4} > 0)$, $j = 1, 2, \ldots, 124$, where $\mathbf{I}_{4\times4}$ is 4-by-4 identity matrix. The fixed effect $\boldsymbol{\gamma}$ follows a uniform distribution $U(-2,2)$. The prior to the variance-covariance matrix of the level-2 ability dimensions is a 4-by-4 identity matrix. The prior to the variance-covariance matrix of the level-3 $\boldsymbol{T}_1$, $\boldsymbol{T}_2$, $\boldsymbol{T}_3$, and $\boldsymbol{T}_2$ are set to non-informative priors based on Fox and Glas (2001)'s paper, where $p(\boldsymbol{T}_q) \propto$constant, $q = 1, 2, 3, 4$.

5.1.2. Convergence Diagnosis

The full conditional distribution of Gibbs sampling is run for 20,000 iterations using real data. The trace plots of parameters stabilize after 5,000 iterations. Thus, the first 5,000 iterations are set as the burn-in period. The average over the drawn parameters is calculated after the burn-in period. Moreover, Four chains started at overdispersed starting values are run for monitoring the convergence. The Brook-Gelman ratios are close to 1.2. Therefore, it can be inferred that the estimated parameters are convergent.

TABLE 4 | Evaluating the accuracy of the two-dimensional fixed effects and variance-covariance components.

Fixed effect	True	Bias	RMSE	Fixed effect	True	Bias	RMSE
γ_{001}	1.000	−0.018	0.082	γ_{002}	−0.350	−0.027	0.169
γ_{011}	0.300	0.026	0.156	γ_{012}	0.300	−0.019	0.096
γ_{101}	0.500	0.021	0.148	γ_{102}	0.500	0.022	0.147
γ_{111}	0.350	−0.025	0.173	γ_{112}	−1.000	0.014	0.121
Level-2 random effect			**True**		**Bias**		**RMSE**
	$\sigma^2_{e_1}$		0.300		0.023		0.098
	$\sigma_{e_1e_2}$		0.075		0.018		0.163
	$\sigma_{e_2e_1}$		0.075		0.018		0.163
	$\sigma^2_{e_2}$		0.500		0.029		0.117
Level-3 T_1	**True**	**Bias**	**RMSE**	**Level-3 T_2**	**True**	**Bias**	**RMSE**
τ_{001}	0.100	0.015	0.164	τ_{002}	0.100	−0.029	0.143
τ_{011}	0	0.013	0.182	τ_{012}	0	0.017	0.187
τ_{101}	0	0.013	0.182	τ_{102}	0	0.017	0.187
τ_{111}	0.100	−0.026	0.139	τ_{112}	0.100	0.019	0.167

TABLE 5 | Evaluating the accuracy of the structure parameters in the simulation 2.

Number of	Number of	Fixed effect γ		Level-2 VC Σ_e		Level-3 VC T	
individuals	items	Bias	RMSE	Bias	RMSE	Bias	RMSE
	40	−0.089	0.031	0.046	0.438	0.064	0.038
1000	100	0.073	0.191	0.078	0.195	−0.037	0.203
	200	0.094	0.174	−0.063	0.160	0.081	0.198
	40	0.056	0.206	0.117	0.319	0.105	0.207
2000	100	0.028	0.167	0.064	0.177	−0.069	0.189
	200	−0.041	0.152	−0.037	0.154	0.021	0.156
	40	0.039	0.231	0.055	0.213	0.032	0.195
3000	100	−0.035	0.189	0.082	0.246	−0.058	0.145
	200	0.017	0.159	0.041	0.147	0.045	0.132

The VC stands for the abbreviation of variance-covariance.

5.2. Model Selection in Real Data

In the real data example, we consider four dimensions of ability: vocabulary cognitive ability, grammar structure diagnosing ability, reading comprehension ability, and table computing ability. These abilities are affected by individual covariates such as socioeconomic status and gender. The individual can be nested into higher group levels (school), which are affected by group covariates such as teacher satisfactions and school climate from the teachers' perspective. In this current study, we only focus on the specific abilities of four dimensions without the general ability, which is different from Huang and Wang (2014, p. 497, Equation 3)'s ability model with hierarchical structure. According to the above-mentioned DIC model selection method, three models are considered in fitting the real data, in which the DIC can be formulated to choose between models that differ in the fixed and/or random part of the structural model to combine with the measurement model. The multidimensional IRT measurement model is identical to the three candidate models. The structural multilevel model 1 consists of the two level-2 background variables SES and Gender and the level-2 random intercept. The effects of the level-2 background variables SES and Gender are fixed across schools. The structural multilevel part is given by

$$\textbf{Model 1} \quad \begin{cases} \theta_{ijq} = \beta_{0jq} + SES_{ij}\beta_{1jq} + Gender_{ij}\beta_{2jq} + e_{ijq}, \\ \beta_{0jq} = \gamma_{00q} + u_{0jq}, \\ \beta_{1jq} = \gamma_{10q}, \\ \beta_{2jq} = \gamma_{20q}. \end{cases} \tag{5.1}$$

Model 2 is extended by including two latent predictors at level 3, Satisfaction and Climate. The effects of the level-2 background variable SES are allowed to vary across schools. The structural multilevel part is given by

TABLE 6 | Estimated DIC values for the three models fitted to the English test data.

	P_D	$\overline{D}$	DIC
Model 1	134,470	1,010,030	1,144,500
Model 2	79,065	891,425	970,490
Model 3	81,607	895,073	976,680

$$\textbf{Model 2}\ \begin{cases} \theta_{ijq} = \beta_{0jq} + SES_{ij}\beta_{1jq} + Gender_{ij}\beta_{2jq} + e_{ijq}, \\ \beta_{0jq} = \gamma_{00q} + Satisfaction_j\gamma_{01q} + Climate_j\gamma_{02q} + u_{0jq}, \\ \beta_{1jq} = \gamma_{10q} + u_{1jq}, \\ \beta_{2jq} = \gamma_{20q}. \end{cases} \tag{5.2}$$

Model 3 captures the effects of the level-2 background variables SES and Gender, which are allowed to vary across schools. The structural multilevel part is given by

$$\textbf{Model 3}\ \begin{cases} \theta_{ijq} = \beta_{0jq} + SES_{ij}\beta_{1jq} + Gender_{ij}\beta_{2jq} + e_{ijq}, \\ \beta_{0jq} = \gamma_{00q} + Satisfaction_j\gamma_{01q} + Climate_j\gamma_{02q} + u_{0jq}, \\ \beta_{1jq} = \gamma_{10q} + u_{1jq}, \\ \beta_{2jq} = \gamma_{20q} + u_{2jq}. \end{cases} \tag{5.3}$$

Question (1): According to the model selection results, which model is the best to fit the data and how can judge the individual-level regression coefficients be judged as fixed effect or random effect?

The estimated DIC values are presented in **Table 6**. Model 2 shows that the smallest effective number of model parameters among the three models, which is preferred given the DIC values of the three models. The DIC values of models 2 and 3 are smaller than those of model 1, which can be attributed to the additional latent predictors at level 3, i.e., Satisfaction and Climate. Note that in model 2, the individual random-effect parameters are modeled as group-specific random effects (level-3 Satisfaction and Climate latent predictors), leading to a serious reduction in the effective number of model parameters, which can be inferred from the P_D value in **Table 6**. The DIC value of model 2 is smaller than that of model 3. The residual u_{2jq} of the random effect β_{2jq} is estimated equal to 0, which is equivalent to fixing the effect of the level-2 background variable Gender across schools.

TABLE 7 | Parameter estimation of the multilevel multidimensional IRT model for vocabulary cognitive ability.

	Vocabulary cognitive ability		
Fixed effects	**EAP**	***SD***	**HPDI**
γ_{001}	0.760	0.186	[0.391, 1.137]
γ_{011} (*ST*)	0.502	0.143	[0.223, 0.788]
γ_{021} (*CT*)	0.225	0.149	[−0.068, 0.520]
γ_{101} (*SES*)	0.642	0.128	[0.390, 0.893]
γ_{201} (*GD*)	0.339	0.160	[0.025, 0.657]
Random effects	**EAP**	**SD**	**HPDI**
τ^2_{001}	0.537	0.124	[0.227, 1.200]
τ^2_{011}	0.004	0.126	[−0.228, 0.241]
τ^2_{021}	−0.006	0.164	[−0.344, 0.383]
τ^2_{111} (*SES*)	0.247	0.134	[0.112, 0.541]
τ^2_{121}	−0.064	0.112	[−0.292, 0.110]
τ^2_{221} (*GD*)	0.030	0.191	[0.015, 0.043]

ST, teacher satisfaction; CT, climate; SES, socioeconomic-status; GD, gender. EAP denotes the expected a posteriori estimation. SD denotes the standard deviation. HPDI is the 95% highest posterior density interval.

TABLE 8 | Parameter estimation of the multilevel multidimensional IRT model for diagnosing ability of grammar structure.

	Vocabulary cognitive ability		
Fixed effects	**EAP**	***SD***	**HPDI**
γ_{001}	0.760	0.186	[0.391, 1.137]
γ_{011} (*ST*)	0.502	0.143	[0.223, 0.788]
γ_{021} (*CT*)	0.225	0.149	[−0.068, 0.520]
γ_{101} (*SES*)	0.642	0.128	[0.390, 0.893]
γ_{201} (*GD*)	0.339	0.160	[0.025, 0.657]
Random effects	**EAP**	**SD**	**HPDI**
τ^2_{001}	0.537	0.124	[0.227, 1.200]
τ^2_{011}	0.004	0.126	[−0.228, 0.241]
τ^2_{021}	−0.006	0.164	[−0.344, 0.383]
τ^2_{111} (*SES*)	0.247	0.134	[0.112, 0.541]
τ^2_{121}	−0.064	0.112	[−0.292, 0.110]
τ^2_{221} (*GD*)	0.030	0.191	[0.015, 0.043]

ST, teacher satisfaction; CT, climate; SES, socioeconomic-status; GD, gender. EAP denotes the expected a posteriori estimation. SD denotes the standard deviation. HPDI is the 95% highest posterior density interval.

5.3. Structural Parameter Analysis

Over the past 40 years, a large number of studies have shown that there is a direct relationship between the individuals' language learning ability and the parents' education. For example, Teachman (1987) made use of high school survey data in the United States to explore the influence of family background on childhood education. The results of this study indicated that the parents' occupations, incomes, and educations have a very important impact on children language academic achievement. Moreover, Stern (1983) shows that language is a social mechanism, which needs to be learned in the social environment, even in the biological basis play an important role of mother tongue acquisition, social factors related to children and their parents also play an important role. However, in our study, whether the parents' educational level (SES) has influence on the four kinds of abilities in English learning; the following question will be considered:

Question (2): How will students from different ends of the socioeconomic-status (SES) score in English performance as

TABLE 9 | Parameter estimation of the multilevel multidimensional IRT model for reading comprehension ability.

	Reading comprehension ability		
Fixed effects	**EAP**	***SD***	**HPDI**
γ_{003}	0.919	0.187	[0.548, 1.293]
γ_{013} (*ST*)	0.332	0.148	[0.041, 0.624]
γ_{023} (*CT*)	0.081	0.168	[−0.249, 0.417]
γ_{103} (*SES*)	0.542	0.118	[0.308, 0.780]
γ_{203} (*GD*)	0.232	0.155	[−0.070, 0.544]
Random effects	**EAP**	***SD***	**HPDI**
τ^2_{003}	0.535	0.111	[0.223, 1.220]
τ^2_{013}	0.040	0.198	[−0.156, 0.275]
τ^2_{023}	−0.024	0.153	[−0.342, 0.264]
τ^2_{113} (*SES*)	0.207	0.133	[0.091, 0.456]
τ^2_{123}	0.004	0.089	[−0.170, 0.182]
τ^2_{223} (*GD*)	0.037	0.177	[0.027, 0.052]

ST, teacher satisfaction; CT, climate; SES, socioeconomic-status; GD, gender. EAP denotes the expected a posteriori estimation. SD denotes the standard deviation. HPDI is the 95% highest posterior density interval.

TABLE 10 | Parameter estimation of the multilevel multidimensional IRT model for table computing ability.

	Table computing ability		
Fixed effects	**EAP**	***SD***	**HPDI**
γ_{004}	0.255	0.130	[−0.003, 0.514]
γ_{014} (*ST*)	0.039	0.104	[−0.165, 0.246]
γ_{024} (*CT*)	0.295	0.101	[0.099, 0.498]
γ_{104} (*SES*)	0.596	0.126	[0.351, 0.849]
γ_{204} (*GD*)	−0.266	0.120	[−0.506, -0.026]
Random effects	**EAP**	***SD***	**HPDI**
τ^2_{004}	0.447	0.144	[0.201, 0.970]
τ^2_{014}	0.082	0.084	[−0.043, 0.269]
τ^2_{024}	−0.041	0.100	[−0.223, 0.098]
τ^2_{114} (*SES*)	0.226	0.106	[0.101, 0.485]
τ^2_{124}	−0.014	0.069	[−0.160, 0.114]
τ^2_{224} (*GD*)	0.022	0.102	[0.015, 0.035]

ST, teacher satisfaction; CT, climate; SES, socioeconomic-status; GD, gender. EAP denotes the expected a posteriori estimation. SD denotes the standard deviation. HPDI is the 95% highest posterior density interval.

tested in four types of latent abilities, based on the level-2 gender (GD), level-3 teacher satisfaction (ST) and school climate (CT).

From **Tables 7–10**, we can find that the estimated fixed effects γ_{10q}(SES) are 0.642, 0.312, 0.542, and 0.596 for $q = 1, 2, 3, 4$, respectively. It can be observed that students with high SES scores perform better than students with low SES scores, where performance is measured by four types of latent abilities when controlling for the level-2 GD individual covariates and the level-3 ST and CT school covariates. That is, their parents' educational level differs by one unit for the male students from the same class and school. In English learning, vocabulary cognitive ability, the ability to diagnose grammar structure, reading comprehension ability and table computing ability have the differences of 0.642, 0312, 0.542, and 0.596, respectively. The rate of increase in grammatical diagnostic ability (0.312) is markedly smaller than that of the other three kinds of abilities. In addition, compared to male students, the differences in the four dimensions of ability are 0.981, 0.706, 0.874, and 0.330 for female students, respectively. In summary, the education of parents (SES) is responsible for students' English learning abilities. The parents with a high SES values have more prospective awareness in English learning based on their own learning experiences, provide more diversified learning ways, and know how to create a better English learning environment for students. In addition, parents with better education can provide more important learning guidance in English. In general, the better the parents' education, the better they will able to tutor student's English learning.

Etaugh and Bridges (2003), Li (2005), and Burstall (1975) found that females were better than males in most of the language tasks (vocabulary, reading, grammar, spelling and writing), and the difference in language ability appeared earlier than other cognitive abilities. In infancy, females show more linguistic advantages than males, and they speak more fluently, and have a richer vocabulary. To about 11 years old, they are not only good at simple spelling, but also are able to do more complicated writing tasks. In schools, teachers have found that females do better in reading comprehension, and they are less likely to have reading problems, including reading barriers. However, whether or not have the above conclusions in this study, next the following issues will be considered:

Question (3): What relationship exists between males and females' performances in different latent abilities by controlling for SES, ST and CT?

Results from **Tables 7–10** show that for male and female students from the same class and school with the same SES scores, female students' performances of vocabulary cognitive ability, the ability to diagnose grammar structure and reading comprehension ability are higher than those of male students 0.339, 0.394, 0.232. However, male students have a 0.266 advantage over female students in table computing ability. This empirical study yields almost identical conclusions for Etaugh and Bridges (2003). That is, male and female students, who have the same SES scores in the same class and school, have a great difference in the acquisition of English proficiency. Moreover, in terms of vocabulary cognition, grammatical structure analysis, reading comprehension it can be seen that females are better than males at vivid memory and mechanical memory is stronger than males. However, compared to females, males are markedly better than females at logical reasoning, deductive induction, and computing ability. In addition, according to gender difference in English learning of middle school students, the improving measure of learning from others' strong points to offset one' own weakness mainly covers: first, either teachers of students should properly understand the gender difference; second, to strengthen female students' training of logical thinking; third, to widen female students' reasoning computing ability; fourth, for the male students, to develop their vivid memory through a

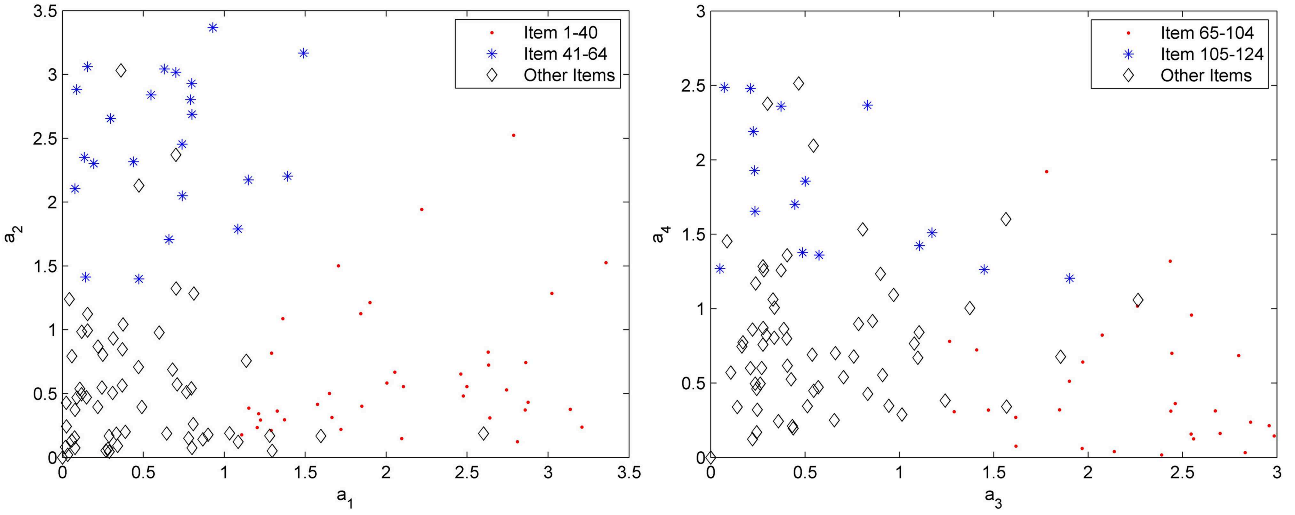


FIGURE 1 | Parameters of estimation a_{k1}, a_{k2}, a_{k3}, and a_{k4} for subscale 1 (items 1–40), subscale 2 (items 41–64), subscale 3 (items 65–104), and subscale 4 (items 105–124).

variety of teaching methods. These four points should be parallel in structure.

Question (4): What effects, if any, are seen with different teachers' or schools' effects (covariates)?

For male students who have the same SES scores from different schools, if the difference in teacher satisfaction is a unit, the difference in vocabulary cognitive ability, the ability to diagnose grammar structure and reading comprehension ability are 0.502, 0.335, and 0.331, respectively. However, the difference in the table computing ability is very small for 0.039. Teachers' factor has an important effect on students' cognitive ability, the ability to diagnose grammar structure and reading ability. On the contrary, the table computing ability has little impact.

This study indicates that the middle school teachers with high teacher satisfactions have a strong sense of responsibility, can be filled with enthusiasm in the work of education and teaching, and inspire students' learning motivation. This results in a great improvement in the students' vocabulary cognitive ability, the ability to analyze grammatical structure and reading comprehension ability owing to teachers' teaching attitude and responsibility. However, the margin of the improvement for the table computing ability is small. It is possible to play a decisive role in the students' internal factors as compared with the teachers' external factors.

As we know, people are the product of the environment. The environment has a great impact on cognition, emotion and behavior intention. Different people live in different environments so that there is a huge difference in cognition, emotion and behavior intention. Similarly, in English teaching, are whether or not the performances identical for different schools' effects (school climate)? If not, what are the effects?

The estimated results for school climate effects γ_{02q} are 0.225, 0.081, 0.086, and 0.295 for $q = 1, 2, 3, 4$, respectively. The performances associated with vocabulary cognitive ability and table computing ability are markedly affected by the level-3 CT covariates, whereas the ability to diagnose grammar structure and reading comprehension ability are not markedly affected when controlling for the level-2 SES and GD individual covariates and the level-3 ST school covariates. Analysis of the level-3 variance components reveals that the values of τ^2_{11q}(SES) are markedly different from 0, and their estimates are 0.247, 0.272, 0.207, and 0.226 for $q = 1, 2, 3, 4$, respectively. This result illustrates that the effect of SES varies from school to school. In addition, the τ^2_{22q}(GD) values are not markedly different from 0. In addition, according to the DIC model selection results, model 2 shows the best fit to the real data when β_{2jq} are defined as fixed effects. The estimation results show that the proportion of females to males does not vary among schools. The estimation covariance between the random effects τ^2_{01q}, τ^2_{02q}, and τ^2_{12q} are all not markedly different from 0. It can be concluded that the random effects are independent of each other for each type of ability. All estimated parameters are shown in **Tables 7–10**.

5.4. Item Test Dimension Evaluation

Question (5): Is it possible to use a measurement tool to determine whether items' factor patterns correlate to the subscales of the test battery? In particular, will the four subtests of the test battery be discernable according to the discrimination parameters on the four dimensions?

A test battery contains four subtests, which consist of items of measuring four dimensional abilities, and a type of latent ability can be measured mainly by a subtest. It can be observed that the EAP estimates of the discrimination parameters are plotted to determine whether the items' factor patterns reflect the subtest of the test battery in **Figure 1**. In the left-hand panel of **Figure 1**, the discrimination parameters of the first two dimensions are plotted for subtest 1 (items marked by a dot) and subtest 2 (items marked by a star), and the other items are marked by a diamond. It can be observed that the items of subtest 1 (1–40 item) have a high factor loading on the first dimension and a low factor loading on the second dimension, and the items of subtest 2 (41–64 item) have a high factor loading on the second dimension and a low factor loading on the first dimension. The other items do not vary appreciably between the two dimensions. The right-hand panel of **Figure 1** shows the pattern of the discrimination parameters of the third and fourth subtests on the third and fourth dimensions. The items of subtest 3 (65–104 item) have a high factor loading on the third dimension and a low factor loading on the fourth dimension, and the items of subtest 4 (105–124 item) have a high factor loading on the fourth dimension and a low factor loading on the third dimension. The overall pattern of the discrimination parameters fit the test battery quite well, demonstrating that each dimension is identified by items of one subtest.

6. CONCLUDING REMARKS

In this study, we mainly focus on constructing a multilevel multidimensional model to fit the hierarchical dataset about a large-scale English achievement test. Particular attention is given to assessing the correlation between multiple latent abilities and covariates.

In view of the characteristics of the test structure (i.e., (1) the students are nested within classes or schools; (2) the binary response consists of several subtests and each subtest measures a distinct latent trait), we extend the measurement model developed by Fox and Glas (2001) and Kamata (2001) to the multidimensional case by replacing their unidimensional IRT model with a multidimensional normal ogive model. The numerical results show that the multidimensional IRT model is appropriate for modeling the measurement model. It can accurately model the item/person interaction and utilize the correlations between subtests to increase the measurement precision of each subtest.

From what has been using the above empirical data, we may safely draw valuable conclusions to provide guidance for the future English teaching. Socioeconomic status (SES) has a positive impact on the abilities of four dimensions. That is, the higher families' SESs, the better performances in the four dimensional abilities. In addition, the study also found that students of different genders do not demonstrate the same level of expertise in English skills are expert in the English skills are not the same. Female students are good at the items related to the memory of the image and mechanical memory, such as the vocabulary, grammar and reading comprehension; but the male students have the advantage in reasoning calculation. Therefore, teachers should adjust the teaching methods based on the gender differences so that he or she can acquire the ability to overcome their own deficiency. Teachers' satisfaction as level 3 teacher covariate markedly impacts English table computing ability. It is possible to play a decisive role in the students' internal factors as compared with the teachers' external factors. Finally, the impact of the school climate factor on students' grammatical structure analysis and reading comprehension is not very obvious, and the specific reasons are to be studied later.

In the future studies, the correlations between schools at the level-3 should be taken into consideration. For example, the different secondary schools which are located in the same district may share a common education resources. In addition, the measurement model can be improved by considering polytomous item response theory model to analyze ordinal response data with more information. As an extension of this paper, the polytomous response model associated with the multilevel models can be used to help evaluate the multiple latent abilities, which may be more suitable for the current complex situation of educational and psychological research. In the field of estimation method, Bayesian estimation method will face serious challenges when the number of examinees or the number of items, or MCMC sample size are substantially increased. Therefore, the proposal of efficient Bayesian algorithm and the development of easy-to-use software package are also important research focus in the later period.

DATA AVAILABILITY STATEMENT

The datasets for this manuscript are not publicly available because Data from NENU Branch, Collaborative Innovation Center of Assessment toward Basic Education Quality at Beijing Normal University has signed a confidentiality agreement. Requests to access the datasets should be directed to taoj@nenu.edu.cn.

AUTHOR CONTRIBUTIONS

FC completed the writing of the article. JL and JT provided key technical support. JZ provided original thoughts and article revisions.

FUNDING

This research was supported by the National Natural Science Foundation of China (Grant 11571069).

ACKNOWLEDGMENTS

We would like to thank the Editor, and two referees for their constructive suggestions and comments, which have led to an improved version of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02387/full#supplementary-material

Figure S1 | Trace plot of $a_{9,1}$.

Figure S2 | Trace plot of $a_{9,2}$.

Figure S3 | Trace plot of b_9.

Figure S4 | Trace plot of $a_{26,1}$.

Figure S5 | Trace plot of $a_{26,2}$.

Figure S6 | Trace plot of b_{26}.

Figure S7 | Trace plots of the fixed effects in the first dimension.

Figure S8 | Trace plots of the fixed effects in the second dimension.

Figure S9 | Trace plot of σ^2_{e1}.

Figure S10 | Trace plot of σ_{e1e2}.

Figure S11 | Trace plot of σ^2_{e2}.

Figure S12 | Trace plot of τ_{001}.

Figure S13 | Trace plot of τ_{002}.

Figure S14 | Trace plot of $\tau_{011}\tau_{101}$.

Figure S15 | Trace plot of $\tau_{012}\tau_{102}$.

REFERENCES

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Appl. Psychol. Meas.* 13, 113–127. doi: 10.1177/014662168901300201

Adams, R. J., Wilson, M., and Wu, M. (1997). Multilevel item response models: an approach to errors in variables regression. *J. Educ. Behav. Stat.* 22, 47–76. doi: 10.3102/10769986022001047

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibb ssampling. *J. Educ. Stat.* 17, 251–269. doi: 10.3102/10769986017003251

Asparouhov, T., and Muthén, B. (2012). *General Random Effect Latent Variable Modeling: Random Subjects, Items, Contexts, and Parameters*. Available online at: https://www.statmodel.com/download/NCME12.pdf

Béguin, A. A., and Glas, C. A. W. (2001). MCMC estimation of multidimensional IRT models. *Psychometrika* 66, 541–561. doi: 10.1007/BF02296195

Bock, R. D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* 46, 443–459. doi: 10.1007/BF02293801

Bock, R. D., and Schilling, S. G. (2003). "IRT based item factor analysis," in *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*, ed M. du Toit (Lincolnwood, IL: Scientific Software International), 584–591.

Box, G. E. P., and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley.

Bradlow, E. T., Wainer, H., and Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika* 64, 153–168. doi: 10.1007/BF02294533

Brooks, S. P., and Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* 7, 434–455. doi: 10.1080/10618600.1998.10474787

Burstall (1975). Factors affecting foreign language learning: a consideration of some recent research findings. *Lang. Teach. Linguist. Abstr.* 29, 132–140. doi: 10.1017/S0261444800002585

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika* 75, 33–57. doi: 10.1007/s11336-009-9136-x

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *J. Educ. Behav. Stat.* 35, 307–335. doi: 10.3102/1076998609353115

Cai, L. (2010c). A two-tier full-information item factor analysis model with applications. *Psychometrika* 75, 33–57. doi: 10.1007/s11336-010-9178-0

Cai, L. (2013). *flexMIRT: Flexible Multilevel Multidimensional Item Analysis and Test Scoring (Version 2) [Computer software]*. Chapel Hill, NC: Vector Psychometric Group.

Chalmers, R. P. (2015). Extended mixed-effects item response models with the MH-RM algorithm. *J. Educ. Meas.* 52, 200–222. doi: 10.1111/jedm.12072

De Jong, M. G., and Steenkamp, J. B. E. M. (2010). Finite mixture multilevel multidimensional ordinal IRT models for large scale cross-cultural research. *Psychometrika* 75, 3–32. doi: 10.1007/s11336-009-9134-z

Embretson, S. E., and Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Etaugh, C., and Bridges, J. S. (2003). *The Psychology of Women: A Lifespan Perspective*. Boston, MA: Allyn & Bacon.

Fox, J. P., and Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika* 66, 271–288. doi: 10.1007/BF02294839

Fraser, C. (1988). *NOHARM: A Computer Program for Fitting Both Unidimensional and Multidimensional Normal Ogive Models of Latent Trait Theory*. Armidale, NSW: University of New England.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis, 3rd Edn*. Boca Raton, FL: CRC Press.

Gelman, A., Meng, X. -L., and Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Stat. Sin.* 6, 733–807.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.* 6, 721–741. doi: 10.1109/tpami.1984.4767596

Goldstein, H. (2003). *Multilevel Statistical Models*, *3rd Edn*. London: Edward Arnold.

Höhler, J., Hartig, J., and Goldhammer, F. (2010). Modeling the multidimensional structure of students' foreign language competence within and between classrooms. *Psychol. Test Assess. Model.* 52, 323–340. Retrieved from: http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2010_20100928/07_Hoehler.pdf

Holtmann, J., Koch, T., Lochner, K., and Eid, M. (2016). A comparison of ml, wlsmv, and bayesian methods for multilevel structural equation models in small samples: a simulation study. *Multivar. Behav. Res.* 51, 661–680. doi: 10.1080/00273171.2016.1208074

Hox, J. (2002). *Multilevel Analysis, Techniques and Applications*. New Jersey: Lawrence Erlbaum Associates.

Huang, H.-Y., and Wang, W.-C. (2014). Multilevel higher-order item response theory models. *Educ. Psychol. Meas.* 73, 495–515. doi: 10.1177/0013164413509628

Huang, H.-Y., Wang, W.-C., Chen, P.-H., and Su, C.-M. (2013). Higher-order item response theory models for hierarchical latent traits. *Appl. Psychol. Meas.* 37, 619–637. doi: 10.1177/0146621613488819

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *J. Educ. Meas.* 38, 79–93. doi: 10.1111/j.1745-3984.2001.tb01117.x

Kelderman, H., and Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika* 59, 149–176. doi: 10.1007/BF02295181

Kim, S. (2001). An evaluation of the Markov chain Monte Carlo method for the Rasch model. *Appl. Psychol. Meas.* 25, 163–176. doi: 10.1177/01466210122031984

Klein Entink, R. H. (2009). *Statistical models for responses and response times* (Ph.D. dissertation). University of Twente, Faculty of Behavioural Sciences, Enschede, Netherlands.

Klein Entink, R. H., Fox, J. P., and van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika* 74, 21–48. doi: 10.1007/s11336-008-9075-y

Li, L. J. (2005). *A Study on Gender Differences and Influencing Factors of High School Students' English Learning*. Fuzhou: Fujian Normal University Press.

Lindley, D. V., and Smith, A. F. M. (1972). Bayes estimates for the linear model. *J. R. Stat. Soc. B* 34, 1–41. doi: 10.2307/2985048

Lu, I. R., Thomas, D. R., and Zumbo, B. D. (2005). Embedding IRT in structural equation models: a comparison with regression based on IRT scores. *Struct. Equat. Model.* 12, 263–277. doi: 10.1207/s15328007sem1202_5

Lu, Y. (2012). *A multilevel multidimensional item response theory model to address the role of response style on measurement of attitudes in PISA 2006*. (Doctoral dissertation). University of Wisconsin, Madison, WI, United States, 164.

McDonald, R. P. (1999). *Test Theory*: *A Unified Treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mulaik, S. A. (1972). "A mathematical investigation of some multidimensional rasch models for psychological tests," in *Paper Presented at the Annual Meeting of the Psychometric Society* (Princeton, NJ).

Muraki, E., and Carlson, J. E. (1993). "Full-information factor analysis for polytomous item responses," in *Paper Presented at the Annual Meeting of the American Educational Research Association* (Atlanta, GA).

Muthén, B. O., and Asparouhov, T. (2013). "Item response modeling in Mplus: a multi-dimensional, multi-level, and multi-time point example," in *Handbook of Item Response Theory: Models, Statistical Tools, and Applications*. Retrieved from: http://www.statmodel.com/download/IRT1Version2.pdf

Muthén, B. O., du Toit, S. H. C., and Spisic, D. (1997). *Robust Inference Using Weighted Least Squares and Quadratic Estimating Equations in Latent Variable Modeling With Categorical and Continuous Outcomes*. Unpublished technical report.

Muthén, L. K., and Muthén, B. O. (1998). (1998–2012). *Mplus User's Guide, 7th Edn.* Los Angeles, CA: Muthén & Muthén.

Pastor, D. A. (2003). The use of multilevel IRT modeling in applied research: an illustration. *Appl. Meas. Educ.* 16, 223–243. doi: 10.1207/S15324818AME1603_4

Patz, R. J., and Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *J. Educ. Behav. Stat.* 24, 146–178. doi: 10.3102/10769986024002146

Patz, R. J., and Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. *J. Educ. Behav. Stat.* 24, 342–366. doi: 10.3102/10769986024004342

Plummer, M. (2003). "JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling," in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, eds K. Hornik, F. Leisch, and A. Zeileis (Vienna). Available online at: http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/

Reckase, M. D. (1972). *Development and application of a multivariate logistic latent trait model*. (Unpublished doctoral dissertation). Syracuse University, Syracuse, NY, United States.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York, NY: Springer Science Business Media, LLC.

Rupp, A. A., Dey, D. K., and Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether to when: applications of Bayesian methodology to modeling. *Struct. Equat. Model.* 11, 424–451. doi: 10.1207/s15328007sem1103_7

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional space. *Psychometrika* 39, 111–121. doi: 10.1007/BF02291580

Shalabi, F. (2002). *Effective schooling in the west bank* (Ph.D. dissertation). University of Twente, Enschede, Netherlands.

Sheng, Y. (2010). A sensitivity analysis of Gibbs sampling for 3PNO IRT models: effects of prior specifications on parameter es timates. *Behaviormetrika* 37, 87–110. doi: 10.2333/bhmk.37.87

Sheng, Y., and Wikle, C. K. (2007). Bayesian multidimensional IRT models with a hierarchical structure. *Educ. Psychol. Meas.* 68, 413–430. doi: 10.1177/0013164407308512

Skrondal, A., and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Boca Raton, FL: Chapman & Hall.

Song, X.-Y., and Lee, S.-Y. (2012). A tutorial on the Bayesian approach for analyzing structural equation models. *J. Math. Psychol.* 56, 135–148. doi: 10.1016/j.jmp.2012.02.001

Spiegelhalter, D. J, Thomas, A., Best, N. G., and Lunn, D. (2003). *WinBUGS Version 1.4 User Manual.* Cambridge: MRC Biostatistics Unit. Available online at: http://www.mrc-bsu.cam.ac.uk/bugs/

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Linde, A. V. D. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B* 64, 583–639. doi: 10.1111/1467-9868.00353

Stern, H. H. (1983). *Fundamental Concepts of Language Teaching*. Oxford: Oxford University Press.

Sympson, J. B. (1978). "A model for testing with multidimensional items," in *Proceedings of the 1977 Computerized Adaptive Testing Conference*, ed D. J. Weiss (Minneapolis, MN: University of Minnesota).

Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* 82, 528–550. doi: 10.1080/01621459.1987.10478458

Teachman, J. D. (1987). Family background, educational resources, and educational attainment. *Am. Sociol. Rev.* 52, 548–557. doi: 10.2307/2095300

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy. *Psychometrika* 72, 287–308. doi: 10.1007/s11336-006-1478-z

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *J. Educ. Behav. Stat.* 33, 5–20. doi: 10.3102/1076998607302626

Wang, C., Xu, G., and Shang, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika* 83, 223–254. doi: 10.1007/s11336-016-9525-x

Way, W. D., Ansley, T. N., and Forsyth, R. A. (1988). The comparative effects of compensatory and non-compensatory two-dimensional data on unidimensional IRT estimates. *Appl. Psychol. Meas.* 12, 239–252. doi: 10.1177/014662168801200303

Whitely, S. E. (1980a). *Measuring Aptitude Processes With Mutlicomponent Latent Trait Models*. Technical Report No. NIE-80-5. Lawrence, KS: University of Kansas.

Whitely, S. E. (1980b). Multicomponent latent trait models for ability tests. *Psychometrika* 45, 479–494. doi: 10.1007/BF02293610

Yao, L., and Schwarz, R. (2006). A multidimensional partial credit model with associated item and test statistics: an application to mixed-format tests. *Appl. Psychol. Meas.* 30, 469–492. doi: 10.1177/0146621605284537

Zhang, X., Tao, J., Wang, C., and Shi, N. Z. (2019). Bayesian model selection methods for multilevel IRT models: a comparison of five DIC-based indices. *J. Educ. Meas.* 56, 3–27. doi: 10.1111/jedm.12197

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

BRIEF RESEARCH REPORT
published: 26 November 2019
doi: 10.3389/fpsyg.2019.02479

Executive Impairment in Alcohol Use Disorder Reflects Structural Changes in Large-Scale Brain Networks: A Joint Independent Component Analysis on Gray-Matter and White-Matter Features

Chiara Crespi[1,2*†], *Caterina Galandra*[2†], *Marina Manera*[3], *Gianpaolo Basso*[4], *Paolo Poggi*[5] *and Nicola Canessa*[1,2]

[1]NEtS Center, Scuola Universitaria Superiore Istituto Universitario di Studi Superiori Pavia, Pavia, Italy, [2]Cognitive Neuroscience Laboratory, Istituti Clinici Scientifici Maugeri IRCCS, Pavia, Italy, [3]Psychology Unit, Istituti Clinici Scientifici Maugeri IRCCS, Pavia, Italy, [4]University of Milano-Bicocca, Milan, Italy, [5]Radiology Unit, Istituti Clinici Scientifici Maugeri IRCCS, Pavia, Italy

OPEN ACCESS

Edited by:
Pietro Cipresso,
Italian Auxological Institute (IRCCS),
Italy

Reviewed by:
Bernd Lachmann,
University of Ulm,
Germany
Gustavo Gonzalez-Cuevas,
Idaho State University,
United States

Correspondence:
Chiara Crespi
chiara.crespi@iusspavia.it

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *16 August 2019*
Accepted: *21 October 2019*
Published: *26 November 2019*

Citation:
Crespi C, Galandra C, Manera M, Basso G, Poggi P and Canessa N (2019) Executive Impairment in Alcohol Use Disorder Reflects Structural Changes in Large-Scale Brain Networks: A Joint Independent Component Analysis on Gray-Matter and White-Matter Features. Front. Psychol. 10:2479. doi: 10.3389/fpsyg.2019.02479

Alcohol Use Disorder (AUD) entails chronic effects on brain structure. Neurodegeneration due to alcohol toxicity is a neural signature of executive impairment typically observed in AUD, previously related to both gray-matter volume/density and white-matter abnormalities. Recent studies highlighted the role of meso-cortico-limbic structures supporting the salience and executive networks, in which the extent of neurostructural damage is significantly related to patients' executive performance. Here we aim to integrate multimodal information on gray-matter and white-matter features with a multivariate data-driven approach (joint Independent Component Analysis, jICA), and to assess the relationship between the extent of damage in the resulting neurostructural superordinate components and executive profile in AUD. Twenty-two AUD patients and 18 matched healthy controls (HC) underwent a Magnetic Resonance Imaging (MRI) protocol, alongside clinical and neuropsychological examinations. We ran jICA on five neurostructural features, including gray-matter density and different diffusion tensor imaging metrics. We extracted 12 Independent Components (ICs) and compared the resulting mixing coefficients in patients vs. HC. Finally, we correlated significant ICs with executive and clinical variables. One out of 12 ICs (IC11) discriminated patients from healthy controls and correlated positively both with executive performance in all subjects, and with lifetime duration of alcohol abuse in patients. In line with previous related evidence, this component involved widespread gray-matter and white-matter patterns including key nodes and fiber tracts of salience, default-mode and central executive networks. These findings highlighted the role of multivariate data integration as a valuable approach revealing superordinate hallmarks of neural changes related to cognition in neurological and psychiatric populations.

Keywords: alcohol use disorder, alcohol chronic consumption, voxel-based morphometry, diffusion tensor MRI, joint independent component analysis, large-scale brain network, rehabilitative applications

INTRODUCTION

Alcohol Use Disorder (AUD) is characterized by prolonged and excessive alcohol consumption, as well as constant concerns about alcoholic drinks despite adverse consequences. This condition can produce relevant alterations at different levels of analysis, from social maladaptation and cognitive impairment (Center for Behavioral Health Statistics and Quality, 2016) to pathological changes affecting anatomo-functional brain regions and networks (De La Monte and Kril, 2014; Fritz et al., 2019). Neuroimaging studies have shown both gray-matter (GM) and white-matter (WM) alterations in AUD (Bühler and Mann, 2011; Yang et al., 2016; Zahr and Pfefferbaum, 2017). Such alcohol-related neurostructural effects involve both decreased GM and WM volume and/or density (Jansen et al., 2015; Xiao et al., 2015; Galandra et al., 2018b), and altered microstructural features (e.g., fractional anisotropy decrease, mean diffusivity increase) in main fiber tracts (Fortier et al., 2014; Chumin et al., 2019). Importantly, neuroimaging studies provided converging evidence about the topological distribution of neurostructural alterations in AUD, showing a diffuse damage pattern that mostly involves fronto-striatal networks alongside frontal WM (De La Monte and Kril, 2014; Suckling and Nestor, 2017). These alterations may represent a neurostructural marker of core cognitive deficits in AUD, including impulsivity and abnormal reward-based choice behavior (see Galandra et al., 2018a). Indeed, major theories proposed to explain cognitive impairment in addiction are related to the dysregulation of either *impulsive vs. reflective* brain systems – the *Control-related deficit theory* (Bechara and Damasio, 2005) – or the *reward vs. stress* systems – the *Reward-related deficit theory* (Koob, 2013). While the latter is more focused on the emotional states associated to craving, seen as a result of the down-regulation of the reward system in favor of the up-regulation of the stress system, the former attributes the emergence of craving to the failure of attention control resources that facilitate impulsive behaviors. Such a view is in line with neuropsychological literature in AUD highlighting the involvement of basic cognitive skills such as memory (Trivedi et al., 2013), processing speed (Sorg et al., 2015), and, more generally, executive functions (Bates et al., 2002; Glass et al., 2009; Le Berre et al., 2017). It is still unknown, however, whether impaired executive profile in AUD reflects a multimodal pattern of neurostructural damage transcending single MRI metrics. Preliminary attempts toward this goal have been pursued by distinct studies relating a global proxy of basic executive functioning (involving psychomotor speed, attention and working memory performances) to the degree of GM atrophy in meso-cortico-limbic structures (Galandra et al., 2018b), and altered functional connectivity in fronto-striatal-limbic networks (Galandra et al., 2019). In line with this evidence, other studies reported an association between attentional/executive deficits and glucose metabolism in the anterior cingulate cortex (ACC) in alcoholics (Goldstein et al., 2004). Altogether, the aforementioned findings consistently suggest that AUD patients' executive impairment might reflect anatomo-functional alterations involving the *salience network* (SN) (Galandra et al., 2018b, 2019). The latter, indeed, underpins the switch from automatic to controlled effortful processing, associated with the activity of the *default-mode network* (DMN) and *central executive network* (CEN), respectively, when relevant stimuli are detected (Smith et al., 2009; Menon and Uddin, 2010; Goulden et al., 2014).

On this ground, we aimed to integrate multimodal information on GM and WM features in AUD *via* a multivariate data-driven approach – joint Independent Component Analysis (jICA) – suitable to identify superordinate patterns at the *network level* (Calhoun, 2018). While this method has been successfully applied to other neurological and psychiatric conditions (Guo et al., 2012; Sui et al., 2013; Teipel et al., 2014; Kim et al., 2015), to the best of our knowledge no previous research has taken a comparable multivariate approach to investigate the ICs discriminating AUD patients from healthy controls (HC), and to assess their relationship with a superordinate proxy of impaired executive profile.

We expected to reconcile separate single-modality findings (Galandra et al., 2018b, 2019) into a unique consistent framework in which the ICs differentiating AUD patients from HC reflect neurostructural alterations of nodes and connections involving the salience network, with their mixing coefficient reflecting the degree of patients' executive impairment.

MATERIALS AND METHODS

Participants

Twenty-two AUD patients and 18 HC took part in the study, including a semi-structured interview about alcohol and nicotine use habits, a neuropsychological assessment, and a multimodal Magnetic Resonance Imaging (MRI) session. AUD patients were consecutively enrolled from the Functional Rehabilitation Unit of ICS Maugeri-Pavia (Italy), while HC were recruited *via* local advertisement. HC were matched for age and education to AUD patients, and groups were also balanced for gender (see **Table 1** for details).

Inclusion criteria for AUD subjects were age between 20 and 60 years and a diagnosis of alcohol dependence according to DSM-V criteria. We excluded HC in the presence and/or history of alcohol abuse. Exclusion criteria for both

TABLE 1 | Demographics and clinical information about substance habits.

	AUD (*n* = 22)	HC (*n* = 18)	*p*
Age (years; mean ± SD)	45.56 ± 7.99	45.11 ± 8.69	0.426
Education (years; mean ± SD)	9.91 ± 2.65	10.11 ± 2.78	0.405
Gender (m:f)	13:9	11:7	0.890
Smoking status (yes:no)	18/4	6/12	0.184
Duration of alcohol use (years; range, mean ± SD)	1–26 (10.11 ± 6.56)	—	—
Average daily alcohol dose (UA; range, mean ± SD)	5–32 (14.34 ± 6.66)	—	—

AUD, AUD patients; HC, healthy controls; Daily UA, units of alcohol.
The table reports demographic data related to age, education, gender and smoking status of AUD patients and healthy controls, alongside clinical information about alcohol use history and daily dose, and nicotine consumption, of AUD patients.

groups were the presence and/or history of neurological/psychiatric disorders other than AUDs, or any comorbid disorder except for smoking dependence, family history of neurological/psychiatric disorders, major medical disorders (e.g., kidney or liver diseases, severe diabetes and/or malnutrition), current use of any psychotropic substance/medication, past brain injury or loss of consciousness, inability to complete the neuropsychological assessment, and presence of contraindications to MRI.

AUD patients were enrolled after being detoxified for at least 10 days by means of medically supported standard treatments. However, they underwent MRI protocol only after at least 8 days without benzodiazepine treatment. HC were asked to be abstinent at least 10 days before the scanning day. We ascertained the abstinence of HC *via* a semi-structured interview about the consumption of alcoholic drinks covering that time. None of the participants received financial incentives to join the research protocol. Each enrolled subject had signed informed consent to the experimental protocol, which was approved by the Ethical Committee of ICS Maugeri (Pavia, Italy). The investigation was conducted in accordance with the latest version of the Declaration of Helsinki.

Clinical Interview, Neuropsychological Evaluation, and Data Analysis

AUD patients underwent a semi-structured interview conducted by an expert clinician about their drinking history, including the type, the amount, and lifetime duration of alcohol use. We used the average number of standard units of alcohol (UA) per day (1 UA = 330 ml beer, 125 ml wine, or 40 ml hard liquor = 12 g of ethanol) as a proxy of alcohol consumption (**Table 1**).

Neuropsychological assessment was performed using the Brief neuropsychological examination (ENB-2, Mondini, 2011), encompassing 15 tests assessing attention (trail making test, i.e., TMT-A and TMT-B), verbal short-term (digit span) and long-term memory (immediate and delayed prose memory), working memory (10″ and 30″ interference memory), executive functions (TMT-B, cognitive estimation, abstract reasoning, phonemic fluency, clock drawing, overlapping pictures), perceptive and praxis skills (praxis abilities, spontaneous drawing, copy drawing task). The ENB returns a global score, as well as different sub-scores for each task. The analysis of the resulting neuropsychological data has been previously described by Galandra and colleagues (Galandra et al., 2018b, 2019) and reported in **Supplementary Table S1**.

Magnetic Resonance Imaging Protocol and Data Acquisition

We use a 3 Tesla General Electrics Discovery MR750 scanner (GE Healthcare), equipped with a 16-channel phased array head coil, to run a multimodal MRI protocol including (1) a high-resolution 3D T1-weighted IR-prepared FSPGR (BRAVO) brain scan acquired along the AC-PC plane (152 slices, FOV = 24 cm, reconstruction matrix = 256 × 256, slice thickness = 1 mm); (2) a diffusion tensor imaging (DTI) scan based on a single-shot echo planar sequence (TR/TE = 8,986/80; FOV = 256 mm2; 56 sections; bandwidth = 250.0, 2 mm isotropic resolution), with diffusion gradients applied along 81 non-collinear directions (b = 1,000 s/mm^2), plus two non-diffusion weighted volumes. We also collected a T2-weighted image in order to explore any possible accidental diagnosis.

Voxel-Based Morphometry Data Pre-Processing

The pre-processing and statistical analysis of T1-weighted anatomical data were based on SPM12[1] and the CAT12 toolbox[2]. Pre-processing included bias-field inhomogeneities correction; spatial normalization using the DARTEL algorithm (Ashburner, 2007); and segmentation into GM, WM, and cerebrospinal fluid (CFS) (Ashburner and Friston, 2005). We did not apply the Jacobian modulation of segmented GM images, which corrects for volume change during spatial normalization, since this procedure has been shown to decrease the sensitivity to morphometric abnormalities (Radua et al., 2014). Our results thus involve GM density, i.e., GM volume relative to WM and CSF volume. Finally, a smoothing kernel of 8 mm (FWHM) was applied to the normalized segmented GM images. The resulting smoothed normalized GM images were fed into joint ICA.

DTI Data Pre-Processing

We performed the pre-processing of DTI data with the FMRIB Software Library tools (FSL; http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/). Single-subject datasets were first corrected for eddy current distortions and motion artifacts, skull-stripped, and finally, as a result of the fitting of the diffusion tensor model at each voxel, maps of diffusion scalar indices were generated. We then carried out DTI group analyses with Tract-Based Spatial Statistics (TBSS) (Smith et al., 2006), involving a voxelwise non-linear registration of all participants' fractional anisotropy (FA) maps that, once aligned, are affine-transformed on a standard space (1 mm × 1 mm × 1 mm MNI152). After co-registration, FA maps are averaged to create a mean FA image, and then used to generate a mean FA tract skeleton, representing all common tracts across subjects. We applied to the mean FA skeleton image a threshold of 0.20 to exclude from further analyses those parts of the skeleton that could not ensure a good correspondence across subjects. Finally, to account for residual misalignments after the initial non-linear registration, all subjects' FA data were projected onto the thresholded mean FA skeleton, creating a 4D dataset of all subjects' FA skeletonized data. In addition, we ran the non-FA TBSS script on maps of mean (MD), radial (RD), and axial (AD) diffusivities. The resulting skeletonized data were then fed into joint ICA.

Joint Independent Component Analysis

We used jICA – a multivariate approach integrating data from different neuroimaging modalities (i.e., features) unveiling

[1] http://www.fil.ion.ucl.ac.uk/spm

[2] http://www.neuro.uni-jena.de/cat/

covariance patterns of signal change across the brain – to estimate maximally independent components (ICs) at the topological level (i.e., spatial maps) for single features, which are then combined by a shared loading (mixing) parameter (Calhoun et al., 2006).

Shared information from GM density and DTI-invariant skeletonized (i.e., FA, AD, MD, RD) images were obtained using the Fusion ICA Toolbox (FIT v2.0c; http://icatb.sourceforge.net). Specific methodological details on this approach have been already described (Calhoun et al., 2006). Briefly, each participant's GM and DTI-invariant skeletonized images were first separately converted into a one-dimensional row vector. The initial data matrix was thus formed by arraying 22 GM, FA, AD, MD, and RD vectors of AUD patients and 18 GM, FA, AD, MD, and RD vectors of HC into a 40-row subjects-by-voxels matrix. Each feature dataset was then combined into a single data (participant × feature) matrix. All feature maps were normalized, resulting in the same average sum-of-square (computed across all voxels and subjects for each modality) and thus in equal data ranges. We used standard PCA to reduce the dimensionality of the data to 12 ICs, with this value being estimated for each feature using the minimum description length (MDL) criterion (Li et al., 2007). The Infomax algorithm (Bell and Sejnowski, 1995) was then used to decompose the reduced feature-matrix to maximally independent component images and subject-specific mixing (loading) coefficients. This jICA approach was repeated 50 times in Icasso[3]. The resulting 12 ICs were clustered to ensure the consistency and reliability of the decomposition, which are quantified using a quality index (QI) ranging from 0 to 1, reflecting the difference between intra-cluster and extra-cluster similarity (Himberg et al., 2004).

Mixing coefficients, reflecting how strongly each participant contributes to the relationship described in a given IC, were fed into a two-sample *t*-test assessing a significant difference between AUD patients and HC. Only significant components reflecting neurostructural changes in AUD patients (i.e., mixing coefficients AUD > mixing coefficients HC) were considered in subsequent analyses.

On this basis, we aimed to investigate whether ICs differentiating patients from HC additionally confirmed the involvement of *salience network* regions as neurostructural markers of the neuro-cognitive impairment associated with AUD (Galandra et al., 2018b, 2019). To this purpose, we finally correlated mixing coefficients with (1) a measure of each participant's executive profile, obtained *via* a multivariate data reduction approach (see Galandra et al., 2018b for detailed information about the statistical procedure) and (2) clinical variables (i.e., lifetime duration of alcohol abuse and daily alcohol consumption) in AUD patients.

The anatomical localization of significant clusters was performed with the JHU White-Matter Tractography Atlas and the JHU ICBM-DTI-81 White-Matter Labels (Wakana et al., 2007; Hua et al., 2008) for DTI features, while the SPM Anatomy toolbox (Eickhoff et al., 2005) was used to localize gray-matter features.

RESULTS

Joint Components of Neurostructural Change in Alcohol Use Disorder Patients Vs. Healthy Controls

We found significantly different mixing coefficients, in AUD patients vs. HC, in three out of 12 joint ICs (IC06, IC08, IC11), with $p = 0.004$ as adjusted significance threshold applied to control for multiple comparisons (Bonferroni correction). All these ICs were associated with a quality index >0.95, indicating a highly stable ICA decomposition. While mixing coefficients of IC06 and IC11 were higher in patients compared with controls [IC06: $t(38) = -5.69$, $p < 0.001$; IC11: $t(38) = -3.82$, $p < 0.001$], IC08 displayed the opposite pattern [IC08: $t(38) = 5.17$, $p < 0.001$]. IC06 involved a widespread GM pattern encompassing the sensorimotor cortex and supplementary motor area, cingulate cortex, and precuneus, subcortical nuclei (bilateral thalamus, left caudate), plus an extensive sector of the occipital cortex (calcarine cortex, cuneus, lingual gyrus), and bilateral cerebellum (crus II). The distribution of DTI indices for this component involved commissural (body of corpus callosum and forceps major), projections (bilateral anterior thalamic radiation and superior corona radiata, as well as the right cerebral peduncle), and associative (fornix plus stria terminalis, as well as the posterior sectors of bilateral superior and inferior longitudinal fasciculi, inferior fronto-occipital fasciculus and cingulum bundle) fibers (**Figure 1**). IC08 was represented by a subcortical GM pattern including bilateral amygdala and left hippocampus, and by a widespread DTI pattern primarily involving all sectors of corpus callosum (genu, body, and splenium), forceps minor and forceps major, anterior thalamic radiations and cerebral peduncles, the anterior limb of internal capsule and the fornix (body and column) (**Figure 2**). IC11 involved bilaterally the middle frontal gyrus, insula, anterior and posterior sectors of the cingulate cortex, distinct sectors of the temporal (superior and middle temporal gyri, supramarginal gyrus), parietal (precuneus, angular gyrus) and occipital (lingual and fusiform gyri) lobes, plus the left hippocampus and the cerebellum (crus I). Here, the overall DTI pattern encompassed commissural (genu and body of corpus callosum, forceps major and forceps minor), projection (anterior limb of internal capsule and thalamic radiations, as well as superior corona radiata), and associative (both fornix body and column, along with superior and inferior longitudinal fasciculi, inferior fronto-occipital fasciculus and cingulum bundle, with a right hemispheric prevalence) fibers (**Figure 3**). Detailed information about localization of significant ICs is reported in **Supplementary Tables S2–S4**.

Relationship Between Independent Components and Executive/Clinical Variables

Among the three ICs differentiating AUD patients from HC, only IC6 and IC11 mixing coefficients presented a difference pattern (i.e., AUD > HC) suggesting a stronger contribution

[3] http://research.ics.aalto.fi/ica/icasso/

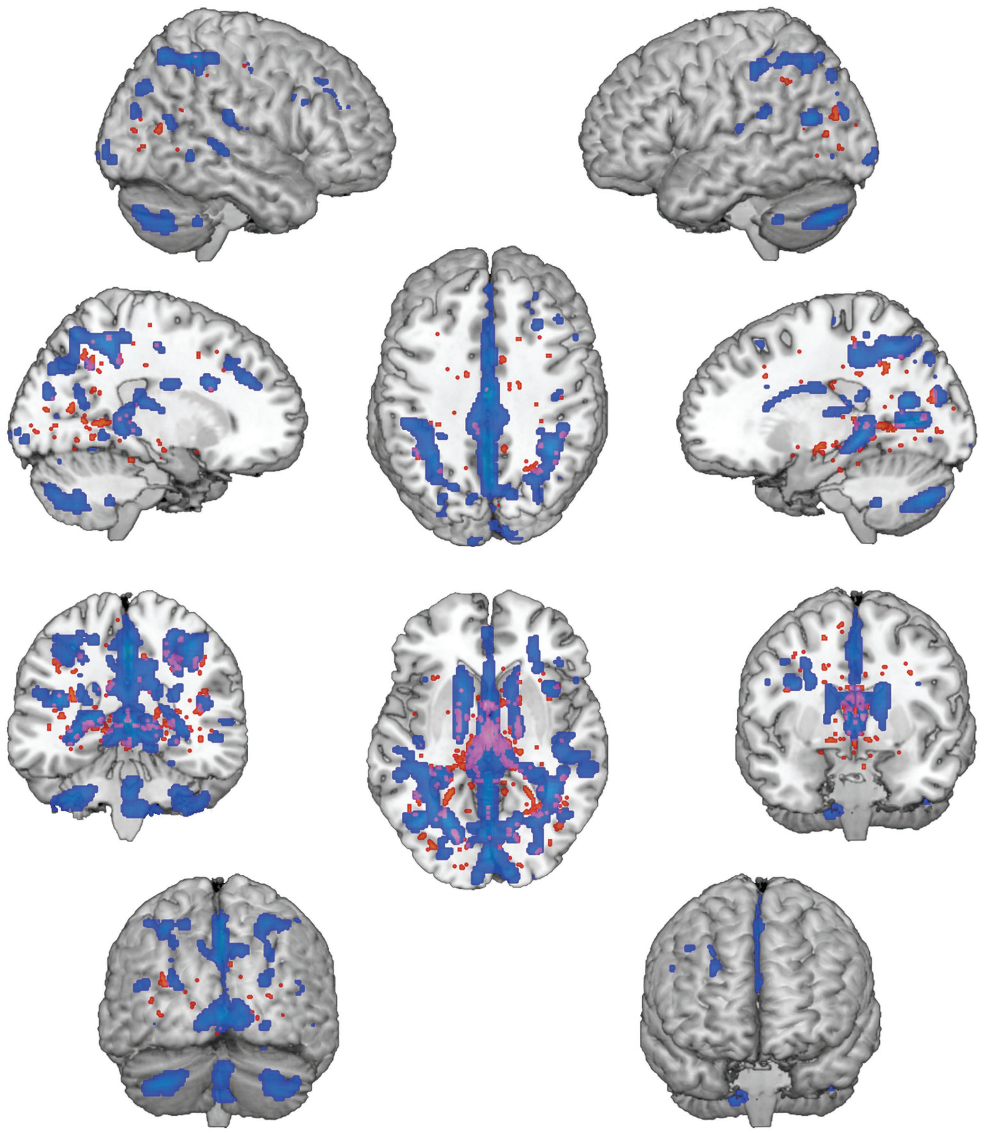


FIGURE 1 | IC06 Pattern. The figure depicts the covariance pattern of IC06 emerging from joint ICA. Statistical maps are thresholded at $z = 2.5$ for visualization purposes. Gray-matter clusters and white-matter clusters including all DTI metrics (i.e., FA, MD, AD, RD) are shown in blue-green and red-yellow colors, respectively.

of signal covariance in patients than HC, and thus reflecting a possible neurostructural alteration characterizing AUD. However, the IC11, but not IC06 mixing coefficients ($r = -0.23$, $p = 0.159$), were significantly correlated with participants' executive profile ($r = -0.54$, $p < 0.001$).

Correlation analyses in the patient group highlighted a significant positive correlation between lifetime duration of alcohol abuse and IC11 ($r = 0.51$, $p = 0.016$), but not with IC6 mixing coefficients ($r = 0.37$, $p = 0.09$). No significant correlations with daily alcohol consumption was observed neither in IC6 ($r = -0.002$, $p = 0.994$) nor in IC11 mixing coefficients ($r = -0.05$, $p = 0.499$). Scatterplots of all correlations are reported in **Supplementary Figures S1–S6**.

DISCUSSION

We used jICA to investigate supramodal patterns of covariance (ICs) reflecting shared information across several neurostructural features including GM density and distinct WM microstructural properties. We then explored the relationship between the ICs discriminating AUD patients from HC and the overall executive profile highlighted by a multivariate analysis of performance in several neuropsychological tasks (Galandra et al., 2018b). This approach aimed to investigate the connection between a superordinate proxy of AUD patients' cognitive impairment transcending single tasks, and spatial maps integrating multimodal MRI information on the underlying neuro-anatomical alterations.

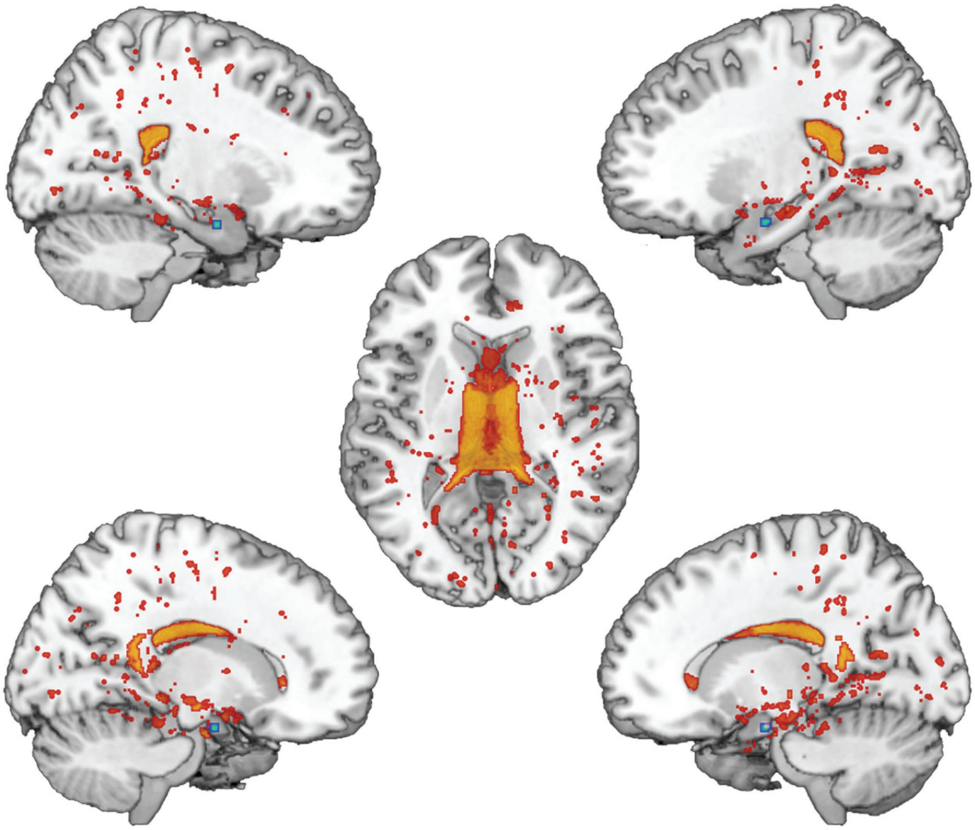


FIGURE 2 | IC08 Pattern. The figure depicts the covariance pattern of IC08 resulting from joint ICA. Statistical maps are thresholded at $z = 2.5$ for visualization purposes. Gray-matter clusters and white-matter clusters including all DTI metrics (i.e., FA, MD, AD, RD) are shown in blue-green and red-yellow colors, respectively.

As far as we know, this is the first research *combining* two parallel data-driven multivariate analyses of neurostructural and behavioral data, and their relationship, in AUD. The resulting evidence confirmed previous reports, from univariate analyses of single MRI modalities, of a defective interplay, in AUD, between large-scale brain networks underlying the salience-based switch from automatic to controlled cognition and behavior (e.g., Galandra et al., 2018b, 2019).

First, three out of the 12 extracted ICs (IC06, IC08, IC11) differentiated AUD patients from HC. The mixing coefficients of IC06 and IC11 were significantly higher in AUD patients compared to controls, while we found the opposite pattern for IC08. Higher mixing coefficients in a given IC are suggestive of a greater contribution to the original features by its constituting regions (Calhoun et al., 2006; Kim et al., 2015). The present evidence seems thus to indicate that the regions included in IC06 and IC11 are more tightly related to neural changes associated with chronic alcohol consumption than IC08. Importantly, IC11 mixing coefficients were also significantly related to executive performance in the whole sample, and to lifetime duration of alcohol abuse in AUD patients.

This component encompassed a set of GM regions including the insula bilaterally and both the anterior (ACC) and posterior (PCC) cingulate cortex, alongside commissural (corpus callosum) and major associative fiber tracts (superior and inferior longitudinal fasciculi, inferior fronto-occipital fasciculus and cingulum bundle). The fronto-insular cortex and the anterior cingulate cortex (ACC) are two interconnected key components of the *salience network* (SN), typically co-activated by behaviorally relevant stimuli (Seeley et al., 2007; Goulden et al., 2014). Both these regions are connected with sensory and motor areas, and their activation is considered to underpin the switch between *default-mode* and *central executive networks* (Goulden et al., 2014). Such a general-purpose function fits with the insula role as a site of multimodal convergence of signals concerning sensory and affective processing (Uddin et al., 2014), likely supporting salience-related top-down mechanisms such as impulse control and self-regulation (see Sullivan et al., 2013). Therefore, it is not surprising to observe neurostructural alterations in a portion of the insula previously associated with abnormal network efficiency and functional connectivity (Wang et al., 2018), hypo-connectivity with precuneus, SMA, postcentral, lingual/vermis, and fusiform gyri (Vergara et al., 2017), and a perfusion deficit (Sullivan et al., 2013), in AUD patients. The possible relationship between such insular dysfunction and a defective interplay among salience, default mode, and executive control networks in AUD (Sullivan et al., 2013) is indirectly supported by the

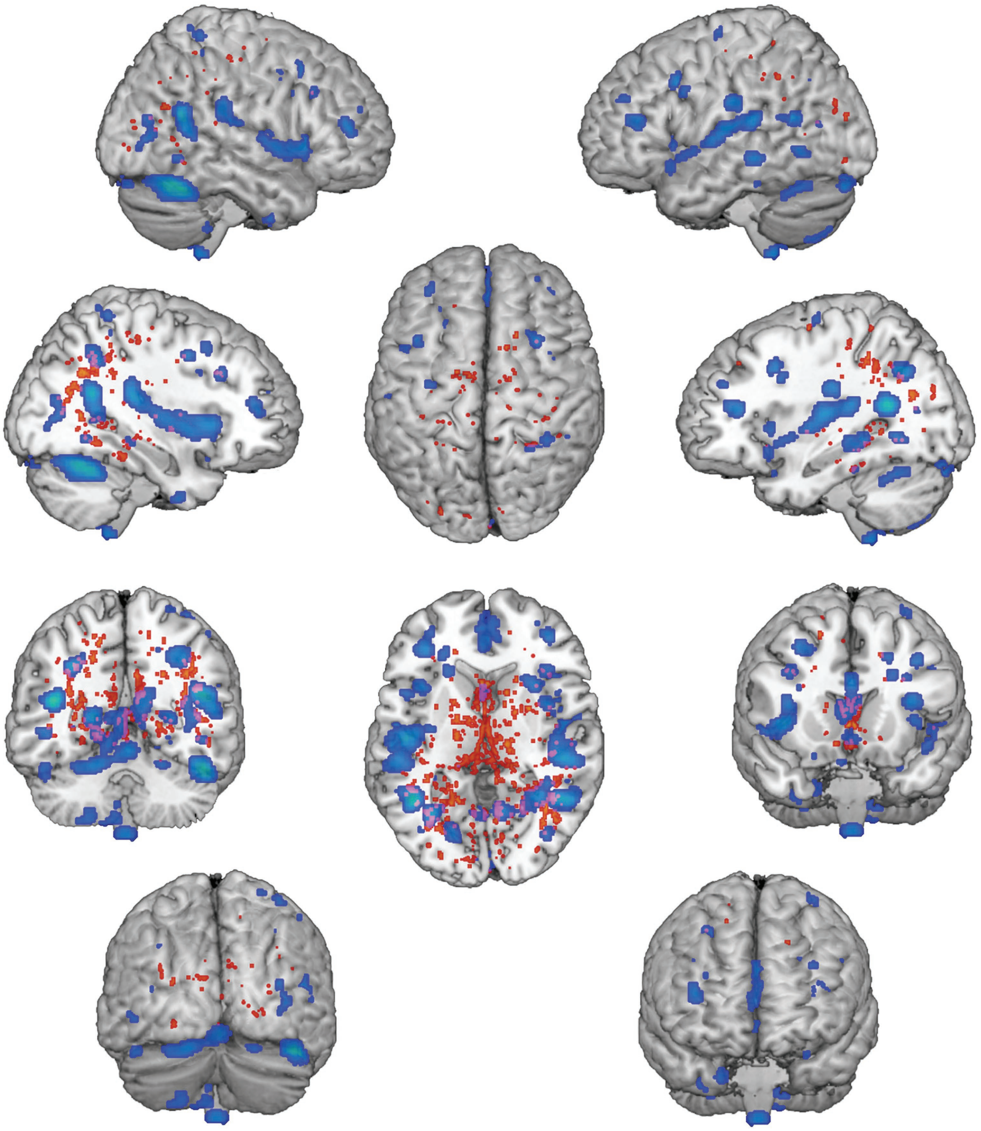


FIGURE 3 | IC11 Pattern. The figure depicts the covariance pattern of IC11 highlighted by joint ICA. Statistical maps are thresholded at $z = 2.5$ for visualization purposes. Gray-matter clusters and white-matter clusters including all DTI metrics (i.e., FA, MD, AD, RD) are shown in blue-green and red-yellow colors, respectively.

present evidence of co-occurrent structural and functional alterations in the ACC (Müller-Oehring et al., 2015; Galandra et al., 2018b). The latter is a key node of the reward pathway (Haber and Knutson, 2010), in which neural mechanisms of performance monitoring signal to the fronto-parietal executive network the need of behavioral adjustments (Ridderinkhof et al., 2004). Thus the ACC is a well-suited region to promote salience-based behavioral adaptations, which require to switch from default to controlled processing. Therefore, the impairment of both these networks in AUD (Chanraud et al., 2011; Sullivan et al., 2013; Galandra et al., 2018b, 2019) fits with previous evidence of a connection between altered ACC activity and craving, likely promoting relapses (Koob and Le Moal, 2008). Further evidence of an altered interplay between salience and default mode networks in AUD is represented by IC11 including the posterior cingulate cortex, a key node of default mode network (Greicius et al., 2009) in which decreased coherence of the spontaneous BOLD fluctuations has long been known as a neural marker of impaired functional connectivity (Chanraud et al., 2011).

Several evidences support the relationship between such alterations and impaired executive functioning. On the one hand, GM atrophy in the insular and anterior cingulate cortex predicts executive deficits, mainly involving attention and working-memory, in AUD patients (Galandra et al., 2018b). Moreover, abstinence seems to reverse alcohol-related morphological alterations in these regions (Fritz et al., 2019) and restore connectivity within and between the salience and executive networks (Kohno et al., 2017), with these changes paralleling an improvement of executive skills (Le Berre et al., 2017).

While the present findings support previous data on the relationship between AUD patients' executive impairment and GM nodes within the salience network, our analytic approach allowed to extend this evidence to WM connections. In particular, the corpus callosum (genu) and the cingulum bundle included in IC11 connect the key nodes of large-scale functional networks (Van Den Heuvel et al., 2009) in which functional alterations have been ascribed both to GM loss in crucial nodes, and to macro- and/or micro-structural impairments in WM tracts connecting them (Peer et al., 2017). It is noteworthy that the genual fibers interconnect homologous prefrontal regions such as the dorsolateral prefrontal cortex (DLPFC) (Voineskos et al., 2010) – an important hub of the central executive network (e.g., Seeley et al., 2007; Chen et al., 2013; Marstaller et al., 2015) – and the ACC (van der Knaap and van der Ham, 2011), and that both the genu microstructure and the DLPFC function have been associated to executive performance (Zahr et al., 2009). All these data converge to suggest that the well-established damage of genual fibers in AUD (Pfefferbaum et al., 2006), *via* DLPFC dysfunction, can decrease the efficiency of the central executive network.

A limitation of our work concerns the small-to-moderate sample size, due to the strictness of inclusion criteria and the accurate control of possible nuisance variables, which highlights the need of confirmatory studies before strong conclusions can be drawn on the multimodal neural bases of executive deficits in AUD. Moreover, the lack of information about brain activity limits our conclusions to the neurostructural level, thus preventing inferences about possible consequences of structural damage in terms of impaired functional connectivity and/or compensatory mechanisms. Future studies might fill these gaps by addressing a more comprehensive view integrating functional and structural connectivity measures in larger samples. Importantly, however, the present results pave the way for such further investigations by starting to unveil the relationship among cognitive impairment in AUD and the topographic properties of multimodal "neurostructural" ICs differentiating AUD patients from HC. Building on the present evidence, longitudinal studies might also benefit from the application of multivariate analytic approaches to explore multimodal changes and their association with cognitive status, in relation to abstinence and relapses, or as a result of rehabilitative interventions.

In conclusion, the present findings confirm and integrate into a coherent framework previously scattered evidence about the involvement of key nodes of salience, default mode, and central executive networks, and their structural connections, as reliable neuroimaging markers of executive impairment in AUD. The relevance of IC11 in discriminating AUD patients from HC and its uniqueness in synthesizing different facets of the neurostructural damage in AUD are further supported by the positive relationship between its associated GM and WM patterns and lifetime duration of alcohol abuse. From the methodological standpoint, the present data confirm the consistency between the outputs of multivariate jICA and standard univariate analyses (see Galandra et al., 2018b), thus supporting the notion that jICA can capture the complexity of the neurostructural impairment in AUD based on unique coefficients expressing covariance patterns of morphometric GM and microstructural WM data.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of IRCCS ICS Maugeri, 27100, Pavia, Italy. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CC, CG, and NC participated in study design and conceptualization and in manuscript drafting and revising. CG, MM, GB, PP, and NC collected data. CC and CG performed data analysis and interpretation. CC, CG, MM, GB, PP, and NC approved the final version of the manuscript.

ACKNOWLEDGMENTS

We wish to thank Dr. Giovanni Vittadini for his contribution in patient recruitment and clinical assessment.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02479/full#supplementary-material

REFERENCES

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage* 38, 95–113. doi: 10.1016/j.neuroimage.2007.07.007

Ashburner, J., and Friston, K. J. (2005). Unified segmentation. *NeuroImage* 26, 839–851. doi: 10.1016/j.neuroimage.2005.02.018

Bates, M. E., Bowden, S. C., and Barry, D. (2002). Neurocognitive impairment associated with alcohol use disorders: implications for treatment. *Exp. Clin. Psychopharmacol.* 10, 193–212. doi: 10.1037/1064-1297.10.3.193

Bechara, A., and Damasio, A. R. (2005). The somatic marker hypothesis: a neural theory of economic decision. *Game Econ. Behav.* 52, 336–372. doi: 10.1016/j.geb.2004.06.010

Bell, A. J., and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* 7, 1129–1159. doi: 10.1162/neco.1995.7.6.1129

Bühler, M., and Mann, K. (2011). Alcohol and the human brain: a systematic review of different neuroimaging methods. *Alcohol. Clin. Exp. Res.* 35, 1771–1793. doi: 10.1111/j.1530-0277.2011.01540.x

Calhoun, V. (2018). Data-driven approaches for identifying links between brain structure and function in health and disease. *Dialogues Clin. Neurosci.* 20, 87–99.

Calhoun, V. D., Adali, T., Giuliani, N. R., Pekar, J. J., Kiehl, K. A., and Pearlson, G. D. (2006). Method for multimodal analysis of independent source differences in schizophrenia: combining gray matter structural and auditory oddball functional data. *Hum. Brain Mapp.* 27, 47–62. doi: 10.1002/hbm.20166

Center for Behavioral Health Statistics and Quality (2016). Key substance use and mental health indicators in the United States: Results from the 2015 National Survey on Drug Use and Health (HHS Publication No. SMA 16-4984, NSDUH Series H-51). Available at: http://www.samhsa.gov/data/

Chanraud, S., Pitel, A. L., Pfefferbaum, A., and Sullivan, E. V. (2011). Disruption of functional connectivity of the default-mode network in alcoholism. *Cereb. Cortex* 21, 2272–2281. doi: 10.1093/cercor/bhq297

Chen, A. C., Oathes, D. J., Chang, C., Bradley, T., Zhou, Z.-W., Williams, L. M., et al. (2013). Causal interactions between fronto-parietal central executive and default-mode networks in humans. *Proc. Natl. Acad. Sci. USA* 110, 19944–19949. doi: 10.1073/pnas.1311772110

Chumin, E. J., Grecco, G. G., Dzemidzic, M., Cheng, H., Finn, P., Sporns, O., et al. (2019). Alterations in white matter microstructure and connectivity in young adults with alcohol use disorder. *Alcohol. Clin. Exp. Res.* 43, 1170–1179. doi: 10.1111/acer.14048

De La Monte, S. M., and Kril, J. J. (2014). Human alcohol-related neuropathology. *Acta Neuropathol.* 127, 71–90. doi: 10.1007/s00401-013-1233-3

Eickhoff, S., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., et al. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage* 25, 1325–1335. doi: 10.1016/j.neuroimage.2004.12.034

Fortier, C. B., Leritz, E. C., Salat, D. H., Lindemer, E., Maksimovskiy, A. L., Shepel, J., et al. (2014). Widespread effects of alcohol on white matter microstructure. *Alcohol. Clin. Exp. Res.* 38, 2925–2933. doi: 10.1111/acer.12568

Fritz, M., Klawonn, A. M., and Zahr, N. M. (2019). Neuroimaging in alcohol use disorder: from mouse to man. *J. Neurosci. Res.* doi: 10.1002/jnr.24423 [Epub ahead of print].

Galandra, C., Basso, G., Cappa, S., and Canessa, N. (2018a). The alcoholic brain: neural bases of impaired reward-based decision-making in alcohol use disorders. *Neurol. Sci.* 39, 423–435. doi: 10.1007/s10072-017-3205-1

Galandra, C., Basso, G., Manera, M., Crespi, C., Giorgi, I., Vittadini, G., et al. (2018b). Salience network structural integrity predicts executive impairment in alcohol use disorders. *Sci. Rep.* 8:14481. doi: 10.1038/s41598-018-32828-x

Galandra, C., Basso, G., Manera, M., Crespi, C., Giorgi, I., Vittadini, G., et al. (2019). Abnormal fronto-striatal intrinsic connectivity reflects executive dysfunction in alcohol use disorders. *Cortex* 115, 27–42. doi: 10.1016/j.cortex.2019.01.004

Glass, J. M., Buu, A., Adams, K. M., Nigg, J. T., Puttler, L. I., Jester, J. M., et al. (2009). Effects of alcoholism severity and smoking on executive neurocognitive function. *Addiction* 104, 38–48. doi: 10.1111/j.1360-0443.2008.02415.x

Goldstein, R. Z., Leskovjan, A. C., Hoff, A. L., Hitzemann, R., Bashan, F., Khalsa, S. S., et al. (2004). Severity of neuropsychological impairment in cocaine and alcohol addiction: association with metabolism in the prefrontal cortex. *Neuropsychologia* 42, 1447–1458. doi: 10.1016/j.neuropsychologia.2004.04.002

Goulden, N., Khusnulina, A., Davis, N. J., Bracewell, R. M., Bokde, A. L., McNulty, J. P., et al. (2014). The salience network is responsible for switching between the default mode network and the central executive network: replication from DCM. *NeuroImage* 99, 180–190. doi: 10.1016/j.neuroimage.2014.05.052

Greicius, M. D., Supekar, K., Menon, V., and Dougherty, R. F. (2009). Resting-state functional connectivity reflects structural connectivity in the default mode network. *Cereb. Cortex* 19, 72–78. doi: 10.1093/cercor/bhn059

Guo, X., Han, Y., Chen, K., Wang, Y., and Yao, L. (2012). Mapping joint grey and white matter reductions in Alzheimer's disease using joint independent component analysis. *Neurosci. Lett.* 531, 136–141. doi: 10.1016/j.neulet.2012.10.038

Haber, S. N., and Knutson, B. (2010). The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology* 35, 4–26. doi: 10.1038/npp.2009.129

Himberg, J., Hyvärinen, A., and Esposito, F. (2004). Validating the independent components of neuroimaging time series via clustering and visualization. *NeuroImage* 22, 1214–1222. doi: 10.1016/j.neuroimage.2004.03.027

Hua, K., Zhang, J., Wakana, S., Jiang, H., Li, X., Reich, D. S., et al. (2008). Tract probability maps in stereotaxic spaces: analyses of white matter anatomy and tract-specific quantification. *NeuroImage* 39, 336–347. doi: 10.1016/j.neuroimage.2007.07.053

Jansen, J. M., Van Holst, R. J., Van Den Brink, W., Veltman, D. J., Caan, M. W. A., and Goudriaan, A. E. (2015). Brain function during cognitive flexibility and white matter integrity in alcohol-dependent patients, problematic drinkers and healthy controls. *Addict. Biol.* 20, 979–989. doi: 10.1111/adb.12199

Kim, S. G., Jung, W. H., Kim, S. N., Jang, J. H., and Kwon, J. S. (2015). Alterations of gray and white matter networks in patients with obsessive-compulsive disorder: a multimodal fusion analysis of structural MRI and DTI using mCCA+jICA. *PLoS One* 10:e0127118. doi: 10.1371/journal.pone.0127118

Kohno, M., Dennis, L. E., McCready, H., and Hoffman, W. F. (2017). Executive control and striatal resting-state network interact with risk factors to influence treatment outcomes in alcohol-use disorder. *Front. Psych.* 8:182. doi: 10.3389/fpsyt.2017.00182

Koob, G. F. (2013). Addiction is a reward deficit and stress surfeit disorder. *Front. Psych.* 4:72. doi: 10.3389/fpsyt.2013.00072

Koob, G. F., and Le Moal, M. (2008). Addiction and the brain antireward system. *Annu. Rev. Psychol.* 59, 29–53. doi: 10.1146/annurev.psych.59.103006.093548

Le Berre, A. P., Fama, R., and Sullivan, E. V. (2017). Executive functions, memory, and social cognitive deficits and recovery in chronic alcoholism: a critical review to inform future research. *Alcohol. Clin. Exp. Res.* 41, 1432–1443. doi: 10.1111/acer.13431

Li, Y. O., Adali, T., and Calhoun, V. D. (2007). Estimating the number of independent components for functional magnetic resonance imaging data. *Hum. Brain Mapp.* 28, 1251–1266. doi: 10.1002/hbm.20359

Marstaller, L., Williams, M., Rich, A., Savage, G., and Burianová, H. (2015). Aging and large-scale functional networks: white matter integrity, gray matter volume, and functional connectivity in the resting state. *Neuroscience* 290, 369–378. doi: 10.1016/j.neuroscience.2015.01.049

Menon, V., and Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Struct. Funct.* 214, 655–667. doi: 10.1007/s00429-010-0262-0

Mondini, S. (2011). *Esame neuropsicologico breve 2: Una batteria di test per lo screening neuropsicologico*. Cortina.

Müller-Oehring, E. M., Jung, Y. C., Pfefferbaum, A., Sullivan, E. V., and Schulte, T. (2015). The resting brain of alcoholics. *Cereb. Cortex* 25, 4155–4168. doi: 10.1093/cercor/bhu134

Peer, M., Nitzan, M., Bick, A. S., Levin, N., and Arzy, S. (2017). Evidence for functional networks within the human brain's white matter. *J. Neurosci.* 37, 6394–6407. doi: 10.1523/JNEUROSCI.3872-16.2017

Pfefferbaum, A., Adalsteinsson, E., and Sullivan, E. V. (2006). Dysmorphology and microstructural degradation of the corpus callosum: interaction of age and alcoholism. *Neurobiol. Aging* 27, 994–1009. doi: 10.1016/j.neurobiolaging.2005.05.007

Radua, J., Canales-Rodríguez, E. J., Pomarol-Clotet, E., and Salvador, R. (2014). Validity of modulation and optimal settings for advanced voxel-based morphometry. *NeuroImage* 86, 81–90. doi: 10.1016/j.neuroimage.2013.07.084

Ridderinkhof, K. R., Van Den Wildenberg, W. P. M., Segalowitz, S. J., and Carter, C. S. (2004). Neurocognitive mechanisms of cognitive control: the role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning. *Brain Cogn.* 56, 129–140. doi: 10.1016/j.bandc.2004.09.016

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., et al. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *J. Neurosci.* 27, 2349–2356. doi: 10.1523/JNEUROSCI.5587-06.2007

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., et al. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proc. Natl. Acad. Sci. USA* 106, 13040–13045. doi: 10.1073/pnas.0905267106

Smith, S. M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T. E., Mackay, C. E., et al. (2006). Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *NeuroImage* 31, 1487–1505. doi: 10.1016/j.neuroimage.2006.02.024

Sorg, S. F., Squeglia, L. M., Taylor, M. J., Alhassoon, O. M., Delano-Wood, L. M., and Grant, I. (2015). Effects of aging on frontal white matter microstructure in alcohol use disorder and associations with processing speed. *J. Stud. Alcohol Drugs* 76, 296–306. doi: 10.15288/jsad.2015.76.296

Suckling, J., and Nestor, L. J. (2017). The neurobiology of addiction: the perspective from magnetic resonance imaging present and future. *Addiction* 112, 360–369. doi: 10.1111/add.13474

Sui, J., He, H., Pearlson, G. D., Adali, T., Kiehl, K. A., Yu, Q., et al. (2013). Three-way (N-way) fusion of brain imaging data based on mCCA+jICA and its application to discriminating schizophrenia. *NeuroImage* 66, 119–132. doi: 10.1016/j.neuroimage.2012.10.051

Sullivan, E. V., Müller-Oehring, E., Pitel, A. L., Chanraud, S., Shankaranarayanan, A., Alsop, D. C., et al. (2013). A selective insular perfusion deficit contributes to compromised salience network connectivity in recovering alcoholic men. *Biol. Psychiatry* 74, 547–555. doi: 10.1016/j.biopsych.2013.02.026

Teipel, S. J., Grothe, M. J., Filippi, M., Fellgiebel, A., Dyrba, M., Frisoni, G. B., et al. (2014). Fractional anisotropy changes in Alzheimer's disease depend on the underlying fiber tract architecture: a multiparametric DTI study using joint independent component analysis. *J. Alzheimers Dis.* 41, 69–83. doi: 10.3233/JAD-131829

Trivedi, R., Bagga, D., Bhattacharya, D., Kaur, P., Kumar, P., Khushu, S., et al. (2013). White matter damage is associated with memory decline in chronic alcoholics: a quantitative diffusion tensor tractography study. *Behav. Brain Res.* 250, 192–198. doi: 10.1016/j.bbr.2013.05.001

Uddin, L. Q., Kinnison, J., Pessoa, L., and Anderson, M. L. (2014). Beyond the tripartite cognition-emotion-interoception model of the human insular cortex. *J. Cogn. Neurosci.* 26, 16–27. doi: 10.1162/jocn_a_00462

Van Den Heuvel, M. P., Mandl, R. C. W., Kahn, R. S., and Hulshoff Pol, H. E. (2009). Functionally linked resting-state networks reflect the underlying structural connectivity architecture of the human brain. *Hum. Brain Mapp.* 30, 3127–3141. doi: 10.1002/hbm.20737

van der Knaap, L. J., and van der Ham, I. J. M. (2011). How does the corpus callosum mediate interhemispheric transfer? *A review. Behav. Brain Res.* 30, 3127–3141. doi: 10.1016/j.bbr.2011.04.018

Vergara, V. M., Liu, J., Claus, E. D., Hutchison, K., and Calhoun, V. (2017). Alterations of resting state functional network connectivity in the brain of nicotine and alcohol users. *NeuroImage* 151, 45–54. doi: 10.1016/j.neuroimage.2016.11.012

Voineskos, A. N., Farzan, F., Barr, M. S., Lobaugh, N. J., Mulsant, B. H., Chen, R., et al. (2010). The role of the corpus callosum in transcranial magnetic stimulation induced interhemispheric signal propagation. *Biol. Psychiatry* 68, 825–831. doi: 10.1016/j.biopsych.2010.06.021

Wakana, S., Caprihan, A., Panzenboeck, M. M., Fallon, J. H., Perry, M., Gollub, R. L., et al. (2007). Reproducibility of quantitative tractography methods applied to cerebral white matter. *NeuroImage* 36, 630–644. doi: 10.1016/j.neuroimage.2007.02.049

Wang, C., Wu, H., Chen, F., Xu, J., Li, H., Li, H., et al. (2018). Disrupted functional connectivity patterns of the insula subregions in drug-free major depressive disorder. *J. Affect. Disord.* 234, 297–304. doi: 10.1016/j.jad.2017.12.033

Xiao, P. R., Dai, Z. Y., Zhong, J. G., Zhu, Y. L., Shi, H. C., and Pan, P. L. (2015). Regional gray matter deficits in alcohol dependence: a meta-analysis of voxel-based morphometry studies. *Drug Alcohol Depend.* 153, 22–28. doi: 10.1016/j.drugalcdep.2015.05.030

Yang, X., Tian, F., Zhang, H., Zeng, J., Chen, T., Wang, S., et al. (2016). Cortical and subcortical gray matter shrinkage in alcohol-use disorders: a voxel-based meta-analysis. *Neurosci. Biobehav. Rev.* 66, 92–103. doi: 10.1016/j.neubiorev.2016.03.034

Zahr, N. M., and Pfefferbaum, A. (2017). Alcohol's effects on the brain: neuroimaging results in humans and animal models. *Alcohol Res.* 38, 183–206.

Zahr, N. M., Rohlfing, T., Pfefferbaum, A., and Sullivan, E. V. (2009). Problem solving, working memory, and motor correlates of association and commissural fiber bundles in normal aging: a quantitative fiber tracking study. *NeuroImage* 44, 1050–1062. doi: 10.1016/j.neuroimage.2008.09.046

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 04 December 2019
doi: 10.3389/fpsyg.2019.02714

A Comparison of Classical and Modern Measures of Internal Consistency

Pasquale Anselmi, Daiana Colledani and Egidio Robusto*

Department of Philosophy, Sociology, Education and Applied Psychology, University of Padua, Padua, Italy

Three measures of internal consistency – Kuder-Richardson Formula 20 (KR20), Cronbach's alpha (α), and person separation reliability (R) – are considered. KR20 and α are common measures in classical test theory, whereas R is developed in modern test theory and, more precisely, in Rasch measurement. These three measures specify the observed variance as the sum of true variance and error variance. However, they differ for the way in which these quantities are obtained. KR20 uses the error variance of an "average" respondent from the sample, which overestimates the error variance of respondents with high or low scores. Conversely, R uses the actual average error variance of the sample. KR20 and α use respondents' test scores in calculating the observed variance. This is potentially misleading because test scores are not linear representations of the underlying variable, whereas calculation of variance requires linearity. Contrariwise, if the data fit the Rasch model, the measures estimated for each respondent are on a linear scale, thus being numerically suitable for calculating the observed variance. Given these differences, R is expected to be a better index of internal consistency than KR20 and α. The present work compares the three measures on simulated data sets with dichotomous and polytomous items. It is shown that all the estimates of internal consistency decrease with the increasing of the skewness of the score distribution, with R decreasing to a larger extent. Thus, R is more conservative than KR20 and α, and prevents test users from believing a test has better measurement characteristics than it actually has. In addition, it is shown that Rasch-based infit and outfit person statistics can be used for handling data sets with random responses. Two options are described. The first one implies computing a more conservative estimate of internal consistency. The second one implies detecting individuals with random responses. When there are a few individuals with a consistent number of random responses, infit and outfit allow for correctly detecting almost all of them. Once these individuals are removed, a "cleaned" data set is obtained that can be used for computing a less biased estimate of internal consistency.

Keywords: internal consistency, reliability, Rasch models, modern test theory, classical test theory, infit, outfit

Edited by:
Pietro Cipresso,
Italian Auxological Institute (IRCCS),
Italy

Reviewed by:
Oscar Lorenzo Olvera Astivia,
University of South Florida, Tampa,
United States
Chester Chun Seng Kam,
University of Macau, China

Correspondence:
Pasquale Anselmi
pasquale.anselmi@unipd.it

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *20 August 2019*
Accepted: *18 November 2019*
Published: *04 December 2019*

Citation:
Anselmi P, Colledani D and Robusto E (2019) A Comparison of Classical and Modern Measures of Internal Consistency. Front. Psychol. 10:2714. doi: 10.3389/fpsyg.2019.02714

INTRODUCTION

The present work deals with internal consistency, which expresses the degree to which the items of a test produce similar scores. Three measures of internal consistency are considered, namely Kuder-Richardson Formula 20 (KR20; Kuder and Richardson, 1937), Cronbach's α (Cronbach, 1951), and person separation reliability (R; Wright and Masters, 1982).

KR20 and α are well-known measures in classical test theory, where they are widely used to evaluate the internal consistency of cognitive and personality tests. The derivations of KR20 and α used continuous random variables for item scores (Sijtsma, 2009). As such, they include dichotomous scoring (e.g., correct/incorrect; yes/no) and ordered polytomous scoring (e.g., never/sometimes/often/always; very difficult/difficult/easy/very easy) as special cases. The formula for the computation of KR20 is suitable for items with dichotomous scores, whereas the formula for the computation of α is suitable for items with dichotomous scores and items with polytomous scores. When all items are scored 1 or 0, the formula for KR20 reduces to that for α (Cronbach, 1951).

Less known than KR20 and α, R develops within modern test theory and, more precisely, within Rasch models. There are several applications of these models to the development and validation of measurement instruments (see, e.g., Duncan et al., 2003; Cole et al., 2004; Vidotto et al., 2006, 2007, 2010; Pallant and Tennant, 2007; Shea et al., 2009; Anselmi et al., 2011, 2013a,b, 2015; Da Dalt et al., 2013, 2015, 2017; Balsamo et al., 2014; Liu et al., 2017; Rossi Ferrario et al., 2019; Sotgiu et al., 2019). Rasch models characterize the responses of persons to items as a function of person and item measures (in the Rasch framework, the terms "person measure" and "item measure" are used to denote the values of the person parameter and item parameter, respectively). These measures pertain to the level of a quantitative latent trait possessed by a person or item, and their specific meaning relies on the subject of the assessment. In educational assessments, for instance, person measures indicate the ability of persons, and item measures indicate the difficulty of items. In health status assessments, person measures indicate the health of persons, and item measures indicate the severity of items. The Rasch model for dichotomous items is the simple logistic model (SLM; Rasch, 1960). This model allows for estimating a measure for each person and a measure for each item. An extension of the SLM to polytomous items is the rating scale model (RSM; Andrich, 1978). In addition to the measures estimated by the SLM, the RSM also estimates measures that describe the functioning of the response scale. These measures, called thresholds, represent the point on the latent variable where adjacent response categories are equally probable. If the thresholds are increasingly ordered, then the response scale functions as expected (i.e., increasing levels of the latent variable in a respondent correspond to increasing probabilities that the respondent will choose the higher response categories; Linacre, 2002a; Tennant, 2004). R can be computed both for the person measures estimated on dichotomous data and for the person measures estimated on polytomous data.

KR20, α, and R are based on the essentially tau-equivalent measurement model, a measurement model that requires a number of assumptions to be met for the estimate to accurately reflect the true reliability. Essential tau-equivalence assumes that each item measures the same latent variable (unidimensionality), on the same scale (similar variances), but with possibly different degrees of precision (different means; Raykov, 1997). Within the framework of factor analysis, essential tau-equivalence is represented by all items having equal factor loadings on a single underlying factor (McDonald, 1999). Graham (2006) provides a nice example to describe this measurement model. The author considers a test designed to measure depression in which each item is measured on a five-point Likert scale from "strongly disagree" to "strongly agree." Responses to items like "I feel sad sometimes" and "I almost always feel sad" are likely to have similar distributions, but with different modes. This might be due to the fact that, though both items measure the same latent variable on the same scale, the second one is worded more strongly than the first. As long as the variances of these items are similar across respondents, they are both measuring depression in the same scale, but with different precision.

KR20, α, and R are all estimates of the ratio between true variance and observed variance, and specify the observed variance as the sum of true and error variance. However, they differ for the way in which these quantities are obtained. Let us consider, for instance, a cognitive test with correct and incorrect item responses. In KR20, the error variance is computed as the sum of the variances of the items. In particular, with p_i denoting the proportion of correct responses to item $i = 1, 2, \ldots, I$, the error variance is $\sum_{i=1}^{I} p_i(1 - p_i)$. For dichotomous responses, p_i corresponds to the sample mean of the responses to item i. Thus, it represents what is expected from an "average" respondent from the sample on item i (Wright and Stone, 1999). When the variances $p_i \left(1 - p_i\right)$ are summed across the items, an error variance is obtained that represents the error variance of an "average" respondent from the sample. Respondents with high or low scores have less error variance than "average" respondents. Thus, the error variance of an "average" respondent used in KR20 overestimates the error variance of respondents with high or low scores. Furthermore, such an error variance is not the same as an average of the error variances of individual respondents. If the score distribution is not symmetric, the two quantities are different (Wright and Stone, 1999). Rasch measurement provides, for each estimate of a respondent's trait level, an accompanying estimate of the precision of the measure, called standard error (*SE*). The lower the *SE*, the higher the precision of trait level estimate. These individual *SE*s are used to compute the average error variance of the sample. In particular, with SE_n denoting the standard error associated with the trait level estimate of respondent $n = 1, 2, \ldots, N$, the average error variance of the sample is given by $\frac{\sum_{n=1}^{N} SE_n^2}{N}$.

KR20 and α use respondents' test scores (each of which being the sum of the responses over all items) in calculating the observed variance. This is potentially misleading. On the one hand, test scores are not linear representations of the variable they are intended to represent. For instance, a compression of the scale is bound to occur near the lower and upper boundaries of the score domain ("floor" and "ceiling" effects, respectively; Fischer, 2003). On the other hand, calculation of mean and variance necessary to obtain the observed variance assumes linearity in the numbers that are used (Wright and Stone, 1999). Thus, the observed variance computed from test scores might be incorrect to some degree. Contrariwise, if the data fit the Rasch model, the measures estimated for each respondent are on a linear scale, thus being numerically suitable for calculating the observed variance (Wright and Stone, 1999; Smith, 2001).

Given the aforementioned differences, classical and modern estimates of internal consistency might differ to some extent. Compared with KR20 and α, R is expected to be a better index of internal consistency as the numerical values are linear rather than non-linear, and the actual average error variance of the sample is used instead on the error variance of an "average" respondent.

The estimates of internal consistency might be affected by particular response behaviors. For instance, Pastore and Lombardi (2013) observed that α decreases with the increasing of the proportion of fake-good responses (i.e., responses aimed at providing a positive self-description) in the data set. The estimates of internal consistency might also be affected by random responding, that is a response set where individuals do not consider the content of the items and randomly choose all response options one by one. Random responding is not uncommon when respondents do not have an intrinsic interest in the investigation, the test is long, and the setting is uncontrolled (such as, e.g., in interned-based surveys; Johnson, 2005; Meade and Craig, 2012).

A method for identifying random responding requires the use of special items and scales. Examples include bogus items (e.g., "the water is wet"), instructed response items (e.g., "respond with a 2 for this item"), lie scales (e.g., MMPI-2 Lie scale), and scales for assessing inconsistent responding (e.g., MMPI-2 VRIN and TRIN scales). A drawback of this method is that testing time is lengthened.

Rasch framework provides methods and procedures for identifying and handling unexpected response behaviors. Mean-square fit statistics are computed for each individual and each item. Their expected value is 1. Values greater than 1 indicate underfit to the model (i.e., the responses are less predictable than the Rasch model expects), whereas values smaller than 1 indicate overfit (i.e., the responses are more predictable than the model expects; Linacre, 2002b). There are two types of mean-square fit statistics: outfit and infit. Outfit is mostly influenced by unexpected responses of high entity, whereas infit is mostly influenced by unexpected responses of small entity. An example of unexpected response is an incorrect response to an item for which a correct response is expected (i.e., an item for which, according to the Rasch model, the probability of a correct response is larger than that of an incorrect response). If the probability of the correct response is much larger than that of the incorrect response, the unexpected response mainly influences outfit. If the probability of the correct response is slightly larger than that of the incorrect response, the unexpected response mainly influences infit.

Infit and outfit allow for detecting individuals with unexpected response behaviors. For instance, they have been used to identify possible fakers to self-report personality tests (Vidotto et al., 2018) and to identify individuals who miss responses to items they are not capable of solving (Anselmi et al., 2018). In the present work, infit and oufit are used for handling random responses in the estimation of internal consistency. Two options are available. The first option implies taking into account random responses in order to compute a more conservative estimate of internal consistency. In the Rasch framework, this is done by enlarging the *SE* of latent trait estimates of those individuals with infit statistic larger than 1. With SE_n denoting the standard error associated with the trait level estimate of respondent $n = 1, 2, \ldots, N$, and infit_n denoting his/her infit statistic, the new infit-inflated standard error is given by $SE_n \times \max(1, \text{infit}_n)$ (see, e.g., Linacre, 1997). Then, this new standard error is used in place of SE_n to compute the average error variance of the sample. In the present work, a modification of this procedure is presented, in which an outfit-inflated standard error is computed as $SE_n \times \max(1, \text{outfit}_n)$. The larger the percentage of random responses, the larger the infit/outfit-inflated standard errors and the lower the estimate of internal consistency.

The second option implies "cleaning" the data set before estimating internal consistency. To this aim, individuals with infit or outfit above a certain, appropriately chosen cut-off are flagged as possible respondents with random responses and removed. A conservative choice for the cut-off is 1.3 (Wright and Linacre, 1994). Such a value indicates that, in the response pattern, there is 30% more randomness than expected by the Rasch models. If most individuals with random responses are correctly identified and removed, the internal consistency estimated on the "cleaned" data set should be less biased than that estimated on the "uncleaned" data set.

The aim of the present work is twofold. Firstly, it attempts to show the conditions in which classical and modern estimates of internal consistency are similar and those in which they are not. To this aim, data sets are simulated that differ for the distribution of test scores. Secondly, it investigates the use of respondents' infit and outfit statistics to compute more conservative estimates of internal consistency or to detect individuals with random responses. To this aim, data sets are simulated that include different percentages of random responses. Tests with dichotomous items and tests with polytomous items are considered.

STUDY 1 – EFFECTS OF SCORE DISTRIBUTION ON INTERNAL CONSISTENCY MEASURES

The present study aims at investigating the effects of score distribution on classical and modern estimates of internal consistency. Data sets are simulated that differ for the skewness of the score distribution. Classical and modern measures are expected to be substantially the same when the score distribution is symmetric, whereas they are expected to differ more and more with the increasing of the skewness of the score distribution. This study largely resembles that described by Linacre (1997). The author has only dealt with the dichotomous case and generated a single data set for each skewness condition. In the present study, both the dichotomous and polytomous cases are considered, and multiple data sets are generated for each skewness condition.

Data Simulation

All the data sets simulated in this study consist of the responses of 100 individuals to tests with 30 items. The polytomous data sets were simulated considering items with four response categories. Different skewed score distributions were obtained using the following three-step procedure:

1. A total of 30 true item measures were randomly drawn from a uniform distribution defined on the interval [−3, 3]. When simulating the polytomous data, three true thresholds were randomly simulated (i.e., the threshold between responses 1 and 2, that between 2 and 3, and that between 3 and 4) that were increasingly ordered and equally distant from each other. A total of 100 true person measures were randomly drawn from a standard normal distribution. This construction results in a sample of simulated respondents that is targeted on the test. This condition is denoted with offset = 0.
2. Four mistargeted samples were obtained by adding one, two, three, or four logits to the true person measures drawn in Step 1 (the logits are the measurement units constructed by Rasch models; Wright, 1993). These conditions are denoted with offset = 1, 2, 3, and 4.
3. Data sets were simulated for each of the five offset conditions. The dichotomous data sets were simulated using the SLM (Rasch, 1960), whereas the polytomous data sets were simulated using the RSM (Andrich, 1978).

It is noted in passing that the use of a uniform distribution for the item measures is a common choice (Linacre, 2007), and depicts the condition of tests measuring the different latent trait levels with the same precision. The use of thresholds that are increasingly ordered and equally distant depicts the condition of a well-functioning response scale (i.e., the response options are equally relevant and their choice appropriately reflects respondents' latent trait levels).

The aforementioned three-step procedure was repeated 100 times. Thus, 100 data sets were simulated for each of five offset conditions.

Results

Results considering the tests with dichotomous items are considered first. For each of the five offset conditions, **Figure 1** displays the score distribution, averaged across the 100 data sets simulated for that condition. When offset = 0 (i.e., the sample is targeted on the test), the score distribution resembles the distribution of person measures. Contrariwise, as offset increases (i.e., the samples are less and less targeted on the tests), the score distributions are more skewed, with high scores becoming more and more frequent. Ceiling effects are observed when offset is 3 or 4. It is worth noting that, in the five offset conditions, the underlying distribution of person measures is always the normal distribution.

Figure 2 plots average internal consistency (and standard deviation) for each of the five offset conditions. There are three lines in the figure. The solid line and the dashed line represent KR20 and R, respectively. The dotted line represents the true-measure-based internal consistency (TMBIC), which is a Rasch measure of internal consistency computed directly from the true person and item measures, without data. In the computation of TMBIC, the true variance is the variance of the true person measures, whereas the *SE*s that are necessary to obtain the error variance are derived from the true person and item measures). TMBIC is taken to be the maximum possible internal consistency under the Rasch model (Linacre, 1997).

When offset = 0, KR20 and R are virtually the same ($M_{\text{KR20}} = M_{\text{R}} = 0.81$; $SD_{\text{KR20}} = SD_{\text{R}} = 0.03$). Both the measures of internal consistency decrease with the increasing of offset, with R decreasing to a larger extent. With offset = 3, KR20 suggests that internal consistency is acceptable (*M* = 0.71, *SD* = 0.04), whereas R does not (*M* = 0.55, *SD* = 0.05). KR20 is larger than TMBIC, whereas R is smaller.

Also in the tests with polytomous items, the score distributions become more and more skewed with the increasing of offset. **Figure 3** plots α (solid line), R (dashed line), and TMBIC (dotted line) against the five offset conditions. As for the dichotomous tests, the two measures of internal consistency decrease with the increasing of offset. The two measures are largely the same when offset $\leq$ 2, whereas they differ when offset is 3 or 4. When offset = 4, α suggests that internal consistency is acceptable (*M* = 0.79, *SD* = 0.05), whereas R does not (*M* = 0.51, *SD* = 0.08). In addition, α is larger than TMBIC, whereas R is smaller. Offset being the same, internal consistency is larger in the polytomous tests than in the dichotomous tests. This result is due to the fact that, the number items being equal, internal consistency increases with the number of response categories (Lozano et al., 2008).

Brief Discussion

When the score distributions are substantially symmetric, classical and modern estimates of internal consistency are largely the same. In the case of a symmetric score distribution, the error variance estimated by KR20 and α largely resembles that resulting from R. Moreover, in the middle of the score domain, the relationship between scores and measures is approximately linear. Thus, when the largest part of the scores belongs to this central region (as it is in a symmetric score distribution), the observed variance obtained from scores is similar to that obtained from measures.

In presence of skewed score distributions, classical and modern estimates of internal consistency differ. Andrich (2016) warns researchers that "distributions skewed artificially because of floor or ceiling effects render the calculation of α essentially meaningless" (Andrich, 2016, p. 29). It is worth noting that R is more conservative than KR20 and α. In addition, R is lower than TMBIC, whereas KR20 and α are larger. Thus, using R in place of the classical measures reduces the changes of test users attributing the test better measurement characteristics than it actually has.

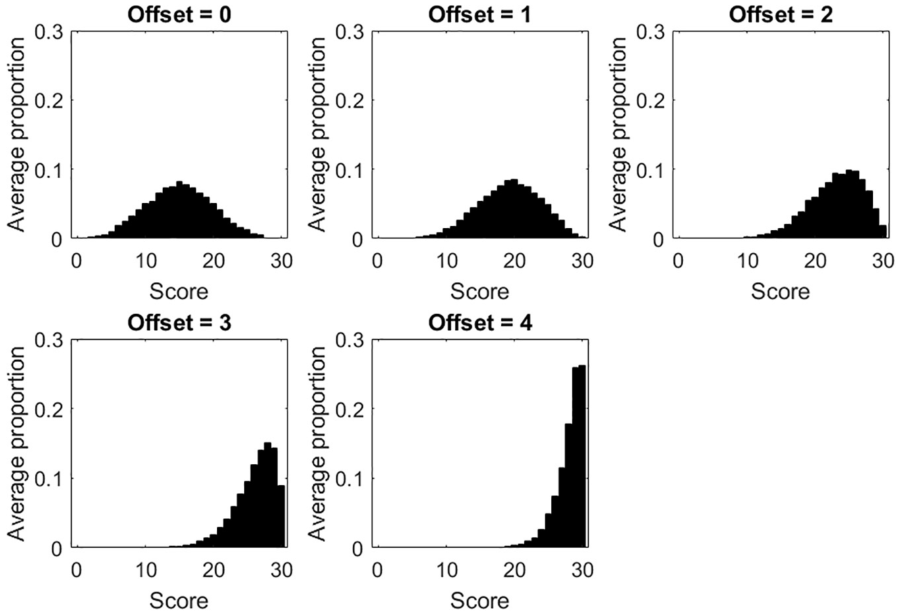


FIGURE 1 | Score distributions for each of the five offset conditions in the tests with dichotomous items.

The dichotomous and polytomous tests are not directly comparable, even if they contain the same number of items. This is due to the fact that internal consistency increases not only with the number of items but also with the number of response categories (Lozano et al., 2008). To this respect, a test with 30 polytomous items each having four response categories is analogous to a test with 90 dichotomous items. Similarly, a test with 30 dichotomous items is analogous to a test with 10 items each having four response categories. This explains why, offset being the same, internal consistency was larger in the polytomous tests than in the dichotomous tests.

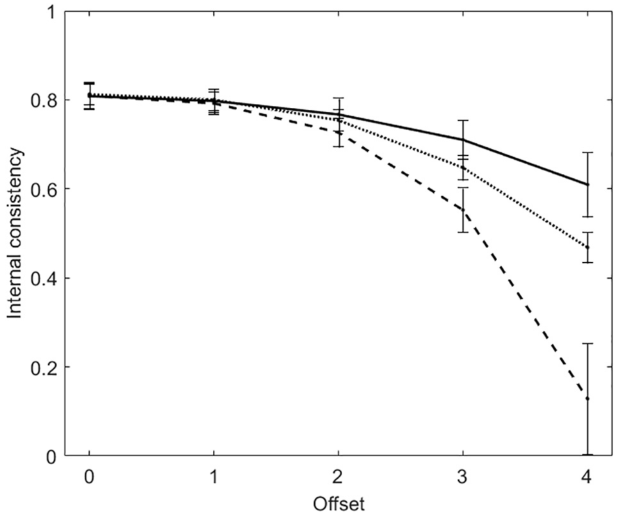


FIGURE 2 | Average internal consistency (and standard deviation) for each of the five offset conditions in the tests with dichotomous items. The solid line represents KR20, the dashed line represents R, the dotted line represents the true-measure-based internal consistency (TMBIC).

STUDY 2 – HANDLING UNEXPECTED RESPONSE BEHAVIORS WHEN COMPUTING INTERNAL CONSISTENCY

The present study aims at investigating the use of infit and outfit statistics to compute more conservative estimates of internal consistency and to detect individuals with random responses. Data sets are simulated that differ for (a) the percentage of respondents with random responses, and (b) the percentage of items with random responses. It is expected that, with the increasing of the two percentages, internal consistency decreases. Moreover, it is expected that, if the respondents with random responses are correctly identified and removed, the internal consistency computed on the cleaned data sets is similar to the true internal consistency.

Data Simulation

All the data sets simulated in this study consist of the responses of 100 individuals to tests with 30 items. The polytomous data sets were simulated considering items with four response categories. The data sets were obtained using the following three-step procedure:

1. A total of 30 true item measures were randomly drawn from a uniform distribution defined on the interval $[-3, 3]$. When simulating the polytomous data, three true thresholds were randomly simulated that were increasingly ordered

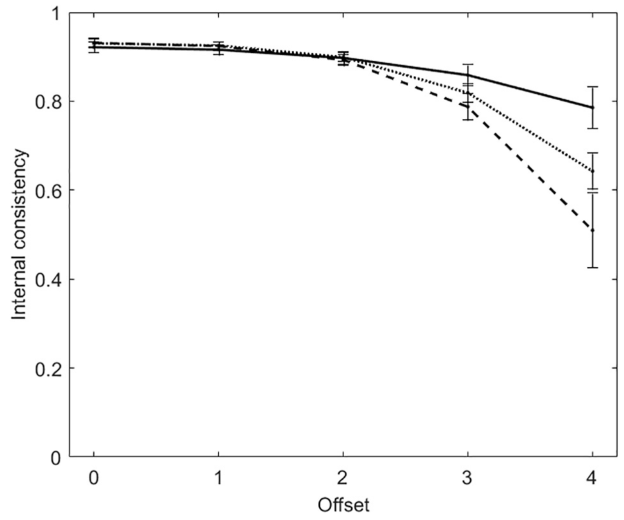


FIGURE 3 | Average internal consistency (and standard deviation) for each of the five offset conditions in the tests with polytomous items. The solid line represents α, the dashed line represents R, the dotted line represents the true-measure-based internal consistency (TMBIC).

and equally distant from each other. A total of 100 true person measures were randomly drawn from a standard normal distribution.

2. Data sets were simulated using the measures drawn in Step 1. The dichotomous data sets were simulated using the SLM (Rasch, 1960), whereas the polytomous data sets were simulated using the RSM (Andrich, 1978).
3. Twenty-five data sets with random responses were obtained from the data sets simulated at Step 2. These data sets differed for the proportion of simulees with random responses (p_{sim} = 0.10, 0.20, 0.30, 0.40, 0.50), and for the proportion of random item responses (p_{resp} = 0.10, 0.20, 0.30, 0.40, 0.50). The condition with p_{sim} = 0.20 and p_{resp} = 0.30 indicates 30% of random responses (i.e., 9 items) for 20% of simulees (i.e., 20 simulees). For each simulee, the items with random responses were randomly selected, and the responses to these items were set to be different to those in the original simulated data set.

The aforementioned three-step procedure was repeated 100 times. This resulted in 100 data sets without random responses, and 100 × 25 data sets with random responses (denoted as "uncleaned" data sets).

Results

Computing More Conservative Estimates of Internal Consistency

Results concerning the tests with dichotomous items are considered first. **Figure 4** displays the average internal consistency for the different proportions of simulees with random responses and the different proportions of items with random responses. There are four lines in each panel. The solid line represents KR20, the (unmarked) dashed line represents R, the +-marked dashed line represents infit-corrected R and the o-marked dashed line represents the outfit-corrected R. Some comments to the figure follows. In all the conditions, uncorrected KR20 and R lead to the same measure of internal consistency (the solid line substantially overlaps the unmarked dashed line). As shown in Study 1, when the samples are well-targeted on the tests (as it is in the case considered here), then KR20 and R lead to virtually the same estimate of internal consistency. As expected, all the internal consistency measures decrease with the increasing of the proportion of simulees with random responses and with the proportion of items in the patterns with random responses. The two underfit-corrected R measures of internal consistency (the two marked lines) are systematically lower than the two uncorrected measures (the two unmarked lines). The outfit-corrected R measure of internal consistency (the o-marked dashed line) is systematically lower than the infit-corrected R measure (the +-marked dashed line).

Figure 5 depicts the results concerning the tests with polytomous items. Results are similar to those observed in the dichotomous case. Given otherwise identical conditions, internal consistency is systematically larger in the polytomous case than in the dichotomous case. As discussed in Study 1, this result is due to the fact that, the number items being equal, internal consistency increases with the number of response categories.

Detection of Simulees With Random Responses

For each data set and each fit statistic (infit, outfit), sensitivity and specificity of the cut-off at 1.3 were computed by creating a 2 × 2 contingency matrix as follows:

		Simulee type		
		With random responses	Without random responses	
Fit statistic	>1.3	*a*	*b*	*a* + *b*
	≤1.3	*c*	*d*	*c* + *d*
		a + *c*	*b* + *d*	

Sensitivity refers to the capacity of correctly detecting simulees with random responses. It is the proportion of simulees with fit statistic larger than 1.3 among those simulees with random responses, that is $a/(a + c)$. Specificity refers to the capacity of correctly ignoring simulees without random responses. It is the proportion of simulees with fit statistic smaller than or equal to 1.3 among those simulees without random responses, that is $d/(b + d)$.

Table 1 shows sensitivity and specificity of infit and outfit statistics in the tests with dichotomous items. Both the proportion of simulees with random responses and the proportion of random responses in the patterns affect sensitivity. Overall, the lower the proportion of simulees with random responses and the higher the proportion of random responses in the patterns, the higher the sensitivity. A cut-off at 1.3 allows for detecting only 13% (infit) or 30% (outfit) of simulees with random responses when these simulees represent 50% of the sample and the random responses concern 10% of the items. Conversely, the same cut-off allows for detecting almost all simulees with random responses when they represent 10% of the sample and the random responses concern 50% of the

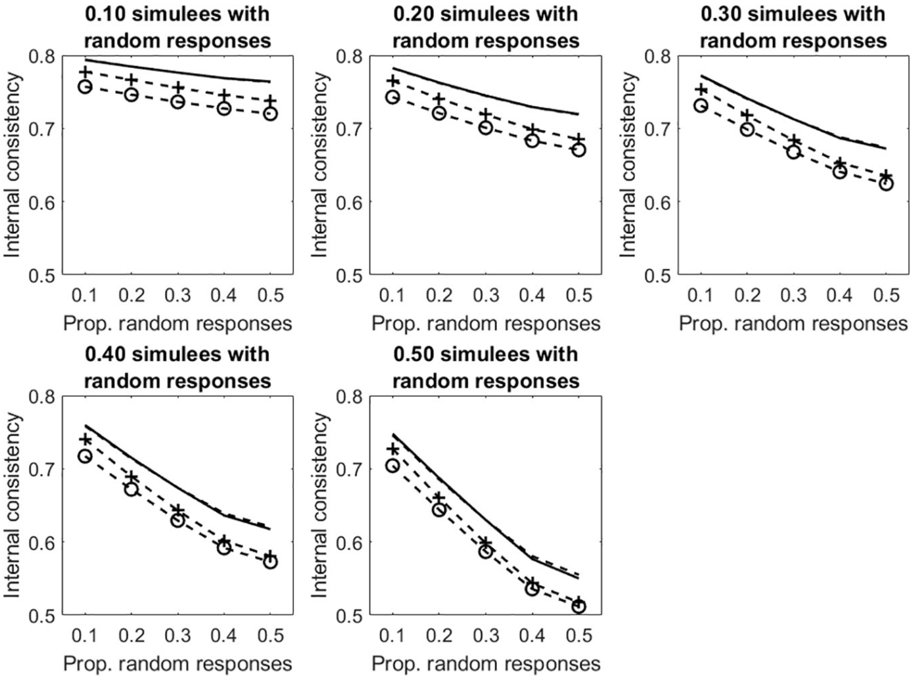


FIGURE 4 | Average internal consistency for the different proportions of simulees with random responses and the different proportions of dichotomous items with random responses. The solid line represents KR20, the unmarked dashed line represents R, the +-marked dashed line represents infit-corrected R, and the o-marked dashed line represents the outfit-corrected R.

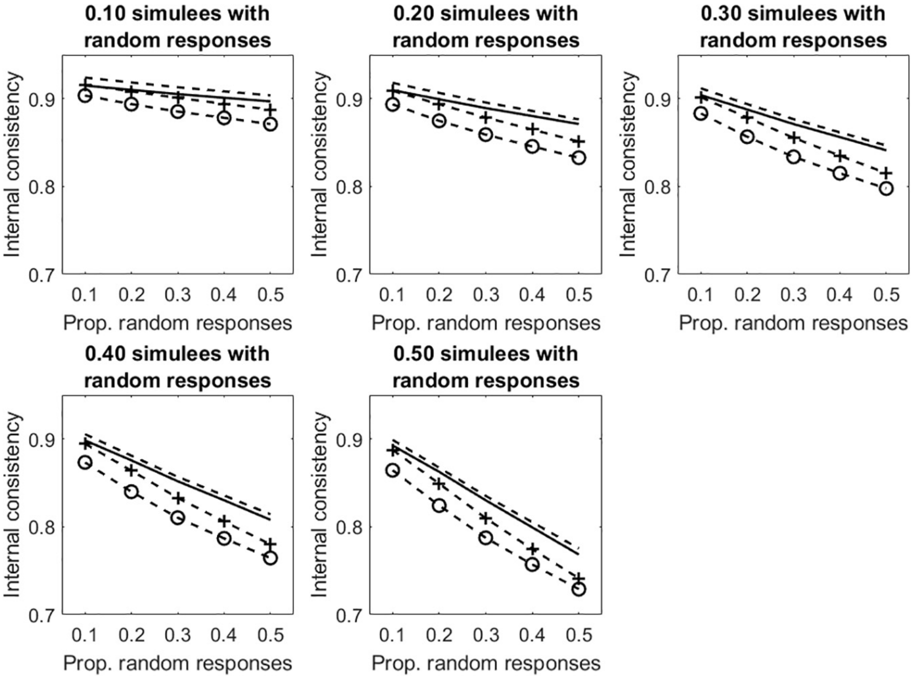


FIGURE 5 | Average internal consistency for the different proportions of simulees with random responses and the different proportions of polytomous items with random responses. The solid line represents α, the unmarked dashed line represents R, the +-marked dashed line represents infit-corrected R, and the o-marked dashed line represents the outfit-corrected R.

TABLE 1 | Sensitivity and specificity of infit and outfit in the tests with dichotomous items.

		Infit		Outfit	
p_{sim}	p_{resp}	Sensitivity	Specificity	Sensitivity	Specificity
0.10	0.10	0.30	0.93	0.51	0.86
0.10	0.20	0.58	0.95	0.76	0.88
0.10	0.30	0.84	0.96	0.90	0.90
0.10	0.40	0.93	0.97	0.97	0.92
0.10	0.50	0.98	0.98	0.99	0.94
0.20	0.10	0.25	0.95	0.46	0.88
0.20	0.20	0.50	0.98	0.66	0.93
0.20	0.30	0.69	0.99	0.80	0.95
0.20	0.40	0.84	0.99	0.90	0.97
0.20	0.50	0.92	1.00	0.96	0.98
0.30	0.10	0.21	0.97	0.41	0.91
0.30	0.20	0.37	0.99	0.56	0.95
0.30	0.30	0.52	1.00	0.67	0.97
0.30	0.40	0.65	1.00	0.76	0.99
0.30	0.50	0.73	1.00	0.83	1.00
0.40	0.10	0.17	0.98	0.35	0.92
0.40	0.20	0.24	0.99	0.44	0.97
0.40	0.30	0.34	1.00	0.50	0.99
0.40	0.40	0.39	1.00	0.53	1.00
0.40	0.50	0.42	1.00	0.55	1.00
0.50	0.10	0.13	0.98	0.30	0.94
0.50	0.20	0.16	1.00	0.33	0.98
0.50	0.30	0.18	1.00	0.32	1.00
0.50	0.40	0.16	1.00	0.28	1.00
0.50	0.50	0.11	1.00	0.20	1.00

p_{sim} = proportion of simulees with random responses; p_{resp} = proportion of random item responses. Cut-off for infit and outfit = 1.3.

items (sensitivity = 0.98, 0.99 for infit and outfit, respectively). Sensitivity of the cut-off on outfit is always larger than that of the cut-off on infit. Specificity remains very high regardless of the proportion of simulees with random responses and the proportion of random responses in the patterns (specificity from 0.93 to 1 for infit; from 0.86 to 1 for outfit). Taken all together, these results suggest that, when there are a few individuals with a consistent number of random responses, a cut-off at 1.3 allows for detecting almost all of them.

Figure 6 displays the average internal consistency for the different proportions of simulees with random responses and the different proportions of random responses in the patterns. The solid lines represent KR20, the dashed lines represent R. The unmarked lines represented the uncleaned data sets. The +-marked lines represent the infit-cleaned data sets. The o-marked lines represent the outfit-cleaned data sets. When simulees with random responses represent 10% of the sample, internal consistency obtained on the uncleaned data sets decreases with the increasing of the proportion of random responses in the patterns, whereas that obtained by removing underfitting simulees does not change. Even if the cut-off allows for identifying only a few of the simulees with random responses on 10% of items (sensitivity = 0.30, 0.51 for infit and outfit, respectively; see **Table 1**), the remaining simulees represent a small part of the sample so that they do not affect internal consistency too much. When the proportion of items with random responses increases to 0.50 (so that the random responses are a substantial threat for internal consistency), almost all of the underfitting simulees are correctly identified and removed (sensitivity = 0.98, 0.99 for infit and outfit, respectively; see **Table 1**). Similar results are observed when the proportion of simulees with random responses is 0.20 or 0.30. When this proportion is 0.40 or larger, the measures of internal consistency obtained by removing the underfitting simulees decrease with the increase with the proportion of missing data in the patterns. This is due to the fact that, when simulees with random responses become a consistent part of the sample, the cut-off fails in identifying a large part of them (with p_{sim} = 0.40, sensitivity $\leq$ 0.42, 0.55 for infit and outfit, respectively; with p_{sim} = 0.50, sensitivity $\leq$ 0.18, 0.33 for infit and outfit, respectively). Therefore, these simulees remain in the sample and affect internal consistency. Since sensitivity is larger for outfit than for infit, internal consistency obtained by removing simulees on the basis of outfit is never lower than that obtained by removing them on the basis of infit.

Similar results are obtained in the tests with polytomous items (see **Figure 7** and **Table 2**).

Brief Discussion

Internal consistency decreases with the increasing of random responses in the data set. Two options for dealing with such responses have been presented that are based on infit and outfit statistics. The first option implies using infit and outfit to compute more conservative estimates of internal consistency. In the presented simulations, the measures based on outfit were found to be more conservative than those based on infit.

The second option implies using infit and outfit to detect individuals with random responses. These statistics are a valid tool for this purpose, especially when there are a few individuals with a consistent number of random responses. Under these conditions, infit and outfit allow for correctly detecting almost all of them. When these individuals are removed, the internal consistency computed on the cleaned data sets is similar to the true internal consistency. In the presented simulations, outfit outperformed infit in identifying individuals with random responses. Consequently, the internal consistency obtained on the outfit-cleaned data sets resembled the true internal consistency more than that obtained on the infit-cleaned data sets.

OVERALL DISCUSSION

The present work compared classical and modern measures of internal consistency, which were computed on data sets with dichotomous and polytomous items. Classical and modern estimates of internal consistency are largely the same when the score distribution is substantially symmetric, whereas they differ when the score distribution is skewed. R is more conservative than KR20 and α, and prevents test users from believing a

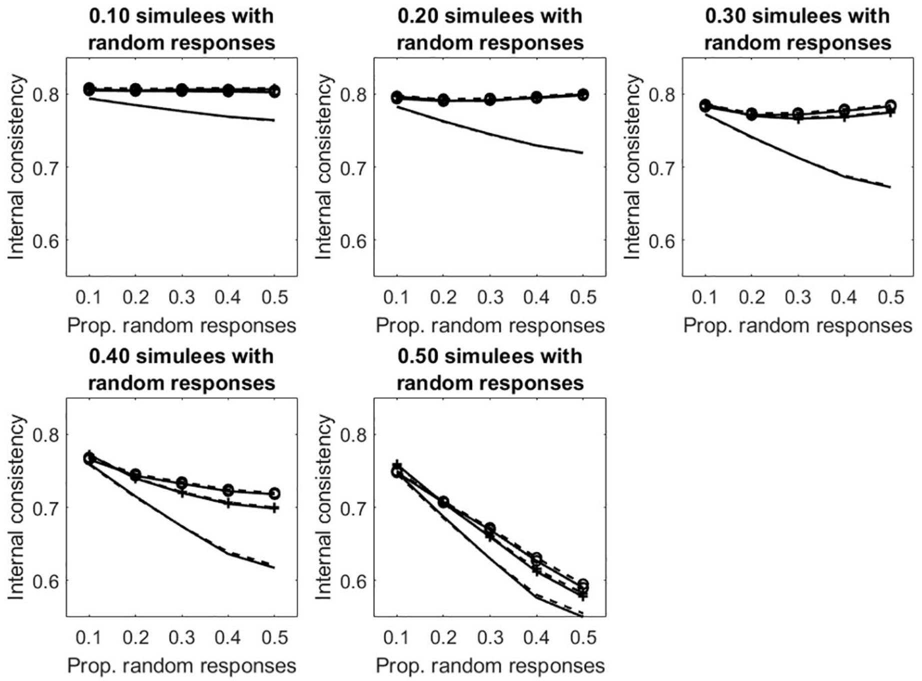


FIGURE 6 | Average internal consistency for the different proportions of simulees with random responses and the different proportions of dichotomous items with random responses. The solid lines represent KR20, the dashed lines represent R. The unmarked lines represented the full, uncleaned data sets. The +-marked lines represent infit-cleaned data sets. The o-marked lines represent the outfit-cleaned data sets.

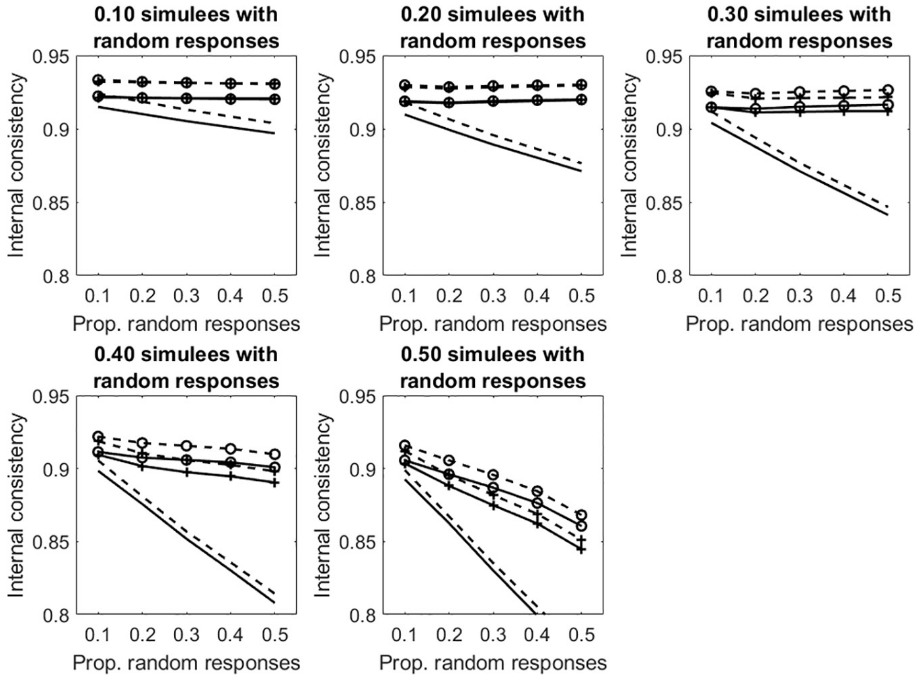


FIGURE 7 | Average internal consistency for the different proportions of simulees with random responses and the different proportions of polytomous items with random responses. The solid lines represent α, the dashed lines represent R. The unmarked lines represented the full, uncleaned data sets. The +-marked lines represent infit-cleaned data sets. The o-marked lines represent the outfit-cleaned data sets.

TABLE 2 | Sensitivity and specificity of infit and outfit in the tests with polytomous items.

		Infit		Outfit	
p_{sim}	p_{resp}	Sensitivity	Specificity	Sensitivity	Specificity
0.10	0.10	0.55	0.92	0.69	0.86
0.10	0.20	0.83	0.95	0.90	0.90
0.10	0.30	0.93	0.97	0.97	0.93
0.10	0.40	0.98	0.98	0.99	0.95
0.10	0.50	1.00	0.99	1.00	0.97
0.20	0.10	0.49	0.95	0.64	0.90
0.20	0.20	0.72	0.98	0.84	0.95
0.20	0.30	0.86	0.99	0.93	0.98
0.20	0.40	0.92	0.99	0.97	0.99
0.20	0.50	0.96	1.00	0.98	1.00
0.30	0.10	0.43	0.97	0.58	0.93
0.30	0.20	0.61	0.99	0.75	0.98
0.30	0.30	0.74	1.00	0.85	0.99
0.30	0.40	0.81	1.00	0.90	1.00
0.30	0.50	0.86	1.00	0.93	1.00
0.40	0.10	0.36	0.98	0.53	0.96
0.40	0.20	0.50	0.99	0.67	0.99
0.40	0.30	0.60	1.00	0.75	1.00
0.40	0.40	0.66	1.00	0.79	1.00
0.40	0.50	0.69	1.00	0.80	1.00
0.50	0.10	0.29	0.99	0.48	0.97
0.50	0.20	0.39	1.00	0.59	1.00
0.50	0.30	0.46	1.00	0.63	1.00
0.50	0.40	0.49	1.00	0.63	1.00
0.50	0.50	0.48	1.00	0.61	1.00

p_{sim} = proportion of simulees with random responses; p_{resp} = proportion of random item responses. Cut-off for infit and outfit = 1.3.

test has better measurement characteristics than it actually has. Compared with KR20 and α, R is expected to be a better index of internal consistency as the numerical values are linear rather than non-linear, and the actual average error variance of the sample is used instead of the error variance of an "average" respondent (Wright and Stone, 1999; Smith, 2001).

Internal consistency decreases with the increasing of random responses in the data set. Two options for dealing with such responses have been presented that are based on Rasch-based infit and outfit statistics. The first option implies using infit and outfit to compute a more conservative estimate of internal consistency. The second option implies using infit and outfit to detect individuals with unexpected responses. When there are a few individuals who gave a consistent number of unexpected responses, infit and outfit allow for correctly detecting almost all of them. The response pattern of each of these individuals can be carefully analyzed to try to discover the reason behind the unexpected responses (Has the individual responded randomly? Does he/she belong to a different population?). Once the individuals with random responses are removed, a cleaned data set is obtained that can be used for computing a less biased estimate of internal consistency.

Limitations and Suggestions for Future Research

In the present study, the data have been simulated under the assumption that the Rasch model was true in the population. Although KR20, α, and R are based on the same measurement model, it is not possible to exclude that the data generating process might have influenced the results. In future studies, the data could be generated using some procedure that puts the different indexes on an equal footing. For instance, the data could be generated from a multivariate normal distribution with the same variance for all items and the same covariance for all pairs of items. Alternatively, they could be generated from a one-factor model with equal factor loadings for all items.

In the present study, Rasch-based R has been shown as an example of modern measure of internal consistency. However, there are other models within modern test theory, which are distinct from Rasch models and pertain to item response theory (IRT). As for the Rasch models, there are several applications of IRT models to the development and validation of measurement scales (see, e.g., Wagner and Harvey, 2006; Thomas, 2011; Zanon et al., 2016; Colledani et al., 2018a,b, 2019a,b). Future studies should investigate the functioning of IRT-based measures of internal consistency, and compare them with classical and Rasch-based measures.

Random responding is only one type of careless responding. Another type of careless responding is identical responding. Individuals with this response behavior may give a certain response (e.g., Strongly agree) to all the items on one page and give the same or another response (e.g., Agree) to all the items on the next page. Future studies should investigate whether infit and outfit statistics allow the identification of individuals with this type of response behavior. Certainly, infit and outfit are unable to detect individuals who choose an *extreme* (minimum or maximum) response option for *all* test items, when there are no reverse-keyed items. Response patterns with extreme scores to all test items always fit the Rasch model perfectly (Linacre, 2019), so infit and outfit are not computed for them. Nevertheless, it is worth noting that these response patterns can be simply identified by looking at the average and standard deviation of the item responses (the former being equal to the minimum or maximum response score; the latter being equal to 0).

DATA AVAILABILITY STATEMENT

The R scripts used for simulating and analyzing the data will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

PA, DC, and ER contributed conception and design of the study, manuscript revision, read and approved the submitted version. PA performed the statistical analyses and wrote the first draft of the manuscript.

REFERENCES

Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika* 43, 561–573. doi: 10.1007/bf02293814

Andrich, D. (2016). Components of variance of scales with a bifactor subscale structure from two calculations of α. *Educ. Meas. Issues Pra.* 35, 25–30. doi: 10.1111/emip.12132

Anselmi, P., Robusto, E., and Cristante, F. (2018). Analyzing missingness at the level of the individual respondent: comparison of five statistical tests. *Test. Psychom. Methodol. Appl. Psychol.* 25, 379–394. doi: 10.4473/TPM25.3.4

Anselmi, P., Vianello, M., and Robusto, E. (2011). Positive associations primacy in the IAT. a many-facet rasch measurement analysis. *Exp. Psychol.* 58, 376–384. doi: 10.1027/1618-3169/a000106

Anselmi, P., Vianello, M., and Robusto, E. (2013a). Preferring thin people does not imply derogating fat people. a Rasch analysis of the implicit weight attitude. *Obesity* 21, 261–265. doi: 10.1002/oby.20085

Anselmi, P., Vianello, M., Voci, A., and Robusto, E. (2013b). Implicit sexual attitude of heterosexual, gay and bisexual individuals: disentangling the contribution of specific associations to the overall measure. *PLoS One* 8:e78990. doi: 10.1371/journal.pone.0078990

Anselmi, P., Vidotto, G., Bettinardi, O., and Bertolotti, G. (2015). Measurement of change in health status with Rasch models. *Health Qual. Life Out.* 13:16. doi: 10.1186/s12955-014-0197-x

Balsamo, M., Giampaglia, G., and Saggino, A. (2014). Building a new Rasch-based self-report inventory of depression. *Neuropsych. Dis. Treat.* 10, 153–165. doi: 10.2147/NDT.S53425

Cole, J. C., Rabin, A. S., Smith, T. L., and Kaufman, A. S. (2004). Development and validation of a Rasch-derived CES-D short form. *Psychol. Assess.* 16, 360–372. doi: 10.1037/1040-3590.16.4.360

Colledani, D., Anselmi, P., and Robusto, E. (2018a). Using item response theory for the development of a new short form of the Eysenck personality questionnaire-revised. *Front. Psychol.* 9:1834. doi: 10.3389/fpsyg.2018.01834

Colledani, D., Robusto, E., and Anselmi, P. (2018b). Development of a new abbreviated form of the Junior Eysenck personality questionnaire-revised. *Pers. Indiv. Differ.* 120, 159–165. doi: 10.1016/j.paid.2017.08.037

Colledani, D., Anselmi, P., and Robusto, E. (2019a). Development of a new abbreviated form of the Eysenck personality questionnaire-revised with multidimensional item response theory. *Pers. Indiv. Differ.* 149, 108–117. doi: 10.1016/j.paid.2019.05.044

Colledani, D., Anselmi, P., and Robusto, E. (2019b). Using multidimensional item response theory to develop an abbreviated form of the Italian version of Eysenck's IVE questionnaire. *Pers. Indiv. Differ.* 142, 45–52. doi: 10.1016/j.paid.2019.01.032

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334. doi: 10.1007/bf02310555

Da Dalt, L., Anselmi, P., Bressan, S., Carraro, S., Baraldi, E., Robusto, E., et al. (2013). A short questionnaire to assess pediatric resident's competencies: the validation process. *Ital. J. Pediatr.* 39:41. doi: 10.1186/1824-7288-39-41

Da Dalt, L., Anselmi, P., Furlan, S., Carraro, S., Baraldi, E., Robusto, E., et al. (2015). Validating a set of tools designed to assess the perceived quality of training of pediatric residency programs. *Ital. J. Pediatr.* 41:2. doi: 10.1186/s13052-014-0106-2

Da Dalt, L., Anselmi, P., Furlan, S., Carraro, S., Baraldi, E., Robusto, E., et al. (2017). An evaluation system for postgraduate pediatric residency programs: report of a 3-year experience. *Eur. J. Pediatr.* 176, 1279–1283. doi: 10.1007/s00431-017-2967-z

Duncan, P. W., Bode, R. K., Lai, S. M., and Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: the stroke impact scale. *Arch. Phys. Med. Rehabil.* 84, 950–963. doi: 10.1016/S0003-9993(03)00035-2

Fischer, G. H. (2003). The precision of gain scores under an item response theory perspective: a comparison of asymptotic and exact conditional inference about change. *Appl. Psych. Meas.* 27, 3–26. doi: 10.1177/0146621602239474

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: what they are and how to use them. *Educ. Psychol. Meas.* 66, 930–944. doi: 10.1177/0013164406288165

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *J. Res. Pers.* 39, 103–129. doi: 10.1016/j.jrp.2004.09.009

Kuder, G. F., and Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika* 2, 151–160. doi: 10.1007/BF02288391

Linacre, J. M. (1997). KR-20 / Cronbach alpha or Rasch person reliability: which tells the "truth"? *Rasch Meas. Trans.* 11, 580–581.

Linacre, J. M. (2002a). Optimizing rating scale category effectiveness. *J. Appl. Meas.* 39, 85–106.

Linacre, J. M. (2002b). What do infit and outfit, mean-square and standardized mean? *Rasch Meas. Trans.* 16:878.

Linacre, J. M. (2007). How to simulate Rasch data. *Rasch Meas. Trans.* 21:1125.

Linacre, J. M. (2019). *Winsteps (Version 4.4.6) [Computer Software]*. Beaverton, OR: Winsteps.com.

Liu, R., Sun, L., Yuan, J., and Bradley, K. (2017). Using the 2006 PISA questionaire to evaluate the measure of educational resources: a Rasch measurement approach. *Int. J. Asst. Tools Educ.* 4, 211–222. doi: 10.21449/ijate.319486

Lozano, L. M., García-Cueto, E., and Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology* 4, 73–79. doi: 10.1027/1614-2241.4.2.73

McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Meade, A. W., and Craig, S. B. (2012). Identifying careless responses in survey data. *Psychol. Methods* 17, 437–455. doi: 10.1037/a0028085

Pallant, J. F., and Tennant, A. (2007). An introduction to the Rasch measurement model: an example using the hospital anxiety and depression scale (HADS). *Brit. J. Clin. Psychol.* 46, 1–18. doi: 10.1348/014466506X96931

Pastore, M., and Lombardi, L. (2013). The impact of faking on Cronbach's alpha for dichotomous and ordered rating scores. *Qual. Quant.* 48, 1191–1211. doi: 10.1007/s11135-013-9829-1

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Test. Copenhagen: Danish Institute for Educational Research. Reprinted, 1980.* Chicago, IL: The University of Chicago Press.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Appl. Psychol. Meas.* 21, 173–184. doi: 10.1177/01466216970212006

Rossi Ferrario, S., Panzeri, A., Anselmi, P., and Vidotto, G. (2019). Development and psychometric properties of a short form of the illness denial questionnaire. *Psychol. Res. Behav. Manag.* 12, 727–739. doi: 10.2147/PRBM.S207622

Shea, T. L., Tennant, A., and Pallant, J. F. (2009). Rasch model analysis of the depression. anxiety and stress scales (DASS). *BMC Psychiatry* 9:21. doi: 10.1186/1471-244X-9-21

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* 74, 107–120. doi: 10.1007/s11336-008-9101-0

Smith, E. V. Jr. (2001). Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. *J. Appl. Meas.* 2, 281–311.

Sotgiu, I., Anselmi, P., and Meneghini, A. M. (2019). Investigating the psychometric properties of the questionnaire for eudaimonic well-being: a Rasch analysis. *Test. Psychom. Methodol. Appl. Psychol.* 26, 237–247. doi: 10.4473/TPM26.2.5

Tennant, A. (2004). Disordered thresholds: an example from the functional independence measure. *Rasch Meas. Trans.* 17, 945–948.

Thomas, M. L. (2011). The value of item response theory in clinical assessment: a review. *Assessment* 18, 291–307. doi: 10.1177/1073191110374797

Vidotto, G., Anselmi, P., Filipponi, L., Tommasi, M., and Saggino, A. (2018). Using overt and covert items in self-report personality tests: susceptibility to faking and identifiability of possible fakers. *Front. Psychol.* 9:1100. doi: 10.3389/fpsyg.2018.01100

Vidotto, G., Bertolotti, G., Carone, M., Arpinelli, F., Bellia, V., Jones, P. W., et al. (2006). A new questionnaire specifically designed for patients affected by chronic obstructive pulmonary disease. the Italian health status questionnaire. *Respir. Med.* 100, 862–870. doi: 10.1016/j.rmed.2005.08.024

Vidotto, G., Carone, M., Jones, P. W., Salini, S., and Bertolotti, G. (2007). Maugeri respiratory failure questionnaire reduced form: a method for improving the questionnaire using the Rasch model. *Disabil. Rehabil.* 29, 991–998. doi: 10.1080/09638280600926678

Vidotto, G., Moroni, L., Burro, R., Filipponi, L., Balestroni, G., Bettinardi, O., et al. (2010). A revised short version of the depression questionnaire. *Eur. J. Cardiovasc. Prev. Rehabil.* 17, 187–197. doi: 10.1097/HJR.0b013e328333edc8

Wagner, T. A., and Harvey, R. J. (2006). Development of a new critical thinking test using item response theory. *Psychol. Assess.* 18, 100–105. doi: 10.1037/1040-3590.18.1.100

Wright, B., and Stone, M. (1999). *Measurement Essentials*. Wilmington, DE: Wide Range, Inc.

Wright, B. D. (1993). Logits? *Rasch Meas. Trans.* 7:288.

Wright, B. D., and Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Meas. Trans.* 8:370.

Wright, B. D., and Masters, J. (1982). *Rating Scale Analysis*. Chicago, IL: MESA Press.

Zanon, C., Hutz, C. S., Yoo, H., and Hambleton, R. K. (2016). An application of item responsetheory to psychological test development. *Psicol. Reflex. Crit.* 29:18. doi: 10.1186/s41155-016-0040-x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Psychology

ORIGINAL RESEARCH
published: 05 December 2019
doi: 10.3389/fpsyg.2019.02693

Tailored Screening for Late-Life Depression: A Short Version of the Teate Depression Inventory (TDI-E)

Michela Balsamo, Aristide Saggino and Leonardo Carlucci*

School of Medicine and Health Sciences, Università degli Studi G. d'Annunzio Chieti e Pescara, Chieti, Italy

A number of assessment instruments have been developed as efficacy measures of geriatric depression in clinical trials but most showed several weaknesses, such as time-consuming administration, development and validation in younger populations, and lack of discrimination between anxiety and depression. Among the extant self-report measures of depression, the 21-item Teate Depression Inventory (TDI; Balsamo and Saggino, 2013), developed *via* Rasch analysis, showed a satisfactory level of diagnostic accuracy, and allowed the reduction of false positives in test scoring in adult population. The present study explored the potential improvement in the psychometric performance of the TDI in the elderly by item refinement through Rasch analysis in a sample of 836 elderly people (49.5% males; mean age = 73.28; SD = 6.56). A resulting shorter version was composed of the best-fitting and discriminative nine items from the full form. The Teate Depression Inventory (TDI-E) (E for elderly) presented good internal construct validity, with unidimensional structure, local dependency, good reliability (person separation index and Cronbach's alpha), and no signs of differential item functioning or measurement bias due to gender and age (65 vs. 75+ years). Cut-off points and normative data provided could enhance the clinical usefulness of the TDI-E, which seems to be a promising valid and reliable tool for the screening of geriatric depression, with less risk of finding false positives due to overlapping of depression in elderly with other comorbid conditions.

Keywords: depression, elderly, late-life, adults, Rasch analysis, item response theory

OPEN ACCESS

Edited by:
Pietro Cipresso,
Italian Auxological Institute
(IRCCS), Italy

Reviewed by:
Chung-Ying Lin,
Hong Kong Polytechnic
University, Hong Kong
Roger Muñoz Navarro,
University of Valencia, Spain

****Correspondence:***
Michela Balsamo
m.balsamo@unich.it;
michela.balsamo@unich.it

Specialty section:
This article was submitted to
Quantitative Psychology and
Measurement,
a section of the journal
Frontiers in Psychology

Received: *11 August 2019*
Accepted: *14 November 2019*
Published: *05 December 2019*

Citation:
Balsamo M, Saggino A and
Carlucci L (2019) Tailored Screening
for Late-Life Depression: A Short
Version of the Teate Depression
Inventory (TDI-E).
Front. Psychol. 10:2693.
doi: 10.3389/fpsyg.2019.02693

INTRODUCTION

Depression in Elderly and Its Measurement

Among older adults, depression is a common with more persistent and debilitating consequences than other forms of psychological distress condition (Arean and Ayalon, 2005; Friedman et al., 2007; Gilchrist and Gunn, 2007; Rodda et al., 2011; Sözeri-Varma, 2012). Among these, diminished cognitive, physical, and social functioning, increasing of risk of morbidity, general self-neglect, dependence by the others and mortality are those mainly noteworthy (Unützer et al., 2000, 2002; Fiske et al., 2009; Grover and Malhotra, 2015; Kennedy et al., 2016).

Late-life depression is characterized by different ways of presentation with respect to depression earlier in the lifespan (Koenig et al., 1993). Elderly depressed people are more likely to be affected by concomitant medical illness and psychiatric problems that can complicate their detection and therapy. For example, the presence of somatic and dementing disorders, the comorbidity with anxiety and physical complaints, may be misattributed to depression or vice versa (Lyness et al., 1995; Beekman et al., 2000; Friedman et al., 2007; Gilchrist and Gunn, 2007; Sözeri-Varma, 2012; Kennedy et al., 2016). For 516 depressed patients aged 70 years and older, suffered from a concomitant medical illness (e.g., weight loss, somatic anxiety, middle insomnia, and work impairment) eight items of the Hamilton Rating Scale for Depression (HAMD, Hamilton, 1960) may be elevated by the concurrent somatic disorder (Linden et al., 1995). Thus, the detection and assessment of elderly depression could be overlooked, misunderstood, or even misattributed because its symptoms can be easily confused with those of medical problems (e.g., fatigue, loss of involvement, pleasure, and interest in sexual activity, trouble sleeping, appetite, or weight change) and/or with natural cognitive functioning decline, including problems of concentration and memory, and/or with the senescence, an irreversible decline in mental and physical capabilities, as well as with some anxiety symptoms, including the hypochondriasis (Clark and Watson, 1991; Lyness et al., 1995). Ideally, depression assessment should be restricted to items that avoid confounding by medicall illness.

Lastly, items tapping pessimism, reduced actvity or interest, thoughts of death, possible suicidal intention, and meaning of the life have a different meaning for those approaching the end of their life, compared to younger individuals (Cusin et al., 2009). Probably, problems of unique interpretation could only be addressed if an experienced interviewer administers the scale of depression, but this turns out to be the case (Balsamo et al., 2018).

Given its costly and wide-ranging implications and the different psychopathological expression, sound and specific measurement of late-life depression is mandatory to improve the recognition and treatment of depressed elderly patients.

Current Self-Report Instruments on Geriatric Depression

A number of self-report measures developed in the adult population have been used to assess the incidence and intensity of depression symptoms and to monitor anti-depressant treatment progresses in the elderly (Andersen, 1999). Indeed, despite the differences in depressive symptoms between adult and geriatric population, the primary outcome measures used for the antidepressant trials in the people aged 65 years or older are still the self-report instruments developed in the adult population, such as the Beck Depression Inventory-II (BDI-II; Beck et al., 1996), the Center for Epidemiological Studies Depression Scale (CES-D, Radloff, 1977), and the HAM-D (Bent-Hansen et al., 2003; Roose et al., 2004; Sheikh et al., 2004; Wohlreich et al., 2004).

Nevertheless, controversies are emerged about their psychometric quality in the elderly because of some relevant shortcomings, such as time-consuming administration, vulnerability to misinterpretation and response biases, questionable structure of their response formats, and dependence of their scores on cultural factors (Balsamo and Saggino, 2007). For example, according to methodological studies, the number of items could be shortened by about 70% without compromising the measurement properties substantially (Moran et al., 2001). For the extant depression scales, it was highlighted that short forms with as few as nine items performed in ways very similar to the full version, while a version composed of only five items had a detectable difference from the full version (Cheung et al., 2007). Shorter form of the extant depression scales currently used for elderly should permit to decrease the overall time testing, in order to reduce the survey fatigue or boredom that older participants may feel, mostly when taking longer measures made up of many similar or repetitive items (Balsamo et al., 2018).

Moreover, some of these scales vary in terms of their primary content focus and their coverage of the core symptoms of depressive symptomatology. This aspect, which cast doubt on their content validity, could result in the under-recognition of depressive symptoms (Faravelli et al., 1986; Balsamo and Saggino, 2007).

Among the scales designed with the specific aim of screening depression in the elderly, the 30-item GDS was the gold standard (Yesavage et al., 1982). However, it has been repeatedly criticized (e.g., Friedman et al., 2005), mainly because of its length (Jongenelis et al., 2005; Chachamovich et al., 2010). The 15-item GDS-SF, extracted by the full-length form based on the base of diagnostic accuracy criteria (Yesavage and Sheikh, 1986), was also criticized (Chiang et al., 2009; Mitchell et al., 2010a,b; Wongpakaran et al., 2019) because of its lacking unidimensional nature. Indeed, two- and three-factor models emerged in different samples of elderly (Incalzi et al., 2003; Brown et al., 2007). Moreover, several items (i.e., #2, #3, and #10) were found to have a low clinical validity since did not contribute to the construct of geriatric depression and to be more related to subjective aspects of depression (e.g., life satisfaction or cognitive impairment) (Tang et al., 2005; Chiang et al., 2009; Wongpakaran et al., 2019). Further, some daunting multidimensional issues, such as differential item function (DIF), item misfit and redundancy, have been highlighted through IRT approach (Tang et al., 2005; Chachamovich et al., 2010; Wongpakaran et al., 2019). The development of the GDS brief forms (GDS-10, -6, -5, -4, and -1; Mitchell et al., 2010a,b) raised supplementary problems, including the difficulty to compare scores across different cultures and languages. In addition, the forced binary (yes or no) response format potentially provides no indication about the relative intensity or frequency of depression symptoms experienced by elderly (Castle and Engberg, 2004). Thus, to avoid these unidimensional and diagnostic problems, the GDS-SF is usually included together with other methods of screening for depression in a wide range clinical assessment for geriatric sample (Chiang et al., 2009).

Summing up, a brief, specific and unidimensional method of assessment of the severity of depressive symptoms in older adults seems to be the answer to the main challenges posed by the measurement of depression in this population (Balsamo et al., 2018). The different presentation of the depressive psychopathology between adults and elderly imposes different and specific measures in these populations. Measures specifically designed to measure depression in older adults result to lack of unidimensionality, i.e., an important requirement for calculating and interpreting a total score of an instrument (Lichtenberg, 2010; Ziegler and Hagemann, 2015). As a result, special emphasis should be laid on the investigation of the unidimensional structure of the scales used in the elderly general population.

Additionally, in most epidemiological studies, more females than males were diagnosed with depression (Albert, 2015), although these reported rates might be due to the use of generic diagnostic criteria and psychometric instruments that are not sensitive to depression in men (Oliffe and Phillips, 2008). As regards, age, there is some concern that older adults can obtain inflated scores on self-report depression instruments, which stem from non-depressive sources (e.g., medical problems) (e.g., Joiner et al., 2005). About this, it is worth noting that there is a difference between young- and old-old subjects groups (Garfein and Herzog, 1995; Mehta et al., 2008). Therefore, a further open question remains whether *bias-free* dimensional assessment of depression, independent of age, somatic morbidity, and gender, is feasible in the elderly general population.

Rasch measurement model is a powerful modern approach to develop unidimensional and bias-free instruments in health sciences. It examines both the scale and individual item performance in depth, leading to measures of depression, which are sample free and item free, and without DIF due to gender and age (Embretson and Reise, 2000). To our knowledge, no Rasch-based self-report measure of geriatric depression was developed. Up to the present, few IRT models have been applied only in the shortening process of extant few measures of depression used in the elderly, developed within classical test theory (CTT). The deriving advantage was to provide an improvement in the psychometric performance by item refinement, e.g., by revealing item redundancy, so that these instruments could be shortened without information loss (Tang et al., 2005; Lamoureux et al., 2009; Chachamovich et al., 2010; Forkmann et al., 2013; Spangenberg et al., 2015).

The Teate Depression Inventory

Among the extant self-report measures of depression used in older people, the 21-item Teate Depression Inventory (TDI; Balsamo and Saggino, 2013) was developed within Rasch logistic approach of responses. The TDI had shown to have an excellent Person Separation Index (PSI), no bias due to item-trait interaction, and control of major response sets (Innamorati et al., 2013, 2014; Balsamo et al., 2013a,b, 2014, 2015a,b,c, 2016, 2019; Saggino et al., 2017; Carlucci et al., 2018a,b). Three cut-off scores were recommended in terms of sensitivity, specificity, and classification accuracy for screening for varying levels (minimal, mild, moderate, and severe) of depression severity in a group of patients diagnosed with major depressive disorder (Balsamo and Saggino, 2014). More recently, applying the Bayes' theorem, the TDI showed to allow significant reduction of false positives in test scoring in clinical and non-clinical samples (Tommasi et al., 2018). Indeed, it was found to overcome the 50% level of diagnostic accuracy, unlike the BDI, the HAMD, the Zung Self-Rating Depression Scale (ZSDS; Zung, 1965), and the CES-D, because of a good procedure to select test items and subjects with clearly defined pathological symptoms.

About the pitfalls in the measurement of the geriatric-specific characteristics of late-life depression, the TDI significantly related to measures of anxiety and depression in expected directions and showed promise discriminating depression from anxiety (Picconi et al., 2018a,b). As such, it displayed significantly (p < 0.01) higher correlation with depression measure (GDS) compared with the anxiety measure, both trait and state, both cognitive and somatic scales, in a sample of 396 community-dwelling middle aged and elderly adults (Balsamo et al., 2015b).

Regarding the sex, the performance of the TDI has been found to be sufficiently insensitive for gender biases in a sample of 529 subjects (229 psychiatric outpatients and 300 healthy community-dwelling adults). Indeed, all items showed no difference due to gender, except for the item #10. It could represent an advantage over the extant depression questionnaires (like the BDI-II), that included several items showing DIF dependent of the respondent's sex since they might substantially interfere with the valid interpretation of instrument's sum score (Santor et al., 1994; Forkmann et al., 2009; da Rocha et al., 2013).

Regarding the impact of the somatic multimorbidity on the measurement of depression, the TDI was a unidimensional screening instrument of depression that included no items referring to somatic complaints (sleep and appetite disturbances). Present in an initial set of items, they did not fit the Rasch model because of no additional information provided to estimate the person's depression level. The lack of these items results to be consistent with the confounding of comorbidity that may be expected when applied to other diagnostic groups and can result in false positives (Thombs et al., 2007; Gibbons et al., 2011; da Rocha et al., 2013), as well as more useful for assessing depression in somatically ill patients, as are most of the elderly. Indeed, total scores of existing depression scales containing somatic items could be biased if those were filled from patients suffering from somatic illnesses because they did not reflect depression severity.

Although these compelling psychometric characteristics, the length of the TDI could be a limitation, which hinders its widespread use in elderly population. Reading and filling out its 21 items can be stressful for some older respondents, as well as not very useful for practitioners interested in measuring multiple constructs or repeated measurement of constructs, in the presence of time constraints. Moreover, even if the TDI is a Rasch-based measure, it is preferable to verify its psychometric functioning in a special population, like that of the elderly. Indeed, although Rasch analysis specifies that item parameters be sample free, constant item parameters imply a constant construct while different item parameters across samples of the relevant population could imply that the construct has changed, as Linacre (1996) outlined. Depressive psychopathology among elderly patients has been shown to be different in some aspects from younger individuals. Thus, given the construct of depression could change in different populations, it is desirable to test the TDI performance in the elderly population, which is different from the adult population, for which the TDI was developed.

Another point worth nothing concerns the availability of age-relevant norms in assessing mental health disorders among older adults (Therrien and Hunsley, 2012). Specific cut-off point represents a point of demarcation along continuum to address clinical decision and to identify good candidates for psychological treatments or protocols by clinicians interested to routinely screen their older patients for depression. This is particularly useful in clinical research, where the number of patients who receive the same intervention is usually limited. Only some depression measures currently used for measuring geriatric depression cut off points were computed. As regards the norms, few self-report instruments showed adequate normative data for elderly, which limited their clinical value (Breeman et al., 2015). With the growing number of older adults who is requiring mental health services, the diagnosis and treatment selection is helped by assessment data; thus, it is mandatory to have measures that are normed for an older population (Edelstein et al., 2007).

The Present Study

The present study aims at shortening and adapting the TDI to the elderly population using Rasch analysis with special emphasis on its unidimensional structure and DIF due to gender and age. Adherence of the brief TDI for elderly to Rasch model assumptions was determined with the analysis of Rasch model and item fit, unidimensionality, local dependency (LD) (principal component factor analysis of the residuals and correlation matrix of residuals), reliability (PSI and Cronbach' alpha), and DIF with regard to participants' age (65 vs. 75+ years) and gender.

A secondary aim was to examine the choice of cut-point to identify older people as depressed for screening and diagnostic purposes.

Finally, norm values were calculated. Based on the individual raw sum scores, each person's latent trait score θ was calculated and transformed linearly into percentiles, z values (mean = 0; SD = 1) and t values (mean = 50; SD = 10).

METHODS

Participants and Procedure

The sample included 836 elderly participants, of whom 49.5% were males. They were, on average, 73.28 (SD =6.56) years. Included in the sample was a subsample of 80 elderly clinical depressed participants (69% males) with an average of 72.60 years (SD = 5.44) years. No statistical differences were found in age variable between clinical vs. nonclinical group ($t_{834} = -1.207$, $p = 0.304$). Non-clinical participants have been enrolled by licensed psychologists at various community centers; groups; associations, senior citizens' Universities in Central and Southern Italy. They were preliminarily screened for psychiatric illness with a short interview. Only individuals evidencing no current psychopathology, no history of psychiatric hospitalization, and no cognitive impairment or neurological diseases (e.g., dementia, Parkinson, and Alzheimer's disease) were included in the non-clinical sample. Depressed participants were extracted from the standardization sample (Balsamo and Saggino, 2013). They were recruited from mental health counseling services and from private centers by clinical psychologists and psychotherapists. Eligible depressed participants were screened for major depressive disorders using the Structured Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders Axis I (SCID-I; First et al., 1997). Only participants who met the Diagnostic and Statistical Manual of Mental Disorders criteria (5th ed.; DSM–5; American Psychiatric Association, 2013) for a primary diagnosis of depression were included in the clinical subsample.

The study was reviewed and approved by the Department of Psychological Sciences, Health and Territory, University of Chieti, Italy, Review Board. All our procedures were in accordance with the ethical standards of the 1964 Declaration of Helsinki and its later amendments. Written informed consent was obtained from all individual participants included in the study.

Measure

Teate Depression Inventory

The Teate Depression Inventory (TDI) was composed of 21 items that aimed to assess symptoms of major depression during the past 2 weeks (Balsamo and Saggino, 2013). Participants responded on a five-point Likert scale ranging from "never" to "always." Total scores were created by first reverse-coding several items and then summing single items. Higher total scores indicate more severe depressive symptoms.

Data Analysis

The analysis plan consisted of two consecutive steps: initial evaluation of unidimensionality of the TDI using Mokken analysis and evaluation of Rasch model assumptions.

Firstly, a Mokken analysis was carried out within the framework of IRT in order to assess the assumption of unidimensionality. Following Sijtsma et al. (2011) recommendations, unidimensionality for polytomous-item measures was investigated through the Automated Item Selection Procedure (AISP) algorithm developed in *Mokken p*ackage of R, using recommended value of $c = 0.3$ and $\alpha = 0.05$ (Molenaar and Sijtsma, 2000). The AISP algorithm aimed at partition a set of items (or a set of unscalable items) into Mokken scales (Mokken, 1971; Sijtsma and Molenaar, 2002). Mokken scale is defined by a set of dichotomously scored items for which, given a lower bound "c," all inter-item covariances are positive and scalability coefficients (H_i/H) were set $\geq c > 0$. This definition can be extended to polytomous scored items. To date, values of $0.3 < H_i/H < 0.4$ identified weak scalability; values of $0.4 < H_i/H < 0.5$ as moderate, and values of $H_i/H > 0.5$ as strong scalability (Mokken, 1971).

Next, in line with the previous literature on the TDI (Balsamo and Saggino, 2013; Balsamo et al., 2014), data were fitted to the Rasch measurement model using RUMM2030 (Andrich et al., 2010). According to the Rasch model, probability of a person endorsing a dichotomic item was a logistic function of the difference between the person's abilities and the item difficulty (Rasch, 1960). Because the Rasch model was originally developed for intelligence and attainment tests, the "item difficulty" (Rasch, 1960) can be "translated" as and the severity of depression expressed by the item, i.e., the probability (expressed in logits) to endorse a high category of an item: for "difficult" items this probability would be lower than for "easy" items, relative to the individual person measure. Similarly, person's abilities are referred to as the latent trait score or person measure. If the Rasch model holds, persons and items can be scaled along a single linear latent continuum (i.e., depression). Since the TDI was conceived as a polytomous scales, an extended parameterization of the Rasch model for dichotomous responses (the Rating Scale Model, RSM) was used to fit the logistic function between the severity of depression and the severity of depression expressed by the items (Andiel, 1995). Like the Rasch model, the RSM and others extended models for polytomous scales (i.e., the Partial Credit Model) can be categorized as generalized linear model (GLM), with random effects modeling for the subject ability (Raju et al., 2014).

Data were found to fit Rasch model when the observed patterns of response are close to the expected model and satisfy a series of assumptions: local independency, response category ordering, lack of item bias or DIF, overall model and individual item fit, and reliability. Rasch analysis represents an iterative process where an initial observed pattern of response was tailored to ensure the overall fit of the data to the model. In this view, a series of sequential steps and fit indices has been estimated. In details:

1. assumption of stochastic ordering of the items along the whole latent trait was determined by a series of fit statistics within adequate ranges (Andrich, 1988): (1) chi-squared statistics (χ^2) and probability (χ^2_{prob}) should be not significant at level α = 0.05 with Bonferroni adjustment (also named item-trait interaction); (2) items with fit residual values >|2.5| (95% CI) should be discharged from the model; (3) summary item and person fit residual statistics should be approximated to the normal z distribution with mean = 0 and SD = |1| (or approximately |1.4|);
2. monotonicity for polytomous items was assessed by the inspection of the ordered items category thresholds. Thresholds represent the transition point between categories. When ordered, the amount of the probability of the category response itself leads to an amount of the latent trait (i.e., depression);
3. assumption of local response independency was assessed performing a Principal Component Factor Analysis of the Residuals (PCFAR; Smith and Miao, 1994; Linacre, 1998). Local independence implies that when controlling for the main Rasch dimension, no high or substantial residual correlations between the items shall remain. Hence, high residual correlation values (higher than the absolute value > 0.2; Marais, 2013) revealed that performance on the items was accounted for by a third trait dimension (Lee, 2004; Baghaei, 2008), displaying LD issue or multidimensionality. In addition, LD inflated reliability and affected parameters estimation (Wright, 1996);
4. DIF for age (60–75 years/over 75) and gender (males/females) person factors was also evaluated for each item by the two-way ANOVA (α = 0.05 with Bonferroni adjustment). DIF or item bias may occur systematically in responses based on characteristics of the respondents (*trait*) (uniform DIF) and varying along the construct (non-uniform DIF). In this study both, the uniform and non-uniform DIF issues were assessed;
5. afterward, strict unidimensionality was tested on the shortened set of items using the Smith' test of unidimensionality implemented in the RUMM 2030. A series of independent t test was performed in order to compare person estimates from two sets of items, composed, respectively, of items with positive and negative factor loadings ($\lambda \geq |0.30|$) on the first principal components analysis of the residuals (Smith, 2002). If more than 5% of these t tests was found to be significant, the resulting scale was labeled as multidimensional;
6. reliability and scale targeting were evaluated in order to assess the measurement validity of the final model. Reliability has been evaluated using the PSI. Values of PSI from 0.70 to 0.85 identified the minimum requirement for group and individual person measurement; a PSI > 0.85 was considered excellent (Nunnally, 1978). In addition, the internal consistency of the scale was examined by Cronbach's α. Targeting was measured by comparing graphically the mean location score obtained for the participants with that of the items: good values should be located in the center of the scale, close to the zero. Targeting of the person-item threshold distribution assesses how well individual item difficulties and individual person abilities can be matched on a common logit scale (Andrich, 1988) and how are the ceiling and floor effects (Tennant et al., 2004).

Next, following Davis et al. (2008), a regression analysis was performed to determine how well the resulting Rasch interval scale predicted the TDI scores, as conventionally computed using Likert interval scale (e.g., the raw summed scores of all the items), by fitting a cubic model.

To facilitate the clinical use of the TDI short version, norms values were computed. Person's latent trait scores (θ, expressed in logits) were transformed to an interval-metric scale using the original TDI 0–4 range scores (Tennant and Conaghan, 2007; Lundgren-Nilsson et al., 2013). This transformation is allowed since "Rasch model is capable of constructing linear measures from counts of qualitatively ordered observations, provided the structure of quantity is present in the data" (Salzberger, 2010). Next, the trait scores (θ) were transformed linearly into percentiles, z and t values.

Further, the receiver operating characteristic (ROC) curve analysis (Gleitman, 1986) with the Youden index (J) method was employed in order to detect the cut-off score, potentially useful in determining clinically depressed elderly. In this case, the optimal cut-off score represents the J function of the difference between true positive rate and false positive rate over all possible cut-point values. In the present sample, the prevalence rate of depression was 10.4% ($N = 80$). The performance of a diagnostic variable was quantified by computing the area under the curve (AUC; Bradley, 1997). Optimal values of AUC ranged from 0 "weak performance" to 1 "perfect performance" (Hanley and McNeil, 1982), with a value of >0.70 as recommended (Swets et al., 2000).

RESULTS

Mokken Analysis

After rescoring those formulated in reversed mode, all the TDI items were submitted to Mokken analysis, in order to test the unidimensionality assumption. The AISP revealed that all the TDI items loaded on a single latent dimension. The inter-item covariances were found positive, thus satisfied the first criterion of the Mokken scale. Next, all the item scalability coefficients (H_i) ranged from 0.350 (weak) to 0.470 (moderate); hence, the second criterion of a Mokken scale has been satisfied. The scalability coefficient for the entire TDI scale (H), equals to 0.409, showed a moderate scalability. Then, the assumption of unidimensionality was met for the 21 items of the TDI.

Rasch Analysis

The initial Rasch model was run with all the 21 items of the TDI exhibiting, an excellent PSI of 0.91. No floor and ceiling effects have been found. However, the initial model showed a poor overall model fit [χ^2 = 309.57(189), $p < 0.001$], and four items displayed disordered thresholds. The mean fit residual was 0.773 (SD = 2.066), indicating that the items did not fit the model properly, with an observed modest local response dependency. Out of the 21 items, six exhibited misfit criteria, including large fit residuals (±2.5) and significant χ^2 probabilities ($p < 0.0001$) with Bonferroni adjustment.

Since our goal was to develop a brief measure of depression for elderly people, attempts were made to improve fit to the initial model, by collapsing categories to achieve sequential order in items with disordered thresholds. The remaining items showed properly ordered thresholds, and all response categories were retained.

After collapsing item thresholds and ordering categories, the results showed non-considerable change (see Model #2 in **Table 1**). Thus, shortening of the TDI has been continued toward a final model, using an iterative strategy. Firstly, LD was pursued by deleting the pairs of items with correlations exceeding 0.3 were taken to indicate dependency. Items misfitting were removing item-by-item if displayed fit residuals outside the acceptable range (±2.5) and/or χ^2 probability value of the individual item fit was significant. Lastly, item bias or DIF for age and gender was also evaluated to determine if it was contributing to the misfit of items.

After removing item by item all misfitting items by the 21-item set, best model fit (with Bonferroni adjustment) was achieved by a final nine-item set, named the Teate Depression Inventory (TDI-E) (E for elderly) (**Table 2**). The final solution showed good fit to model expectations, with a not significant item-trait interaction index [χ^2 = 97.53(81), p = *0*.101]. Its item mean was 0.00 and SD = 0.264. No local response dependency was observed within the nine-item TDI model, as revealed by the inspection of the PCA residual correlations matrix. All item thresholds were found ordered, excepting for item #6. For achieving its sequential order, the "rarely" and "sometimes" response categories were collapsed. An inspection of the category response frequencies revealed that elderly participants chosen these two categories with the same probability (rarely = 14.64%; sometimes = 15.24%).

There was no DIF for both gender or age, based on Bonferroni adjusted *p*'s. Strict unidimensionality test (Smith, 2002) performed on the TDI-E showed that only the 5% (CI: 3.5–6.5%) of the paired t tests fell outside the 95% confidence interval, hence the assumption of unidimensionality held. The PSI of 0.83 indicated an adequate person separation reliability (Andrich, 1982) and also suggested that the power to detect items that do not fit the model was good. The TDI-E also showed high internal consistency (Cronbach's α = 0.85).

The shortened scale displayed an unbalanced person-item targeting to the left side of the person threshold distribution plot (easier questions or greater severity of depression to endorse the item), with a percentage of extreme scores <5%. No floor and ceiling effects have been found. However, the TDI-E was well targeted to the clinical sample, with the means of the person being 0.435 (SD = 1.01) on the logit scale (**Figure 1**).

Given the drastic scale reduction of the TDI (leading from 21 to 9 items), it was evaluated how the Rasch scale predicted the summed score of the selected nine items. Results from regression analysis supported the appropriateness of the cubic

TABLE 1 | Summary fit statistics for Rasch analyses.

Model	# Items	Items				Persons				Item-trait interaction		PSI
		Location		Fit residual		Location		Fit residual				
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	χ^2	χ^2 prob	
Initial	21	0.000	0.389	0.772	2.066	−8.836	0.969	−0.402	1.783	309.57	0.0000	0.91
1	21	0.000	0.425	0.539	1.880	−8.741	1.019	−0.403	1.767	315.68	0.0000	0.91
Final	9	0.000	0.264	0.410	0.959	−0.714	0.969	−0.411	1.386	97.53	0.1017	0.83

PSI, pearson separation index (person/item).

TABLE 2 | Final model with nine items.

Item/content	DSM diagnostic criterion	Location	SE	FitResid	χ^2	χ^2 prob	*F*-stat	*p*
TDI1\feeling blue	VII	−0.465	0.041	0.50	12.892	0.1676	1.630	0.1024
TDI13\fatigability	VI	−0.063	0.037	0.246	8.952	0.4417	1.073	0.3804
TDI18\loss of interest	II	−0.056	0.036	−0.519	9.978	0.3522	1.148	0.3259
TDI8\ concentration ability*	VIII	−0.044	0.041	1.756	5.904	0.7495	0.592	0.8041
TDI15\ enjoyment*	I	−0.025	0.04	0.998	6.086	0.7313	0.653	0.7520
TDI11\loss of self-confidence	VII	−0.009	0.039	1.561	12.395	0.1919	1.588	0.1145
TDI2\ concentration difficulty	VIII	0.000	0.039	−0.92	12.854	0.1693	1.644	0.0987
TDI14\lack of energy*	VI	0.089	0.04	0.722	8.826	0.4535	0.94	0.4895
TDI6\withdrawal	IX	0.572	0.06	−0.647	19.643	0.0202	2.534	0.0072

*χ^2 prob with Bonferroni adj. = 0.0055; DSM, diagnostic and statistical manual of mental disorders.*Reverse scored items.*

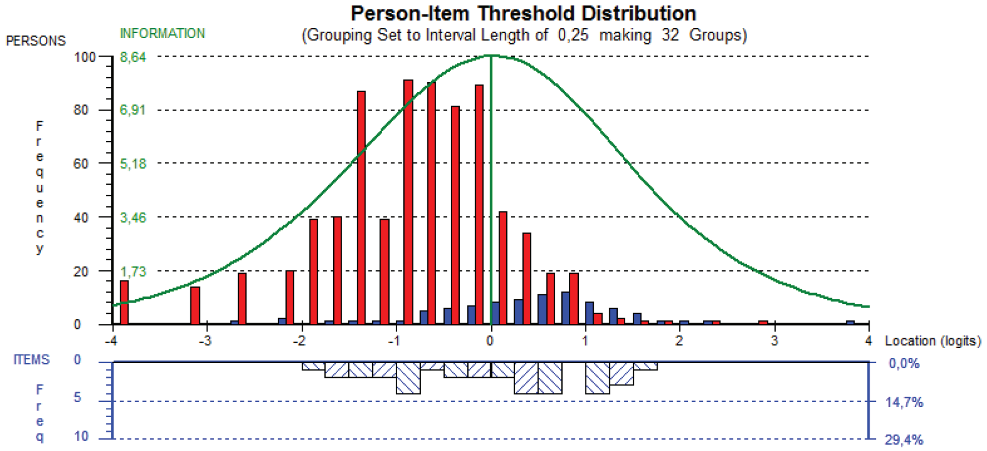


FIGURE 1 | Targeting of person and item. Red bars, non-clinical; blue bars, clinical.

function into predicted Rasch-based scores in relation to the summed score. Summary and coefficient estimates for the raw scores were displayed in **Table 3**.

Next, since Rasch model conformity of the TDI-E has been confirmed, norm values were determined. As no DIFs for gender and age groups were found, the score of the TDI-E resulted to be independent from gender and age. Norms values were displayed in **Table 4**. Rasch-based scores for all the raw summed score shave been estimated by transforming the Person's latent trait scores (θ) to their interval scale equivalent scores (or Rasch interval scale). This transformation is valid if no missing value was observed in the TDI items. Practically, a raw summed score of 10 ($\theta = -0.86$) on the TDI-E is equivalent to a Rasch interval score of 1.58, with a Z value of −0.49 (31st percentile) and a T score of 45.

Receiver Operating Characteristic Curve Analysis

A ROC curve analysis was performed to compare the non-depressed elderly group vs. the depressed group. Results indicated that the nine-item TDI scale was able to discriminate the two groups being examined. In details, the optimal cut-off point useful for the screening and diagnostic purposes was detected. The AUC for the TDI-E total score was 0.833 (95% CI of 0.806–0.858), suggesting good discrimination between the groups. The Youden index (0.54, CI 0.42–0.62) for the TDI-E total score was observed at a score of 18 points, corresponding to a sensitivity of 69% and specificity of 85%. Positive and negative predictive power were 35.5 and 96%, respectively, and overall diagnostic efficiency was 84%. Alternative cut-off values (see **Table 5**), were also estimated *via* BCa

TABLE 3 | Logits scores regressed by raw summed score for the nine items model.

	Model estimates			
	Coefficient	**SE**	***t***	***p***
Constant	−3.4184	0.0159	−214.815	<0.0001
(Raw summed score)^3– cubic trend	0.0003	0.0001	51.448	<0.0001
(Raw summed score)^2– quadratic trend	−0.0176	0.0003	−58.630	<0.0001
Raw summed score – linear trend	0.4026	0.0041	98.605	<0.0001
	Model fit			
	R	*R*^{2}	Adjusted *R*^{2}	
	0.995	0.989	0.989	

bootstrapped 95% CI (Efron and Tibshirani, 1993; Zhou et al., 2002). For instance, a cut-off of >11 could be employed for the screening purpose, corresponding to a sensitivity of 90.8% and specificity of 57.3%. Positive and negative predictive powers were 18.4 and 98%.

DISCUSSION

An appropriate answer to the several issues posed by challenges to measurement of late-life depression could reside in a self-reported measurement late-life depression, with the characteristics of brevity, specificity, and unidimensionality (Balsamo et al., 2018).

Concerning the brevity, it is well known that brief tools in primary care would be very useful for general practitioners, who are scarce of time and their high frequent patients may be elderly (Luber et al., 2001; Frank et al., 2018).

Several briefer versions of the GDS, the gold standard measure for depression in the elderly, have been developed. However, they have not been shown to be exempt from weakness. For example, in a meta-analytic study on their diagnostic accuracy, there was inconsistency in the items that contributed to these briefer versions and there are no standardized cut-off scores. This cast doubt on the validity of their scores, as well as on their diagnostic performance (Pocklington et al., 2016).

Concerning the unidimensionality, extant scales currently used in the elderly general population has been found lacking because some items related to a different latent trait, such as physical illness, were included (Osman et al., 2004; Storch et al., 2004; Crockett et al., 2005). As a result, using a single total score could result in its unclear interpretation. For example, two patients with the same summed score might differ in terms of the relative severity and frequency of different components of depression; thus, a treatment targeting only one of these aspects would be harder to detect in its effect. By applying the Rasch analysis, it is possible to develop unidimensional and *bias-free* measures of depression in the elderly general population.

TABLE 4 | Transformation of raw score to Rasch-based scores.

Raw scores	*θ*	Rasch interval scale (0–4)	Z	%	*T*
0	−3.90	0.00	−2.27	1	27
1	−3.11	0.41	−1.81	4	32
2	−2.57	0.69	−1.49	7	35
3	−2.21	0.88	−1.28	10	37
4	−1.93	1.03	−1.11	13	39
5	−1.69	1.15	−0.97	17	40
6	−1.49	1.25	−0.86	20	41
7	−1.31	1.35	−0.75	23	43
8	−1.15	1.43	−0.66	26	43
9	−1.00	1.51	−0.57	29	44
10	−0.86	1.58	−0.49	31	45
11	−0.73	1.65	−0.41	34	46
12	−0.60	1.72	−0.33	37	47
13	−0.48	1.78	−0.26	40	47
14	−0.36	1.84	−0.19	42	48
15	−0.25	1.90	−0.13	45	49
16	−0.13	1.96	−0.06	48	49
17	−0.02	2.02	0.01	50	50
18	0.09	2.08	0.07	53	51
19	0.21	2.14	0.14	56	51
20	0.32	2.20	0.21	58	52
21	0.44	2.26	0.28	61	53
22	0.56	2.32	0.35	64	53
23	0.68	2.38	0.42	66	54
24	0.81	2.45	0.50	69	55
25	0.94	2.52	0.57	72	56
26	1.09	2.59	0.66	75	57
27	1.24	2.67	0.75	77	57
28	1.41	2.76	0.85	80	58
29	1.61	2.87	0.97	83	60
30	1.83	2.98	1.10	86	61
31	2.11	3.13	1.26	90	63
32	2.47	3.31	1.47	93	65
33	3.00	3.59	1.78	96	68
≥34	3.79	4.00	2.25	99	72

θ, estimated Pearson's latent trait score for depression; %, percentiles; Z (M = 0, SD = 1); T (M = 50, SD = 10).

The TDI is a newly developed Rasch-based measure of depression. Given the necessity of brevity of measurement in older adults, Rasch analysis was employed to develop a briefer measure of geriatric depression from the Rasch-based 21-item TDI. Given the differences in depressive symptoms between geriatric and adult populations, this study aimed at evaluating its performance in this specific population.

Mokken and Rasch Analyses

In line with the previous literature, Mokken analysis of the TDI items showed that they mapped on to the depression trait, with medium scalability coefficients. To select items from the 21-item TDI with best measurement properties for composing a briefer, homogeneous, and unidimensional scale of geriatric depression, a Rasch analysis was performed. A shortened measure with nine items was derived. The newly developed TDI-E included items covering a wide range of diagnostic criteria of the DSM-5 for the major depressive episode (for a comprehensive review of the criteria, see Balsamo et al., 2014). Like the TDI,

TABLE 5 | Alternative cut-off values for the TDI-E.

Cut-off	Youden	Sensitivity	95% CI	Specificity	95% CI	+LR	95% CI	−LR	95% CI	+PV	95% CI	−PV	95% CI
>6	0.204	95.4	88.6–98.7	24.97	21.9–28.2	1.27	1.2–1.4	0.18	0.07–0.5	12.9	12.2–13.6	97.9	94.7–99.2
>11	0.441	90.8	82.7–95.9	53.27	49.6–56.9	1.94	1.8–2.2	0.17	0.09–0.3	18.4	16.9–20.0	98	96.3–99.0
>13	0.471	82.76	73.2–90.0	64.35	60.8–67.8	2.32	2.0–2.7	0.27	0.2–0.4	21.2	19.1–23,6	97	95.3–98.1
>18	**0.544**	**68.97**	**58.1–78.5**	**85.45**	**82.7–87.9**	**4.74**	**3.8–5.9**	**0.36**	**0.3–0.5**	**35.5**	**30.6–40.8**	**96**	**94.5–97.0**
>20	0.518	62.07	51.0–72.3	89.72	87.3–91.8	6.04	4.6–7.9	0.42	0.3–0.6	41.2	34.9–47.8	95.3	94.0–96.4
>23	0.387	43.68	33.1–54.7	95.06	93.3–96.5	8.84	6.0–13.1	0.59	0.5–0.7	50.7	40.9–60.4	93.6	92.3–94.6
>25	0.267	28.74	19.5–39.4	98.00	96.7–98.9	14.35	7.9–26.2	0.73	0.6–0.8	62.5	47.8–75.2	92.2	91.2–93.1

In bold, the recommended cut-off value.
+LR, positive likelihood ratio; −LR, negative likelihood ratio; +PV, positive predictive values; −PV, negative predictive values.

the TDI-E covered the same patterns of difficulties, within the range ± 1 logit. Item #1 ("I felt down") resulted to be the easiest to endorse, while item #6 ("I felt the desire to retire and disappear") was found the most difficult to endorse. This result was in accordance with previous literature (e.g., Lewinsohn et al., 2003), according to which depressed mood is the common symptom of depression, more so than anhedonia and other symptoms. Similarly, wish to die was considered a component of suicidal desire, an extremely important indicator of dangerousness across categories of mental disorders, including depression (Joiner et al., 2005). All the TDI-E items displayed no significant differences in the thresholds distances, suggesting that respondents discriminated properly between response options. Only item #6 showed two collapsed categories to achieve sequential order. As suggested by Bode (1997), ambivalent categories in rating scale (e.g., "do not know") often share more noise than information and should be threatened as missing data, so the pivot point for collapsing categories may be in the middle of uncertain categories. Notably, elderly respondents with reduced working memory capacity were more prone to answer "do not know" or to choose ambivalent categories in difficult questions, compared to respondents with higher cognitive abilities (Knäuper et al., 1997, 2016).

Like the TDI, the TDI-E demonstrated no DIF with regard to participants' age (65 vs. 75+ years) and gender. This means that all the TDI-E items performed equivalently for males and females, and for young old and old-old subjects (Garfein and Herzog, 1995; Mehta et al., 2008).

Prior evidence demonstrated that females showed an elevated risk of major depressive episode, and this risk increase in elderly females (+65 years) (Angst et al., 2002; Kessler et al., 2010). Potentially, this unbiased version of the TDI could allow an easy and efficient assessment of depression among elderly, thus avoiding the extensive use of differentiated norms (e.g., by gender or age) that are complex and may be difficult to communicate to general audiences or within a multidisciplinary team of experts.

The present study supported unidimensional construct of geriatric depression of the TDI-E. As revealed by the strict test of unidimensionality, neither subset of item from the factorial analysis of the residuals showed a significant difference on person estimates from the nine-item measure. Reliability, as measured by the PSI, was 0.83, an acceptable level especially for individual level data, which indicated not too large reduction from the PSI of 21-item TDI (0.96). A significant reduction of PSI values in short self-report measures derived from long self-report measures was expected (Davis et al., 2008; Shea et al., 2009). Unlike coefficient Cronbach's alpha, the PSI was not affected (or inflated) by the test length (Mallinson et al., 2004). Nevertheless, a limited and homogenous range of items, e.g., items with a close range of abilities, potentially resulted in decreasing of variability detected or in an increasing the amount of error, leading to a decrease in reliability (Mallinson et al., 2004). The reliability issue could represent a limitation for the present study, since a small set of items has been selected from a homogenous sample of participants (mostly healthy), which potentially weakens the ability of the scale to differentiate people.

Teate Depression Inventory Cut-off Scores and Diagnostic Utility

Results from regression analysis also revealed the measurement precision of the TDI-E. The raw summed scores for the nine items of the TDI-E seemed to predict the Rasch-based scores expressed in logits and the appropriateness of the cubic function (Lin et al., 2019). In other words, there is a substantial equivalence on the precision of the TDI-E score as measure of depression, whether it is computed as raw summed score or as Rasch-based interval score.

The diagnostic performance of the TDI-E in detecting elderly people who meet clinical thresholds for depressive symptoms, analyzed by the ROC curves, identified the cut-off point of 18 for differentiating non-depressed and depressed respondents. This value could facilitate researchers and clinicians into maximizing the clinical utility of the TDI-E when using in an applied way. For example, in clinical setting, a cut-off score easily allows to differentiate potential cases of clinical depression (True Positive) from probable "non-cases" (False Positive) or make decisions about who to treat and what treatments to provide (Widiger and Samuel, 2005; Van Dam et al., 2013). However, for clinicians who use the TDI-E as a screening instrument in clinical settings, where a higher sensitivity may be required, sensitivities and specificities corresponding to alternative cut off points were provided (**Table 5**).

Finally, although it may very tempting, to use a cut-off score on a self-report inventory as the single means of deriving, a diagnosis is a practice that should be avoided (Nezu et al., 2000). Rather, respondents scoring above the established cut-off level should be interviewed to assess for the depressive disorders criteria found in the DSM5.

Teate Depression Inventory Norms

The presented normative data could offer important advancements for the interpretation of the self-report measure scores and enhance its usefulness for clinical and research applications. For example, the *z* and *t* scores, set out here, makes it possible to compare TDI-E scores with the distribution of summed scores arising from convergent/divergent measures of depression and anxiety (e.g., the GDS or the Geriatric Anxiety Inventory), both in the clinical and general population. Thus, researchers and clinicians could benefit from these data in order to estimate significant changes across treatment (especially in repeated assessment) and/or to perform a brief assessment of the patient's depression severity. Moreover, the norms table provided makes the TDI-E scores comparable to the scores derived from other geriatric measures, even developed within the CTT.

Limitations

These results were based in a convenience sample almost exclusively composed of healthy and cognitive intact older people. They may not might be different in a depressed and/or cognitive impaired older population. Another limitation raises from the choice to use the Rasch model to shorten the TDI. Within the IRT models, analysis of Rasch is a fairly straightforward model and showed advantages and limitations. One limitation concerns the Rasch assumption of equal measurement error for each item (no discrimination parameters were provided, like in the 2PL model), as well as the possibility that a simple model may not fit the data. However, as Ryan outlined (Ryan, 1983), the inclusion of adjunctive parameters, i.e., the discrimination or guessing parameters (in the 2PL and 3PL models, respectively) could make potentially difficult and ambiguous the interpretation of item difficulties because all parameters are estimates simultaneously (Andrich, 2004, 2011; Han, 2012). Far from others IRT models, the Rasch model estimated a single person and item parameters; thus, the total score represents a sufficient statistic for the person parameter (Andrich and Marais, 2019). Further limitation concerns the lack of the investigation on test-retest reliability of this instrument and on the correlations with external measures for assessing its concurrent and discriminant validity.

Future investigations will be devoted (1) to verify if it displays validity coefficients with well-known depression and anxiety questionnaires currently used in the elderly; (2) to define its responsiveness to different contexts and different clinical samples (i.e., elderly with cognitive impairment or dementia); and (3) to examine if the TDI-E is composed of cultural-invariant items, which could then be applied in transcultural investigations free of bias.

The TDI has been translated in English and Portuguese, in order to be used as an outcome measure in internationally based longitudinal studies and clinical trials.

CONCLUSION

The present study explored the potential improvement in the psychometric performance of the 21-item TDI in the elderly by item refinement *via* Rasch analysis. This resulted in a short version of nine items, which was unidimensional, showed good internal construct validity, good reliability, and no signs of DIF due to gender and age. A specific cut-off point provided here could be more meaningful for screening purpose, as well as its normative data. To sum up, the TDI-E seems to be a valid and reliable tool for the screening of geriatric depression, with less risk of finding false positives due to overlapping of depression in elderly with other comorbid conditions. Its brevity could improve feasibility and compliance of older adults, mostly when several self-report measures are being used in a multidimensional psychological assessment in late life.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The study was reviewed and approved by the Department of Psychological Sciences, Health and Territory, University of Chieti, Italy, Review Board. Data were collected during 2018-2019. All our procedures were in accordance with the ethical standards of the 1964 Declaration of Helsinki and its later amendments. Written informed consent was obtained from all individual participants included in the study.

AUTHOR CONTRIBUTIONS

MB and LC designed the study. LC conducted the statistical analyses. MB, AS, and LC interpreted the data. MB and LC drafted the manuscript. All authors contributed toward drafting and revising the paper and agreed to be accountable for all aspects of the work.

REFERENCES

Albert, P. R. (2015). Why is depression more prevalent in women? *J. Psychiatry Neurosci.* 40, 219–221. doi: 10.1503/jpn.150205

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (*DSM-5*®). Arlington, VA: American Psychiatric Association.

Andersen, P. A. (1999). *Nonverbal communication: Forms and functions*. CA: Mayfield Mountain View.

Andiel, C. (1995). Rasch analysis: a description of the model and related issues. *Can. J. Rehabil.* 9, 17–26.

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR. 20 index, and the Guttman scale response pattern. *Educ. Res. Persp.* 9, 95–104.

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, California: Sage.

Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Med. Care* 42, I7–I16. doi: 10.1097/01.mlr.0000103528.48582.7c

Andrich, D. (2011). Rating scales and Rasch measurement. *Exper. Rev. Pharmacocon. Outcomes Res.* 11, 571–585. doi: 10.1586/erp.11.59

Andrich, D., Lyne, A., Sheridan, B., and Luo, G. (2010). *RUMM 2030*. Perth, Western Australia, Australia: RUMM Laboratory Pty Ltd.

Andrich, D., and Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social and health sciences*. Singapore: Springer.

Angst, J., Gamma, A., Gastpar, M., Lépine, J.-P., Mendlewicz, J., and Tylee, A. (2002). Gender differences in depression. *Eur. Arch. Psy. Clin. N.* 252, 201–209. doi: 10.1007/s00406-002-0381-6

Arean, P. A., and Ayalon, L. (2005). Assessment and treatment of depressed older adults in primary care. *Clin. Psychol. Sci. Pr.* 12, 321–335. doi: 10.1093/clipsy.bpi034

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Meas. Trans.* 22, 1145–1146.

Balsamo, M., Carlucci, L., Sergi, M. R., Murdock, K. K., and Saggino, A. (2015a). The mediating role of early maladaptive schemas in the relation between co-rumination and depression in young adults. *PLoS One* 10:e0140177. doi: 10.1371/journal.pone.0140177

Balsamo, M., Carlucci, L., Sergi, M., and Saggino, A. (2016). Validazione della versione italiana del co-rumination questionnaire [validation of the Italian version of the co-rumination questionnaire]. *Ital. J. Cogn. Behav. Psych.* 22, 13–34.

Balsamo, M., Cataldi, F., Carlucci, L., Padulo, C., and Fairfield, B. (2018). Assessment of late-life depression via self-report measures: a review. *Clin. Interv. Aging* 13, 2021–2044. doi: 10.2147/CIA.S178943

Balsamo, M., Giampaglia, G., and Saggino, A. (2014). Building a new Rasch-based self-report inventory of depression. *Neuropsychiatr. Dis. Treat.* 10, 153–165. doi: 10.2147/NDT.S53425

Balsamo, M., Imperatori, C., Sergi, M. R., Belvederi Murri, M., Continisio, M., Tamburello, A., et al. (2013a). Cognitive vulnerabilities and depression in young adults: an ROC curves analysis. *Depress. Res. Treat.* 2013, 1–8. doi: 10.1155/2013/407602

Balsamo, M., Innamorati, M., and Lamis, D. A. (2019). Editorial in the research topic clinical psychometrics: old issues and new perspectives. *Front. Psychol.* 10:947. doi: 10.3389/fpsyg.2019.00947

Balsamo, M., Innamorati, M., Van Dam, N. T., Carlucci, L., and Saggino, A. (2015b). Measuring anxiety in the elderly: psychometric properties of the state trait inventory of cognitive and somatic anxiety (STICSA) in an elderly Italian sample. *Int. Psychogeriatr.* 27, 999–1008. doi: 10.1017/S1041610214002634

Balsamo, M., Macchia, A., Carlucci, L., Picconi, L., Tommasi, M., Gilbert, P., et al. (2015c). Measurement of external shame: an inside view. *J. Pers. Assess.* 97, 81–89. doi: 10.1080/00223891.2014.947650

Balsamo, M., Romanelli, R., Innamorati, M., Ciccarese, G., Carlucci, L., and Saggino, A. (2013b). The state-trait anxiety inventory: shadows and lights on its construct validity. *J. Psychopathol. Behav. Assess.* 35, 475–486. doi: 10.1007/s10862-013-9354-5

Balsamo, M., and Saggino, A. (2007). Test per l'assessment della depressione nel contesto italiano: un'analisi critica. *Ital. J. Cogn. Behav. Psych.* 13, 167–199.

Balsamo, M., and Saggino, A. (2013). *Il Teate depression inventory-Manuale [Teate depression inventory, manual]*. Florence: Hoegrefe.

Balsamo, M., and Saggino, A. (2014). Determining a diagnostic cut-off on the Teate Depression Inventory. *Neuropsychiatr. Dis. Treat.* 10, 987–995. doi: 10.2147/NDT.S55706

Beck, A. T., Steer, R. A., and Brown, G. K. (1996). Beck depression inventory-II. *San Antonio* 78, 490–498.

Beekman, A. T., de Beurs, E., van Balkom, A. J., Deeg, D. J., van Dyck, R., and van Tilburg, W. (2000). Anxiety and depression in later life: co-occurrence and communality of risk factors. *Am. Psychiat.* 157, 89–95. doi: 10.1176/ajp.157.1.89

Bent-Hansen, J., Lunde, M., Klysner, R., Andersen, M., Tanghøj, P., Solstad, K., et al. (2003). The validity of the depression rating scales in discriminating between citalopram and placebo in depression recurrence in the maintenance therapy of elderly unipolar patients with major depression. *Pharmacopsychiatry* 36, 313–316. doi: 10.1055/s-2003-45120

Bode, R. K. (1997). Pivoting items for construct definition. *Rasch Meas. Trans.* 11, 576–577.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn.* 30, 1145–1159. doi: 10.1016/S0031-3203(96)00142-2

Breeman, S., Cotton, S., Fielding, S., and Jones, G. T. (2015). Normative data for the hospital anxiety and depression scale. *Qual. Life Res.* 24, 391–398. doi: 10.1007/s11136-014-0763-z

Brown, P. J., Woods, C. M., and Storandt, M. (2007). Model stability of the 15-item geriatric depression scale across cognitive impairment and severe depression. *Psychol. Aging* 22, 372–379. doi: 10.1037/0882-7974.22.2.372

Carlucci, L., D'Ambrosio, I., Innamorati, M., Saggino, A., and Balsamo, M. (2018a). Co-rumination, anxiety, and maladaptive cognitive schemas: when friendship can hurt. *Psychol. Res. Behav. Manag.* 2018, 133–144. doi: 10.2147/PRBM.S144907

Carlucci, L., Watkins, M. W., Sergi, M. R., Cataldi, F., Saggino, A., and Balsamo, M. (2018b). Dimensions of anxiety, age, and gender: assessing dimensionality and measurement invariance of the state-trait for cognitive and somatic anxiety (STICSA) in an Italian sample. *Front. Psychol.* 9:2345. doi: 10.3389/fpsyg.2018.02345

Castle, N. G., and Engberg, J. (2004). Response formats and satisfaction surveys for elders. *Gerontologist* 44, 358–367. doi: 10.1093/geront/44.3.358

Chachamovich, E., Fleck, M. P., and Power, M. (2010). Is geriatric depression scale-15 a suitable instrument for measuring depression in Brazil? Results of a Rasch analysis. *Psychol. Health Med.* 15, 596–606. doi: 10.1080/13548506.2010.487108

Cheung, Y. B., Liu, K. Y., and Yip, P. S. (2007). Performance of the CES-D and its short forms in screening suicidality and hopelessness in the community. *Suicide Life Threat. Behav.* 37, 79–88. doi: 10.1521/suli.2007.37.1.79

Chiang, K. S., Green, K. E., and Cox, E. O. (2009). Rasch analysis of the geriatric depression scale–short form. *Gerontologist* 49, 262–275. doi: 10.1093/geront/gnp018

Clark, L. A., and Watson, D. (1991). Tripartite model of anxiety and depression: psychometric evidence and taxonomic implications. *J. Abnorm. Psychol.* 100:316. doi: 10.1037/0021-843X.100.3.316

Crockett, L. J., Randall, B. A., Shen, Y. L., Russell, S. T., and Driscoll, A. K. (2005). Measurement equivalence of the center for epidemiological studies depression scale for Latino and Anglo adolescents: a national study. *J. Consult. Clin. Psychol.* 73, 47–58. doi: 10.1037/0022-006X.73.1.47

Cusin, C., Yang, H., Yeung, A., and Fava, M. (2009). "Rating scales for depression" in *Handbook of clinical rating scales and assessment in psychiatry and mental health*. eds. L. Baer and M. Blais (Totowa, NJ: Humana Press), 7–35.

da Rocha, N. S., Chachamovich, E., de Almeida Fleck, M. P., and Tennant, A. (2013). An introduction to Rasch analysis for psychiatric practice and research. *J. Psychiatr. Res.* 47, 141–148. doi: 10.1016/j.jpsychires.2012.09.014

Davis, A., Perruccio, A., Canizares, M., Tennant, A., Hawker, G., Conaghan, P., et al. (2008). The development of a short measure of physical function for hip OA HOOS-physical function Shortform (HOOS-PS): an OARSI/OMERACT initiative. *Osteoarthr. Cartil.* 16, 551–559. doi: 10.1016/j.joca.2007.12.016

Edelstein, B. A., Woodhead, E. L., Segal, D. L., Heisel, M. J., Bower, E. H., Lowery, A. J., et al. (2007). Older adult psychological assessment: current instrument status and related considerations. *Clin. Gerontol.* 31, 1–35. doi: 10.1080/07317110802072108

Efron, B., and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall.

Embretson, S. E., and Reise, S. P. (2000). *Multivariate applications books series. Item response theory for psychologists*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Faravelli, C., Albanesi, G., and Poli, E. (1986). Assessment of depression: a comparison of rating scales. *J. Affect. Disord.* 11, 245–253. doi: 10.1016/0165-0327(86)90076-5

First, M. B., Spitzer, R. L., Gibbon, M., and Williams, J. B. W. (1997). *Structured clinical interview for DSM-IV axis I disorders* (*SCID I*). New York: Biometric Research Department.

Fiske, A., Wetherell, J. L., and Gatz, M. (2009). Depression in older adults. *Annu. Rev. Clin. Psychol.* 5, 363–389. doi: 10.1146/annurev.clinpsy.032408.153621

Forkmann, T., Boecker, M., Wirtz, M., Eberle, N., Westhofen, M., Schauerte, P., et al. (2009). Development and validation of the Rasch-based depression screening (DESC) using Rasch analysis and structural equation modelling. *J. Behav. Ther. Exp. Psychiatry* 40, 468–478. doi: 10.1016/j.jbtep.2009.06.003

Forkmann, T., Gauggel, S., Spangenberg, L., Brähler, E., and Glaesmer, H. (2013). Dimensional assessment of depressive severity in the elderly general population: psychometric evaluation of the PHQ-9 using Rasch analysis. *J. Affect. Disord.* 148, 323–330. doi: 10.1016/j.jad.2012.12.019

Frank, C. C., Feldman, S., and Wyman, R. (2018). Caring for older patients in primary care: wisdom and innovation from Canadian family physicians. *Can. Fam. Physician* 64, 416–418.

Friedman, B., Conwell, Y., and Delavan, R. L. (2007). Correlates of late-life major depression: a comparison of urban and rural primary care patients. *Am. J. Geriatr. Psychiatry* 15, 28–41. doi: 10.1097/01.JGP.0000224732.74767.ad

Friedman, B., Heisel, M. J., and Delavan, R. L. (2005). Psychometric properties of the 15-item geriatric depression scale in functionally impaired, cognitively intact, community-dwelling elderly primary care patients. *J. Am. Geriatr. Soc.* 53, 1570–1576. doi: 10.1111/j.1532-5415.2005.53461.x

Garfein, A. J., and Herzog, A. R. (1995). Robust aging among the young-old, old-old, and oldest-old. *J. Gerontol. Ser. B Psychol. Sci. Soc. Sci.* 50, S77–S87.

Gibbons, C. J., Mills, R. J., Thornton, E. W., Ealing, J., Mitchell, J. D., Shaw, P. J., et al. (2011). Rasch analysis of the hospital anxiety and depression scale (HADS) for use in motor neurone disease. *Health Qual. Life Outcomes* 9:82. doi: 10.1186/1477-7525-9-82

Gilchrist, G., and Gunn, J. (2007). Observational studies of depression in primary care: what do we know? *BMC Fam. Pract.* 8:28. doi: 10.1186/1471-2296-8-28

Gleitman, H. (1986). *Psychology*. New York, NY: Norton.

Grover, S., and Malhotra, N. (2015). Depression in elderly: a review of Indian research. *J. Geriatr. Ment. Health* 2, 4–15. doi: 10.4103/2348-9995.161376

Hamilton, M. (1960). A rating scale for depression. *J. Neurol. Neurosurg. Psychiatry* 23, 56–62.

Han, K. T. (2012). Fixing the c parameter in the three-parameter logistic model. *Pract. Assess. Res. Eval.* 17, 1–24. Retrieved from: http://pareonline.net/getvn.asp?v=17&n=1

Hanley, J. A., and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36. doi: 10.1148/radiology.143.1.7063747

Incalzi, R. A., Cesari, M., Pedone, C., and Carbonin, P. U. (2003). Construct validity of the 15-item geriatric depression scale in older medical inpatients. *J. Geriatr. Psychiatry Neurol.* 16, 23–28. doi: 10.1177/0891988702250532

Innamorati, M., Lester, D., Balsamo, M., Erbuto, D., Ricci, F., Amore, M., et al. (2014). Factor validity of the Beck Hopelessness Scale in Italian medical patients. *J. Psychopathol. Behav.* 36, 300–307. doi: 10.1007/s10862-013-9380-3

Innamorati, M., Tamburello, S., Contardi, A., Imperatori, C., Tamburello, A., Saggino, A., et al. (2013). Psychometric properties of the attitudes toward self-revised in Italian young adults. *Depress. Res. Treat.* 2013, 1–6. doi: 10.1155/2013/209216

Joiner, T. E. Jr., Walker, R. L., Pettit, J. W., Perez, M., and Cukrowicz, K. C. (2005). Evidence-based assessment of depression in adults. *Psychol. Assess.* 17, 267–277. doi: 10.1037/1040-3590.17.3.267

Jongenelis, K., Pot, A. M., Eisses, A. M. H., Gerritsen, D. L., Derksen, M., Beekman, A. T. F., et al. (2005). Diagnostic accuracy of the original 30-item and shortened versions of the geriatric depression scale in nursing home patients. *Int. J. Geriatr. Psychiatry* 20, 1067–1074. doi: 10.1002/gps.1398

Kennedy, G. J., Castro, J., Chang, M., Chauhan-James, J., and Fishman, M. (2016). Psychiatric and medical comorbidity in the primary care geriatric patient—an update. *Curr psychiatry Rev* 18:62. doi: 10.1007/s11920-016-0700-7

Kessler, R. C., Birnbaum, H., Bromet, E., Hwang, I., Sampson, N., and Shahly, V. (2010). Age differences in major depression: results from the National Comorbidity Survey Replication (NCS-R). *Psychol. Med.* 40, 225–237. doi: 10.1017/S0033291709990213

Knäuper, B., Belli, R. F., Hill, D. H., and Herzog, A. R. (1997). Question difficulty and respondents' cognitive ability: the effect on data quality. *J. Off. Stat.* 13, 181–199.

Knäuper, B., Carrière, K., Chamandy, M., Xu, Z., Schwarz, N., and Rosen, N. O. (2016). How aging affects self-reports. *Eur. J. Ageing* 13, 185–193. doi: 10.1007/s10433-016-0369-0

Koenig, H. G., Cohen, H. J., Blazer, D. G., Krishnan, K. R. R., and Sibert, T. E. (1993). Profile of depressive symptoms in younger and older medical inpatients with major depression. *J. Am. Geriatr. Soc.* 41, 1169–1176. doi: 10.1111/j.1532-5415.1993.tb07298.x

Lamoureux, E. L., Tee, H. W., Pesudovs, K., Pallant, J. F., Keeffe, J. E., and Rees, G. (2009). Can clinicians use the PHQ-9 to assess depression in people with vision loss? *Optom. Vis. Sci.* 86, 139–145. doi: 10.1097/OPX.0b013e318194eb47

Lee, Y.-W. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Lang. Test.* 21, 74–100. doi: 10.1191/0265532204lt260oa

Lewinsohn, P. M., Petit, J. W., Joiner, T. E. Jr., and Seeley, J. R. (2003). The symptomatic expression of major depressive disorder in adolescents and young adults. *J. Abnorm. Psychol.* 112, 244–252. doi: 10.1037/0021-843X.112.2.244

Lichtenberg, P. A. (2010). *Handbook of assessment in clinical gerontology*. Burlington, MA: Academic Press.

Lin, C.-Y., Hwang, J.-S., Wang, W.-C., Lai, W.-W., Su, W.-C., Wu, T.-Y., et al. (2019). Psychometric evaluation of the WHOQOL-BREF, Taiwan version, across five kinds of Taiwanese cancer survivors: Rasch analysis and confirmatory factor analysis. *J. Formos. Med. Assoc.* 118, 215–222. doi: 10.1016/j.jfma.2018.03.018

Linacre, J. M. (1996). The Rasch model cannot be "disproved". *Rasch Meas. Trans.* 10, 512–514.

Linacre, J. M. (1998). Structure in Rasch residuals: why principal components analysis. *Rasch Meas. Trans.* 12:636

Linden, M., Borchelt, M., Barnow, S., and Geiselmann, B. (1995). The impact of somatic morbidity on the Hamilton depression rating scale in the very old. *Acta Psychiatr. Scand.* 92, 150–154. doi: 10.1111/j.1600-0447.1995.tb09559.x

Luber, M. P., Meyers, B. S., Williams-Russo, P. G., Hollenberg, J. P., DiDomenico, T. N., Charlson, M. E., et al. (2001). Depression and service utilization in elderly primary care patients. *Am. J. Geriat. Psych.* 9, 169–176. doi: 10.1097/00019442-200105000-00009

Lundgren-Nilsson, Å., Jonsdottir, I. H., Ahlborg, G., and Tennant, A. (2013). Construct validity of the psychological general well being index (PGWBI) in a sample of patients undergoing treatment for stress-related exhaustion: a Rasch analysis. *Health Qual. Life Outcomes* 11:2. doi: 10.1186/1477-7525-11-2

Lyness, J. M., Cox, C., Curry, J., Conwell, Y., King, D. A., and Caine, E. D. (1995). Older age and the underreporting of depressive symptoms. *J. Am. Geriatr. Soc.* 43, 216–221. doi: 10.1111/j.1532-5415.1995.tb07325.x

Mallinson, T., Stelmack, J., and Velozo, C. (2004). A comparison of the separation ratio and coefficient α in the creation of minimum item sets. *Med. Care*, I17–I24. doi: 10.1097/01.mlr.0000103522.78233.c3

Marais, I. (2013). "Local dependence" in *Rasch Models in Health*. eds. K. B. Christensen, S. Kreiner, and M. Mesbah (London, UK and Hoboken, NJ: Wiley-ISTE Ltd), 111–130.

Mehta, M., Whyte, E., Lenze, E., Hardy, S., Roumani, Y., Subashan, P., et al. (2008). Depressive symptoms in late life: associations with apathy, resilience and disability vary between young-old and old-old. *Int. J. Geriat. Psych.* 23, 238–243. doi: 10.1002/gps.1868

Mitchell, A. J., Bird, V., Rizzo, M., and Meader, N. (2010a). Diagnostic validity and added value of the geriatric depression scale for depression in primary care: a meta-analysis of GDS30 and GDS15. *J. Affect. Disord.* 125, 10–17. doi: 10.1016/j.jad.2009.08.019

Mitchell, A. J., Bird, V., Rizzo, M., and Meader, N. (2010b). Which version of the geriatric depression scale is most useful in medical settings and nursing homes? Diagnostic validity meta-analysis. *Am. J. Geriatr. Psychiatry* 18, 1066–1077. doi: 10.1097/jgp.0b013e3181f60f81

Mokken, R. J. (1971). *A theory and procedure of scale analysis: With applications in political research*. Berlin, Germany: Walter de Gruyter.

Molenaar, I. W., and Sijtsma, K. (2000). *User's manual MSP5 for Windows*. Groningen: iecProGAMMA.

Moran, L. A., Guyatt, G. H., and Norman, G. R. (2001). Establishing the minimal number of items for a responsive, valid, health-related quality of life instrument. *J. Clin. Epidemiol.* 54, 571–579. doi: 10.1016/S0895-4356(00)00342-5

Nezu, A. M., Ronan, G. F., Meadows, E. A., and McClure, K. S. (2000). *Practitioner's guide to empirically-based measures of depression*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Nunnally, J. (1978). *Psychometric methods*. New York: McGraw-Hill.

Oliffe, J. L., and Phillips, M. J. (2008). Men, depression and masculinities: A review and recommendations. *J. Men's Health* 5, 194–202. doi: 10.1016/j.jomh.2008.03.016

Osman, A., Kopper, B. A., Barrios, F., Gutierrez, P. M., and Bagge, C. L. (2004). Reliability and validity of the Beck depression inventory--II with adolescent psychiatric inpatients. *Psychol. Assess.* 16, 120–132. doi: 10.1037/1040-3590.16.2.120

Picconi, L., Balsamo, M., Palumbo, R., and Fairfield, B. (2018a). Testing factor structure and measurement invariance across gender with Italian geriatric anxiety scale. *Front. Psychol.* 9:1164. doi: 10.3389/fpsyg.2018.01164

Picconi, L., Jackson, C. J., Balsamo, M., Tommasi, M., and Saggino, A. (2018b). Factor structure and measurement invariance across groups of the Italian Eysenck Personality Questionnaire-Short Form (EPP-S). *Pers. Ind. Differ.* 123, 76–80. doi: 10.1016/j.paid.2017.11.013

Pocklington, C., Gilbody, S., Manea, L., and McMillan, D. (2016). The diagnostic accuracy of brief versions of the geriatric depression scale: a systematic review and meta-analysis. *Int J. Geriatr. Psychiatry* 31, 837–857. doi: 10.1002/gps.4407

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Appl. Psychol. Meas.* 1, 385–401.

Raju, D., Su, X., and Patrician, P. A. (2014). Using item response theory models to evaluate the practice environment scale. *J. Nurs. Meas.* 22, 323–341. doi: 10.1891/1061-3749.22.2.323

Rasch, G. (1960). *Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests*. Oxford, England: Nielsen & Lydiche.

Rodda, J., Walker, Z., and Carter, J. (2011). Depression in older adults. *BMJ* 343:d5219. doi: 10.1136/bmj.d5219

Roose, S. P., Sackeim, H. A., Krishnan, K. R. R., Pollock, B. G., Alexopoulos, G., Lavretsky, H., et al. (2004). Antidepressant pharmacotherapy in the treatment of depression in the very old: a randomized, placebo-controlled trial. *Am. J. Psychiatr.* 161, 2050–2059. doi: 10.1176/appi.ajp.161.11.2050

Ryan, J. P. (1983). "Introduction to latent trait analysis and item response theory" in *Testing in the schools new directions for testing and measurement*. ed. W. E. Hathaway (San Francisco, CA: Jossey-Bass), 48–64.

Saggino, A., Carlucci, L., Sergi, M. R., D'Ambrosio, I., Fairfield, B., Cera, N., et al. (2017). A validation study of the psychometric properties of the other as Shamer scale–2. *SAGE Open* 7, 1–10. doi: 10.1177/2158244017704241

Salzberger, T. (2010). Does the Rasch model convert an ordinal scale into an interval scale. *Rasch Meas. Trans.* 24, 1273–1275.

Santor, D. A., Ramsay, J., and Zuroff, D. C. (1994). Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychol. Assess.* 6:255.

Shea, T. L., Tennant, A., and Pallant, J. F. (2009). Rasch model analysis of the Depression, Anxiety and Stress Scales (DASS). *BMC Psychiatry* 9:21. doi: 10.1186/1471-244X-9-21

Sheikh, J. I., Cassidy, E. L., Doraiswamy, P. M., Salomon, R. M., Hornig, M., Holland, P. J., et al. (2004). Efficacy, safety, and tolerability of sertraline in patients with late-life depression and comorbid medical illness. *J. Am. Geriatr. Soc.* 52, 86–92. doi: 10.1111/j.1532-5415.2004.52015.x

Sijtsma, K., Meijer, R. R., and van der Ark, L. A. (2011). Mokken scale analysis as time goes by: an update for scaling practitioners. *J. Pers. Ind. Differ.* 50, 31–37. doi: 10.1016/j.paid.2010.08.016

Sijtsma, K., and Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Smith, J. E. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *J. Appl. Meas.* 3, 205–231.

Smith, R. M., and Miao, C. Y. (1994). Assessing unidimensionality for Rasch measurement. *Obj. Meas.* 2, 316–327.

Sözeri-Varma, G. (2012). Depression in the elderly: clinical features and risk factors. *Aging Dis.* 3, 465–471.

Spangenberg, L., Glaesmer, H., Boecker, M., and Forkmann, T. (2015). Differences in patient health questionnaire and Aachen depression item Bank scores between tablet versus paper-and-pencil administration. *Qual. Life Res.* 24, 3023–3032. doi: 10.1007/s11136-015-1040-5

Storch, E. A., Roberti, J. W., and Roth, D. A. (2004). Factor structure, concurrent validity, and internal consistency of the beck depression inventory—second edition in a sample of college students. *Depress. Anxiety* 19, 187–189. doi: 10.1002/da.20002

Swets, J. A., Dawes, R. M., and Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychol. Sci. Public Interest* 1, 1–26. doi: 10.1111/1529-1006.001

Tang, W. K., Wong, E., Chiu, H. F., Lum, C., and Ungvari, G. S. (2005). The Geriatric Depression Scale should be shortened: results of Rasch analysis. *Int. J. Geriatr. Psychiatry* 20, 783–789. doi: 10.1002/gps.1360

Tennant, A., and Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care Res.* 57, 1358–1362. doi: 10.1002/art.23108

Tennant, A., McKenna, S. P., and Hagell, P. (2004). Application of Rasch analysis in the development and application of quality of life instruments. *Value Health* 7, S22–S26. doi: 10.1111/j.1524-4733.2004.7s106.x

Therrien, Z., and Hunsley, J. (2012). Assessment of anxiety in older adults: a systematic review of commonly used measures. *Aging Ment. Health* 16, 1–16. doi: 10.1080/13607863.2011.602960

Thombs, B. D., Magyar-Russell, G., Bass, E. B., Stewart, K. J., Tsilidis, K. K., Bush, D. E., et al. (2007). Performance characteristics of depression screening instruments in survivors of acute myocardial infarction: review of the evidence. *Psychosomatics* 48, 185–194. doi: 10.1176/appi.psy.48.3.185

Tommasi, M., Ferrara, G., and Saggino, A. (2018). Application of Bayes' theorem in valuating depression tests performance. *Front. Psychol.* 9:1240. doi: 10.3389/fpsyg.2018.01240

Unützer, J., Katon, W., Callahan, C. M., Williams, J. W. Jr., Hunkeler, E., Harpole, L., et al. (2002). Collaborative care management of late-life depression in the primary care setting: a randomized controlled trial. *JAMA* 288, 2836–2845. doi: 10.1001/jama.288.22.2836

Unützer, J., Patrick, D. L., Diehr, P., Simon, G., Grembowski, D., and Katon, W. (2000). Quality adjusted life years in older adults with depressive symptoms and chronic medical disorders. *Int. Psychogeriatr.* 12, 15–33. doi: 10.1017/S1041610200006177

Van Dam, N. T., Gros, D. F., Earleywine, M., and Antony, M. M. (2013). Establishing a trait anxiety threshold that signals likelihood of anxiety disorders. *Anxiety Stress Coping* 26, 70–86. doi: 10.1080/10615806.2011.631525

Widiger, T. A., and Samuel, D. B. (2005). Diagnostic categories or dimensions? A question for the diagnostic and statistical manual of mental disorders. *J. Abnorm. Psychol.* 114, 494–504. doi: 10.1037/0021-843X.114.4.494

Wohlreich, M. M., Mallinckrodt, C. H., Watkin, J. G., and Hay, D. P. (2004). Duloxetine for the long-term treatment of major depressive disorder in patients aged 65 and older: an open-label study. *BMC Geriatr.* 4:11. doi: 10.1186/1471-2318-4-11

Wongpakaran, N., Wongpakaran, T., and Kuntawong, P. (2019). Evaluating hierarchical items of the geriatric depression scale through factor analysis and item response theory. *Heliyon* 5:e02300. doi: 10.1016/j.heliyon.2019.e02300

Wright, B. D. (1996). Local dependency, correlations and principal components. *Rasch Meas. Trans.* 10, 509–511.

Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., Adey, M., et al. (1982). Development and validation of a geriatric depression screening scale: a preliminary report. *J. Psychiatr. Res.* 17, 37–49.

Yesavage, J. A., and Sheikh, J. I. (1986). 9/geriatric depression scale (GDS) recent evidence and development of a shorter version. *Clin. Gerontol.* 5, 165–173.

Zhou, X. H., Obuchowski, N. A., and McClish, D. K. (2002). *Statistical methods in diagnostic medicine*. New York, NY: Wiley & Sons.

Ziegler, M., and Hagemann, D. (2015). Testing the unidimensionality of items. *EJPA* 31, 231–237. doi: 10.1027/1015-5759/a000309

Zung, W. W. (1965). A self-rating depression scale. *Arch. Gen. Psychiatry* 12, 63–70. doi: 10.1001/archpsyc.1965.01720310065008

Conflict of Interest: We declare a potential conflict of interests for some authors, those who have published the handbook for one of the tests analyzed in the present report (AS and MB–please see "References" section).

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 10 December 2019
doi: 10.3389/fpsyg.2019.02767

Indices of Effect Existence and Significance in the Bayesian Framework

Dominique Makowski[1]*, Mattan S. Ben-Shachar[2], S. H. Annabel Chen[1,3,4]*† and Daniel Lüdecke[5]†

[1] School of Social Sciences, Nanyang Technological University, Singapore, Singapore, [2] Department of Psychology, Ben-Gurion University of the Negev, Beersheba, Israel, [3] Centre for Research and Development in Learning, Nanyang Technological University, Singapore, Singapore, [4] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore, [5] Department of Medical Sociology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Turmoil has engulfed psychological science. Causes and consequences of the reproducibility crisis are in dispute. With the hope of addressing some of its aspects, Bayesian methods are gaining increasing attention in psychological science. Some of their advantages, as opposed to the frequentist framework, are the ability to describe parameters in probabilistic terms and explicitly incorporate prior knowledge about them into the model. These issues are crucial in particular regarding the current debate about statistical significance. Bayesian methods are not necessarily the only remedy against incorrect interpretations or wrong conclusions, but there is an increasing agreement that they are one of the keys to avoid such fallacies. Nevertheless, its flexible nature is its power and weakness, for there is no agreement about what indices of "significance" should be computed or reported. This lack of a consensual index or guidelines, such as the frequentist *p*-value, further contributes to the unnecessary opacity that many non-familiar readers perceive in Bayesian statistics. Thus, this study describes and compares several Bayesian indices, provide intuitive visual representation of their "behavior" in relationship with common sources of variance such as sample size, magnitude of effects and also frequentist significance. The results contribute to the development of an intuitive understanding of the values that researchers report, allowing to draw sensible recommendations for Bayesian statistics description, critical for the standardization of scientific reporting.

Keywords: Bayesian, significance, NHST, *p*-value, Bayes factors

OPEN ACCESS

Edited by:
Pietro Cipresso,
Istituto Auxologico Italiano (IRCCS),
Italy

Reviewed by:
Richard S. John,
University of Southern California,
United States
Jose D. Perezgonzalez,
Massey University Business School,
New Zealand

Correspondence:
Dominique Makowski
dmakowski@ntu.edu.sg;
dom.makowski@gmail.com
S. H. Annabel Chen
annabelchen@ntu.edu.sg

†These authors share senior authorship

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *18 September 2019*
Accepted: *25 November 2019*
Published: *10 December 2019*

Citation:
Makowski D, Ben-Shachar MS, Chen SHA and Lüdecke D (2019) Indices of Effect Existence and Significance in the Bayesian Framework. Front. Psychol. 10:2767. doi: 10.3389/fpsyg.2019.02767

INTRODUCTION

The Bayesian framework is quickly gaining popularity among psychologists and neuroscientists (Andrews and Baguley, 2013), for reasons such as flexibility, better accuracy in noisy data and small samples, less proneness to type I errors, the possibility of introducing prior knowledge into the analysis and the intuitiveness and straightforward interpretation of results (Kruschke, 2010; Kruschke et al., 2012; Etz and Vandekerckhove, 2016; Wagenmakers et al., 2016, 2018; Dienes and Mclatchie, 2018). On the other hand, the frequentist approach has been associated with the focus on *p*-values and null hypothesis significance testing (NHST). The misinterpretation and misuse of *p*-values, so called "p-hacking" (Simmons et al., 2011), has been shown to critically contribute to the reproducibility crisis in psychological science (Chambers et al., 2014; Szucs and Ioannidis, 2016). The reliance on *p*-values has been criticized for its association with inappropriate inference, and effects can be drastically overestimated, sometimes even in the wrong direction, when estimation is tied to statistical significance in highly variable data (Gelman, 2018). Power calculations allow researchers to control the probability of falsely rejecting the null hypothesis, but do not completely solve this problem. For instance, the "false-alarm probability" of even very small *p*-values can be much higher than expected (Nuzzo, 2014). In response, there is an increasing belief that the generalization and utilization of the Bayesian framework is one way of overcoming these issues (Maxwell et al., 2015; Etz and Vandekerckhove, 2016; Marasini et al., 2016; Wagenmakers et al., 2017; Benjamin et al., 2018; Halsey, 2019).

The tenacity and resilience of the *p*-value as an index of significance is remarkable, despite the long-lasting criticism and discussion about its misuse and misinterpretation (Gardner and Altman, 1986; Cohen, 1994; Anderson et al., 2000; Fidler et al., 2004; Finch et al., 2004). This endurance might be informative on how such indices, and the accompanying heuristics applied to interpret them (e.g., assigning thresholds like 0.05, 0.01, and 0.001 to certain levels of significance), are useful and necessary for researchers to gain an intuitive (although possibly simplified) understanding of the interactions and structure of their data. Moreover, the utility of such an index is most salient in contexts where decisions must be made and rationalized (e.g., in medical settings). Unfortunately, these heuristics can become severely rigidified, and meeting significance has become a goal unto itself rather than a tool for understanding the data (Cohen, 1994; Kirk, 1996). This is particularly problematic given that *p*-values can only be used to reject the null hypothesis and not to accept it as true, because a statistically non-significant result does not mean that there is no difference between groups or no effect of a treatment (Wagenmakers, 2007; Amrhein et al., 2019).

While significance testing (and its inherent categorical interpretation heuristics) might have its place as a complementary perspective to effect estimation, it does not preclude the fact that improvements are needed. For instance, one possible advance could focus on improving the understanding of the values being used, for instance, through a new, simpler, index. Bayesian inference allows making intuitive probability statements of an effect, as opposed to the less straightforward mathematical definition of the *p*-value, that contributes to its common misinterpretation. Another improvement could be found in providing an intuitive understanding (e.g., by visual means) of the behavior of the indices in relationship with main sources of variance, such as sample size, noise, or effect presence. Such better overall understanding of the indices would hopefully act as a barrier against their mindless reporting by allowing the users to nuance the interpretations and conclusions that they draw.

The Bayesian framework offers several alternative indices for the *p*-value. To better understand these indices, it is important to point out one of the core differences between Bayesian and frequentist methods. From a frequentist perspective, the effects are fixed (but unknown) and data are random. On the other hand, instead of having single estimates of some "true effect" (for instance, the "true" correlation between *x* and *y*), Bayesian methods compute the probability of different effects values *given* the observed data (and some prior expectation), resulting in a distribution of possible values for the parameters, called the posterior distribution. The description of the posterior distribution (e.g., through its centrality, dispersion, etc.) allows to draw conclusions from Bayesian analyses.

Bayesian "significance" testing indices could be roughly grouped into three overlapping categories: Bayes factors, posterior indices and Region of Practical Equivalence (ROPE)-based indices. Bayes factors are a family of indices of relative evidence of one model over another (e.g., the null vs. the alternative hypothesis; Jeffreys, 1998; Ly et al., 2016). Aside from having a straightforward interpretation ("given the observed data, is the null hypothesis of an absence of an effect more, or less likely?"), they allow to quantify the evidence in favor of the null hypothesis (Dienes, 2014; Jarosz and Wiley, 2014). However, its use for parameters description in complex models is still a matter of debate (Wagenmakers et al., 2010; Heck, 2019), being highly dependent on the specification of priors (Etz et al., 2018; Kruschke and Liddell, 2018). On the contrary, "posterior indices" reflect objective characteristics of the posterior distribution, for instance the proportion of strictly positive values. They also allow to derive legitimate statements that indicate the probability of an effect falling in a given range similar to the misleading conclusions related to frequentist confidence intervals. Finally, ROPE-based indices are related to the redefinition of the null hypothesis from the classic point-null hypothesis to a range of values considered negligible or too small to be of any practical relevance (the Region of Practical Equivalence – ROPE; Kruschke, 2014; Lakens, 2017; Lakens et al., 2018), usually spread equally around 0 (e.g., [−0.1; 0.1]). The idea behind this index is that an effect is almost never exactly zero, but instead can be very tiny, with no practical relevance. It is interesting to note that this perspective unites significance testing with the focus on effect size (involving a discrete separation between at least two categories: negligible and non-negligible), which finds an echo in recent statistical recommendations (Ellis and Steyn, 2003; Sullivan and Feinn, 2012; Simonsohn et al., 2014).

Despite the richness provided by the Bayesian framework and the availability of multiple indices, no consensus has yet emerged on which ones to be used. Literature continues to bloom in a raging debate, often polarized between proponents of the Bayes factor as the supreme index and its detractors (Spanos, 2013; Robert, 2014, 2016; Wagenmakers et al., 2019), with strong theoretical arguments being developed on both sides. Yet no practical, empirical and direct comparison between these indices has been done. This might be a deterrent for scientists interested in adopting the Bayesian framework. Moreover, this gray area can increase the difficulty of readers or reviewers unfamiliar with the Bayesian framework to follow the assumptions and conclusions, which could in turn generate unnecessary doubt upon an entire study. While we think that such indices of significance and their interpretation guidelines (in the form of rules of thumb) are useful in practice, we also strongly believe that they should be accompanied with the understanding of their "behavior" in relationship with major sources of variance, such as sample size, noise or effect presence. This knowledge is important for people to implicitly and intuitively appraise the meaning and implication of the mathematical values they report. Such an understanding could prevent the crystallization of the possible heuristics and categories derived from such indices, as has unfortunately occurred for the *p*-values.

Thus, based on the simulation of linear and logistic regressions (arguably some of the most widely used models in the psychological sciences), the present work aims at comparing several indices of effect "significance," provide visual representations of the "behavior" of such indices in relationship with sample size, noise and effect presence, as well as their relationship to frequentist *p*-values (an index which, beyond its many flaws, is well known and could be used as a reference for Bayesian neophytes), and finally draw recommendations for Bayesian statistics reporting.

MATERIALS AND METHODS

Data Simulation

We simulated datasets suited for linear and logistic regression and started by simulating an independent, normally distributed *x* variable (with mean 0 and SD 1) of a given sample size. Then, the corresponding *y* variable was added, having a perfect correlation (in the case of data for linear regressions) or as a binary variable perfectly separated by *x*. The case of no effect was simulated by creating a *y* variable that was independent of (i.e., not correlated to) *x*. Finally, a Gaussian noise (the error) was added to the x variable before its standardization, which in turn decreases the standardized coefficient (the effect size).

The simulation aimed at modulating the following characteristics: *outcome type* (linear or logistic regression), *sample size* (from 20 to 100 by steps of 10), *null hypothesis* (original regression coefficient from which data is drawn prior to noise addition, 1 – presence of "true" effect, or 0 – absence of "true" effect) and *noise* (Gaussian noise applied to the predictor with SD uniformly spread between 0.666 and 6.66, with 1000 different values), which is directly related to the absolute value of the coefficient (i.e., the effect size). We generated a dataset for each combination of these characteristics, resulting in a total of 36,000 (2 model types × 2 presence/absence of effect × 9 sample sizes × 1,000 noise variations) datasets. The code used for data generation is available on GitHub[1]. Note that it takes usually several days/weeks for the generation to complete.

Indices

For each of these datasets, Bayesian and frequentist regressions were fitted to predict *y* from *x* as a single unique predictor. We then computed the following seven indices from all simulated models (see **Figure 1**), related to the effect of *x*.

[1] https://github.com/easystats/easystats/tree/master/publications/makowski_2019_bayesian/data

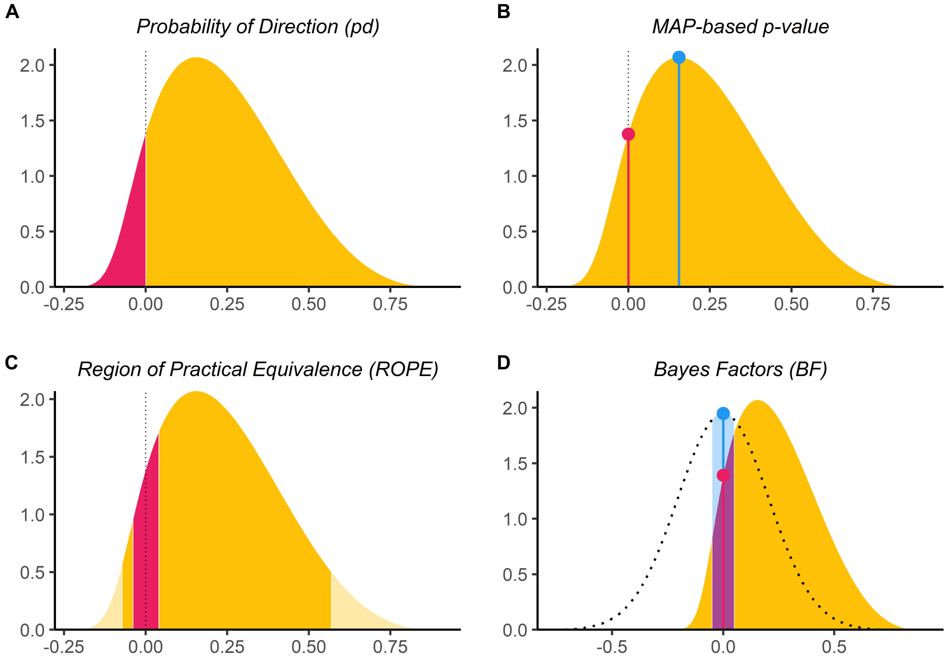


FIGURE 1 | Bayesian indices of effect existence and significance. **(A)** The probability of Direction (*pd*) is defined as the proportion of the posterior distribution that is of the median's sign (the size of the yellow area relative to the whole distribution). **(B)** The MAP-based *p*-value is defined as the density value at 0 – the height of the red lollipop, divided by the density at the Maximum *A Posteriori* (MAP) – the height of the blue lollipop. **(C)** The percentage in ROPE corresponds to the red area relative to the distribution [with or without tails for ROPE (*full*) and ROPE (95%), respectively]. **(D)** The Bayes factor (vs. 0) corresponds to the point-null density of the prior (the blue lollipop on the dotted distribution) divided by that of the posterior (the red lollipop on the yellow distribution), and the Bayes factor (vs. ROPE) is calculated as the odds of the prior falling within vs. outside the ROPE (the blue area on the dotted distribution) divided by that of the posterior (the red area on the yellow distribution).

Frequentist *p*-Value

This was the only index computed by the frequentist version of the regression. The *p*-value represents the probability that for a given statistical model, when the null hypothesis is true, the effect would be greater than or equal to the observed coefficient (Wasserstein and Lazar, 2016).

Probability of Direction (*pd*)

The *Probability of Direction (pd)* varies between 50 and 100% and can be interpreted as the probability that a parameter (described by its posterior distribution) is strictly positive or negative (whichever is the most probable). It is mathematically defined as the proportion of the posterior distribution that is of the median's sign (Makowski et al., 2019).

MAP-Based *p*-Value

The *MAP-based p-value* is related to the odds that a parameter has against the null hypothesis (Mills and Parent, 2014; Mills, 2017). It is mathematically defined as the density value at 0 divided by the density at the Maximum *A Posteriori* (MAP), i.e., the equivalent of the mode for continuous distributions.

ROPE (95%)

The *ROPE* (95%) refers to the percentage of the 95% Highest Density Interval (HDI) that lies within the ROPE. As suggested by Kruschke (2014), the Region of Practical Equivalence (ROPE) was defined as range from −0.1 to 0.1 for linear regressions and its equivalent, −0.18 to 0.18, for logistic models (based on the $\pi/\sqrt{3}$ formula to convert log odds ratios to standardized differences; Cohen, 1988). Although we present the "95% percentage" because of the history of this index and of its widespread use, the reader should note that this value was recently challenged due to its arbitrary nature (McElreath, 2018).

ROPE (Full)

The *ROPE (full)* is similar to *ROPE (95%)*, with the exception that it refers to the percentage of the *whole* posterior distribution that lies within the ROPE.

Bayes Factor (vs. 0)

The Bayes Factor (*BF*) used here is based on prior and posterior distributions of a single parameter. In this context, the Bayes factor indicates the degree by which the mass of the posterior distribution has shifted further away from or closer to the null value (0), relative to the prior distribution, thus indicating if the null hypothesis has become less or more likely given the observed data. The *BF* was computed as a Savage-Dickey density ratio, which is also an approximation of a Bayes factor comparing the marginal likelihoods of the model against a model in which the tested parameter has been restricted to the point-null (Wagenmakers et al., 2010).

Bayes Factor (vs. ROPE)

The *Bayes factor* (vs. *ROPE*) is similar to the *Bayes factor* (vs. 0), but instead of a point-null, the null hypothesis is a range of negligible values (defined here same as for the ROPE indices). The *BF* was computed by comparing the prior and posterior odds of the parameter falling within vs. outside the ROPE (see *Non-overlapping Hypotheses* in Morey and Rouder, 2011). This measure is closely related to the *ROPE (full)*, as it can be formally defined as the ratio between the *ROPE (full)* odds for the posterior distribution and the *ROPE (full)* odds for the prior distribution:

$$BF_{\mathrm{ROPE}} = \frac{\mathrm{odds}(\mathrm{ROPE}_{\text{full posterior}})}{\mathrm{odds}(\mathrm{ROPE}_{\text{full prior}})}$$

Data Analysis

In order to achieve the two-fold aim of this study; (1) comparing Bayesian indices and (2) provide visual guides for an intuitive understanding of the numeric values in relation to a known frame of reference (the frequentist *p*-value), we will start by presenting the relationship between these indices and main sources of variance, such as sample size, noise and null hypothesis (true if absence of effect, false if presence of effect). We will then compare Bayesian indices with the frequentist *p*-value and its commonly used thresholds (0.05, 0.01, 0.001). Finally, we will show the mutual relationship between three recommended Bayesian candidates. Taken together, these results will help us outline guides to ease the reporting and interpretation of the indices.

In order to provide an intuitive understanding of values, data processing will focus on creating clear visual figures to help the user grasp the patterns and variability that exists when computing the investigated indices. Nevertheless, we decided to also mathematically test our claims in cases where the graphical representation begged for a deeper investigation. Thus, we fitted two regression models to assess the impact of sample size and noise, respectively. For these models (but not for the figures), to ensure that any differences between the indices are not due to differences in their scale or distribution, we converted all indices to the same scale by normalizing the indices between 0 and 1 (note that *BF*s were transformed to posterior probabilities, assuming uniform prior odds) and reversing the *p*-values, the MAP-based *p*-values and the ROPE indices so that a higher value corresponds to stronger "significance."

The statistical analyses were conducted using R (R Core Team, 2019). Computations of Bayesian models were done using the *rstanarm* package (Goodrich et al., 2019), a wrapper for Stan probabilistic language (Carpenter et al., 2017). We used Markov Chain Monte Carlo sampling (in particular, Hamiltonian Monte Carlo; Gelman et al., 2014) with 4 chains of 2000 iterations, half of which used for warm-up. Mildly informative priors (a normal distribution with mean 0 and SD 1) were used for the parameter in all models. The Bayesian indices were calculated using the *bayestestR* package (Makowski et al., 2019).

RESULTS

Impact of Sample Size

Figure 2 shows the sensitivity of the indices to sample size. The *p*-value, the *pd* and the MAP-based *p*-value are sensitive to sample size only in case of the presence of a true effect (when the null hypothesis is false). When the null hypothesis is true, all three indices are unaffected by sample size. In other words, these indices reflect the amount of observed evidence (the sample

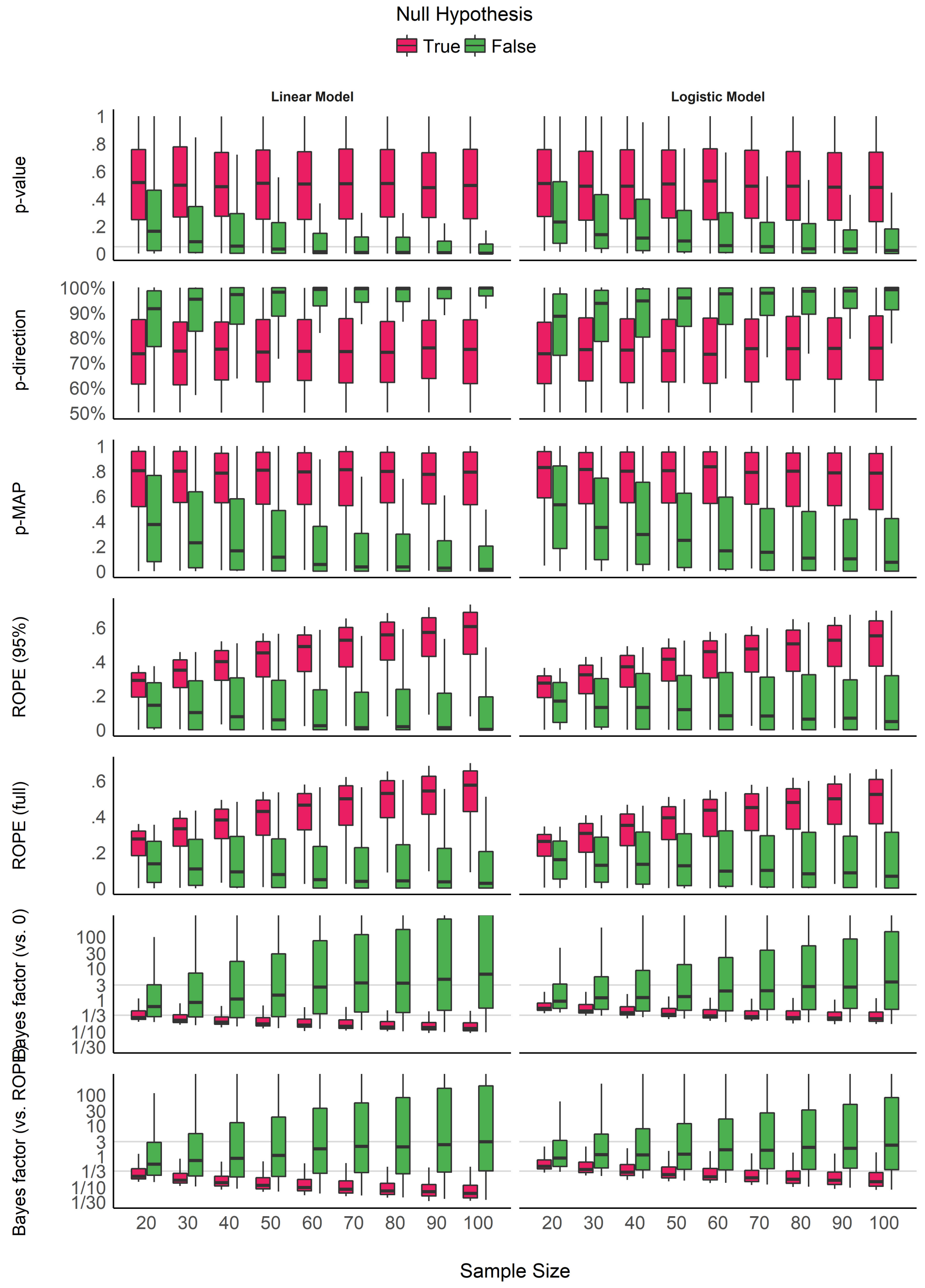


FIGURE 2 | Impact of sample size on the different indices, for linear and logistic models, and when the null hypothesis is true or false. Gray vertical lines for *p*-values and Bayes factors represent commonly used thresholds.

size) for the presence of an effect (i.e., against the null hypothesis being true), but not for the absence of an effect. The *ROPE* indices, however, appear as strongly modulated by the sample size when there is no effect, suggesting their sensitivity to the amount of evidence for the absence of effect. Finally, the figure suggests that *BFs* are sensitive to sample size for both presence and absence of true effect.

Consistently with **Figure 2** and **Table 1**, the model investigating the sensitivity of sample size on the different indices suggests that *BF* indices are sensitive to sample size both when an effect is present (null hypothesis is false) and absent (null hypothesis is true). *ROPE* indices are particularly sensitive to sample size when the null hypothesis is true, while *p*-value, *pd* and MAP-based *p*-value are only sensitive to sample size when the null hypothesis is false, in which case they are more sensitive than *ROPE* indices. These findings can be related to the concept of consistency: as the number of data points increases, the statistic converges toward some "true" value. Here, we observe that *p*-value, *pd* and the MAP-based *p*-value are consistent only when the null hypothesis is false. In other words, as sample size increases, they tend to reflect more strongly that the effect is present. On the other hand, *ROPE* indices appear as consistent when the effect is absent. Finally, *BFs* are consistent both when the effect is absent and when it is present, and *BF* (vs. *ROPE*), compared to *BF* (vs. 0), is more sensitive to sample size when the null hypothesis is true, and *ROPE (full)* is overall slightly more consistent than *ROPE* (95%).

Impact of Noise

Figure 3 shows the indices' sensitivity to noise. Unlike the patterns of sensitivity to sample size, the indices display more similar patterns in their sensitivity to noise (or magnitude of effect). All indices are unidirectional impacted by noise: as noise increases, the observed coefficients decrease in magnitude, and the indices become less "pronounced" (respectively to their direction). However, it is interesting to note that the variability of the indices seems differently impacted by noise. For the *p*-values, the *pd* and the ROPE indices, the variability increases as the noise increases. In other words, small variation in small observed coefficients can yield very different values. On the contrary, the variability of BFs decreases as the true effect tends toward 0. For the MAP-based *p*-value, the variability appears to be the highest for moderate amount of noise. This behavior seems consistent across model types.

Consistently with **Figure 3** and **Table 2**, the model investigating the sensitivity of noise when an effect is present (as there is only noise in the absence of effect), adjusted for sample size, suggests that BFs (especially vs. ROPE), followed by the MAP-based *p*-value and percentages in *ROPE*, are the most sensitive to noise. As noise is a proxy of effect size (linearly related to the absolute value of the coefficient of the parameter), this result highlights the fact that these indices are sensitive to the magnitude of the effect. For example, as noise increases, evidence for an effect becomes weak, and data seems to support the absence of an effect (or at the very least the presence of a negligible effect), which is reflected in *BFs* being consistently smaller than 1. On the other hand, as the *p*-value and the *pd* quantify evidence only for the presence of an effect, as noise increases, they are become more dependent on larger sample size to be able to detect the presence of an effect.

Relationship With the Frequentist *p*-Value

Figure 4 suggests that the *pd* has a 1:1 correspondence with the frequentist *p*-value (through the formula $p_{\text{two-sided}} = 2 \times (1 - p_d)$). *BF* indices still appear as having a severely non-linear relationship with the frequentist index, mostly due to the fact that smaller *p*-values correspond to stronger evidence in favor of the presence of an effect, but the reverse is not true. *ROPE*-based percentages appear to be only weakly related to *p*-values. Critically, their relationship seems to be strongly dependent on sample size.

Figure 5 shows equivalence between *p*-value thresholds (0.1, 0.05, 0.01, 0.001) and the Bayesian indices. As expected, the *pd* has the sharpest thresholds (95, 97.5, 99.5, and 99.95%, respectively). For logistic models, these threshold points appear as more conservative (i.e., Bayesian indices have to be more "pronounced" to reach the same level of significance). This sensitivity to model type is the strongest for BFs (which is possibly related to the difference in the prior specification for these two types of models).

Relationship Between ROPE (Full), *pd*, and *BF* (vs. ROPE)

Figure 6 suggests that the relationship between the *ROPE (full)* and the *pd* might be strongly affected by the sample size, and subject to differences across model types. This seems to echo the relationship between *ROPE (full)* and *p*-value, the latter having a 1:1 correspondence with *pd*. On the other hand, the *ROPE (full)* and the *BF* (vs. *ROPE*) seem very closely related within the same model type, reflecting their formal relationship [see definition of *BF* (vs. *ROPE*) above]. Overall, these results help to demonstrate *ROPE (full)* and *BF (vs. ROPE)*'s consistency both in case of presence and absence of a true effect, whereas the *pd*, being equivalent to the *p*-value, is only consistent when the true effect is absent.

DISCUSSION

Based on the simulation of linear and logistic models, the present work aimed to compare several Bayesian indices of effect "significance" (see **Table 3**), providing visual representations of the "behavior" of such indices in relationship with important sources of variance such as sample size, noise and effect presence, as well as comparing them with the well-known and widely used frequentist *p*-value.

The results tend to suggest that the investigated indices could be separated into two categories. The first group, including the *pd* and the MAP-based *p*-value, presents similar properties to those of the frequentist *p*-value: they are sensitive only to the amount of evidence for the alternative hypothesis (i.e., when an effect is truly present). In other words, these indices are not able to reflect the amount of evidence in favor of the null hypothesis (Rouder et al., 2009; Rouder and Morey, 2012). A high value suggests that the effect exists, but a low value indicates *uncertainty* regarding its existence (but not certainty that it is non-existent). The second group, including ROPE and Bayes factors, seem sensitive to both presence and absence of effect, accumulating evidence as the sample size increases. However, ROPE seems particularly suited to provide evidence in favor of the null hypothesis. Consistent with this, combining Bayes factors with ROPE (BF vs. ROPE), as compared to Bayes factors against the point-null (BF vs. 0), leads

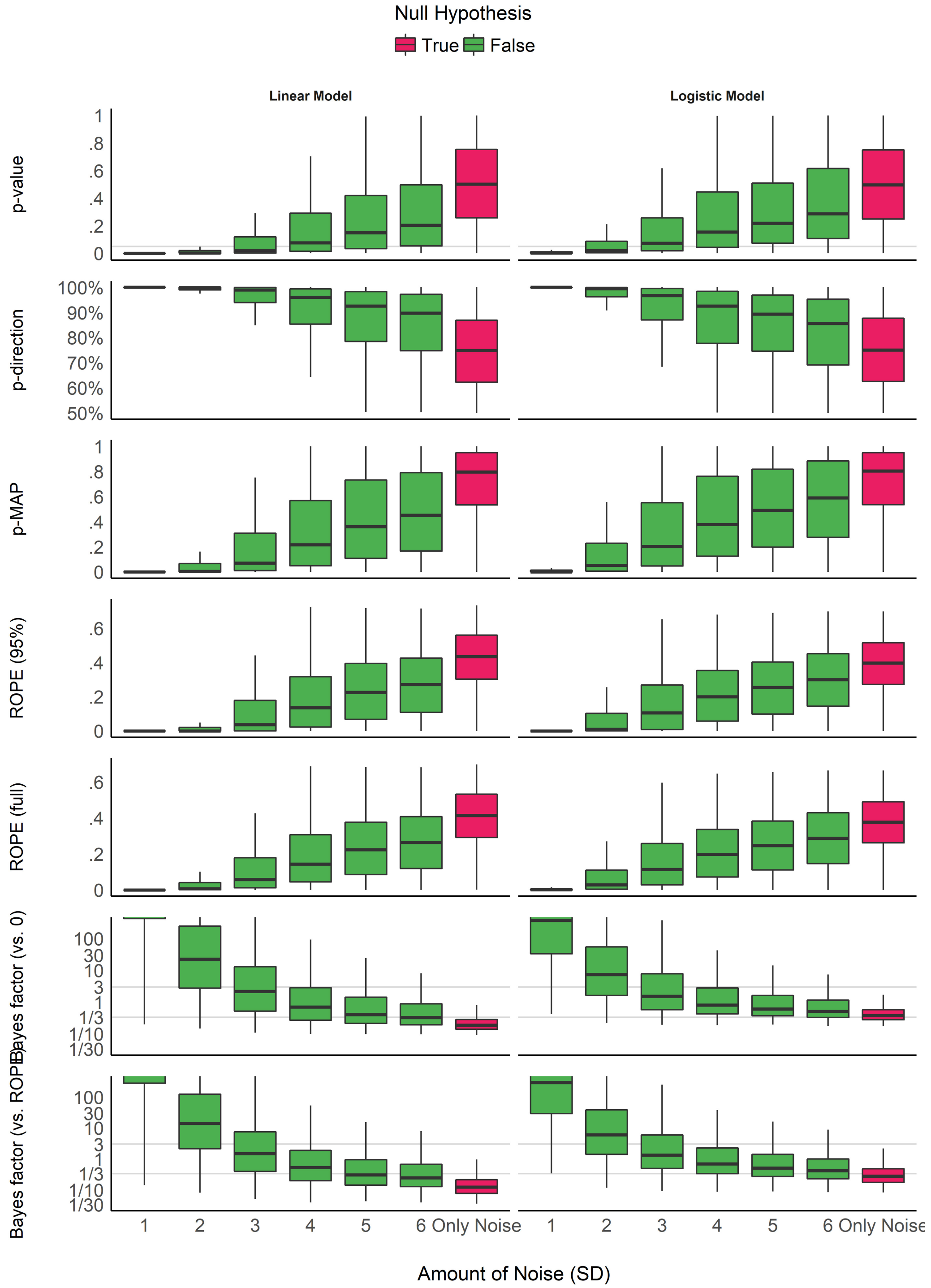


FIGURE 3 | Impact of noise. The noise corresponds to the standard deviation of the Gaussian noise that was added to the generated data. It is related to the magnitude of the parameter (the more noise there is, the smaller the coefficient). Gray vertical lines for *p*-values and Bayes factors represent commonly used thresholds. The scale is capped for the Bayes factors as these extend to infinity.

to a higher sensitivity to null-effects (Morey and Rouder, 2011; Rouder and Morey, 2012).

We also showed that besides sharing similar properties, the *pd* has a 1:1 correspondence with the frequentist *p*-value, being its Bayesian equivalent. Bayes factors, however, appear to have a severely non-linear relationship with the frequentist index, which is to be expected from their mathematical definition and their sensitivity when the null hypothesis is true. This in turn

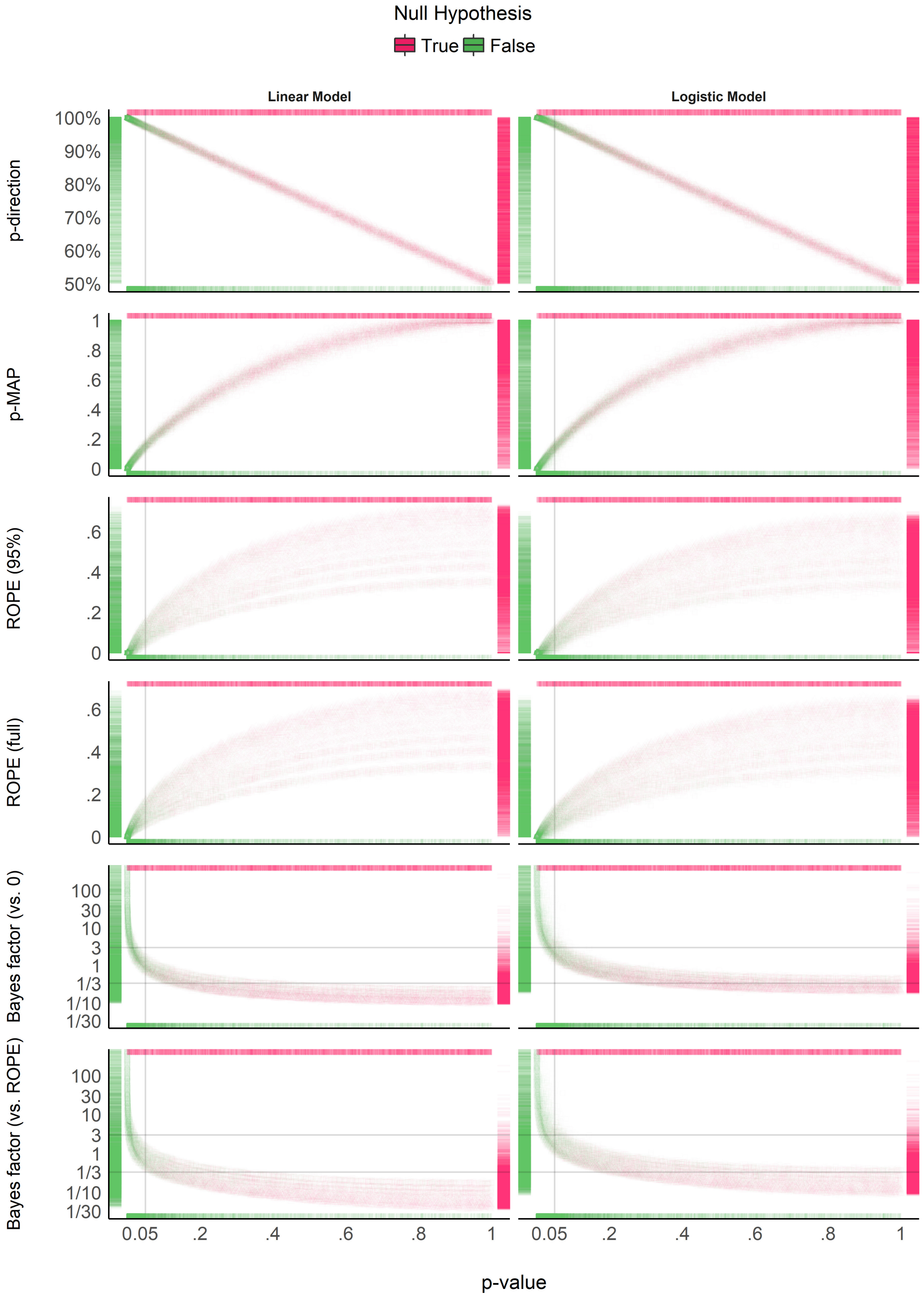


FIGURE 4 | Relationship with the frequentist *p*-value. In each plot, the *p*-value densities are visualized by the marginal **top** (absence of true effect) and **bottom** (presence of true effect) markers, whereas on the **left** (presence of true effect) and **right** (absence of true effect), the markers represent the density of the index of interest. Different point shapes, representing different sample sizes, specifically illustrate its impact on the percentages in ROPE, for which each "curve line" is associated with one sample size (the bigger the sample size, the higher the percentage in ROPE).

can lead to surprising conclusions. For instance, Bayes factors lower than 1, which are considered as providing evidence *against* the presence of an effect, can still correspond to a "significant" frequentist *p*-value (see **Figures 3**, **4**). ROPE indices are more closely related to the *p*-value, as their relationship appears dependent on another factor: the sample size. This suggests

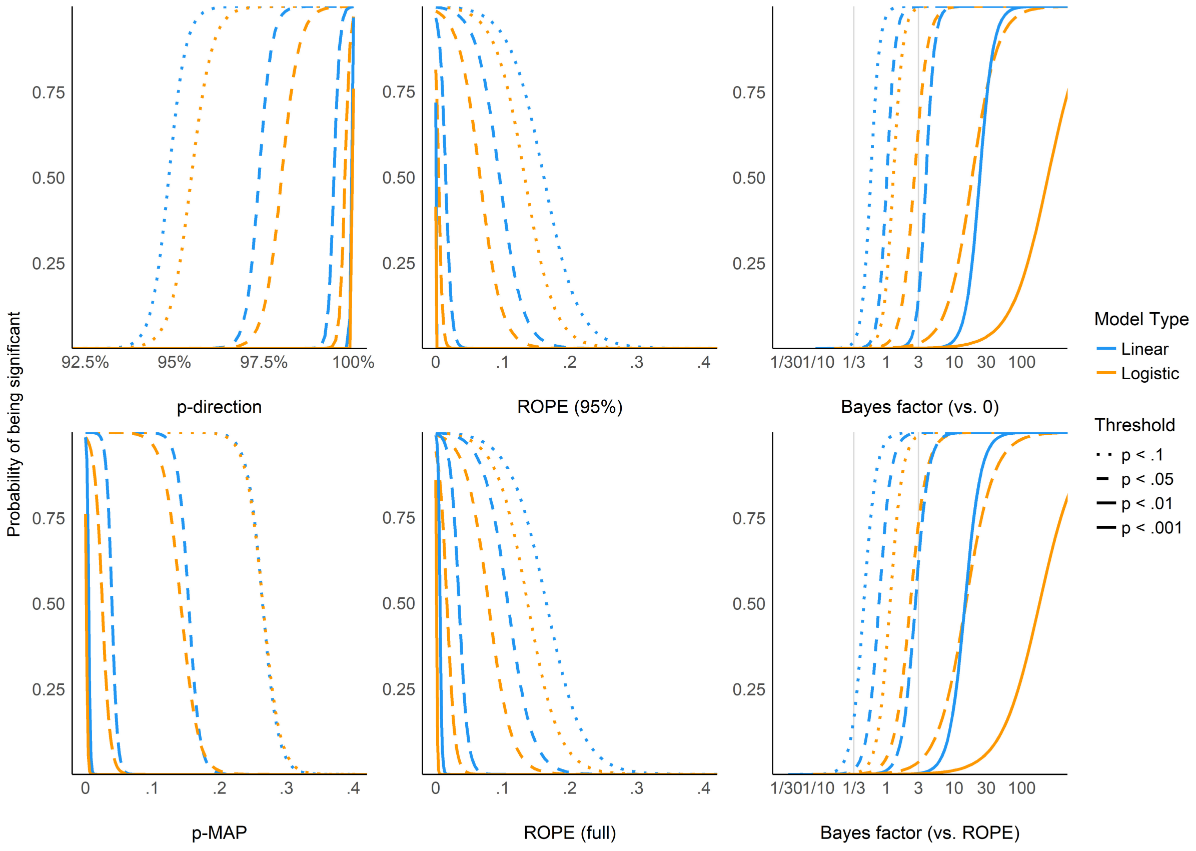


FIGURE 5 | The probability of reaching different *p*-value based significance thresholds (0.1, 0.05, 0.01, 0.001 for solid, long-dashed, short-dashed, and dotted lines, respectively) for different values of the corresponding Bayesian indices.

that the ROPE encapsulates additional information about the strength of evidence.

What is the point of comparing Bayesian indices with the frequentist *p*-value, especially after having pointed out its many flaws? While this comparison may seem counter-intuitive (as Bayesian thinking is intrinsically different from the frequentist framework), we believe that this juxtaposition is interesting for didactic reasons. The frequentist *p*-value "speaks" to many and can thus be seen as a reference and a way to facilitate the shift toward the Bayesian framework. Thus, pragmatically documenting such bridges can only foster the understanding of the methodological issues that our field is facing, and in turn act against dogmatic adherence to a framework. This does not preclude, however, that a change in the general paradigm of significance seeking and "p-hacking" is necessary, and that Bayesian indices are fundamentally different from the frequentist *p*-value, rather than mere approximations or equivalents.

Critically, while the purpose of these indices was solely referred to as *significance* until now, we would like to emphasize the nuanced perspective of existence-significance testing as a dual-framework for parameter description and interpretation. The idea supported here is that there is a conceptual and practical distinction, and possible dissociation to be made, between an effect's existence *and* its significance. In this context, *existence* is simply defined as the consistency of an effect in one particular direction (i.e., positive or negative), without any assumptions or conclusions as to its size, importance, relevance or meaning. It is an objective feature of an estimate (tied to its uncertainty). On the other hand, *significance* would be here re-framed following its original literally definition such as "being worthy of attention" or "importance." An effect can be considered significant if its magnitude is higher than some given threshold. This aspect can be explored, to a certain extent, in an objective way with the concept of *practical equivalence* (Kruschke, 2014; Lakens, 2017; Lakens et al., 2018), which suggests the use of a range of values assimilated to the absence of an effect (ROPE). If the effect falls within this range, it is considered to be non-significant *for practical reasons*: the magnitude of the effect is likely to be too small to be of high importance in real-world scenarios or applications. Nevertheless, *significance* also withholds a more subjective aspect, corresponding to its contextual meaningfulness and relevance. This, however, is usually dependent on the literature, priors, novelty, context or field, and thus cannot be objectively or neutrally assessed using a statistical index alone.

While indices of existence and significance can be numerically related (as shown in our results), the former is conceptually independent from the latter. For example, an effect for which the whole posterior distribution is concentrated within the

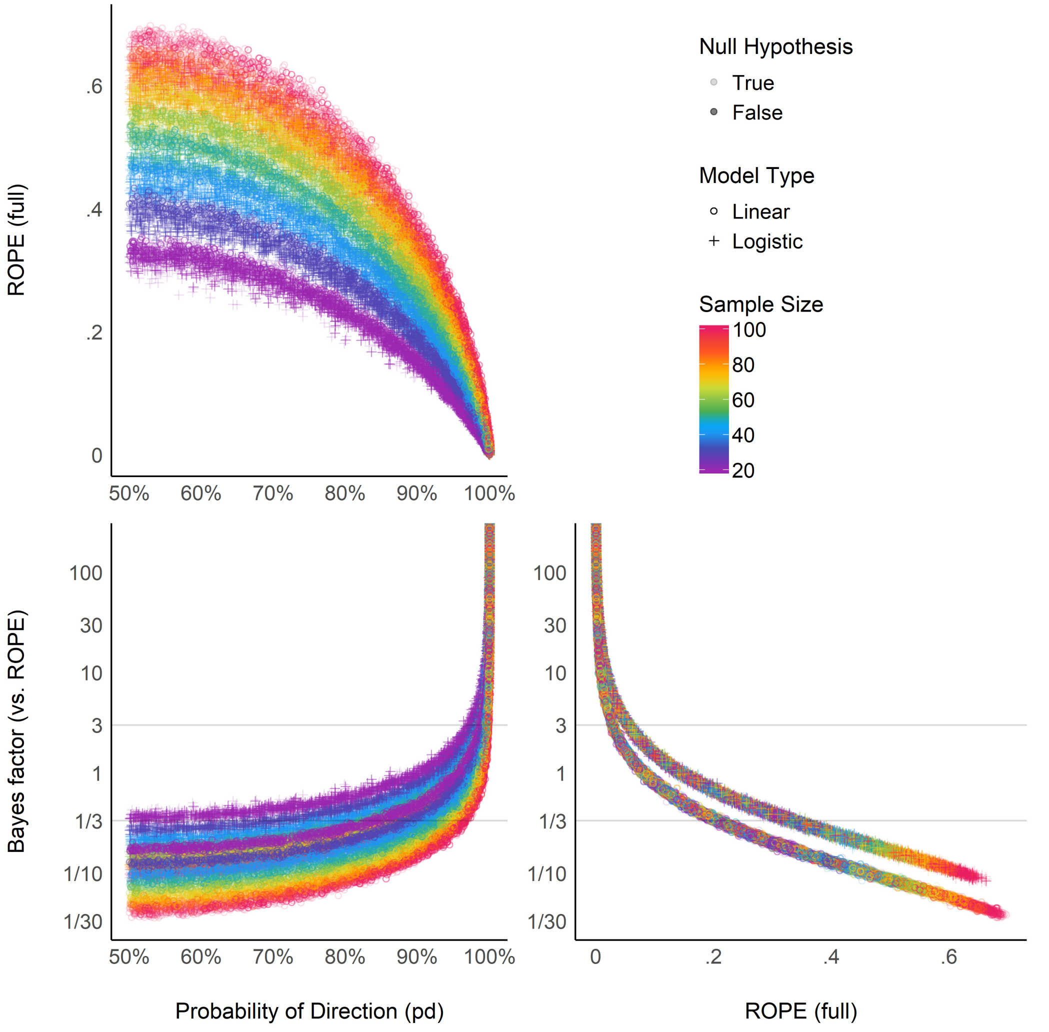


FIGURE 6 | Relationship between three Bayesian indices: the probability of direction (*pd*), the percentage of the full posterior distribution in the ROPE, and the Bayes factor (vs. ROPE).

TABLE 1 | Sensitivity to sample size.

Index	Linear models/presence of effect	Linear models/absence of effect	Logistic models/presence of effect	Logistic models/absence of effect
p-value	0.166	0.008	0.157	0.020
p-direction	0.171	0.013	0.154	0.024
p-MAP	0.239	0.002	0.238	0.032
ROPE (95%)	0.033	0.359	0.008	0.310
ROPE (full)	0.025	0.363	0.016	0.315
Bayes factor (vs. 0)	0.198	0.116	0.116	0.141
Bayes factor (vs. ROPE)	0.152	0.136	0.078	0.180

This table shows the standardized coefficient between the sample size and the value of each index, adjusted for error, and stratified by model type and presence of true effect. The stronger the coefficient is, the stronger the relationship with sample size.

[0.0001, 0.0002] range would be considered to be positive with a high level of certainty (and thus, *existing* in that direction), but also not significant (i.e., too small to be of any practical relevance). Acknowledging the distinction and complementary nature of these two aspects can in turn enrich the information and usefulness of the results reported in psychological science (for practical reasons, the implementation of this dual-framework of existence-significance testing is made straightforward through

TABLE 2 | Sensitivity to noise.

Index	Linear models/presence of effect	Logistic models/presence of effect
p-value	0.35	0.40
p-direction	0.36	0.40
p-MAP	0.55	0.60
ROPE (95%)	0.45	0.45
ROPE (full)	0.46	0.45
Bayes factor (vs. 0)	0.79	0.65
Bayes factor (vs. ROPE)	0.81	0.67

This table shows the standardized coefficient between noise and the value of each index when the true effect is present, adjusted for sample size and stratified by model type. The stronger the coefficient is, the stronger the relationship with noise.

the *bayestestR* open-source package for R; Makowski et al., 2019). In this context, the *pd* and the MAP-based *p*-value appear as indices of effect existence, mostly sensitive to the certainty related to the direction of the effect. ROPE-based indices and Bayes factors are indices of effect significance, related to the magnitude and the amount of evidence in favor of it (see also a similar discussion of statistical significance vs. effect size in the frequentist framework; e.g., Cohen, 1994).

The inherent subjectivity related to the assessment of significance is one of the practical limitations of ROPE-based indices (despite being, conceptually, an asset, allowing for contextual nuance in the interpretation), as they require an explicit definition of the non-significant range (the ROPE). Although default values have been reported in the literature (for instance, half of a "negligible" effect size reference value; Kruschke, 2014), it is critical to reproducibility and transparency that the researcher's choice is explicitly stated (and, if possible, justified). Beyond being arbitrary, this range also has hard limits (for instance, contrary to a value of 0.0499, a value of 0.0501 would be considered non-negligible if the range ends at 0.05). This reinforces a categorical and clustered perspective of what is by essence a continuous space of possibilities. Importantly, as this range is fixed to the scale of the response (it is expressed in the unit of the response), ROPE indices are sensitive to changes in the scale of the predictors. For instance, negligible results may change into non-negligible results when predictors are scaled up (e.g., reaction times expressed in seconds instead of milliseconds), which one inattentive or malicious researcher could misleadingly present as "significant" (note that indices of existence, such as the *pd*, would not be affected by this). Finally, the ROPE definition is also dependent on the model type, and selecting a consistent or homogeneous range for all the families of models is not straightforward. This can make comparisons between model types difficult, and an additional burden when interpreting ROPE-based indices. In summary, while a well-defined ROPE can be a powerful tool to give a different and new perspective, it also requires extra caution on the paets of authors and readers.

As for the difference between ROPE (95%) and ROPE (full), we suggest reporting the latter (i.e., the percentage of the whole posterior distribution that falls within the ROPE instead of a given proportion of CI). This bypasses the use of another arbitrary range (95%) and appears to be more sensitive to delineate highly significant effects). Critically, rather than using the percentage in ROPE as a dichotomous, all-or-nothing decision criterion, such as suggested by the original equivalence test (Kruschke, 2014), we recommend using the percentage as a continuous index of significance (with explicitly specified cut-off points if categorization is needed, for instance 5% for significance and 95% for non-significance).

Our results underline the Bayes factor as an interesting index, able to provide evidence in favor or against the presence of an effect. Moreover, its easy interpretation in terms of odds in favor or against one hypothesis or another makes it a compelling index for communication. Nevertheless, one of the main critiques of Bayes factors is its sensitivity to priors (shown in our results here through its sensitivity to model types, as priors' odds for logistic and linear models are different). Moreover, while the BF appears even better when compared with a ROPE than when compared with a point-null, it also carries all the limitations related to ROPE specification mentioned above. Thus, we recommend using Bayes factors (preferentially vs. a ROPE) if the user has explicitly specified (and has a rationale for) informative priors (often called "subjective" priors; Wagenmakers, 2007). In the end, there is a relative proximity between Bayes factors (vs. ROPE) and the percentage in ROPE (full), consistent with their mathematical relationship.

Being quite different from the Bayes factor and ROPE indices, the Probability of Direction (*pd*) is an index of effect existence representing the certainty with which an effect goes in a particular direction (i.e., is positive or negative). Beyond its simplicity of interpretation, understanding and computation, this index also presents other interesting properties. It is independent from the model, i.e., it is solely based on the posterior distributions and does not require any additional information from the data or the model. Contrary to ROPE-based indices, it is robust to the scale of both the response variable and the predictors. Nevertheless, this index also presents some limitations. Most importantly, the *pd* is not relevant for assessing the size or importance of an effect and is not able to provide information *in favor* of the null hypothesis. In other words, a high *pd* suggests the presence of an effect but a small *pd* does not give us any information about how plausible the null hypothesis is, suggesting that this index can only be used to eventually reject the null hypothesis (which is consistent with the interpretation of the frequentist *p*-value). In contrast, BFs (and to some extent the percentage in ROPE) increase or decrease as the evidence becomes stronger (more data points), in both directions.

Much of the strengths of the *pd* also apply to the MAP-based *p*-value. Although possibly showing some superiority in terms of sensitivity as compared to it, it also presents an important limitation. Indeed, the MAP is mathematically dependent on the density at 0 and at the mode. However, the density estimation of a continuous distribution is a statistical problem on its own and many different methods exist. It is possible that changing the density estimation may impact the MAP-based *p*-value, with unknown results. The *pd*, however, has a linear relationship with the frequentist *p*-value, which is in our opinion an asset.

After all the criticism regarding the frequentist *p*-value, it may appear contradictory to suggest the usage of its

TABLE 3 | Summary of Bayesian indices of effect existence and significance.

Index	Interpretation	Definition	Strengths	Limitations
Probability of Direction (pd)	Probability that an effect is of the same sign as the median's	Proportion of the posterior distribution of the same sign than the median's	Straightforward computation and interpretation. Objective property of the posterior distribution. 1:1 correspondence with the frequentist *p*-value	Limited information favoring the null hypothesis
MAP-based *p*-value	Relative odds of the presence of an effect against 0	Density value at 0 divided by the density value at the mode of the posterior distribution	Straightforward computation. Objective property of the posterior distribution	Limited information favoring the null hypothesis. Relates on density approximation. Indirect relationship between mathematical definition and interpretation
ROPE (95%)	Probability that the credible effect values are not negligible	Proportion of the 95% CI inside of a range of values defined as the ROPE	Provides information related to the practical relevance of the effects	A ROPE range needs to be arbitrarily defined. Sensitive to the scale (the unit) of the predictors. Not sensitive to highly significant effects
ROPE (full)	Probability that the effect possible values are not negligible	Proportion of the posterior distribution inside of a range of values defined as the ROPE	Provides information related to the practical relevance of the effects	A ROPE range needs to be arbitrarily defined. Sensitive to the scale (the unit) of the predictors
Bayes factor (vs. 0)	The degree by which the probability mass has shifted away from or toward the null value, after observing the data	Ratio of the density of the null value between the posterior and the prior distributions	An unbounded continuous measure of relative evidence. Allows statistically supporting the null hypothesis	Sensitive to selection of prior distribution shape, location and scale
Bayes factor (vs. ROPE)	The degree by which the probability mass has into or outside of the null interval (ROPE), after observing the data	Ratio of the odds of the posterior vs. the prior distribution falling inside of the range of values defined as the ROPE	An unbounded continuous measure of relative evidence. Allows statistically supporting the null hypothesis. Compared to the BF (vs. 0), evidence is accumulated faster for the null when the null is true	Sensitive to selection of prior distribution shape, location and scale. Additionally, a ROPE range needs to be arbitrarily defined, which is sensitive to the scale (the unit) of the predictors

Bayesian empirical equivalent. The subtler perspective that we support is that the *p*-value is not an intrinsically bad, or wrong, index. Instead, it is its misuse, misunderstanding and misinterpretation that fuels the decay of the situation into the crisis. Interestingly, the proximity between the *pd* and the *p*-value follows the original definition of the latter (Fisher, 1925) as an index of effect existence *rather than* significance (as in "worth of interest"; Cohen, 1994). Addressing this confusion, the Bayesian equivalent has an intuitive meaning and interpretation, contributing to making more obvious the fact that all thresholds and heuristics are arbitrary. In summary, the mathematical and interpretative transparency of the *pd*, and its conceptualization as an index of effect existence, offer valuable insight into the characterization of Bayesian results, and its practical proximity with the frequentist *p*-value makes it a perfect metric to ease the transition of psychological research into the adoption of the Bayesian framework.

Our study has some limitations. First, our simulations were based on simple linear and logistic regression models. Although these models are widespread, the behavior of the presented indices for other model families or types, such as count models or mixed effects models, still needs to be explored. Furthermore, we only tested continuous predictors. The indices may behave differently when varying the type of predictor (binary, ordinal) as well. Finally, we limited our simulations to small sample sizes, for the reason that data is particularly noisy in small samples, and experiments in psychology often include only a limited number of subjects. However, it is possible that the indices converge (or diverge) for larger samples. Importantly, before being able to draw a definitive conclusion about the qualities of these indices, further studies should investigate the robustness of these indices to sampling characteristics (e.g., sampling algorithm, number of iterations, chains, warm-up) and the impact of prior specification (Kass and Raftery, 1995; Vanpaemel, 2010; Kruschke, 2011), all of which are important parameters of Bayesian statistics.

REPORTING GUIDELINES

How can the current observations be used to improve statistical good practices in psychological science? Based on the present comparison, we can start outlining the following guidelines. As *existence* and *significance* are complementary perspectives, we suggest using at minimum one index of each category. As an objective index of effect existence, the *pd* should be reported, for its simplicity of interpretation, its robustness and its numeric proximity to the well-known frequentist *p*-value; As an index of significance either the *BF* (vs. *ROPE*) or the *ROPE (full)* should be reported, for their ability to discriminate between presence and absence of effect (De Santis, 2007) and the information they provide related to evidence of the size of the effect. Selection between the *BF* (vs. *ROPE*) or the *ROPE (full)* should depend on the informativeness of the priors used – when uninformative priors are used, and there is little prior knowledge regarding the expected size of the effect, the *ROPE (full)* should be reported as it reflects only the posterior distribution and is not sensitive to the width of a wide-range of prior scales (Rouder et al., 2018). On the other hand, in cases where informed priors are used, reflecting prior knowledge regarding the expected size of the effect, *BF* (vs. *ROPE*) should be used.

Defining appropriate heuristics to aid in interpretation is beyond the scope of this paper, as it would require testing them on more natural datasets. Nevertheless, if we take the frequentist framework and the existing literature as a reference point, it seems that 95, 97, and 99% may be relevant reference points (i.e., easy-to-remember values) for the *pd*. A concise, standardized, reference template sentence to describe the parameter of a model including an index of point-estimate, uncertainty, existence, significance and effect size (Cohen, 1988) could be, in the case of *pd* and *BF*:

> "There is moderate evidence ($BF_{\text{ROPE}} = 3.44$) [*BF* (vs. *ROPE*)] in favor of the presence of effect of X, which has a probability of 98.14% [*pd*] of being negative (Median $= -5.04$, 89%CI$[-8.31, 0.12]$), and can be considered to be small (Std. Median $= -0.29$) [*standardized coefficient*]."

And if the user decides to use the percentage in ROPE instead of the *BF*:

> "The effect of X has a probability of 98.14% [*pd*] of being negative (Median $= -5.04$, 89%CI$[-8.31, 0.12]$), and can be considered to be small (Std.Median $= -0.29$) [*standardized coefficient*] and significant (0.82% in ROPE) [*ROPE (full)*]."

DATA AVAILABILITY STATEMENT

The full R code used for data generation, data processing, figures creation, and manuscript compiling is available on GitHub at https://github.com/easystats/easystats/tree/master/publications/makowski_2019_bayesian.

AUTHOR CONTRIBUTIONS

DM conceived and coordinated the study. DM, MB-S, and DL participated in the study design, statistical analysis, data interpretation, and manuscript drafting. DL supervised the manuscript drafting. SC performed a critical review of the manuscript, assisted with the manuscript drafting, and provided funding for publication. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

This study was made possible by the development of the *bayestestR* package, itself part of the *easystats* ecosystem (Lüdecke et al., 2019), an open-source and collaborative project created to facilitate the usage of R. Thus, there is substantial evidence in favor of the fact that we thank the masters of easystats and all the other padawan following the way of the Bayes.

REFERENCES

Amrhein, V., Greenland, S., and McShane, B. (2019). Scientists rise up against statistical significance. *Nature* 567, 305–307. doi: 10.1038/d41586-019-00857-9

Anderson, D. R., Burnham, K. P., and Thompson, W. L. (2000). Null hypothesis testing: problems, prevalence, and an alternative. *J. Wildlife Manag.* 64, 912–923.

Andrews, M., and Baguley, T. (2013). Prior approval: the growth of bayesian methods in psychology. *Br. J. Math. Statist. Psychol.* 66, 1–7. doi: 10.1111/bmsp.12004

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., et al. (2018). Redefine statistical significance. *Nat. Hum. Behav.* 2, 6–10.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: a probabilistic programming language. *J. Statist. Softw.* 76 1–32. doi: 10.18637/jss.v076.i01

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., and Etchells, P. (2014). Instead of 'playing the game' it is time to change the rules: registered reports at aims neuroscience and beyond. *AIMS Neurosci.* 1, 4–17. doi: 10.3934/neuroscience.2014.1.4

Cohen, J. (1988). *Statistical Power Analysis for the Social Sciences*. New York, NY: Academic Publishers.

Cohen, J. (1994). The earth is round ($p < .05$). *Am. Psychol.* 49, 997–1003. doi: 10.1037/0003-066x.49.12.997

De Santis, F. (2007). Alternative bayes factors: sample size determination and discriminatory power assessment. *Test* 16, 504–522. doi: 10.1007/s11749-006-0017-7

Dienes, Z. (2014). Using bayes to get the most out of non-significant results. *Front. Psychol.* 5:781. doi: 10.3389/fpsyg.2014.00781

Dienes, Z., and Mclatchie, N. (2018). Four reasons to prefer bayesian analyses over significance testing. *Psychon. Bull. Rev.* 25, 207–218. doi: 10.3758/s13423-017-1266-z

Ellis, S., and Steyn, H. (2003). Practical significance (effect sizes) versus or in combination with statistical significance (p-values): research note. *Manag. Dyn. J. South. Afr. Instit. Manag. Sci.* 12, 51–53.

Etz, A., Haaf, J. M., Rouder, J. N., and Vandekerckhove, J. (2018). Bayesian inference and testing any hypothesis you can specify. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/wmf3r

Etz, A., and Vandekerckhove, J. (2016). A bayesian perspective on the reproducibility project: psychology. *PLoS One* 11:e0149794. doi: 10.1371/journal.pone.0149794

Fidler, F., Thomason, N., Cumming, G., Finch, S., and Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: statistical reform lessons from medicine. *Psychol. Sci.* 15, 119–126. doi: 10.1111/j.0963-7214.2004.01502008.x

Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., et al. (2004). Reform of statistical inference in psychology: the case ofMemory & cognition. *Behav. Res. Methods Instru. Comput.* 36, 312–324. doi: 10.3758/bf03195577

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver.

Gardner, M. J., and Altman, D. G. (1986). Confidence intervals rather than p values: estimation rather than hypothesis testing. *Br. Med. J.* 292, 746–750. doi: 10.1136/bmj.292.6522.746

Gelman, A. (2018). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Pers. Soc. Psychol. Bull.* 44, 16–23. doi: 10.1177/0146167217729162

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*, 3rd Edn, Boca Raton: CRC Press.

Goodrich, B., Gabry, J., Ali, I., and Brilleman, S. (2019). *Rstanarm: Bayesian Applied Regression Modeling Via Stan*. Available at: http://mc-stan.org/ (accessed November 29, 2019).

Halsey, L. G. (2019). The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? *Biol. Lett.* 15:20190174. doi: 10.1098/rsbl.2019.0174

Heck, D. W. (2019). A caveat on the savage–dickey density ratio: the case of computing bayes factors for regression parameters. *Br. J. Math. Statist. Psychol.* 72, 316–333. doi: 10.1111/bmsp.12150

Jarosz, A. F., and Wiley, J. (2014). What are the odds? A practical guide to computing and reporting bayes factors. *J. Probl. Solving* 7:2.

Jeffreys, H. (1998). *The Theory of Probability*. Oxford: Oxford University Press.

Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *J. Am. Statist. Assoc.* 90, 773–795.

Kirk, R. E. (1996). Practical significance: a concept whose time has come. *Educ. Psychol. Measur.* 56, 746–759. doi: 10.1177/0013164496056005002

Kruschke, J. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, Jags, and Stan*. Cambridge, MA: Academic Press.

Kruschke, J. K. (2010). What to believe: bayesian methods for data analysis. *Trends Cogn. Sci.* 14, 293–300. doi: 10.1016/j.tics.2010.05.001

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspect. Psychol. Sci.* 6, 299–312. doi: 10.1177/1745691611406925

Kruschke, J. K., Aguinis, H., and Joo, H. (2012). The time has come: bayesian methods for data analysis in the organizational sciences. *Organ. Res. Methods* 15, 722–752. doi: 10.1177/1094428112457829

Kruschke, J. K., and Liddell, T. M. (2018). The bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychon. Bull. Rev.* 25, 178–206. doi: 10.3758/s13423-016-1221-4

Lakens, D. (2017). Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Soc. Psychol. Pers. Sci.* 8, 355–362. doi: 10.1177/1948550617697177

Lakens, D., Scheel, A. M., and Isager, P. M. (2018). Equivalence testing for psychological research: a tutorial. *Adv. Methods Pract. Psychol. Sci.* 1, 259–269. doi: 10.1177/2515245918770963

Lüdecke, D., Waggoner, P., and Makowski, D. (2019). Insight: a unified interface to access information from model objects in R. *J. Open Source Softw.* 4:1412. doi: 10.21105/joss.01412

Ly, A., Verhagen, J., and Wagenmakers, E.-J. (2016). Harold jeffreys's default bayes factor hypothesis tests: explanation, extension, and application in psychology. *J. Math. Psychol.* 72, 19–32. doi: 10.1016/j.jmp.2015.06.004

Makowski, D., Ben-Shachar, M., and Lüdecke, D. (2019). Bayestestr: describing effects and their uncertainty, existence and significance within the bayesian framework. *J. Open Source Softw.* 4:1541. doi: 10.21105/joss.01541

Marasini, D., Quatto, P., and Ripamonti, E. (2016). The use of p-values in applied research: Interpretation and new trends. *Statistica* 76, 315–325.

Maxwell, S. E., Lau, M. Y., and Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *Am. Psychol.* 70, 487–498. doi: 10.1037/a0039400

McElreath, R. (2018). *Statistical Rethinking*. London: Taylor and Francis Group.

Mills, J. A. (2017). *Objective Bayesian Precise Hypothesis Testing*. Ohio: University of Cincinnati.

Mills, J. A., and Parent, O. (2014). "Bayesian mcmc estimation," in *Handbook of Regional Science*, eds M. M. Fischer, and P. Nijkamp, (Berlin: Springer), 1571–1595. doi: 10.1007/978-3-642-23430-9_89

Morey, R. D., and Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychol. Methods* 16, 406–419. doi: 10.1037/a0024377

Nuzzo, R. (2014). Scientific method: statistical errors. *Nature* 506, 150–152. doi: 10.1038/506150

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Core Team.

Robert, C. P. (2014). On the jeffreys-lindley paradox. *Philos. Sci.* 81, 216–232. doi: 10.1086/675729

Robert, C. P. (2016). The expected demise of the bayes factor. *J. Math. Psychol.* 72, 33–37. doi: 10.1016/j.jmp.2015.08.002

Rouder, J. N., Haaf, J. M., and Vandekerckhove, J. (2018). Bayesian inference for psychology, part iv: Parameter estimation and bayes factors. *Psychon. Bull. Rev.* 25, 102–113. doi: 10.3758/s13423-017-1420-7

Rouder, J. N., and Morey, R. D. (2012). Default bayes factors for model selection in regression. *Multivar. Behav. Res.* 47, 877–903. doi: 10.1080/00273171.2012.734737

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* 16, 225–237. doi: 10.3758/pbr.16.2.225

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22, 1359–1366. doi: 10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2014). P-curve and effect size: correcting for publication bias using only significant results. *Perspect. Psychol. Sci.* 9, 666–681. doi: 10.1177/1745691614553988

Spanos, A. (2013). Who should be afraid of the jeffreys-lindley paradox? *Philos. Sci.* 80, 73–93. doi: 10.1086/668875

Sullivan, G. M., and Feinn, R. (2012). Using effect size–or why the p value is not enough. *J. Grad. Med. Educ.* 4, 279–282. doi: 10.4300/jgme-d-12-00156.1

Szucs, D., and Ioannidis, J. P. (2016). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *BioRxiv* [Preprint]. doi: 10.1101/071530

Vanpaemel, W. (2010). Prior sensitivity in theory testing: an apologia for the bayes factor. *J. Math. Psychol.* 54, 491–498. doi: 10.1016/j.jmp.2010.07.003

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychon. Bull. Rev.* 14, 779–804. doi: 10.3758/bf03194105

Wagenmakers, E.-J., Lee, M., Rouder, J., and Morey, R. (2019). *Another Statistical Paradox*. Available at: http://www.ejwagenmakers.com/submitted/AnotherStatisticalParadox.pdf (accessed November 29, 2019).

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., and Grasman, R. (2010). Bayesian hypothesis testing for psychologists: a tutorial on the savage–dickey method. *Cogn. Psychol.* 60, 158–189. doi: 10.1016/j.cogpsych.2009.12.001

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., et al. (2018). Bayesian inference for psychology. Part I: theoretical advantages and practical ramifications. *Psychon. Bull. Rev.* 25, 35–57. doi: 10.3758/s13423-017-1343-3

Wagenmakers, E.-J., Morey, R. D., and Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Curr. Dir. Psychol. Sci.* 25, 169–176. doi: 10.1177/0963721416643289

Wagenmakers, E.-J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., et al. (2017). "The need for bayesian hypothesis testing in psychological science," in *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions*, eds S. O. Lilienfeld, and I. D. Waldman, (Chichester: JohnWiley & Sons), 123–138. doi: 10.1002/9781119095910.ch8

Wasserstein, R. L., and Lazar, N. A. (2016). The asa's statement on p-values: context, process, and purpose. *Am. Statist.* 70, 129–133. doi: 10.1080/00031305.2016.1154108

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 17 December 2019
doi: 10.3389/fpsyg.2019.02793

New Perspectives in Computing the Point of Subjective Equality Using Rasch Models

Giulio Vidotto[1]*, *Pasquale Anselmi*[2] *and Egidio Robusto*[2]

[1]Department of General Psychology, University of Padua, Padova, Italy, [2]Department of Philosophy, Sociology, Education and Applied Psychology, University of Padua, Padova, Italy

In psychophysics, the point of subject equality (PSE) is any of the points along a stimulus dimension at which a variable stimulus (visual, tactile, auditory, and so on) is judged by an observer to be equal to a standard stimulus. Rasch models have been found to offer a valid solution for computing the PSE when the method of constant stimuli is applied in the version of the method of transitions. The present work provides an overview of the procedures for computing the PSE using Rasch models and proposes some new developments. An adaptive procedure is described that allows for estimating the PSE of an observer without presenting him/her with all stimuli pairs. This procedure can be particularly useful in those situations in which psychophysical conditions of the individuals require that the number of trials is limited. Moreover, it allows for saving time that can be used to scrutinize the results of the experiment or to run other experiments. Also, the possibility of using Rasch-based fit statistics for identifying observers who gave unexpected judgments is explored. They could be individuals who, instead of carefully evaluating the presented stimuli pairs, gave random, inattentive, or careless responses, or gave the same response to many consecutive stimuli pairs. Otherwise, they could be atypical and clinically relevant individuals who deserve further investigation. The aforementioned developments are implemented using procedures and statistics that are well established in the framework of Rasch models. In particular, computerized adaptive testing procedures are used for efficiently estimating the PSE of the observers, whereas infit and outfit mean-squares statistics are used for detecting observers who gave unexpected judgments. Results of the analyses carried out on simulated data sets suggest that the proposed developments can be used in psychophysical experiments.

Keywords: method of constant stimuli, method of transitions, point of subjective equality, Rasch models, computerized adaptive testing, infit, outfit

Edited by:
Pietro Cipresso,
Italian Auxological Institute (IRCCS), Italy

Reviewed by:
Givago Silva Souza,
Federal University of Pará, Brazil
Fabio Lucidi,
Sapienza University of Rome, Italy

Correspondence:
Giulio Vidotto
giulio.vidotto@unipd.it

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *21 August 2019*
Accepted: *27 November 2019*
Published: *17 December 2019*

Citation:
Vidotto G, Anselmi P and Robusto E (2019) New Perspectives in Computing the Point of Subjective Equality Using Rasch Models. Front. Psychol. 10:2793. doi: 10.3389/fpsyg.2019.02793

INTRODUCTION

In psychophysics, the point of subject equality (PSE) is any of the points along a stimulus dimension at which a variable stimulus (visual, tactile, auditory, and so on) is judged by an observer to be equal to a standard stimulus. When the method of constant stimuli (see, e.g., Laming and Laming, 1992) is used to measure the PSE, the observer is presented with a number I of variable stimuli, each of which is denoted by VS_i, $i = 1, 2, \ldots, I$. The variable stimuli are placed at equal intervals along the physical continuum, and are chosen in such a way that the stimulus at the inferior extreme is perceived little more than 0–5% of the times it is presented, whereas a stimulus at the superior extreme is perceived a little less than 95–100% of the times. The variable stimuli are presented, one at a time and in random order, paired with a standard stimulus (*SS*). The number of presentations for each pair (VS_i, *SS*) typically varies from 20 to 200. The observer judges each pair (VS_i, *SS*) and says which of the two stimuli has a greater (or a fewer) quantity of the attribute under consideration (e.g., volume, roughness, loudness, and so on). The PSE is the value of a comparison stimulus that, for a particular observer, is equally likely to be judged as higher or lower than that of a standard stimulus (Guilford, 1954; Bock and Jones, 1968).

As an example of method of constant stimuli, let us consider an experiment of sound perception in which *SS* is a 50-decibel sound and the variable stimuli are $I = 9$ sounds from 30 to 70 decibels, at the distance of 5 decibels one from the next (i.e., VS = 30, 35, 40, 45, 50, 55, 60, 65, 70 decibels). Pairs of sounds are presented in succession, the former sound being the *SS* and the latter sound being the *VS*. The subject is asked to report whether or not the second sound (the *VS*) is louder than the first sound (the *SS*). In the experiment at hand, sound loudness is the target attribute. The PSE is the level (in decibel) of a comparison stimulus at which this stimulus is judged by the observer to be as loud as *SS*.

When the method of constant stimuli is used, the classical solution for obtaining the PSE is the least square method (Müller, 1879). The proportion $P(VS_i > SS)$ of times in which VS_i is judged higher than *SS* is computed for each VS_i. Then, each $P(VS_i > SS)$ is transformed in the corresponding *z*-score z_i by using the inverse of the cumulative normal function. Alternative and more recent solutions for obtaining the PSE are the weighted least square method (Urban, 1908) and the maximum likelihood procedure (Whittaker and Robinson, 1967).

In some cases, the experimenter cannot use the method of constant stimuli in the classical form. This is particularly true when effects of adaptation, habituation, and sensitization may occur. The greater the number of presentations, the higher the probability that these effects will influence the judgments. In these situations, the method of constant stimuli would be unsuitable. On the one hand, a drastic reduction in the presentation of stimuli would be necessary to reduce biases. On the other hand, a high number of presentations is necessary (especially when the number of observers is small) for the method of constant stimuli to produce good results.

One solution is to present each pair (VS_i, *SS*) to each observer only once, as it happens in the method of transitions (Masin and Cavedon, 1970; Masin and Vidotto, 1982, 1984). A transition occurs when the comparative judgment of the pair (VS_i, *SS*) is different from that of the pair (VS_{i+1}, *SS*). In this case, it is possible to assume that the PSE of the observer takes place between VS_i and VS_{i+1}. More details about the method of transitions, as well as examples of application can be found in Masin and Vidotto (1984) and Burro et al. (2011).

Rasch models have been found to offer a valid solution for computing the PSE when the method of constant stimuli is applied in the version of the method of transitions (Vidotto et al., 1996; Burro et al., 2011). Rasch models represent a family of psychometric models for creating measurements from categorical data. In these models, the probability of observing specified responses (e.g., correct/incorrect; yes/no; never/sometimes/often/always) is modeled as a function of person and item parameters. These parameters pertain to the level of a quantitative latent trait possessed by a person or item, and their specific meaning relies on the subject of the assessment. In educational assessments, for instance, person parameters indicate the ability (or attainment level) of persons, and item parameters indicate the difficulty of items. In health status assessments, person parameters indicate the health of persons, and item parameters indicate the severity of items. The higher the ability of a person relative to the difficulty of an item, the higher the probability that the person will give a correct response to the item. The higher the health of a person relative to the severity of an item, the higher the probability that that person will give to the item a response that is indicative of health (e.g., a response "no" to an item like "I have trouble falling asleep"). Because of their general applicability, Rasch models have been used in several areas, including personality and health assessment, education, and market research (see, e.g., Bechtel, 1985; Vidotto et al., 1998, 2006, 2007, 2010a,b; Duncan et al., 2003; Cole et al., 2004; Bezruczko, 2005; Pallant and Tennant, 2007; Shea et al., 2009; Anselmi et al., 2011, 2013a,b, 2015; Da Dalt et al., 2013, 2015, 2017; Balsamo et al., 2014; Rossi Ferrario et al., 2019; Sotgiu et al., 2019).

When applied to psychophysics, Rasch models allow for identifying two aspects linked to the perceptive judgments. The first one deals with the ability of observers to discriminate the variable stimuli (parameters β). The second one deals with the difficulty of discriminating the variable stimuli from the standard stimulus (parameters δ). These two types of parameters are placed on the same linear scale and can be compared (see, e.g., Andrich, 1988; Wright and Stone, 1999). The comparison between the discriminative ability of an observer and the discriminability of a variable stimulus allows for computing the probability that the observer will judge the variable stimulus in a certain way. It is worth noting that, within the Rasch framework, the estimates of observers' discriminative abilities do not depend on the specific collection of stimuli the observers have been presented with, as well as the estimates of stimuli' discriminability do not depend on the particular sample of observers who have been presented with the stimuli (Rasch, 1960; Bond and Fox, 2001).

There are algorithms that allow for estimating the parameters β and δ from experimental data (see, e.g., Wright, 1977; Linacre, 1999; Wright and Stone, 1999), as well as procedures for deriving the PSE of an observer from his/her parameter β (Vidotto et al., 1996; Burro et al., 2011). Moreover, there are Rasch models for simple judgments (the variable stimulus can only be considered to be higher or lower than the standard stimulus) and for more complex judgments (the variable stimulus can also be considered as not different from the standard stimulus). In particular, the simple logistic model (SLM, Rasch, 1960) is suitable in the first case, whereas the rating scale model (RSM; Andrich, 1978) is suitable in the second case. An application of the RSM for computing the PSE in a psychophysical experiment with three response categories is described in Burro et al. (2011).

The present work provides an overview of the procedures for computing the PSE using Rasch models. Besides, it proposes two new developments that are based on Rasch models and that pertain to the efficient estimation of the PSE and the identification of observers with unexpected judgments. Concerning the first development, a computerized adaptive testing (CAT) procedure is described that allows for estimating the PSE of an observer without presenting him/her with all stimuli pairs. This procedure can be particularly useful in those situations in which psychophysical conditions of individuals require that the number of trials is limited. Moreover, it allows for saving time that can be used to scrutinize the results of the experiment or to run other experiments. Concerning the second development, the possibility of using fit statistics for identifying observers who gave unexpected judgments is explored. They could be individuals who, instead of carefully evaluating the presented stimuli pairs, gave random, inattentive, or careless responses, or gave the same response to many consecutive stimuli pairs. Otherwise, they could be atypical and clinically relevant individuals for whom a further investigation is needed. The aforementioned developments are implemented using procedures and statistics that are well established in the framework of Rasch models and their functioning is illustrated *via* simulated data.

COMPUTING THE POINT OF SUBJECTIVE EQUALITY USING RASCH MODELS

Vidotto et al. (1996) and Burro et al. (2011) proposed to use Rasch models for computing the PSE of observers when the method of constant stimuli is applied in the version of the method of transitions. The authors focused on two models, namely the SLM and the RSM. The former is meant for dichotomous outcomes. As such, it is suitable for psychophysical experiments with two response categories (i.e., in which the variable stimulus can only be considered to be higher or lower than the standard stimulus). The RSM is an extension of the SLM meant for polytomous outcomes. As such, it is suitable for psychophysical experiments with more than two response categories (i.e., in which the variable stimulus can also be considered as not different from the standard stimulus).

Let x_{ni} be the perceptive outcome obtained by observer n in relation to the comparison between VS_i and SS. If the observer n can only report which of the two stimuli has a greater or a smaller quantity of the target attribute, then x_{ni} assumes value 1 if VS_i is perceived higher than SS, and value 0 if it is perceived lower than SS. If the observer n is allowed to say that the two stimuli have the same quantity of the target attribute, then x_{ni} assumes value 2 if VS_i is perceived higher than SS, value 1 if VS_i and SS are perceived as equal, and value 0 if VS_i is perceived lower than SS.

For instance, let us still consider the experiment of sound perception in which pairs of sounds are presented in succession, and the subject is asked to report whether or not the second sound (VS = 30, 35, 40, 45, 50, 55, 60, 65, 70) is louder than the first sound (SS = 50 decibels). **Table 1** shows possible perceptive outcomes for experimental situations with two or three response options. In the former situation, the variable stimuli of 30, 35, 40, 45, 50, 60 decibels are judged to be less loud than *SS*, and those of 55, 65, and 70 decibels are judged to be louder than *SS*. In the latter situation, the variable stimuli of 30, 35, 40, 45, 55 decibels are judged to be less loud than *SS*; those of 50 and 60 decibels are judged to be as loud as *SS*; and those of 65 and 70 decibels are judged to be louder than *SS*.

It is worth noting that sometimes the response option of equal judgments does not actually mean that the two stimuli are perceived as having the same quantity of target attribute but it takes the meaning of "I do not know," "I am uncertain about," or "It seems to me that they are different but I am not sure which one is the greatest."

The SLM and the RSM describe the probability of observing the perceptive outcome x_{ni} as:

$$P\left(X_{ni} = x_{ni} | \beta_n, \delta_i\right) = \frac{\exp\left(x_{ni}\left(\beta_n - \delta_i\right)\right)}{1 + \exp\left(\beta_n - \delta_i\right)},$$

and

$$P\left(X_{ni} = x_{ni} | \beta_n, \delta_i, \tau_k\right) = \frac{\exp\sum_{k=0}^{x}\left(\beta_n - \left(\delta_i - \tau_k\right)\right)}{\sum_{j=0}^{m}\exp\sum_{k=0}^{j}\left(\beta_n - \left(\delta_i - \tau_k\right)\right)},$$

where:

1. β_n is the discriminative ability of observer n;
2. δ_i is the difficulty of discriminating the variable stimulus VS_i from the standard stimulus SS;

TABLE 1 | Example of perceptive outcomes in an experiment of sound perception with *SS* of 50 decibels.

	Perceptive outcome	
***VS_i* (decibels)**	**Two response options**	**Three response options**
30	0	0
35	0	0
40	0	0
45	0	0
50	0	1
55	1	0
60	0	1
65	1	2
70	1	2

In the condition with two response options, the perceptive outcome takes the values 0 or 1 if VS_i is perceived to be less loud or louder than SS, respectively. In the condition with three response options, the perceptive outcome takes the values 0, 1, or 2 if VS_i is perceived to be less loud than, as loud as, or louder than SS, respectively.

3. τ_k is the k-th threshold and expresses the passage from one response category to the next one (thus, if the measurement criterion includes three response categories, there will be two thresholds).

Once parameters β and δ have been estimated, the PSE of observer n is obtained through the following steps:

1. The difficulties of stimuli (δ_i) are put in relation to the relative physical values φ_i. This determines the intercept and the slope of the regression line (i.e., $\varphi_i = a\delta_i + b$).
2. The obtained values of intercept and slope are used to derive the PSEs of observers from their discriminative abilities (i.e., $\text{PSE}_n = a\beta_n + b$).

AN ADAPTIVE PROCEDURE FOR ESTIMATING THE POINT OF SUBJECTIVE EQUALITY

One of the most prominent applications of Rasch models is in CAT. CAT procedures allow for accurately estimating the latent trait level of individuals by presenting them with only a minimum number of items (Linacre, 2000). Typically, the adaptive tests reach the same level of accuracy of the conventional fixed-length tests using about 50% of the items (Embretson and Reise, 2000; van der Linden, 2008). Moreover, the adaptive tests can be a better experience for individuals, as they are only presented with items targeted at their level (Deville, 1993). This section describes the functioning of a CAT procedure that aims at estimating the PSE of an observer.

CAT is preceded by a preliminary phase in which the psychophysical experiment is run on a suitable calibration sample, and an appropriate Rasch model (either the SLM or the RSM) is estimated on the collected data. This phase aims to arrive at an accurate estimate of the parameters δ (if the SLM is estimated) or δ and τ (if the RSM is estimated), so that they can be considered as known during CAT. When the latter begins, the only unknown parameters are the discriminative abilities β of observers under evaluation.

Figure 1 depicts the functioning of the CAT procedure. An initial estimate is determined for the discriminative ability β of the observer. The first pair (VS_i, SS) is selected based on this starting point and presented to the observer. The pair is judged and scored, and the estimate of β is updated accordingly. The stopping criterion is then evaluated. If it is not yet satisfied, another pair (VS_i, SS) is selected based on the current estimate β. The observer judges this new pair, and the estimate of β is updated again. The procedure iterates the aforementioned steps until the stopping criterion is satisfied.

There are several methods and algorithms for implementing each of the steps in a CAT procedure. A brief overview of the main ones is presented here. Readers interested in a more comprehensive discussion are referred to, for instance, Linacre (2000), van der Linden and Glas (2000), Wainer et al. (2000), van der Linden and Pashley (2010), and Thompson and Weiss (2011).

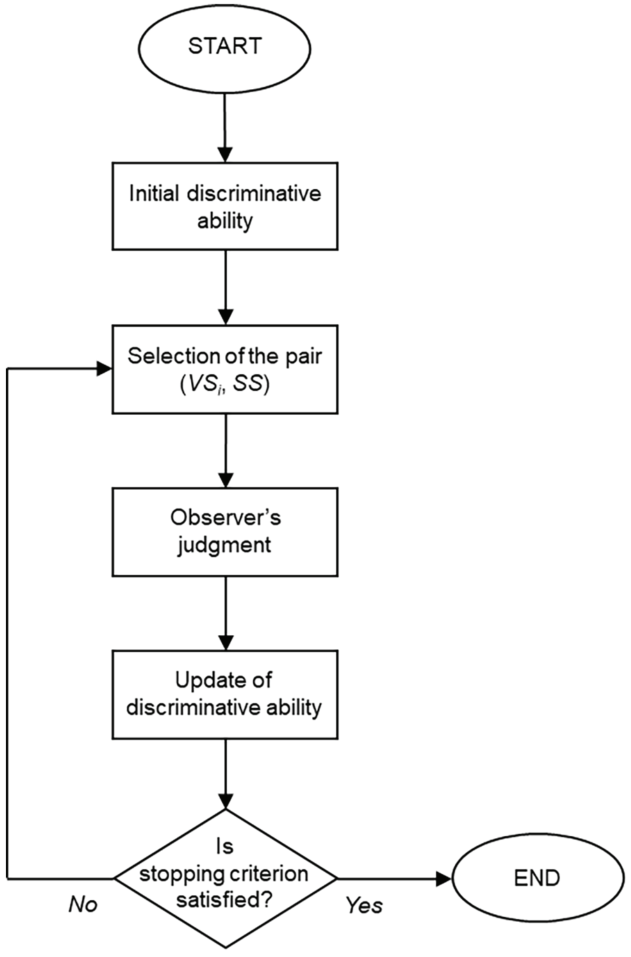

START
Initial discriminative ability
Selection of the pair (VS_i, SS)
Observer's judgment
Update of discriminative ability
Is stopping criterion satisfied?
No
Yes
END

FIGURE 1 | Diagram of the CAT procedure.

Determination of the initial estimate for the discriminative ability: Different options are available for this purpose. The most straightforward one is to use, as an initial estimate of observer's discriminative ability, the mean of the β distribution obtained on the calibration sample. Otherwise, if the information on the observer is available (e.g., results of a previous psychophysical experiment, familiarity of the observer with the perceptive task under consideration), this information can be used to determine a more appropriate initial estimate.

Selection of the pair (VS_i, SS) to be presented: The idea is to select the pair (VS_i, SS) according to the observer's estimated discriminative ability. A method very common in traditional CAT would imply to select the pair that maximizes Fisher information at the current estimate of discriminative ability. This method allows for estimating observer's discriminative ability by presenting him/her with a minimum number of stimuli pairs.

Update of observer's discriminative ability: The current estimate of the observer's discriminative ability is updated based on his/her response to the latest administered stimuli pair. Common methods are maximum-likelihood and Bayesian methods such as expected *a posteriori* (EAP, Bock and Mislevy, 1988) and maximum a posteriori (MAP, Mislevy, 1986).

Stopping criterion: CAT can be designed to be either fixed-length or variable-length. In the former case, the procedure stops when a specified number of stimuli pairs has been presented. In the second case, the procedure can stop when observer's β estimate changes below a certain small amount from one iteration to the other or has reached a certain level of precision, or when no stimuli pairs are left that provide at least some minimal level of information.

Method

Data Simulation

A psychophysical experiment with 11 variable stimuli was considered (i.e., $I = 11$). The stimuli were placed at the distance of one unit along the physical continuum. The smallest variable stimulus was five units smaller than the *SS*, whereas the largest variable stimulus was five units larger than the *SS*. A condition was simulated in which the observers judged each pair (VS_i, *SS*) and reported which of the two stimuli of the pair was the highest (two response options).

Two data samples of 100 observers each were simulated. One sample was used as a calibration sample, the other sample was used for running the CAT procedure (CAT sample). For both samples, 100 PSE values were randomly drawn from a normal distribution with mean = −1.5 and standard deviation = 1.

Calibration and Computerized Adaptive Testing

The SLM was estimated on the calibration sample. Model parameters were estimated using the EAP method.

The CAT procedure was run on the CAT sample using the estimates of parameters δ that were obtained on the calibration sample. The mean of the β distribution obtained on the calibration sample was used as initial estimate of observer's discriminative ability in the CAT procedure. Maximum Fisher information was used for selecting the stimuli pair to the administered. The responses to the selected stimuli pairs were extracted from the CAT sample. The EAP method was used for updating the estimates of β. For each observer in the CAT sample, the estimates of β and PSE were computed for the first five stimuli pairs that were presented.

The performance of the CAT procedure was compared with that of a procedure in which, at each iteration, the stimuli pair to be presented was randomly chosen (random procedure).

Results

Table 2 shows the estimates of parameters δ that were obtained on the calibration sample.

TABLE 2 | Estimates of parameters δ obtained on the calibration sample.

Difference between VS_i and SS	δ	*SE*
−5	−2.45	0.33
−4	−2.82	0.37
−3	−2.34	0.32
−2	−1.97	0.29
−1	−0.67	0.23
0	0.85	0.24
1	1.38	0.25
2	2.33	0.32
3	2.33	0.32
4	2.55	0.34
5	1.97	0.29

Figure 2 depicts the results of the CAT and random procedures. The left diagram depicts the average absolute difference between the parameters β estimated after the presentation of a certain number of stimuli pairs (from 1 to 5 pairs) and those estimated on all stimuli pairs (11 pairs). The right diagram depicts the average absolute difference between the PSEs estimated after the presentation of a certain number of stimuli pairs and those estimated on all stimuli pairs. In both diagrams, the solid line denotes the CAT procedure, the dashed line denotes the random procedure. The bars denote 95% confidence intervals. For both CAT and random procedures, with the increasing of the number of presented stimuli pairs, the estimates of β and PSE approach those obtained on all stimuli pairs. However, the number of presented pairs being equal, the CAT procedure outperforms the random procedure in approximating the estimates obtained on all stimuli pairs. The differences between the estimates β and PSE obtained on 4 or 5 stimuli pairs and those obtained on all stimuli pairs are significantly smaller when stimuli pairs are selected by the CAT procedure, rather than by the random procedure.

Figure 3 depicts the correlation between the PSEs estimated after the presentation of a certain number of stimuli pairs and those estimated on all stimuli pairs. The solid line denotes the CAT procedure, the dashed line denotes the random procedure. For both CAT and random procedures, the strength of the correlation between the PSEs estimated on the presented stimuli pairs and those estimated on all stimuli pairs increases with the number of presented stimuli pairs. On the whole, the number of presented stimuli pairs being equal, the correlation is significantly stronger when PSEs are estimated on the stimuli pairs selected by the CAT procedure than on those selected by the random procedure ($z \geq 1.98$, $p < 0.05$ when 1, 3, 4, or 5 stimuli pairs are presented; $z = 1.21$, $p = 0.23$ when 2 stimuli pairs are presented).

Results of this simulation study suggest that a Rasch-based CAT procedure can be used for estimating the PSE of observers without presenting them with all stimuli pairs.

IDENTIFICATION OF OBSERVERS WHO GAVE UNEXPECTED JUDGMENTS

Rasch framework provides infit and outfit mean-square statistics that allow for detecting individuals with unexpected response behaviors. For instance, these statistics have been used to identify possible fakers to self-report personality tests (Vidotto et al., 2018) and to identify individuals who miss items they are not capable of solving (Anselmi et al., 2018). This section explores the use of these statistics in psychophysical experiments

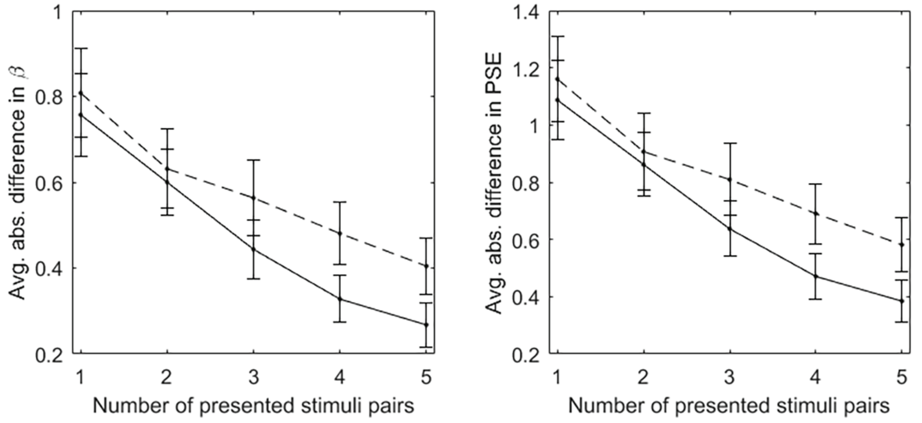


FIGURE 2 | Results of CAT (solid line) and random (dashed line) procedures. The left diagram depicts the average absolute difference between the parameters β estimated after the presentation of a certain number of stimuli pairs and those estimated on all stimuli pairs. The right diagram depicts the average absolute difference between the PSEs estimated after the presentation of a certain number of stimuli pairs and those estimated on all stimuli pairs. The bars denote 95% confidence intervals.

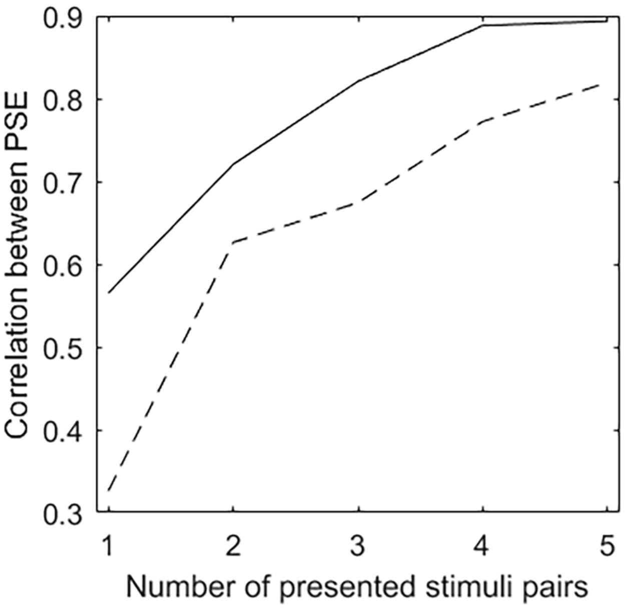


FIGURE 3 | Correlation between the PSEs estimated after the presentation of a certain number of stimuli pairs and those estimated on all stimuli pairs. The solid line denotes the CAT procedure, the dashed line denotes the random procedure.

to identify observers who gave unexpected judgments. They could be individuals who, instead of carefully evaluating the presented stimuli pairs, gave random, inattentive or careless responses, or gave the same response to many consecutive stimuli pairs. Otherwise, they could be atypical and clinically relevant individuals who deserve further investigation.

Infit and outfit mean-square statistics are χ^2 statistics divided by their degrees of freedom, with an expected value of 1. Values greater than 1 for an observer indicate that his/her judgments are less predictable than the Rasch model expects. Infit is influenced more by slightly unexpected judgments (i.e., those observed when the discriminative ability of the observer is similar to the difficulty of the variable stimulus to be discriminated). Outfit is influenced more by highly unexpected judgments (i.e., those observed when the discriminative ability of the observer is quite different from the difficulty of the variable stimulus to be discriminated). Observers with infit or outfit above a certain, appropriately chosen cut-off are flagged as possible observers with careless or random judgments and removed from the data set. A common choice for the cut-off is 1.5 (Wright and Linacre, 1994; Linacre, 2002).

Methods

Data Simulation

A psychophysical experiment with 11 variable stimuli at the distance of one unit from each other was considered. The smallest variable stimulus was five units smaller than the *SS*, whereas the largest variable stimulus was five units larger than the *SS*. A condition was simulated in which the observers reported which of the two stimuli of each pair (VS_i, *SS*) was the highest.

One data sample of 100 observers was simulated, by randomly drawing 100 PSE values from a normal distribution with mean = −1.5 and standard deviation = 1. This data set is denoted as the original data set. Ten of the observers in the original data set were randomly selected and their judgments to six stimuli pairs, randomly chosen among the 11 pairs, were set to be different from those in the original data set. This data set is denoted as the noisy data set.

The SLM was estimated on the two data sets. EAP estimates of the parameters β and δ were computed.

Results

The PSEs were estimated with the Rasch model and with the method of transitions (Masin and Vidotto, 1984; Burro et al., 2011). In what follows, the former are denoted as Rasch-PSEs and the latter are denoted as transition-PSEs.

The Rasch-PSEs estimated on the original data set ($M = -1.30$, $s = 1.69$) do not differ from the randomly drawn true PSEs ($M = -1.50$; $s = 1.00$) [$t(99) = -1.95$, $p = 0.05$, Cohen's $d = -0.15$, Pearson's $r = 0.78$], whereas the transition-PSEs ($M = -1.27$; $s = 1.48$) differ [$t(99) = -2.60$, $p < 0.05$, Pearson's $r = 0.78$] although the effect size is small (Cohen's $d = -0.19$).

Both Rasch-PSEs and transition-PSEs estimated on the noisy data set differ from the randomly drawn true PSEs [Rasch-PSEs: $M = -1.02$, $s = 1.78$, $t(99) = -3.49$, $p < 0.001$, Cohen's $d = -0.33$, Pearson's $r = 0.63$; transition-PSEs: $M = -1.03$, $s = 1.58$, $t(99) = -3.85$, $p < 0.001$, Cohen's $d = -0.35$, Pearson's $r = 0.62$].

Sensitivity and specificity of the cut-off at 1.5 were computed for both fit statistics (infit, outfit) that were obtained for each of the 100 observers in the noisy data set. Sensitivity refers to the capacity of correctly detecting observers with random judgments. It is the proportion of observers with fit statistic larger than 1.5 among those observers with random judgments. Specificity refers to the capacity of correctly ignoring observers without random judgments. It is the proportion of observers with fit statistic smaller than or equal to 1.5 among those observers without random judgments.

As regards outfit, the cut-off allowed for correctly identifying 8 of the 10 observers with random judgments (sensitivity = 0.80) and for correctly ignoring 86 of the 90 observers without random responses (specificity = 0.96). As regards infit, the cut-off allowed for correctly identifying 7 of the 10 observers with random judgments (sensitivity = 0.70) and for correctly ignoring 87 of the 90 observers without random responses (specificity = 0.97).

A "cleaned" data set has been obtained by removing from the noisy data set the observers with the outfit above the cut-off. Both Rasch-PSEs and transition-PSEs estimated on the cleaned data set differ from the randomly drawn true PSEs (Rasch-PSEs: $M = -1.11$, $s = 1.76$, $t(87) = -2.73$, $p < 0.01$, Cohen's $d = -0.25$, Pearson's $r = 0.70$; transition-PSEs: $M = -1.10$, $s = 1.59$, $t(87) = -3.85$, $p < 0.01$, Cohen's $d = -0.28$, Pearson's $r = 0.70$). However, the effect size of the difference between the true PSEs and those estimated on the cleaned data set is slightly smaller than that of the difference between the true PSEs and those estimated on the noisy data set (Rasch-PSEs: Cohen's $d = -0.25$, -0.33, respectively; transition-PSEs: Cohen's $d = -0.28$, -0.35, respectively). A similar result is obtained if the observers with the infit above the cut-off are removed [Rasch-PSEs: $M = -1.06$, $s = 1.79$, $t(89) = -3.12$, $p < 0.01$, Pearson's $r = 0.68$, Cohen's $d = -0.29$ vs. -0.33; transition-PSEs: $M = -1.09$, $s = 1.62$, $t(89) = -3.22$, $p < 0.01$, Pearson's $r = 0.68$, Cohen's $d = -0.30$ vs. -0.35].

In all aforementioned conditions, correlations between Rasch-PSEs and transition-PSEs are very strong (Pearson's $r \geq 0.97$) and effect sizes of the differences are small (Cohen's $d \leq 0.19$).

Results of this simulation study suggest that Rasch-based infit and outfit statistics might allow the detection of observers with unexpected judgments. If these observers are removed from the data set, a more accurate estimate of the overall PSE is obtained.

DISCUSSION

The present work provided an overview of the procedures for computing the PSE using Rasch models and proposed two new developments that are based on procedures and statistics well-established in the framework of Rasch models.

A CAT procedure has been described that allows for estimating the PSE of observers without presenting them with all stimuli pairs. Each observer is asked to judge only those stimuli pairs that are most informative about his/her PSE. The method of transitions requires presenting all stimuli pairs. As such, it cannot be used for adaptively estimating the PSE of observers. Other procedures are available in psychophysical research that can be used for this purpose. The adaptive procedures that currently enjoy widespread use may be placed into three general categories, called parameter estimation by sequential testing, maximum-likelihood adaptive procedures, and staircase procedures (Treutwein, 1995; Leek, 2001). These procedures and that described in the present work share the goal of preserving the accuracy of measurement while maximizing efficiency and minimizing observer and experimenter time.

Infit and outfit have been shown to allow the identification of observers with unexpected judgments. The judgments expressed by each of these observers must be carefully analyzed to try to find out if they are clinically relevant individuals or people who simply performed the task without due attention. Individuals may be distracted during the experiment and forget about the intensity of the stimuli after the presentation, or completely miss them, resulting in biased or random responses (Rinderknecht et al., 2018). In psychophysical experiments, inattentive responses can be identified in at least two ways. Experienced experimenters may be able to potentially detect courses of performance being visibly influenced by inattention, based on sudden performance level decreases for a certain period. However, this way of analyzing the data is not reproducible (Rinderknecht et al., 2018). Physiological signals such as electrodermal activity could potentially be used to detect inattention intervals, as arousal has been found to be a strong predictor for attention (Prokasy and Raskin, 1973). However, the measurement of electrodermal activity requires additional equipment and may not be applicable in some experimental settings. The method described in this study might allow the identification of inattentive or random responses. The strengths of this method are its reproducibility and the fact that it is based solely on the responses recorded during the experiment. Within the method of transitions, no procedure has been developed for identifying observers with unexpected judgments. A possibility in this direction could be sorting the perceptive outcomes according to the physical levels of the variable stimuli and then counting the number of runs (each of which being a sequence of equal perceptive outcomes). A large number of runs might be indicative of observers with unexpected judgments.

It is worth noting that, once the Rasch model has been estimated and validated on a suitable sample of observers, it can be used for adaptively estimating the PSE of new observers, as well as for computing their infit and outfit statistics without having to re-estimate the model parameters.

Limitations and Suggestions for Future Research

In the present work, the adaptive estimation of observers' PSEs and the detection of observers with unexpected judgments have been investigated *via* simulated data. A definitive advantage of using simulated data lies in the full knowledge of the data under consideration. Future works should investigate the usefulness of the proposed developments on real data resulting from psychophysical experiments.

In the present work, a basic Rasch-based CAT procedure has been implemented. However, the literature on CAT is rich in alternative methods that could be used for determining the starting point, selecting the stimuli pairs to be presented, updating the estimate of observer's discriminative ability, and stopping the procedure (see, e.g., Linacre, 2000; van der Linden and Glas, 2000; Wainer et al., 2000; van der Linden and Pashley, 2010; Thompson and Weiss, 2011). Future works should investigate the usefulness of these methods in psychophysical experiments and compare them with the adaptive procedures that are commonly used in psychophysical research (i.e., parameter estimation by sequential testing, maximum-likelihood adaptive procedures, staircase procedures).

In the present work, unexpected judgments have been simulated by randomly modifying the responses of some observers to some stimuli pairs. Other unexpected behaviors could be observed in psychophysical experiments (e.g., some observers could give the same response to many consecutive stimuli pairs). Moreover, in the present work, a single cut-off at 1.5 has been used. Future work could explore the usefulness of infit and outfit statistics to detect different types of response behaviors when various cut-offs are employed.

DATA AVAILABILITY STATEMENT

The R scripts used for simulating and analyzing the data will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

GV and PA contributed to the conception and design of the study. PA performed the statistical analyses and wrote the first draft of the manuscript. All authors contributed to manuscript revision, read and approved the submitted version.

FUNDING

The present work was carried out within the scope of the research program "Dipartimenti di Eccellenza," which is supported by a grant from MIUR to the Department of General Psychology, University of Padua.

REFERENCES

Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika* 43, 561–573. doi: 10.1007/BF02293814

Andrich, D. (1988). *Rasch models for measurement*. Beverly Hills: Sage.

Anselmi, P., Robusto, E., and Cristante, F. (2018). Analyzing missingness at the level of the individual respondent: comparison of five statistical tests. *Test. Psychom. Methodol. Appl. Psychol.* 25, 379–394. doi: 10.4473/TPM25.3.4

Anselmi, P., Vianello, M., and Robusto, E. (2011). Positive associations primacy in the IAT. A many-facet Rasch measurement analysis. *Exp. Psychol.* 58, 376–384. doi: 10.1027/1618-3169/a000106

Anselmi, P., Vianello, M., and Robusto, E. (2013a). Preferring thin people does not imply derogating fat people. A Rasch analysis of the implicit weight attitude. *Obesity* 21, 261–265. doi: 10.1002/oby.20085

Anselmi, P., Vianello, M., Voci, A., and Robusto, E. (2013b). Implicit sexual attitude of heterosexual, gay and bisexual individuals: disentangling the contribution of specific associations to the overall measure. *PLoS One* 8:e78990. doi: 10.1371/journal.pone.0078990

Anselmi, P., Vidotto, G., Bettinardi, O., and Bertolotti, G. (2015). Measurement of change in health status with Rasch models. *Health Qual. Life Outcomes* 13:16. doi: 10.1186/s12955-014-0197-x

Balsamo, M., Giampaglia, G., and Saggino, A. (2014). Building a new Rasch-based self-report inventory of depression. *Neuropsychiatr. Dis. Treat.* 10, 153–165. doi: 10.2147/NDT.S53425

Bechtel, G. G. (1985). Generalizing the Rasch model for consumer rating scales. *Mark. Sci.* 4, 62–73. doi: 10.1287/mksc.4.1.62

Bezruczko, N. (2005). *Rasch measurement in health sciences*. Maple Grove, MN: Jam Press.

Bock, R. D., and Jones, L. V. (1968). *The measurement and prediction of judgment and choice*. San Francisco: Holden-Day.

Bock, R. D., and Mislevy, R. J. (1988). Adaptive EAP estimation of ability in a microcomputer environment. *Appl. Psychol. Meas.* 6, 431–444. doi: 10.1177/014662168200600405

Bond, T. G., and Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah: Lawrence Erlbaum.

Burro, R., Sartori, R., and Vidotto, G. (2011). The method of constant stimuli with three rating categories and the use of Rasch models. *Qual. Quant.* 45, 43–58. doi: 10.1007/s11135-009-9282-3

Cole, J. C., Rabin, A. S., Smith, T. L., and Kaufman, A. S. (2004). Development and validation of a Rasch-derived CES-D short form. *Psychol. Assess.* 16, 360–372. doi: 10.1037/1040-3590.16.4.360

Da Dalt, L., Anselmi, P., Bressan, S., Carraro, S., Baraldi, E., Robusto, E., et al. (2013). A short questionnaire to assess pediatric resident's competencies: the validation process. *Ital. J. Pediatr.* 39:41. doi: 10.1186/1824-7288-39-41

Da Dalt, L., Anselmi, P., Furlan, S., Carraro, S., Baraldi, E., Robusto, E., et al. (2015). Validating a set of tools designed to assess the perceived quality of training of pediatric residency programs. *Ital. J. Pediatr.* 41:2. doi: 10.1186/s13052-014-0106-2

Da Dalt, L., Anselmi, P., Furlan, S., Carraro, S., Baraldi, E., Robusto, E., et al. (2017). An evaluation system for postgraduate pediatric residency programs: report of a 3-year experience. *Eur. J. Pediatr.* 176, 1279–1283. doi: 10.1007/s00431-017-2967-z

Deville, C. (1993). Flow as a testing ideal. *Rasch Meas. Trans.* 7:308.

Duncan, P. W., Bode, R. K., Lai, S. M., and Perera, S.Glycine Antagonist in Neuroprotection Americas Investigators (2003). Rasch analysis of a new stroke-specific outcome scale: the stroke impact scale. *Arch. Phys. Med. Rehabil.* 84, 950–963. doi: 10.1016/S0003-9993(03)00035-2

Embretson, S., and Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.

Laming, D., and Laming, J. (1992). F. Hegelmaier: on memory for the length of a line. *Psychol. Res.* 54, 233–239. doi: 10.1007/BF01358261

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Percept. Psychophys.* 63, 1279–1292. doi: 10.3758/BF03194543

Linacre, J. M. (1999). Understanding Rasch measurement: estimation methods for Rasch measures. *J. Outcome Meas.* 3, 382–405.

Linacre, J. M. (2000). “Computer-adaptive testing: a methodology whose time has come” in *Development of computerized middle school achievement test*. eds. S. Chae, U. Kang, E. Jeon, and J. M. Linacre (Seoul, South Korea: Komesa Press). Available at: https://www.rasch.org/memo69.pdf

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Meas. Trans.* 16, 878.

Masin, S. C., and Cavedon, A. (1970). The estimation of the point of subjective equality and its standard error by averaging equals and transitions from ‘greater’ of ‘less’ in the method of constant stimuli. A preliminary investigation. *Ita. J. Psychol.* 7, 183–186.

Masin, S. C., and Vidotto, G. (1982). A review of the formulas for the standard error of a threshold from the method of constant stimuli. *Percept. Psychophys.* 31, 585–588. doi: 10.3758/BF03204194

Masin, S. C., and Vidotto, G. (1984). The method of transitions. *Percept. Psychophys.* 36, 593–594. doi: 10.3758/BF03207521

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika* 51, 177–195. doi: 10.1007/BF02293979

Müller, G. E. (1879). Über die massbestimmungen des ortssinmes der haut mittels der methode der richtigen und falschen fälle. *Pflugers Arch.* 19, 191–235. doi: 10.1007/BF01639850

Pallant, J. F., and Tennant, A. (2007). An introduction to the Rasch measurement model: an example using the hospital anxiety and depression scale (HADS). *Br. J. Clin. Psychol.* 46, 1–18. doi: 10.1348/014466506X96931

Prokasy, W. F., and Raskin, D. C. (eds) (1973). *Electrodermal activity in psychological research*. New York: Academic Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment test*. Copenhagen: Danish Institute for Educational Research. Reprinted, 1980. Chicago: The University of Chicago Press.

Rinderknecht, M. D., Ranzani, R., Popp, W. L., Lambercy, O., and Gassert, R. (2018). Algorithm for improving psychophysical threshold estimates by detecting sustained inattention in experiments using PEST. *Atten. Percept. Psychophys.* 80, 1629–1645. doi: 10.3758/s13414-018-1521-z

Rossi Ferrario, S., Panzeri, A., Anselmi, P., and Vidotto, G. (2019). Development and psychometric properties of a short form of the illness denial questionnaire. *Psychol. Res. Behav. Manag.* 12, 727–739. doi: 10.2147/PRBM.S207622

Shea, T. L., Tennant, A., and Pallant, J. F. (2009). Rasch model analysis of the depression, anxiety and stress scales (DASS). *BMC Psychiatry* 9:21. doi: 10.1186/1471-244X-9-21

Sotgiu, I., Anselmi, P., and Meneghini, A. M. (2019). Investigating the psychometric properties of the questionnaire for Eudaimonic well-being: a Rasch analysis. *Test. Psychom. Methodol. Appl. Psychol.* 26, 237–247. doi: 10.4473/TPM26.2.5

Thompson, N. A., and Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Pract. Assess. Res. Eval.* 16, 1–9.

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vis. Res.* 35, 2503–2522. doi: 10.1016/0042-6989(95)00016-X

Urban, F. M. (1908). *The application of statistical methods to the problems of psychophysics*. Philadelphia, PA: The Psychological Clinic Press.

van der Linden, W. J. (2008). Some new developments in adaptive testing technology. *J. Psychol.* 216, 3–11. doi: 10.1027/0044-3409.216.1.3

van der Linden, W. J., and Glas, C. A. W. (eds) (2000). *Computerized adaptive testing: Theory and practice*. Dordrecht, The Netherlands: Kluwer.

van der Linden, W. J., and Pashley, P. J. (2010). “Item selection and ability estimation in adaptive testing” in *Elements of adaptive testing*. eds. W. J. van der Linden and C. A. W. Glas (New York: Springer), 3–30.

Vidotto, G., Anselmi, P., Filipponi, L., Tommasi, M., and Saggino, A. (2018). Using overt and covert items in self-report personality tests: susceptibility to faking and identifiability of possible fakers. *Front. Psychol.* 9:1100. doi: 10.3389/fpsyg.2018.01100

Vidotto, G., Bertolotti, G., Carone, M., Arpinelli, F., Bellia, V., Jones, P. W., et al. (2006). A new questionnaire specifically designed for patients affected by chronic obstructive pulmonary disease; The Italian Health Status Questionnaire. *Respir. Med.* 100, 862–870. doi: 10.1016/j.rmed.2005.08.024

Vidotto, G., Carone, M., Jones, P. W., Salini, S., and Bertolotti, G. (2007). Maugeri respiratory failure questionnaire reduced form: a method for improving the questionnaire using the Rasch model. *Disabil. Rehabil.* 29, 991–998. doi: 10.1080/09638280600926678

Vidotto, G., Ferrario, S. R., Bond, T. G., and Zotti, A. M. (2010a). Family strain questionnaire - short form for nurses and general practitioners. *J. Clin. Nurs.* 19, 275–283. doi: 10.1111/j.1365-2702.2009.02965.x

Vidotto, G., Moroni, L., Burro, R., Filipponi, L., Balestroni, G., Bettinardi, O., et al. (2010b). A revised short version of the depression questionnaire. *Eur. J. Cardiovasc. Prev. Rehabil.* 17, 187–197. doi: 10.1097/HJR.0b013e328333edc8

Vidotto, G., Pegoraro, S., and Argentero, P. (1998). Modelli di Rasch e modelli di equazioni strutturali nella validazione del locus of control in ambito lavorativo. *BPA Appl. Psychol. Bull.* 227-228, 49–63.

Vidotto, G., Robusto, E., and Zambianchi, E. (1996). I modelli simple logistic e rating scale nella determinazione del punto di eguaglianza soggettivo: Una nuova prospettiva per il metodo degli stimuli costanti. *Psychom. Methodol. Appl. Psychol.* 3, 227–235.

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B., and Mislevy, R. et al. (eds) (2000). *Computerized adaptive testing: A primer.* 2nd Edn. Mahwah, NJ: Erlbaum.

Whittaker, E. T., and Robinson, G. (1967). *The calculus of observations: An introduction to numerical analysis*. New York: Dover.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *J. Educ. Meas.* 14, 97–116. doi: 10.1111/j.1745-3984.1977.tb00031.x

Wright, B. D., and Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Meas. Trans.* 8:370.

Wright, B., and Stone, M. (1999). *Measurement essentials*. 2nd Edn. Wilmington, DE: Wide Range, Inc.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 17 December 2019
doi: 10.3389/fpsyg.2019.02716

A State Space Approach to Dynamic Modeling of Mouse-Tracking Data

Antonio Calcagnì [1], Luigi Lombardi [2], Marco D'Alessandro [2] and Francesca Freuli [2]*

[1] Department of Developmental and Social Psychology, University of Padova, Padova, Italy, [2] Department of Psychology and Cognitive Science, University of Trento, Trento, Italy

Mouse-tracking recording techniques are becoming very attractive in experimental psychology. They provide an effective means of enhancing the measurement of some real-time cognitive processes involved in categorization, decision-making, and lexical decision tasks. Mouse-tracking data are commonly analyzed using a two-step procedure which first summarizes individuals' hand trajectories with independent measures, and then applies standard statistical models on them. However, this approach can be problematic in many cases. In particular, it does not provide a direct way to capitalize the richness of hand movement variability within a consistent and unified representation. In this article we present a novel, unified framework for mouse-tracking data. Unlike standard approaches to mouse-tracking, our proposal uses stochastic state-space modeling to represent the observed trajectories in terms of both individual movement dynamics and experimental variables. The model is estimated via a Metropolis-Hastings algorithm coupled with a non-linear recursive filter. The characteristics and potentials of the proposed approach are illustrated using a lexical decision case study. The results highlighted how dynamic modeling of mouse-tracking data can considerably improve the analysis of mouse-tracking tasks and the conclusions researchers can draw from them.

Keywords: mouse tracking, state space modeling, dynamic systems, categorization task, aimed movements, Bayesian filtering

OPEN ACCESS

Edited by:
Pietro Cipresso,
Italian Auxological Institute (IRCCS), Italy

Reviewed by:
Raydonal Ospina,
Federal University of Pernambuco, Brazil
Petri Ala-Laurila,
University of Helsinki, Finland
Luca Martino,
Rey Juan Carlos University, Spain

****Correspondence:***
Antonio Calcagnì
antonio.calcagni@unipd.it

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *05 August 2019*
Accepted: *18 November 2019*
Published: *17 December 2019*

Citation:
Calcagnì A, Lombardi L, D'Alessandro M and Freuli F (2019) A State Space Approach to Dynamic Modeling of Mouse-Tracking Data. Front. Psychol. 10:2716. doi: 10.3389/fpsyg.2019.02716

1. INTRODUCTION

Over the last decades, the study of computer-mouse trajectories has brought to light new perspectives into the investigation of a wide range of cognitive processes [e.g., for a recent review see Freeman (2017)]. Unlike traditional behavioral measures, such as reaction times and accuracies, mouse trajectories may offer a valid and cost-effective way to measure the real-time evolution of ongoing cognitive processes during experimental tasks (Friedman et al., 2013). This has also been supported by recent researches investigating mouse-tracking in association to more consolidated experimental devices, such as eye-tracking and fMRI (e.g., Quétard et al., 2016; Stolier and Freeman, 2017). In a typical mouse-tracking experiment, participants are presented with a computer-based interface showing the stimulus at the bottom of the screen and two competing categories on the left and right top corners. Participants are asked to select the most appropriate label given the task instruction and stimulus while the x-y trajectories are instantaneously recorded. The main idea is that trajectories of reaching movements can unfold the decision process underlying the hand movement behavior. For instance, the curvature of computer-mouse trajectories might reveal competing processes activated in discriminating the two categories. Mouse-tracking has been successfully applied in several cognitive research studies, including lexical decision (Incera et al., 2017; Ke et al., 2017), social categorization (Carraro et al., 2016; Freeman et al., 2016), numerical cognition (Faulkenberry, 2014, 2016), memory (Papesh and Goldinger, 2012), moral decision (Koop, 2013), and lie detection (Monaro et al., 2017). Moreover, the availability of specialized and freely-available software for mouse-tracking experiments have strongly contributed to the wide-spread application of such a methodology in the more general psychological domain (Freeman and Ambady, 2010; Kieslich and Henninger, 2017). Recently, the debate on the nature of cognitive processes tracked by this type of reaching trajectories have also received attention from the motor control literature (Van Der Wel et al., 2009; Friedman et al., 2013).

So far, mouse-tracking data have been analyzed using simple strategies based on the conversion of x-y trajectories into summary measures, such as maximum deviation, area under the curve, response time, initiation time (Hehman et al., 2015). Although these steps are still meaningful in case of simple and well-behaved x-y trajectories, they can also provide biased results if applied to more complex and possibly noisy data. To circumvent these problems, other approaches have been proposed more recently (Cox et al., 2012; Calcagnì et al., 2017; Krpan, 2017; Zgonnikov et al., 2017). However, also the more recent proposals require modeling empirical trajectories before the data-analysis. Although these approaches potentially provide informative results in many empirical cases, they can also suffer from a number of issues, which revolve around the reduction of x-y data to simple scalar measures. For instance, problems may arise in the case of trajectories showing multiple phases, averaging with non-homogeneous curves, and signal-noise discrimination (Calcagnì et al., 2017). As far as we know, a proper framework to simultaneously model and analyse mouse-tracking data in a unified way is still lacking.

In this paper we describe an alternative perspective based on a state-space approach with the aim to simultaneously model and analyse mouse-tracking data. State-space models are very general time-series methods that allow estimating unobserved dynamics which gradually evolve over discrete time. As for diffusion models, which are widely used in modeling the temporal evolution of cognitive decision processes (Smith and Ratcliff, 2004), they belong to the general family of stochastic processes and offer optimal discrete approximation to many continuous differential systems used to represent dynamics with autoregressive patterns (Cox, 2017). In particular, we used a non-linear and discrete-time model that represents mouse trajectories as a function of some typical experimental manipulations. The model is estimated under a Bayesian framework, using a conjunction of a non-linear recursive filter and a Metropolis-Hastings algorithm. Data analyses is then performed using posterior distributions of model parameters (Gelman et al., 2014).

The reminder of this article is organized as follows. In section 2 we motivate our proposal by reviewing the main issues of a typical mouse-tracking experiment. In section 3 we present our proposal and describe its main characteristics. In section 4 we describe the application of our method to a psycholinguistic case study. Section 5 provides a general discussion of the results, comments and suggestions for further investigations. Section 6 concludes the article by summarizing its main findings.

2. A MOTIVATING EXAMPLE

To begin with, consider a two-choice semantic categorization task (Dale et al., 2007), in which participants have to classify semantic stimuli (e.g., name of animals) into their corresponding categories (e.g., mammal, fish). In the most typical implementation of a mouse-tracking task, participants would sit in front of a computer screen showing a resting frame (see **Figure 1A**). They start a trial by clicking a starting button at the bottom-center of the screen, after which they are presented with a given stimulus (e.g., hen, see **Figure 1B**). To finalize the trial, participants move the cursor on the screen by means of a well-tuned computer-mouse in order to reach and select one of the two labels presented on the top-left and top-right corners of the screen (e.g., mammal vs. bird, see **Figure 1C**). In the meanwhile, x-y coordinates, initiation time, and final clicking time are registered for each participant and trial. The basic idea is that x-y trajectories reflect the extent to which the real-time categorization response is affected by the experimental manipulation. More precisely, as a result of the assumption that co-activation of competing processes continuously drive the explicit hand response (Spivey and Dale, 2006), one would suppose to see more curved—or generally irregular—trajectories in association with stimuli showing higher ambiguity. In our case, for instance, it would be expected that atypical exemplars, such as hen, dolphin, and penguin, globally produce more curved or irregular trajectories than typical exemplars like dog, rabbit, and lion (see **Figures 1D,E**).

In the mouse-tracking literature, data analysis commonly proceeds summarizing the recorded trajectories by means of few indices, which are then used as input to standard statistical techniques. In the current example, for instance, the typicality manipulation could be tested by assessing whether the distribution of maximum deviations (i.e., the maximum curvature showed by trajectories) over trials and participants is bimodal or not (Freeman and Dale, 2013). In a similar way, linear models could be employed to test whether the typicality effect varies as a function of external covariates, such as psycholinguistic variables.

However, the two-step approach does have some issues. For instance, it lacks a way to represent both the experimental variability—that is induced by task manipulations—and individual variability—that is instead produced by individual-specific motor programs. Likewise, in some cases, it might neglect relevant characteristics of x-y data, with the consequence that similar classes of trajectories are treated as if they were different. Still, a two-step approach does ignore the data generation process underlying observed trajectories. This does not allow, for example, making predictions or simulations on new data given the experimental settings.

In the next section, we will present a dynamic probabilistic model that handles mouse-tracking data in a unified way.

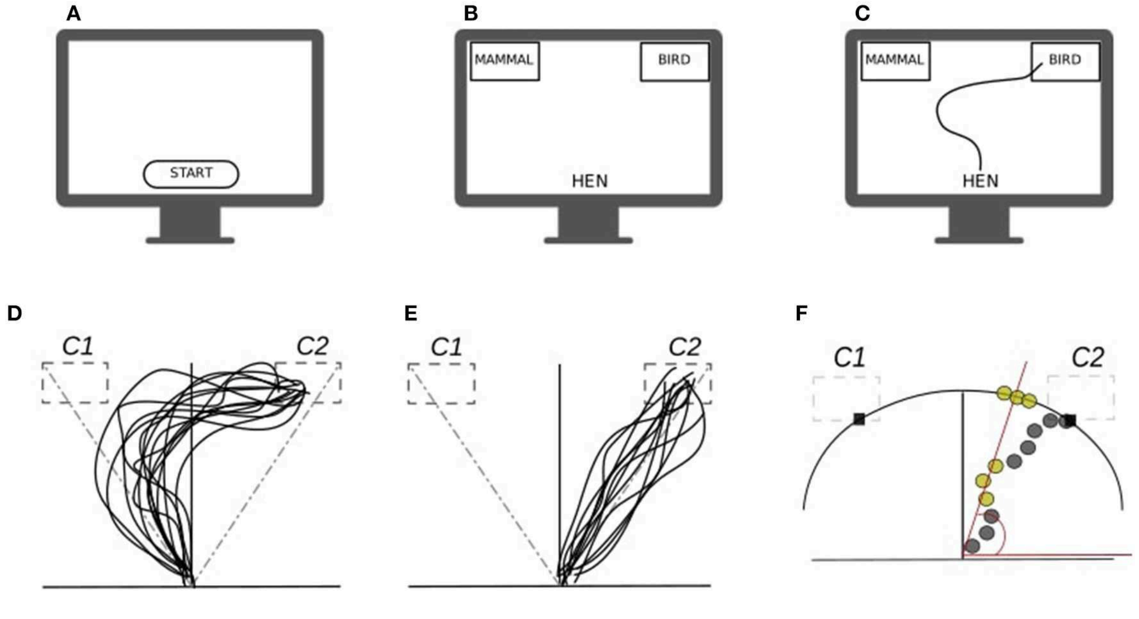


FIGURE 1 | (A–C) Conceptual diagram of a typical mouse-tracking task: **(A,B)** stimulus presentation, **(B)** participant's response. **(D,E)** Prototypical mouse-tracking trajectories collapsed over participants and trials as a function of manipulation task: **(D)** case where the manipulation has an effect—as revealed by the curvature of the trajectories, **(E)** case where the manipulation has no effect. **(F)** Conceptual diagram for the atan2 conversion: gray circles represent the sampled x-y trajectories, yellow circles represent those x-y pairs projected onto the circumference outer the Cartesian plane, whereas red lines represent the projection direction. Note that in a two-choice categorization task, the correct category C2 is presented on the top-right label (*target*) whereas the competing category C1 is presented in the opposite top-left label (*distractor*).

Our proposal is based upon a Bayesian non-linear state space approach, which offers a good compromise between model flexibility and model simplicity while overcoming many drawbacks of the standard mouse-tracking analyses.

3. STATE-SPACE MODELING OF MOUSE-TRACKING DATA

A state-space model is a mathematical description used to represent linear or generally non-linear dynamic models. In their general form, state-space systems consist of (i) a measurement density $f_y(\mathbf{y}_n; \mathbf{z}_n, \boldsymbol{\theta}_y)$ that describes how the observed vector of data $\mathbf{y}_n$ at time step n is linked to a possibly underlying process $\mathbf{z}_n$ and (ii) a state density $f_z(\mathbf{z}_n; \boldsymbol{\theta}_z)$ describing the transition dynamics that drive the vector of states $\mathbf{z}_n$. Temporal dynamics can be discrete or continuous and, in the latter case, stochastic differential equations are used to model the transition dynamics. By and large, there are two aims of any analysis involving state-space models. The first is to infer the unobserved process $\widetilde{\mathbf{Z}} = (\mathbf{z}_0, \ldots, \mathbf{z}_N)$ given the data $\mathbf{Y} = (\mathbf{y}_0, \ldots, \mathbf{y}_N)$. This task is usually accomplished by means of filtering and smoothing procedures (Jazwinski, 2007). The second aim regards estimating the parameters $(\boldsymbol{\theta}_y, \boldsymbol{\theta}_z)$ given the complete set of data $(\widetilde{\mathbf{Z}}, \mathbf{Y})$. This is commonly performed using gradient-based methods on the likelihood of the model (Shumway and Stoffer, 1982). Although state space models were originally used in the area of aerospace modeling (Kalman, 1960), they are now applied in a wide variety of domains, including control theory, remote sensing, economics, and statistics (Hamilton, 1994; Shumway and Stoffer, 2006). Recently, there has also been an increasing interest in psychology, where state-space models have been used to analyse, for example, dyadic interactions (Song and Ferrer, 2009), affective dynamics (Lodewyckx et al., 2011; Bringmann et al., 2017), facial electromyography data (Yang and Chow, 2010), individual differences (Hamaker and Grasman, 2012; Chow and Zhang, 2013), and path analysis (Gu et al., 2014).

In line with this, we developed a state-space representation to simultaneously model and analyse mouse-tracking data. In particular, our proposal is to represent the empirical collection of computer-mouse trajectories as a function of two independent sub-models, one representing the experimental manipulations (*stimuli equation*) and the other capturing the main features of the mouse movement process (*states equation*). Thus, the goal of our analysis is 2-fold: (i) to determine the states equation for each participant over a set of experimental trials, (ii) to estimate the parameters governing the stimuli equation. The first goal will provide information on how participants differ from each other in terms of movement dynamics. By contrast, the second goal will find out to what extent the experimental manipulations affect the individual variations in producing mouse-tracking responses.

3.1. Data

Let $\mathbf{S}$ be a I (individuals) × J (trials) array representing the observed data. The element $\mathbf{s}_{ij}$ of $\mathbf{S}$ defines the sub-array containing the streaming of Cartesian coordinates of the computer mouse movements:

$$\mathbf{s}_{ij} = \big((\tilde{x}_0, \tilde{y}_0), \ldots, (\tilde{x}_n, \tilde{y}_n), \ldots, (\tilde{x}_{N_{ij}}, \tilde{y}_{N_{ij}})\big)$$

with 0 and N_{ij} being the first and the last coordinates for the i-th participant in the j-th trial. The coordinates in $\mathbf{s}_{ij}$ are temporally ordered ($0 < \ldots < n < \ldots < N_{ij}$) because they are usually collected while the computer-mouse is moving along its surface with a constant sampling rate. Further, to make the observed data comparable, we rescale and normalize $\mathbf{s}_{ij}$ as a function of a common ordered scale, which indicates the cumulative amount of progressive time from 0% to N = 100% (e.g., Tanawongsuwan and Bobick, 2001; Ramsay and Silverman, 2007). Thus, the final trajectories $\mathbf{s}_{ij}$ lie on the real plane defined by the hyper-rectangles $[-1,1]^N \times [0,1]^N$, with the first movement being equal to $(\tilde{x}_{0i}, \tilde{y}_{0i}) = (0,0)$ by convention. Since we are interested in studying the co-activation of competing processes as reflected in some spatial properties of the response—such as *location*, *direction*, and *amplitude* of the action dynamics (Spivey and Dale, 2006; Freeman, 2017)—we need to simplify the original data structure so that these properties can easily emerge. Inspired by earlier work on this problem (Gowayyed et al., 2013; Kapsouras and Nikolaidis, 2014; Calcagnì et al., 2017), we reduce the dimensionality of the data by projecting $\mathbf{s}_{ij}$ in a proper lower-dimensional subspace of movement via the restricted four-quadrant inverse tangent mapping [atan2, see Burger and Burge (2010)] from the real coordinates to the interval $[0,\pi]^N$ as follows:

$$\underbrace{(y_0, \ldots, y_n, \ldots, y_N)}_{\mathbf{y}_{ij}} = \text{atan2}\big(\underbrace{(\tilde{x}_0, \tilde{y}_0), \ldots, (\tilde{x}_n, \tilde{y}_n), \ldots, (\tilde{x}_{N_{ij}}, \tilde{y}_{N_{ij}})}_{\mathbf{s}_{ij}}\big)$$

where y_0 is the angle at the beginning of the process whereas y_N is the angle at the end of the process. **Figure 1F** shows a graphical example of the atan2 function for a hypothetical movement path. Finally, the array of angles $\mathbf{y}_{ij}$ is the input for our state-space model.

3.2. Model Representation

The unobserved *states equation* of the model is a AR(1) Gaussian model $Z_{i,n}|Z_{i,n-1}$ with transition density equal to:

$$f(z_{i,n}|z_{i,n-1}, \theta) = \big(\sigma_i^2 \sqrt{2\pi}\big)^{-1} \cdot \exp\big(-(z_{i,n} - z_{i,n-1})^2 / 2\sigma_i^2\big) \quad (1)$$

which models how the movement process of the i-th subject changes from the step $n-1$ to the next step n. The stochastic dynamics for the i-th subject is constrained by the variance parameter $\sigma_i^2 \in \mathbb{R}^+$ that represents the uncertainty about the future location $z_{i,n+1}$ given the current state $z_{i,n}$.

The *measurement equation* for the observations $\mathbf{y}_{ij} = (y_0, \ldots, y_n, \ldots, y_N)$ is modeled by means of a two-component von-Mises mixture distribution with density equal to:

$$f(y_{ijn}|\pi_{ijn}, \theta) = f(y_{ijn}|\mu_1, \kappa_1)\pi_{ijn} + f(y_{ijn}|\mu_2, \kappa_2)(1 - \pi_{ijn}) \quad (2)$$

where the generic density is the standard von-Mises law:

$$f(y_{ijn}|\mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\Big(\cos(y_{ijn} - \mu)^\kappa\Big)$$

In the density formula, the term $I_0(.) \triangleq (2\pi)^{-1} \int_0^{2\pi} e^{\kappa \cos x} dx$ is the exponentially scaled Bessel function of order zero (Abramowitz and Stegun, 1972). The parameters of the mixture density are mapped to the experimental interface of the two-choice categorization task (see **Figures 1D,E**). In particular, the means $\{\mu_1, \mu_2\} \in [-3.14, 3.14]^2$ are mapped to the label categories *C1* and *C2* whereas the concentrations $\{\kappa_1, \kappa_2\} \in \mathbb{R}_+^2$ indicate how the observations are spread around the means. Since $\{\mu_1, \mu_2\}$ are determined by the fixed and known positions of the labels *C1* and *C2* on the screen, they are not treated as parameters to be estimated. Finally, the terms π_{ijn} and $(1 - \pi_{ijn})$ are the probabilities to activate the first and second density of the von-Mises components and are expressed as function of the latent states $\mathbf{z}_{i,0:N}$ and some additional covariates. The model is Markovian, in the sense that the unobserved states $\{Z_n; n > 1\}$ form a Markov sequence and the measurements $\{Y_n; n > 1\}$ are conditionally independent given the unobserved states.

To further characterize our state-space representation, the probability π_{ijn} is defined according to a logistic function:

$$\pi_{ijn} \triangleq \big(1 + \exp(-\beta_j - z_{i,n})\big)^{-1} \quad (3)$$

with $\beta_j \in \mathbb{R}$ being the intercept of the model. Equation (3) can be interpreted as the probability for the i-th subject at step n to categorize the j-th stimuli as belonging to *C1* (π_{ijn} tends to 1) or *C2* (π_{ijn} tends to 0). In addition, when the categories *C1* and *C2* are expressed in terms of distractor and target (Freeman, 2017), the sequences $\boldsymbol{\pi}_{ij,0:N}$ can be interpreted as the *attraction probability* that the distractor has exerted on the trajectory process $\mathbf{z}_{i,0:N}$.

The state-space representation is completed by linearly expanding the intercept term β_j as follows:

$$\beta_j \triangleq \sum_{k=1}^{K} d_{jk}\gamma_k + x_j\Big(\eta + \sum_{k=1}^{K} d_{jk}\delta_k\Big) \quad (4)$$

where $\{\gamma_k, \eta, \delta_k\} \in \mathbb{R}^3$, x_j is an element of the array $\mathbf{x} \in \mathbb{R}^J$, whereas d_{jk} is an element of the (Boolean) partition matrix $\mathbf{D}_{J \times K}$, with $d_{jk} = 1$ indicating whether the j-th stimulus belongs to the k-th level of the variable D. Note that the matrix $\mathbf{D}$ satisfies the property $\sum_{k=1}^K d_{jk} = 1$, for all $j = 1, \ldots, J$. In our model representation, Equation (4) is the *stimuli equation* and conveys information about the experiment. It consists of three main terms. (i) A categorical term $\sum_{k=1}^K d_{jk}\gamma_k$ describing how the stimuli $\mathcal{J} = \{1, \ldots, j, \ldots, J\}$ have been arranged into $K < J$ distinct levels of a categorical variable D. (ii) A continuous term $x_j\eta$ that expresses the stimuli as a function of a continuous variable X weighted by the coefficient η. (iii) An interaction term $x_j(\eta + \sum_{k=1}^K d_{jk}\delta_k)$ between the levels of D and X, where $\delta_k \in \mathbb{R}$ and $\delta_1 = 0$. This definition allows for modeling all the cases implied by an univariate experimental design with at most one covariate variable. Indeed, for $\eta = 0$ and $\boldsymbol{\delta}_K = \mathbf{0}_K$ this formulation boils down to the simplest experimental case with a single categorical variable D. By contrast, for $\boldsymbol{\delta}_K = \mathbf{0}_K$ and $\boldsymbol{\gamma}_K = \mathbf{0}_K$ it reduces to the case where stimuli are simply paired with a continuous predictor X. Finally, when $\mathbf{D}_{J \times K} = \mathbf{I}_{J \times J}$,

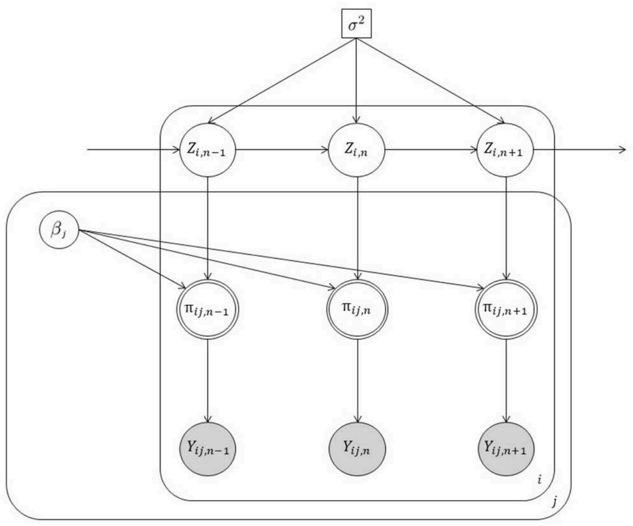


FIGURE 2 | Graphical representation of our state-space model. Note that white circles represent unobserved random variables, white double circles indicate transformed random variables, gray circles are observed random variables. Finally, square objects depict scalar quantities. Loop over individuals i, trials j, and time steps n are represented by outer squares.

the stimuli equation reduces to the most simple case where we have as many parameters as trials[1]. **Figure 2** shows a graphical representation of state-space model whereas **Figure 3** illustrates the inner-working of the model for the simplest design with a two-level experimental factor.

In our model representation, the observed movement angles $\mathbf{y}_{ij,0:N}$ are sampled from *C1* (resp. *C2*) with probabilities equal to $\boldsymbol{\pi}_{ij,0:N}$ (resp. $\boldsymbol{\pi}^c_{ij,0:N} = 1 - \boldsymbol{\pi}_{ij,0:N}$), which in turn are expressed as a function of the AR(1) latent trajectory $\mathbf{z}_{i,0:N}$. Hence, an increase in the latent process $z_{i,n} > 0$ will also increase the probability that y_{ijn} is sampled from the hemispace *C1* (e.g., $\pi_{ijn} > 0.5$). By contrast, a decrease in the latent process $z_{i,n} < 0$ will increase the chance to sample y_{ijn} from *C2* (e.g., $\pi_{ijn} < 0.5$). As a result of Equation (4) such an increasing (or decreasing) pattern can be modulated by the stimuli component $\boldsymbol{\beta}_J$. Moreover, as the coefficients $\boldsymbol{\beta}_J$ are decomposed as a function of η, $\boldsymbol{\gamma}_K$, and $\boldsymbol{\delta}_K$, we can also analyse the effect of $\boldsymbol{\beta}_J$ on $\boldsymbol{\pi}_{ij,0:N}$ in terms of the experimental manipulation D, the covariate X, or the interaction term DX. **Figure 3** shows a conceptual representation of the modeling steps involved by our approach. Panel (A) shows an example of the random-walk used to represent the movement process (Equation 3). Instead, panel (B) shows the logistic function used to form the *stimuli equation* (Equation 3) for two typical cases of $\boldsymbol{\beta}_J$. Panel (C) represents the probability $\boldsymbol{\pi}_{ij,0:N}$ to activate the distractor *C1* (upper panel) and the probability $\boldsymbol{\pi}^c_{ij,0:N}$ to activate the target *C2* (lower panel) as a function of $\mathbf{z}_{i,0:N}$ and $\boldsymbol{\beta}_J$. Finally, panel (D) depicts two cases of observed radians that are associated to $\boldsymbol{\pi}_{ij,0:N}$ and $\boldsymbol{\pi}^c_{ij,0:N}$. In particular, the upper panel shows an example of data with a pronounced attraction toward *C1*, which is in turn reflected by the blue probability curve of the panels (B,C). By contrast, the lower panel represents data with little attraction toward *C1*, as also reflected by the red probability curve of the panels (B,C). In this sense, as Equation (3) represents an intercept model, the parameter $\boldsymbol{\beta}_J$ does not affect the shape of the movement dynamics $\mathbf{z}_{i,0:N}$. On the contrary, it acts by shifting the movement dynamics upward ($\boldsymbol{\beta} < 0$) or downward ($\boldsymbol{\beta} > 0$) toward the *C1* or *C2* hemispaces, respectively.

3.3. Model Identification

State-space model identification consists of inferring the unobserved sequence of states by means of filtering and smoothing algorithms and estimating the model's parameters via Likelihood-based approximations (Shumway and Stoffer, 2006; Särkkä, 2013). For instance, in the simplest linear gaussian case, where both the states and measurement equations are linear with additive gaussian noise, inference of latent states is usually performed via Kalman filter whereas parameter's estimation is realized with the Expectation-Maximization algorithm. In our case, as Equations (3) and (4) describe a more complex non-linear model, we adopted a recursive *Gaussian approximation filter* for the inference problem (Smith and Brown, 2003), coupled with a *marginal Metropolis-Hastings* MCMC for the parameters estimation (Andrieu et al., 2010)[2].

To formulate the problem more precisely, let:

$$\boldsymbol{\Theta} = \big((\beta_1, \dots, \beta_j, \dots, \beta_J), (\kappa_1, \kappa_2)\big) \tag{5}$$

$$\mathbf{Z} = \big((z_{1,0}, \dots, z_{1,N}), \dots, (z_{i,0}, \dots, z_{i,N}), \dots, (z_{I,0}, \dots, z_{I,N})\big) \tag{6}$$

be the arrays representing all the $J \times 2$ unknown parameters and $I \times N$ unobserved states that characterize the model's behavior. In this context, $\boldsymbol{\sigma}^2_I$ can be set to $\mathbf{1}_I$ without loss of model adequacy[3]. The joint log-density of the complete-data given the array of parameters and the observed data is defined as follows:

[1] To understand the meaning of the stimuli equation, consider the case of an experiment with a two-level manipulated factor A and B, each with twenty stimuli. In this case, $K = 2$ and $J = 20$ whereas $\mathbf{D}_{20\times 2}$ is the design matrix codifying which stimulus belongs to level A ($d_{j1} = 1, d_{j2} = 0$) or level B ($d_{j1} = 0, d_{j2} = 1$). In the simple additive case, the stimuli equation is $\boldsymbol{\beta}_{20\times 1} = \mathbf{D}_{20\times 2}\boldsymbol{\gamma}_{2\times 1}$ where $\boldsymbol{\gamma}$ contains the coefficients associated to the experimental levels A and B. If we also have an external covariate $\mathbf{x}$ on the stimuli, we can include this information in the stimuli equation in two ways: (i) as additive component $\boldsymbol{\beta}_{20\times 1} = \mathbf{D}_{20\times 2}\boldsymbol{\gamma}_{2\times 1} + \eta\mathbf{x}_{20\times 1}$, (ii) by including an interaction term $\boldsymbol{\beta}_{20\times 1} = \mathbf{D}_{20\times 2}\boldsymbol{\gamma}_{2\times 1} + \eta\mathbf{x}_{20\times 1} + (\mathbf{x}_{20\times 1} \odot \mathbf{D}_{20\times 2}\boldsymbol{\delta}_{2\times 1})$, where $\boldsymbol{\delta}$ now codifies the interaction between the covariate and the levels A and B included in $\mathbf{D}$ (note that $\odot$ is the element-wise product). For further information on how codify categorical and continuous variables, see Fox (1997).

[2] Interestingly, this version of the MCMC algorithm can be subsumed into the more general family of particle-Metropolis Hasting (PMH) which, in turns, is a special case of Multiple Try Metropolis (MTM) techniques. For a broader review of these connections, see Martino (2018).

[3] Indeed, the constraint $\boldsymbol{\sigma}^2_I = \mathbf{1}_I$ still guarantees the mapping $\pi_{ijn} : \mathbb{R} \to [0, 1]$ to cover the needed time-to-time variability of the random walk, as Equation (3) acts as a shrinkage operator on the support of the r.vs $\{Z_{i,0}, \dots, Z_{i,n}\}$. This has also been confirmed by several pilot simulations we ran on our model. Note that this assumption is not overly limiting, since our state-space representation is built under the *smoothness assumption* on the movement behavior of the hand, according to which large abrupt changes in the small interval $[n, n+1]$ are not allowed (Yu et al., 2007).

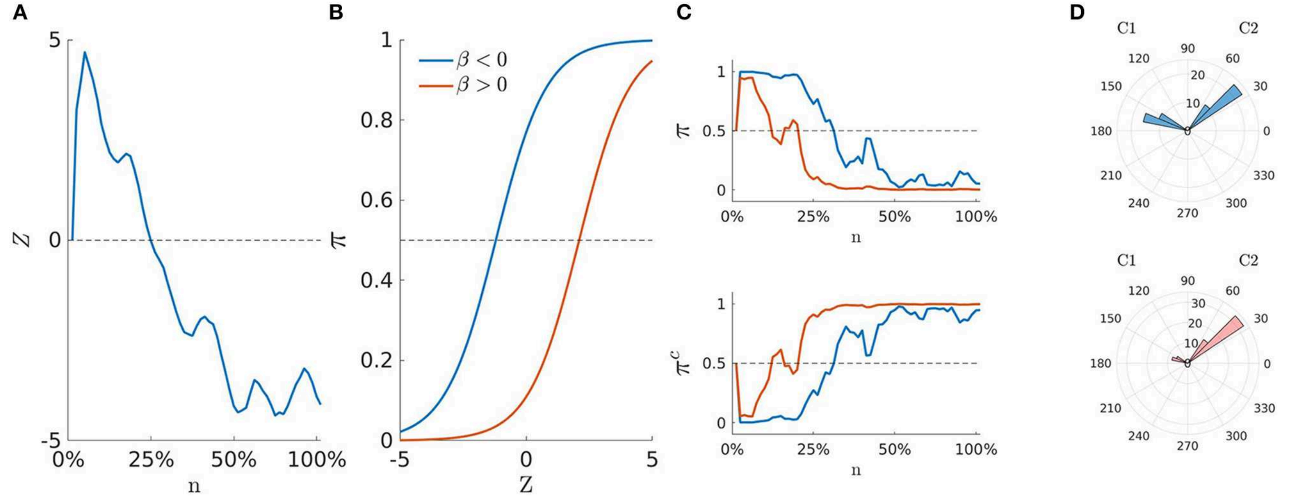


FIGURE 3 | Conceptual diagram of the state-space representation for two hypothetical sequences of mouse-trajectories. **(A)** Latent movement process $\mathbf{z}_{0:N}$. **(B)** Logistic curves π for two cases of $\boldsymbol{\beta}_J$. **(C)** Probability to activate the cue *C1* $\boldsymbol{\pi}_{0:N}$ (upper panel) and probability to activate the cue *C2* $\boldsymbol{\pi}^c_{0:N} = \mathbf{1} - \boldsymbol{\pi}_{0:N}$ (lower panel) for both $\beta < 0$ and $\beta > 0$ cases. **(D)** Measurements $\mathbf{y}_{0:N}$ as a function of their frequency (rose diagram): A case of higher attraction (upper panel) and a case of lower attraction (lower panel).

$$\log f(\mathbf{Z},\mathbf{Y}|\boldsymbol{\Theta}) = \log f(\mathbf{Y}|\mathbf{Z},\boldsymbol{\Theta}) + \log f(\mathbf{Z}|\boldsymbol{\Theta}) \tag{7}$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{J} \log f(\mathbf{z}_{i,0:N}|\boldsymbol{\Theta}) + \log f(\mathbf{y}_{ij,0:N}|\mathbf{z}_{i,0:N},\boldsymbol{\Theta}) \tag{8}$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{J}\Bigg(\log f(z_{i,0}|\theta_{Z_0}) + \sum_{n=1}^{N} \log f(z_{i,n}|z_{i,n-1},\theta_Z) + \sum_{n=1}^{N} \log f(y_{ijn}|z_{i,n},\theta_Y)\Bigg) \tag{9}$$

and the state and measurement equations are given as in (1) and (2) whereas the term $f(z_{ij0}|\theta_{Z_0})$ is the a-priori density function for the initial state of the process. Note that the factorization (9) is due to the Markovian properties of the model. By adopting the Bayesian perspective, we perform inference conditional on the observed sample of angles $\mathbf{Y}$, with $\boldsymbol{\Theta}$ being an unknown term. The posterior density $f(\mathbf{Z},\boldsymbol{\Theta}|\mathbf{Y})$ of hidden states and parameters is as follows:

$$\log f(\mathbf{Z},\boldsymbol{\Theta}|\mathbf{Y}) \propto \log f(\boldsymbol{\Theta}|\mathbf{Y}) + \log f(\mathbf{Z}|\mathbf{Y}) + \log f(\boldsymbol{\Theta}) \tag{10}$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{J} \log f(\boldsymbol{\Theta}|\mathbf{y}_{ij,0:N}) + {} + \sum_{i=1}^{I}\sum_{j=1}^{J} \log f(\mathbf{z}_{i,0:N}|\mathbf{y}_{ij,0:N}) + \log f(\boldsymbol{\Theta}) \tag{11}$$

where $f(\boldsymbol{\Theta})$ is a prior density ascribed on the vector of model's parameters $\boldsymbol{\Theta}$. Note that Equation (10) comes from the standard conditional definition $f(\mathbf{Z},\boldsymbol{\Theta}|\mathbf{Y}) = f(\mathbf{Z},\boldsymbol{\Theta},\mathbf{Y})\big/f(\mathbf{Y})$, where the joint density $f(\mathbf{Z},\boldsymbol{\Theta},\mathbf{Y})$ is re-arranged by factorization using the Markovian properties of the model (e.g., see Andrieu et al., 2010). Since our aim is to get samples from the posterior $f(\mathbf{Z},\boldsymbol{\Theta}|\mathbf{Y})$, we proceed by jointly updating $\boldsymbol{\Theta}$ and $\mathbf{Z}$ using a marginal Metropolis-Hastings. This alternates between proposing a candidate sample $\boldsymbol{\Theta}^{(t)}$ given $\boldsymbol{\Theta}^{(t-1)}$ and filtering the sequences $\mathbf{Z}^{(t)}$ conditioned on $\boldsymbol{\Theta}^{(t)}$. Finally, the candidate couple $\big(\boldsymbol{\Theta}^{(t)},\mathbf{Z}^{(t)}\big)$ is jointly evaluated by the Metropolis-Hasting ratio.

The evaluation of both the densities $f(\mathbf{Z}|\mathbf{Y})$ and $f(\boldsymbol{\Theta}|\mathbf{Y})$ involve computing the expression in Equation (11). To do so, we derived the first term by means of filtering and smoothing procedures (Jazwinski, 2007) whereas the second term was evaluated by implementing a Metropolis-Hasting algorithm. All the technical steps for the model identification are included in **Appendices A–C** whereas all the algorithms are freely available at https://github.com/antcalcagni/SSM_mousetracking.

3.4. Model Evaluation

The state-space model formulated can be evaluated in many ways under the Bayesian framework of analysis (Shiffrin et al., 2008; Gelman et al., 2014). For instance, adequacy of the algorithm can be assessed via standard diagnostic measures, such as traceplot of the chains, autocorrelation measures, and the Gelman-Rubin statistics whereas the recovery of the true model structure can be done by simulations from the priors ascribed to the model (Gelman et al., 2014). Similarly, the adequacy of the model to reproduce the observed data can be assessed by means of simulation-based procedures (e.g., posterior predictive checks) where the fitted model is used to generate new simulated datasets that are then compared to the observed data (e.g., see Gelman et al., 1996; Cook et al., 2006). In our context, the robustness of the model formulation in recovering the true model structure as well as the goodness of fit to the observed data are assessed by

TABLE 1 | Model summary: observed and latent variables, parameters, and equations of the state-space model formulated for the analysis of mouse-tracking trajectory.

$i \in \{1, \ldots, I\}$	Set index for individuals
$j \in \{1, \ldots, J\}$	Set index for trials
$n \in \{0, \ldots, N\}$	Set index for (discrete) time points
$\mathbf{y}_{ij} \in (0, \pi]^N$	Observed $N \times 1$ array of mouse-tracking data
$\mathbf{z}_{i,0:N} \in \mathbb{R}^N$	$N \times 1$ Array of latent states to be inferred
$\mathbf{x} \in \mathbb{R}^J$	Observed covariate of the experiment
$Y_{ijn} \sim \text{mix-vonMises}\ (\mu_1, \mu_2, \kappa_1, \kappa_2)$	Random variable governing the realization of y_{ijn}
$Z_{in} \vert Z_{i,n-1} \sim \mathrm{N}(Z_{i,n-1}, \sigma_i)$	AR(1) random process governing the realization of z_{in}
$\{\mu_1, \mu_2\} \in (-\pi, \pi)^2$	Fixed parameters of the mixture von-Mises law (true locations of the stimuli)
$\{\kappa_1, \kappa_2\} \in \mathbb{R}^2$	Parameters of the von-Mises law (precision)
$\sigma_i \in \mathbb{R}_0^+$	Fixed parameter of the Gaussian law (standard deviation)
$\boldsymbol{\pi}_{ij,0:N} \in [0, 1]^N$	$N \times 1$ array of attraction probability (i.e., probability to activate distractor vs. target hemi-space)
$\boldsymbol{\beta} \in \mathbb{R}^J$	$J \times 1$ array of coefficients (intercepts) modeling the experimental design
$\boldsymbol{\beta} = \mathbf{D}\gamma + \mathbf{x}(\eta + \mathbf{D}\gamma)$	Linear equation governing the experimental design in terms of additive and interaction effects
$\mathbf{D} \in \mathbb{R}^{J \times K}$	$J \times K$ Boolean partition matrix for the experimental design
$\boldsymbol{\gamma} \in \mathbb{R}^K$	$K \times 1$ array of coefficients for the additive components of the experimental design
$(\eta, \boldsymbol{\delta}) \in \mathbb{R}^{K+1}$	$(K + 1) \times 1$ array of coefficients for the interaction components of the experimental design
$\boldsymbol{\Theta} = (\boldsymbol{\gamma}, \boldsymbol{\delta}, \eta, \kappa_1, \kappa_2) \in \mathbb{R}^{2K+1} \times \mathbb{R}_0^{+^2}$	Complete array of parameters to be estimated (some of them can be set to zero, depending on the experimental design)

adopting a simulation-based approach. Technical details on this procedure are available in **Supplementary Materials**.

3.5. Model Summary

Table 1 shows a summary of the components of the complete state-space model used throughout the paper, including observed and latent variables, parameters and their support spaces.

4. APPLICATION

In this section, we will present an application of the model to the analysis of an already published lexical decision dataset (Barca and Pezzulo, 2012). The state-space modeling framework will be evaluated via three different instances of model representation with an increasing level of complexity. Note that the application we report here has only an illustrative purpose with the main goal to introduce and highlight the interpretation of the model's parameters and the flexibility of its representation with dynamic data. All the models were estimated using 20 (chains) × 10,000 (iterations), with a burning-in period of 2500 iterations. Starting values $\boldsymbol{\theta}_0$ for the MH algorithm were determined by maximizing the observed likelihood of the model in Equation (2). Similarly, the starting covariance matrix $\boldsymbol{\Sigma}^{(0)}$ was computed by using the Hessian of the observed likelihood at $\boldsymbol{\theta}_0$. The adaptive phase of the MH algorithm was performed at fixed interval $t + H$ (with $H = 25$) to prevent the degeneracy of the adaptation. For each model, the prior densities were defined as $f(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu} = \mathbf{0}, \mathbf{1}\sigma^2 = 25)$, where the variance was sufficiently large to cover the natural range of the model parameters. The adequacy of the model to reproduce the data was evaluated with a simulation-based approach, where a series of $M = 5,000$ new datasets $(\mathbf{Y}_1^*, \ldots, \mathbf{Y}_M^*)$ were generated through the fitted model and compared with the observed data $\mathbf{Y}$ (e.g., see Cook et al., 2006). The goodness of fit was evaluated *overall* (i.e., the adequacy of the model to reproduce the complete observed matrix $\mathbf{Y}$) and *subject-based* (i.e., the adequacy of the model to reproduce for each subject $i = 1, \ldots, I$ the observed matrix $\mathbf{Y}_i$). Comparisons were computed by means of 0–100% normalized measures, with 0% indicating bad fit and 100% optimal fit. Technical details as well as extended graphical results are included in **Supplementary Materials**.

4.1. General Context and Motivation

Lexical decision is one of the most known and widely used task to study visual world recognition and reading in the cognitive psycholinguistic literature (Norris and Kinoshita, 2008; Yap et al., 2008; Hawkins et al., 2012). Generally, this task is very simple and versatile and provides an ideal context for applying the state-space modeling framework when lexical decision data are collected via the mouse tracking paradigm. In this application, we evaluated the extent to which the parameters of the state-space model reflect eventual differences associated with the manipulation of a stimulus type factor composed by words (with either high-frequency or low-frequency) and random strings (i.e., random sequence of letters that are phonotactically illegal in the language) in the lexical decision task. Moreover, we will take advantage of this psycholinguistic case study to show how our state-space model can deal with both categorical and (pseudo)quantitative predictive variables considered either individually or in interaction in the model. In particular, the first model instance will illustrate the application of our modeling framework when a simple categorical variable (stimulus type factor) is considered to affect the observed mouse-tracking trajectories collected using the lexical decision task. By contrast, the second model will be based on a simple regression-type model with a single quantitative independent variable (bigram frequency) used to predict the attraction toward the distractor category. Finally, the third model will integrate these two

TABLE 2 | Application: adequacy of the model to reproduce the observed matrices **Y** (*overall fit*) and $\mathbf{Y}_i$ (*by-subject fit*).

	Overall (%)	By-subject (%)
Model 1	84	78
Model 2	73	70
Model 3	75	71

All the measures are normalized in the range 0% (bad fit)–100% (optimal fit). See ***Supplementary Materials*** *for technical details.*

TABLE 3 | Application: posterior means (μ), 95% posterior intervals ($[q_{0.05}, q_{0.975}]$), and Gelman-Rubin $\hat{R}$ index for the estimated parameters of Models 1–3.

		$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\eta}$	$\hat{\delta}_1$	$\hat{\delta}_2$
Model 1	$q_{0.05}$	1.224	1.234	1.211			
	μ	1.323	1.337	1.310			
	$q_{0.975}$	1.443	1.457	1.432			
	$\hat{R}$	1.003	1.002	1.003			
Model 2	$q_{0.05}$				0.063		
	μ				0.078		
	$q_{0.975}$				0.091		
	$\hat{R}$				1.001		
Model 3	$q_{0.05}$	0.083	1.130	1.217	0.305	−0.505	−0.437
	μ	0.341	1.300	1.314	0.402	−0.385	−0.336
	$q_{0.975}$	0.605	1.468	1.411	0.500	−0.269	−0.235
	$\hat{R}$	1.008	1.013	1.012	1.011	1.013	1.010
All the models	$\hat{\kappa}_1 = 22.31$						
	$\hat{\kappa}_2 = 44.96$						

variables (stimulus type factor and bigram frequency) into a unified model including the main effects of the two variables as well as their interaction. In our context, the first two models will be considered as simple toy examples to illustrate the main features of the state-space model representation when applied to real data, whereas the third model will be discussed in more details according to a group analysis evaluation as well as an individual analysis representation.

4.2. Model 1

4.2.1. Data Structure and Variables

In the original work by Barca and Pezzulo (2012), the lexical decision experiment was run in Italian and based only on one stimulus type factor with four different levels: Words of high written frequency (HF, e.g., acqua "water"), words of low written frequency (LF, e.g., cervo "deer"), pseudowords (PW, e.g., "dorto"), and strings of letters that are orthographically illegal in Italian (NW, e.g., "btfpr"). In their study, participants saw a total of 96 stimuli, one at the time, and were required to categorize each stimulus as either a word or a non-word by using the mouse-tracking paradigm. Trajectories were recorded using the Mouse Tracker software (Freeman and Ambady, 2010) with sampling rate of ∼70 Hz (Barca and Pezzulo, 2012). As usual, raw trajectories were normalized into $N = 101$ time steps using linear interpolation with equal spaces between coordinate samples. However, for our analysis we preferred to select only three of the four levels of the experimental factor (that is to say, HF,LF, and NW) for a total of 72 stimuli equally distributed within each level[4]. Finally, the dependent variable **Y** of Model 1 consisted of the movement angles array associated with the mouse-movement trajectory recorded for each distinct stimulus in the stimulus set.

4.2.2. Data Analysis and Results

In this first model the term β_j in the stimuli equation boils down to the simple expression:

$$\beta_j = \sum_{k=1}^{3} d_{jk}\gamma_k$$

where the indices $k = 1, 2, 3$ refer to HF, LF, and NW stimuli. The MCMC convergences of the algorithm are reported in **Supplementary Materials**. The model fitted the data very satisfactorily, with overall fit of 84% and subject-based fit of 74% (see **Table 2**). The posterior quantiles (5, 50, and 95%) are reported in **Table 3** whereas **Figure 4A** shows the probability graph, that is to say, the probability to activate the distractor cue for each of the three levels HF, LF, NW as a function of the latent variable Z.

The results of this first analysis clearly show that the dynamics of the state-space model were unaffected by the different categories represented in the recoded experimental factor. This pattern finds further support in the *post-hoc* comparisons between the three experimental conditions (**Figure 4B**). In sum, these findings indicate that for a dynamic model represented according to a state-space modeling framework, the three stimulus categories (HF, LF, and NW) were all processed in a very similar way, as the original trajectories were not sufficiently different among the three stimulus categories. In substantive terms, the results of the categorical model showed how the attraction probability toward the distractor was definitively modest in all the three experimental conditions. This is evident from a direct inspection of **Figure 4B** where the probability activation function (logistic function) is shifted toward right ($Z > 0$) which in turn means that the average activation of the distractor category was relatively poor in HF, LF, and NW items. In this respect, the results of our simple spatial model were partially at odds with the outcomes observed using temporal measures (response time variables) (Barca and Pezzulo, 2012).

4.3. Model 2

4.3.1. Data Structure and Variables

Also for the second model, the dependent variable was represented by the movement angles array **Y**. However, unlike model 1, in model 2 the original independent categorical variable (stymulus type factor) was replaced with a quantitative

[4] The motivation for this selection was due to some technical reasons regarding the lack of design balance in the original dataset, as the PW level showed a large number of errors when compared with the other three categories. In addition, the three-level representation of the stimulus type factor simplifies the interpretation of the results when we consider the full model with interaction.

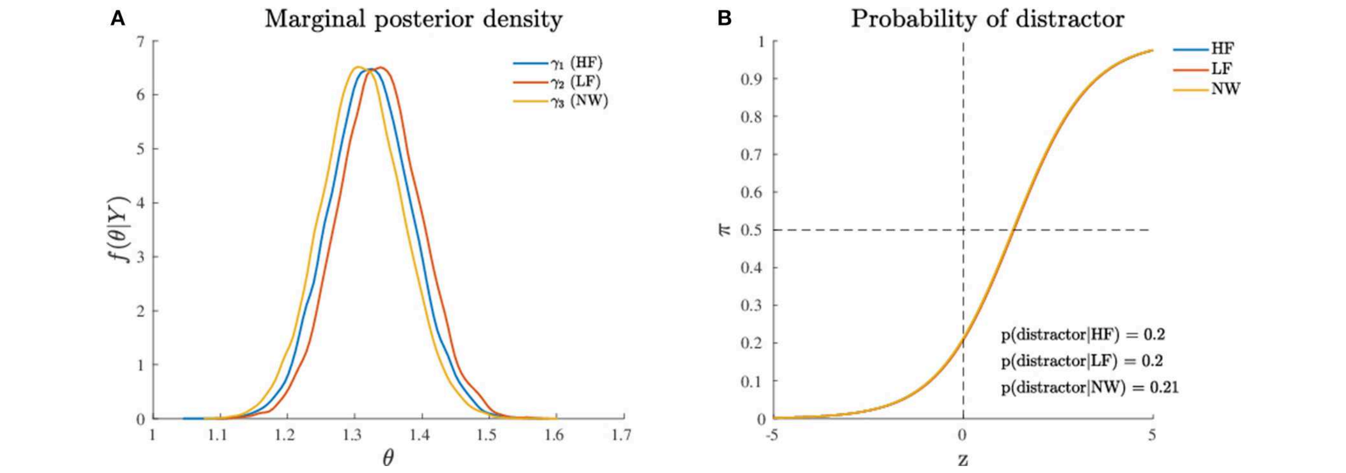


FIGURE 4 | Application (Model 1): **(A)** Marginal posterior densities for the model parameters and **(B)** Probability to activate the distractor cue as a function of the levels (HF, LF, NW) of the categorical variable. Note that the densities in **(A)** are shown together for the sake of comparison.

psycholinguistic variable called *bigram frequency*. Bigram frequency is defined as the frequency with which adjacent pairs of letters (bigrams) occur in printed texts; for its characteristics, it may be considered as a measure of orthographic typicality (e.g., see Hauk et al., 2006). In this second application, only bigram frequency was used as quantitative variable, since it was the only psycholinguistic variable that could be computed for all the 72 stimuli in the stimulus set. This second model instance nicely provides a simple but effective example of application of our state-space model when a continuous variable is considered to predict the attraction toward distractor.

4.3.2. Data Analysis and Results

In model 2 the term β_j simply reduces to:

$$\beta_j = x_j \eta$$

as the first and third terms in formula (4) cancel out. In this case, the variable x_j denotes the value of the bigram frequency for stimulus j in the stimulus set. For the model results, the posterior quantiles are reported in **Table 3** whereas MCMC convergences of the algorithm are reported in **Supplementary Materials**. Also in this case, the model fitted the data very well, with overall fit of 73% and subject-based fit of 70% (see **Table 2**). **Figure 5** shows the probability graph for model 2. This graph represents the probability to activate the distractor hemispace at three representative levels of the variable, i.e., the lowest, the medium, and the highest values of bigram. As evident from the graph, bigram frequency affects the probability to activate the target, with a higher probability for stimuli with low bigram frequency.

In substantive terms, the results of the quantitative model supported the evidence that the attraction probability toward the distractor was slightly affected by the specific value of the quantitative predictor (bigram frequency). In particular, low-level bigram frequencies were characterized by an average larger activation probability (0.55) for the distractor, whereas medium or large frequencies were associated with a logistic function slightly shifted toward positive values of the latent space $Z > 0$, thus reflecting a lower chance for the distractor category (average activation probability of 0.45). Moreover, by an inspection of the contingency table for the joint representation of bigram frequency (as a transformed categorical variable) and stimulus type, we noted that low bigram frequency values were mainly characterized by string letters (NW: 94%) whereas high bigram frequency values were predominantly associated with high frequency words (HF: 55%) or low frequency words (LF: 44%).

4.4. Model 3

4.4.1. Data Structure and Variables

The final and more complex model included both the three-level categorical predictor (stimulus type factor: HF, LF, STR) and the continuous predictor (bigram frequency) as well as the interaction term between these two variables. The dependent variable was the movement angles array **Y**.

4.4.2. Data Analysis and Results

The stimuli equation which characterizes the third model is defined as follows:

$$\beta_j = \sum_{k=1}^{3} d_{jk}\gamma_k + x_j\left(\eta + \sum_{k=1}^{3} d_{jk}\delta_k\right)$$

The MCMC diagnostics together with the estimated marginal posterior densities for the model's parameters are reported in **Supplementary Materials**. The model fit was good, with an overall fit of 75% whereas the subject-based fit was equal to 71% (see **Table 2**). The posterior quantiles are reported in **Table 3**. **Figure 6** shows the probability graph for model 3. This graph represents the probability to activate the distractor hemispace for each of the three levels HF, LF, NW of the categorical factor as a function of the latent variable Z and three distinct levels (high, medium, and low) for bigram frequency. The inspection of **Figure 6** shows a clear interaction between stimulus

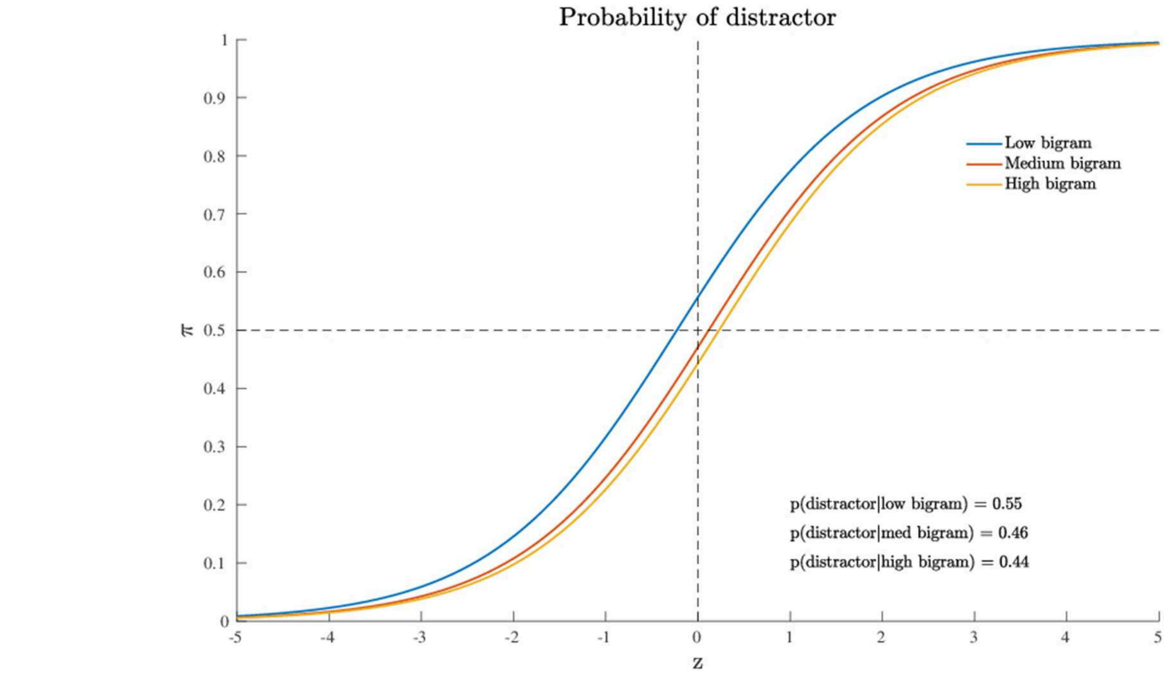


FIGURE 5 | Application (Model 2): probability to activate the distractor cue as a function of the continuous variable. For Note that just three representative levels (low, middle, high) are represented for the sake of graphical interpretation.

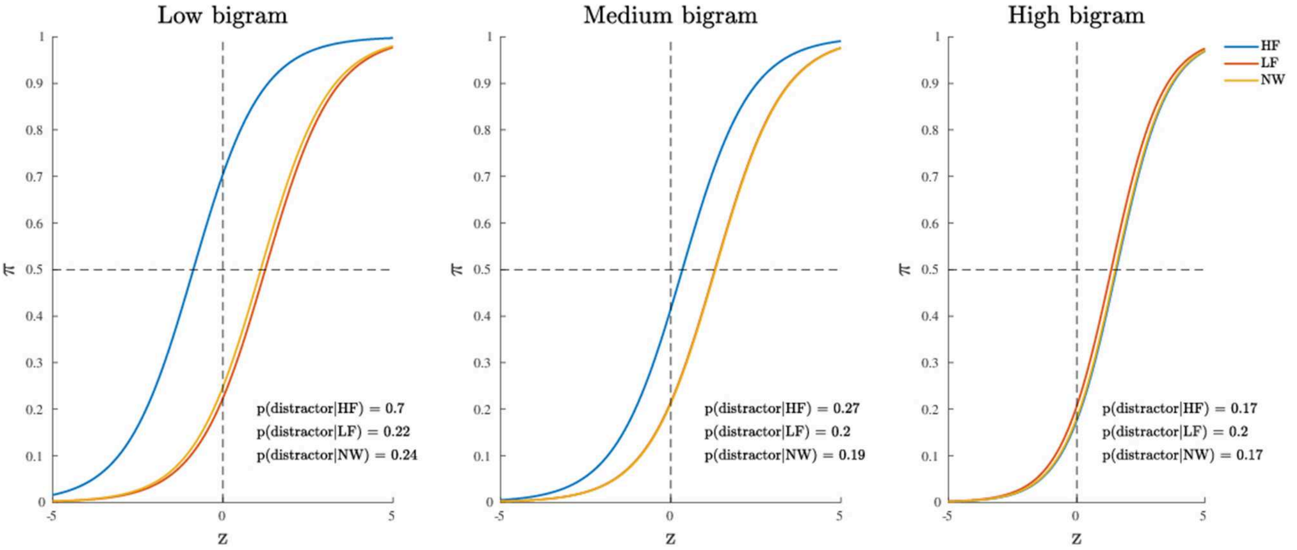


FIGURE 6 | Application (Model 3): probability to activate the distractor cue as a function of the categorical variable (within panels) and three representative levels of the continuous variable (between panels).

type factor and bigram frequency indicating that the impact of stimulus category, in particular word frequency, increases with the decrease of stimulus bigram frequency. In other words, at high level of bigram frequency, the probability to activate the distractor is similarly low in all conditions ($0.17 \leq p$-distractor ≤ 0.2). By contrast, when bigram frequency decreases—that is stimuli become orthographically atypical—the probability of distractor activation increases, but only for the more lexically-familiar stimuli, i.e., words of high frequency (e.g., p-distractor raises from 0.17 to 0.70, in low and high bigram frequency condition, respectively).

Finally, it is also worth mentioning the emergence of the main effect of stimulus category which was instead missing in model 1. By a quick inspection of **Figure 7**, one may clearly

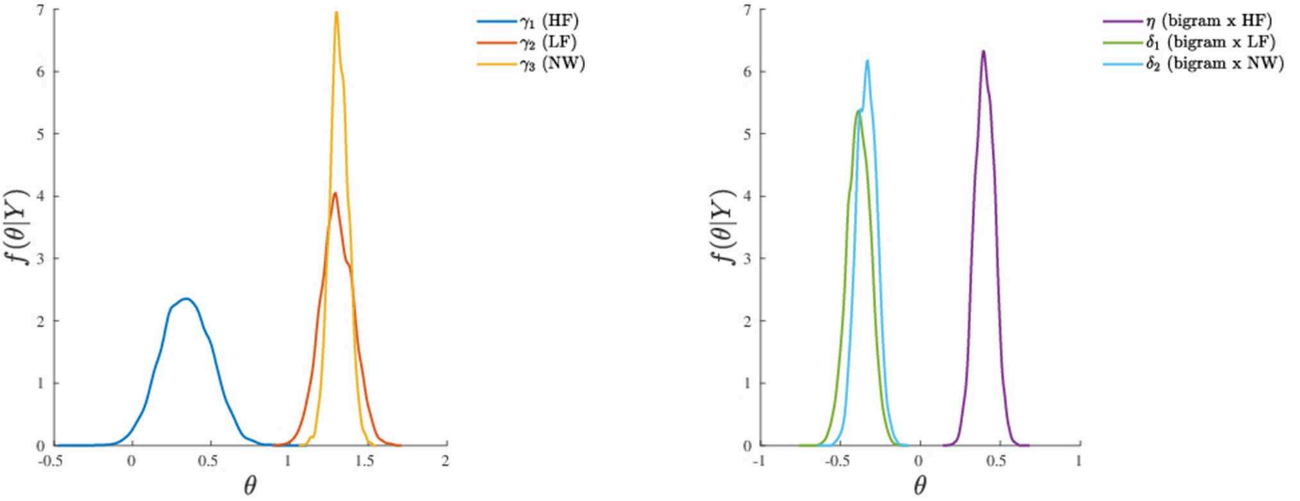


FIGURE 7 | Application (Model 3): marginal posterior densities for the model parameters. **(Left)** Parameters associated to the categorical variable. **(Right)** Parameters associated to the continuous variable and its interaction with the categorical variable. Note that the densities are shown together for the sake of comparison.

observe that HF words differ from both LF words and letter strings (NW), whereas LF words and letter strings do not differ with respect to the probability of activation of the distractor hemispace. Interestingly, the addition of the covariate bigram frequency in the model allowed the main effect of stimulus category to show up. Indeed, while at the medium and high levels of bigram frequency the results are in line with those observed at a sample level in the original study (see Figures 1, 2, 5 in the paper Barca and Pezzulo, 2012) and in a recent re-analysis (see Table 2 in the paper Calcagnì et al., 2017), in the case of low bigram the probability to activate the distractor increases with respect to high frequency words (HF). This might be somewhat related to a moderate difficulty in the orthographic processing of low frequency bigram words (e.g., see Rastle and Davis, 2008) even in the case of stimuli with richer lexical representation.

4.5. Profiles Analysis

To further investigate the dynamic characteristics involved in the lexical decision task, we extend here the results of the third model to include also a profiles analysis. **Figure 8** shows the estimated latent movement states $\mathbf{Z}_{I\times N}$ for all the participants involved in the study. The profiles appear regular, as they evolve smoothly toward the target cue (T). We grouped the dynamics into four well-separated clusters (**Figure 8**, smallest panels on the right) according to their functional similarities (Ramsay and Silverman, 2007). Particularly, the first group is characterized by a higher exploration of the distractor's hemispace, especially in the first 30% of the process. The same applies to the third and fourth groups, although they show a gradual activation of the distractor. Finally, the second group clearly represents those profiles with no uncertainty in the categorization process, as they show no activation of the distractor's hemispace at all. Although well-separated among them, these clusters still show some level of inner heterogeneity (for example, see group 1 and 4). To study this latter issue in terms of experimental manipulations, we focused on group 1 and considered the low vs. high frequency conditions (HF vs. LF). We also selected the middle phase of the process ($\Delta = 30 - 50\%$), where it is expected to observe larger cognitive competitions in the categorization (Barca and Pezzulo, 2012). **Figure 9** shows the participants' profiles in terms of attraction probability $\boldsymbol{\pi}_{4\times N}$ for the two lexical conditions. As expected, the profiles differ between these conditions, with LF eliciting higher attraction probability. This is in line with the fact that low frequency words have a weaker lexical representation than high frequency stimuli and consequently they are more difficult to process (Barca and Pezzulo, 2012). Interestingly, the individual profiles also differ in the way they activate the distractor. For instance, the participant 6 had higher probability in both LF ($p_\Delta(D) = 0.67$) and HF ($p_\Delta(D) = 0.54$) conditions whereas the participant 7 had a more pronounced activation just in the LF condition ($p_\Delta(D) = 0.57$) than HF ($p_\Delta(D) = 0.43$). Similarly, participants 6 and 7 seemed to prolong the competing dynamics up to the 50% of the process, by contrast participants 8 and 15 seemed to resolve the lexical competition earlier as showed by the abrupt decreasing of their curves. We complete our analysis by evaluating how individual profiles are linked to empirical measurements. **Figure 10** represents this scenario for two stimuli belonging to HF and LF conditions. As we can notice, the curves present the same dynamics (due to the latent states $\mathbf{z}_{i,0:N}$) although they clearly differ in terms of attraction exerted by the stimulus (due to the β component of the model). In this case, the LF stimulus produced larger conflict than HF in the lexical categorization. This is evident when we turn back to the observed data: as expected, the rose diagrams of LF showed larger directions in the distractor's hemispace.

5. DISCUSSION

We have described a new approach to model and analyse dynamic data coming from mouse-tracking experiments. Our

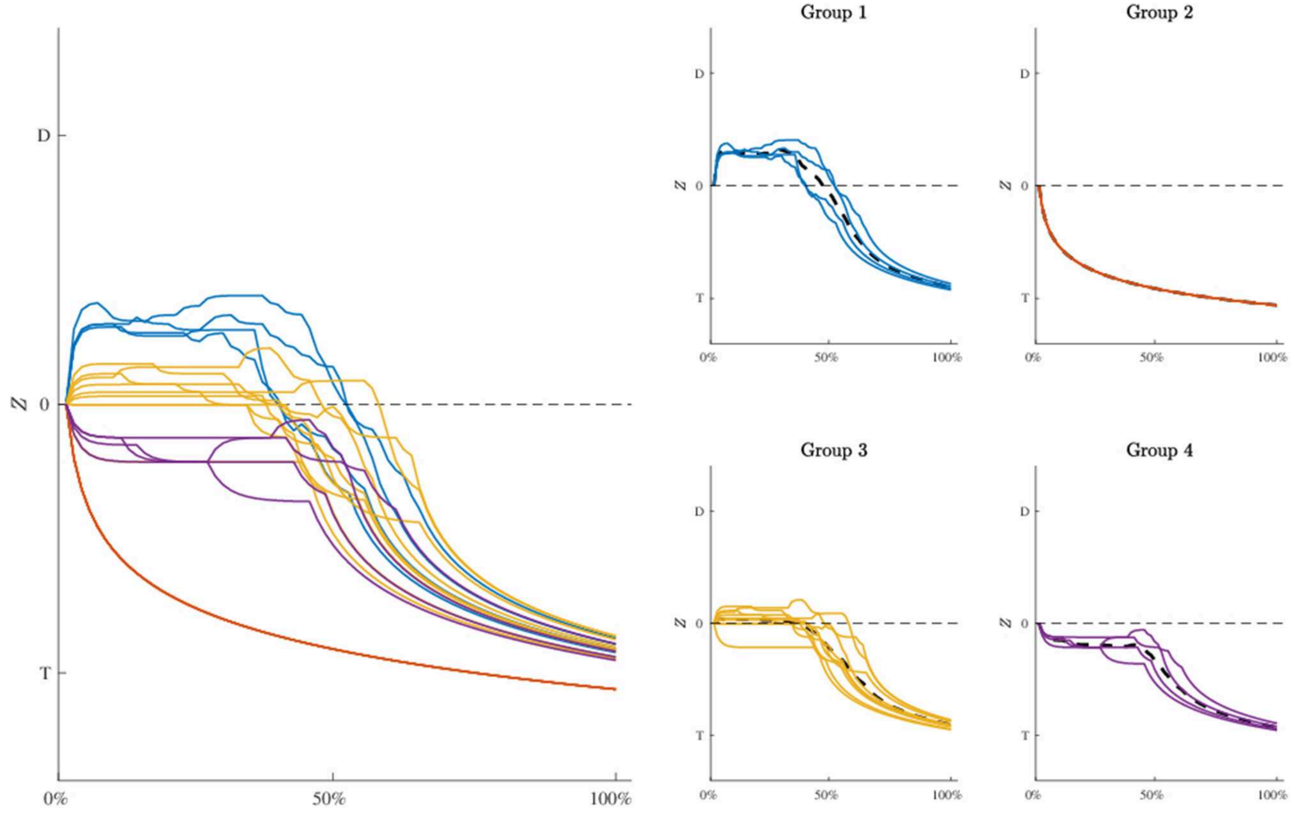


FIGURE 8 | Application: estimated movement dynamics $\mathbf{z}_{i,0:N}$ of each participant (biggest panel, **Left**) and clusters of profiles in terms of their functional similarity (smallest panels, **Right**). Note that averaged profiles are represented as dashed lines whereas D and T in all the panels indicate distractor and target, respectively. Groups' composition: participants 6, 7, 8, 15 (group 1), 1, 4, 19, 21 (group 2), 2, 3, 5, 12, 13, 16, 17, 20, 22 (group 3), 10, 11, 14, 18 (group 4).

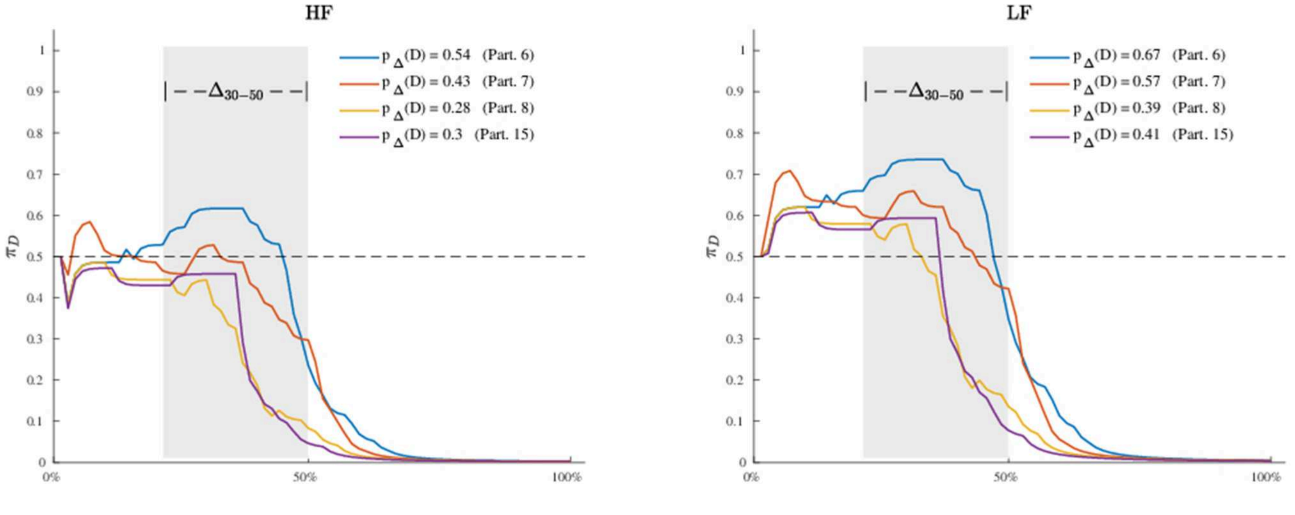


FIGURE 9 | Application: estimated attraction probabilities $\boldsymbol{\pi}_{i,0:N}$ of participants in Group 1 for the high frequency **(Left)** and low frequency **(Right)** lexical conditions. Note that the probability curves are computed with respect to the distractor (D), the gray area in both panels indicates a selected window of processing ($\Delta = 30 - 50\%$), whereas the terms $p_\Delta(D)$ are computed using a normalized discrete approximation of the integral of the probability curves in the selected process window Δ.

proposal took the advantages of a state-space representation, in which the observed data $\mathbf{Y}$ were thought as being function of two independent sub-models, one representing the movement process and its properties ($\mathbf{Z}$) and the second modeling the two-choice experimental task ($\boldsymbol{\beta}$) according to which the data were collected. These sub-models were integrated by means of an

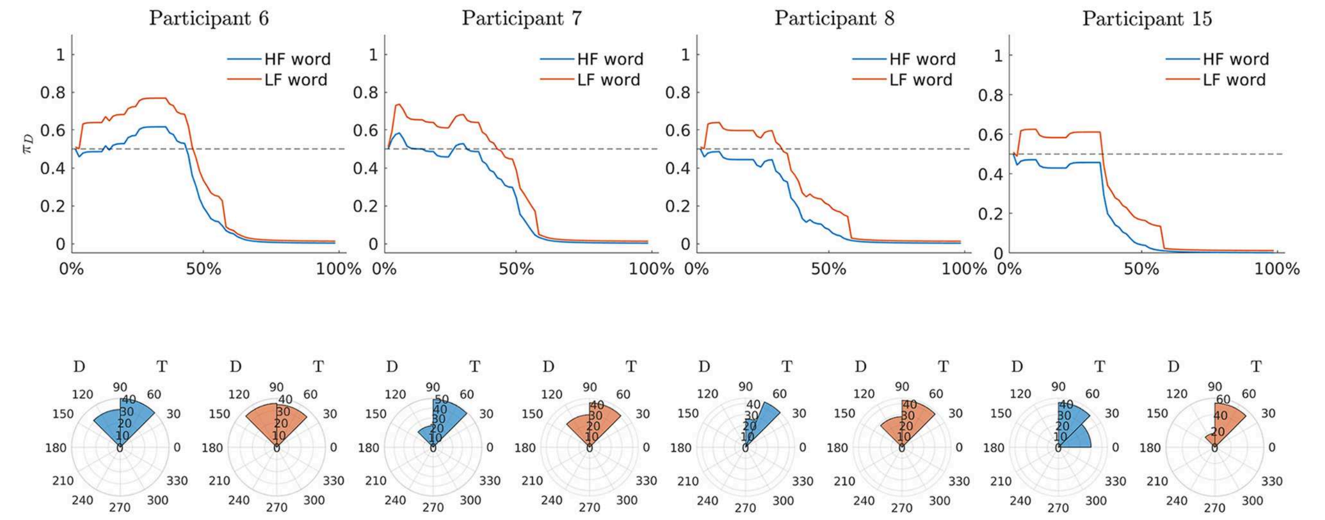


FIGURE 10 | Application: estimated attraction probabilities $\boldsymbol{\pi}_{i,0:N}$ of participants in Group 1 and rose diagrams of observed radians for two stimuli (HF: *epoca*, epoch. LF: *zampa*, paw). Note that D and T in all the panels indicate distractor and target, respectively.

inverse-logit function (π) that expressed how the experimental manipulations acted on the movement processes in selecting the final correct response against the competing one. This formulation was flexible enough to take into account the complexity of some dynamic behaviors showed by the reaching trajectories. Moreover, it allowed for separately accounting for the motor heterogeneity and experimental variability in $\mathbf{Y}$. Indeed, when $\boldsymbol{\beta} = \mathbf{1}\beta_0$ our state-space representation simply reduced to a model where the experimental manipulations had no relevant effect in reproducing the observed data. This instance has been illustrated in section 4.3 (Model 1). In this case, as $\mathbf{Z} = \mathbf{0}$ was not allowed in our model formulation, all the variability of $\mathbf{Y}$ can be ascribed to $\mathbf{Z}$. This is relevant in view of the fact that movement variability may reflect only individual motor executions in absence of any experimental manipulations (Yu et al., 2007). The movement component $\mathbf{Z}$ was modeled to be Markovian with gaussian transition density.

Although more complex models can be used to represent movement dynamics, simple random walks still allows a great deal of flexibility in modeling reaching trajectories under weak assumptions on the movement behavior (e.g., see Yu et al., 2007; Paninski et al., 2010). In particular, in the case of mouse-tracking tasks, they allow representations of the following three properties: (i) Each movement is goal-oriented as individuals have to finalize the action by clicking on one of the two categories shown on the screen. (ii) Mouse-tracking trajectories generally start at rest, proceed out in the movement space, and end at rest. (iii) Hand trajectories tend to be smooth during the reaching process, i.e., small changes in the interval $[n, n+1]$ are more likely than large and abrupt changes (Brockwell et al., 2004; Spivey et al., 2010). The stimuli component $\boldsymbol{\beta}$ was defined to be a linear combination of information typically involved in a univariate design, namely a categorical variable D containing the levels of the experimental factor and a continuous covariate X. This gave researchers the opportunity to additionally analyse which component of the experimental design is relevant in producing the effect of $\boldsymbol{\beta}$ on $\mathbf{Y}$. The data-generation process was defined according to a mixture of two von-Mises distributions representing the categories of a two-choice categorization task. Among others, we chose this distribution because it provides a flexible representation for angular ordered data, especially because it simplifies mathematical computations involved in the model's derivation (e.g., see McClintock et al., 2012; Mulder and Klugkist, 2017).

There are other existing methods that offer alternative ways to model mouse-tracking data. For instance, Van Der Wel et al. (2009) proposed the use of the *movement superposition model* (Henis and Flash, 1995) to model and analyse the typical two-choice lexical decision task. In particular, they modeled mouse-tracking trajectories by representing the complete hand movement as a summation of sub-movements, which were obtained by the solution of the minimum-jerk equation for the standard reaching trajectory (i.e., a movement characterized by a bell-shaped speed profile that minimizes the sum of the squared rates of jerks over the movement duration). Similarly, Friedman et al. (2013) discussed how an intermittent model of arms movement can be used for reaching trajectories in random-dot experiments. They used both Wiener's diffusion process and Flash and Hogan's movement equation to predict reaction times (RTs) and movement data. Their goal was to assess the link between movement trajectories and underlying cognitive processing. Our model differs in some respects from these works. With regards to Van Der Wel et al. (2009), for instance, we used a stochastic state-space approach to model the movement trajectories instead of deterministic equations. Instead, with respect to Friedman et al. (2013), we tailor-made our model to a typical two-choice categorization task, making use of few assumptions on the nature of the movement process [as those implied by the Gaussian AR(1) process]. By and large, our goal was not to provide a mathematical representation of the cognitive components underpinning mouse-trajectories since the model does not describe the cognitive processing *per se*. By contrast, we simply provided a *statistical model* for the analysis of mouse-tracking data, which can offer a good compromise between data modeling and data analysis.

5.1. Model's Advantages and Limitations

Our non-linear state-space model has several advantages. For instance, when comparing with the standard approaches, our proposal provides a unified analytic framework to simultaneously model and analyse trajectories data. By modeling movement heterogeneity and task variability together, we can evaluate how experimental variables directly act on the observed series of trajectories, with no need to use any kind of summary measures. An additional advantage of our model concerns the study of individual differences in terms of latent dynamics. While this is impractical in standard two-step approaches, in our proposal researchers can assess individual variations by studying the movement profiles $\tilde{\mathbf{Z}}$ once they are estimated. For instance, they can be analyzed in terms of similarity/dissimilarity with regards to external individual covariates (e.g., vocabulary knowledge and bilingualism in psycholinguistic experiments; IQ, risk-taking propensity, or more generally clinical variables in decision-making tasks). Still, individual dynamics can be compared each other qualitatively in terms of chance to activate the distractor or target cues. As the dynamics are normalized on a common cumulative scale, researchers can assess whether the chance to activate the distractor cue at a certain percentage of the process and for an experimental manipulation, is particularly higher in a sub-group of participants (this case, for example, has been described in section 4.6).

As for any modeling approach, also the current proposal can potentially suffer from some limitations. A first limitation concerns the only-intercept model $\pi(\mathbf{Z}, \boldsymbol{\beta})$ used to integrate individual dynamics and experimental information. Although this was enough to represent whether or not certain stimuli can increase the probability to select the distractor cue, we may want to known whether some experimental manipulations can modify the individual dynamics as well. However, this would particularly pronounce the computational costs required for the model identification (especially with regards to filtering), as we need to appropriately generalize Equation (3) to include more parameters. Lastly, in the current study we used univariate non-linear state-space models to represent individual dynamics for the sake of parsimony. However, more complicated situations may require models including further movement characteristics like step-length, velocity, acceleration, and jerk (Kulkarni and Paninski, 2008), which may be modeled as statistical constraints of the model (Ciavolino and Calcagnì, 2014; Calcagnì et al., 2017).

5.2. Further Extensions

Our non-linear state-space model can be improved in many aspects. For instance, the stimuli equation (4) can be generalized to cope with more complex experimental designs, like those involving multiple factors and covariates together with high-order interaction terms. Likewise, the current model restrictions can be relaxed to allow changes in slopes of $\pi(\mathbf{Z}, \boldsymbol{\beta})$ as a function of the experimental stimuli. Further, the development of a hierarchical representation of the model, with a random-effect component in the state Equation (3), would offer a way to model the inter-individual variations as resulting from an underlying common population. Still, the development of a multivariate state-space model to include other movement components will surely be considered a future extension of the present work. Further studies may lead to generalize the AR(1) process used for the movement dynamics to include former knowledge on the deterministic constraints of the hand movement as those used, for instance, by Van Der Wel et al. (2009) and Friedman et al. (2013). Moreover, further studies may also lead to generalize the AR(1) process used for the movement dynamics to include former knowledge on the deterministic constraints of the hand movement as those used, for instance, by Van Der Wel et al. (2009) and Friedman et al. (2013). Finally, an open issue which deserves greater consideration in future investigations is the need for a formal comparative framework with which we may eventually contrast and compare spatial modeling perspectives (like the one presented in this contribution) and currently used methods for analyzing mouse tracking data based on descriptive statistics (e.g., see Freeman, 2017).

6. CONCLUSIONS

In this paper we presented a novel and comprehensive analytic framework for modeling and analyse mouse-tracking trajectories. In particular, a non-linear state-space approach was used to model the observed trajectories as a function of both individual movement dynamics and experimental variables. Model identification was performed under the umbrella of Bayesian methods, in which a Metropolis-Hastings algorithm was coupled with a recursive gaussian approximation filter to get posterior distributions of model parameters. For the sake of illustration, we applied our new approach to a real mouse-tracking dataset concerning a two-choice lexical categorization task. The results indicated how our proposal can provide valuable insights to assess the dynamics involved in the decision task and identify how the experimental variables significantly contributed to the observed movement heterogeneity. Moreover, the analysis of individual profiles allowed for comprehensive and reliable identification of individual and group-based differences in the dynamics of decision making.

In conclusion, we think that this work yielded interesting findings in the development of computational models able to capture the unfolding high-level cognitive processes as reflected by motor executions which are typically involved in mouse-tracking tasks. To our knowledge, this is the first time that mouse-tracking data are fully modeled and analyzed within a process-oriented approach. We believe our contribution will offer a novel strategy that may help cognitive researchers to understand the roles of cognition and action in mouse-tracking based experiments.

DATA AVAILABILITY STATEMENT

The datasets analyzed in this manuscript are not publicly available. Requests to access the datasets should be directed to Laura Barca, laurabarcapst@gmail.com.

AUTHOR CONTRIBUTIONS

AC and LL contributed to the conceptualization and statistical modelization. AC, MD'A, and LL performed the main statistical analyses. AC and MD'A performed the supplementary statistical analyses. FF contributed to the statistical analyses and revised supplementary statistical analyses. MD'A contributed to the section Model Evaluation. FF contributed to the section Application. AC wrote the first draft of the study. AC, LL, MD'A, and FF revised the final manuscript.

FUNDING

This work was supported by CARITRO, Fondazione Cassa di Risparmio di Trento e Rovereto (ref. grant 2016.0189).

ACKNOWLEDGMENTS

The authors thank Dr. Simone Sulpizio for many helpful discussions on earlier versions of the manuscript and Dr. Laura Barca for providing the dataset used in the case study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02716/full#supplementary-material

REFERENCES

Abramowitz, M., and Stegun, I. A. (1972). *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*. New York, NY: Dover.

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle markov chain monte carlo methods. *J. R. Stat. Soc. B Stat. Methodol.* 72, 269–342. doi: 10.1111/j.1467-9868.2009.00736.x

Andrieu, C., and Roberts, G. O. (2009). The pseudo-marginal approach for efficient monte carlo computations. *Ann. Stat.* 37, 697–725. doi: 10.1214/07-AOS574

Ansley, C. F., and Kohn, R. (1982). A geometrical derivation of the fixed interval smoothing algorithm. *Biometrika* 69, 486–487.

Banerjee, A., Dhillon, I. S., Ghosh, J., and Sra, S. (2005). Clustering on the unit hypersphere using von mises-fisher distributions. *J. Mach. Learn. Res.* 6, 1345–1382.

Barca, L., and Pezzulo, G. (2012). Unfolding visual lexical decision in time. *PLoS ONE* 7:e35932. doi: 10.1371/journal.pone.0035932

Bringmann, L. F., Hamaker, E. L., Vigo, D. E., Aubert, A., Borsboom, D., and Tuerlinckx, F. (2017). Changing dynamics: time-varying autoregressive models using generalized additive modeling. *Psychol. Methods* 22, 409–425. doi: 10.1037/met0000085

Brockwell, A. E., Rojas, A. L., and Kass, R. (2004). Recursive bayesian decoding of motor cortical signals by particle filtering. *J. Neurophysiol.* 91, 1899–1907. doi: 10.1152/jn.00438.2003

Burger, W., and Burge, M. J. (2010). *Principles of Digital Image Processing: Fundamental Techniques*. London: Springer Science & Business Media.

Cabras, S., Nueda, M. E. C., and Ruli, E. (2015). Approximate bayesian computation by modelling summary statistics in a quasi-likelihood framework. *Bayesian Anal.* 10, 411–439. doi: 10.1214/14-BA921

Calcagnì, A., Lombardi, L., and Sulpizio, S. (2017). Analyzing spatial data from mouse tracker methodology: an entropic approach. *Behav. Res. Methods* 49, 2012–2030. doi: 10.3758/s13428-016-0839-5

Carraro, L., Castelli, L., and Negri, P. (2016). The hand in motion of liberals and conservatives reveals the differential processing of positive and negative information. *Acta Psychol.* 168, 78–84. doi: 10.1016/j.actpsy.2016.04.006

Chow, S.-M., and Zhang, G. (2013). Nonlinear regime-switching state-space (RSSS) models. *Psychometrika* 78, 740–768. doi: 10.1007/s11336-013-9330-8

Ciavolino, E., and Calcagnì, A. (2014). A generalized maximum entropy (GME) approach for crisp-input/fuzzy-output regression model. *Qual. Quant.* 48, 3401–3414. doi: 10.1007/s11135-013-9963-9

Cook, S. R., Gelman, A., and Rubin, D. B. (2006). Validation of software for Bayesian models using posterior quantiles. *J. Comput. Graph. Stat.* 15, 675–692. doi: 10.1198/106186006X136976

Cox, D. R. (2017). *The Theory of Stochastic Processes*. New York, NY: Routledge.

Cox, G., Kachergis, G., and Shiffrin, R. (2012). "Gaussian process regression for trajectory analysis," in *Proceedings of the Cognitive Science Society*, Austin, TX: Cognitive Science Society, Vol. 34.

Dale, R., Kehoe, C., and Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Mem. Cogn.* 35, 15–28. doi: 10.3758/BF03195938

De Jong, P. (1988). The likelihood for a state space model. *Biometrika* 75, 165–169.

Faulkenberry, T. J. (2014). Hand movements reflect competitive processing in numerical cognition. *Can. J. Exp. Psychol.* 68, 147–151. doi: 10.1037/cep0000021

Faulkenberry, T. J. (2016). Testing a direct mapping versus competition account of response dynamics in number comparison. *J. Cogn. Psychol.* 28, 825–842. doi: 10.1080/20445911.2016.1191504

Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods.* Thousand Oaks, CA: Sage Publications, Inc.

Freeman, J. B. (2017). Doing psychological science by hand. *Curr. Dir. Psychol. Sci.* 27, 315–323. doi: 10.1177/0963721417746793

Freeman, J. B., and Ambady, N. (2010). Mousetracker: software for studying real-time mental processing using a computer mouse-tracking method. *Behav. Res. Methods* 42, 226–241. doi: 10.3758/BRM.42.1.226

Freeman, J. B., and Dale, R. (2013). Assessing bimodality to detect the presence of a dual cognitive process. *Behav. Res. Methods* 45, 83–97. doi: 10.3758/s13428-012-0225-x

Freeman, J. B., Pauker, K., and Sanchez, D. T. (2016). A perceptual pathway to bias: interracial exposure reduces abrupt shifts in real-time race perception that predict mixed-race bias. *Psychol. Sci.* 27, 502–517. doi: 10.1177/0956797615627418

Friedman, J., Brown, S., and Finkbeiner, M. (2013). Linking cognitive and reaching trajectories via intermittent movement control. *J. Math. Psychol.* 57, 140–151. doi: 10.1016/j.jmp.2013.06.005

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*, Vol. 2. Boca Raton, FL: CRC Press.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Stat. Sinica* 6, 733–760.

Gowayyed, M. A., Torki, M., Hussein, M. E., and El-Saban, M. (2013). "Histogram of oriented displacements (HOD): describing trajectories of human joints for action recognition," in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence* (AAAI Press), 1351–1357.

Gu, F., Preacher, K. J., and Ferrer, E. (2014). A state space modeling approach to mediation analysis. *J. Educ. Behav. Stat.* 39, 117–143. doi: 10.3102/1076998614524823

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive metropolis algorithm. *Bernoulli* 7, 223–242. doi: 10.2307/3318737

Hamaker, E., and Grasman, R. (2012). Regime switching state-space models applied to psychological processes: handling missing data and making inferences. *Psychometrika* 77, 400–422. doi: 10.1007/s11336-012-9254-8

Hamilton, J. D. (1994). State-space models. *Handb. Econometr.* 4, 3039–3080.

Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., and Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *Neuroimage* 30, 1383–1400. doi: 10.1016/j.neuroimage.2005.11.048

Hawkins, G., Brown, S. D., Steyvers, M., and Wagenmakers, E.-J. (2012). Decision speed induces context effects in choice. *Exp. Psychol.* 59, 206–215. doi: 10.1027/1618-3169/a000145

Hehman, E., Stolier, R. M., and Freeman, J. B. (2015). Advanced mouse-tracking analytic techniques for enhancing psychological science. *Group Process. Intergr. Relat.* 18, 384–401. doi: 10.1177/1368430214538325

Henis, E. A., and Flash, T. (1995). Mechanisms underlying the generation of averaged modified trajectories. *Biol. Cybernet.* 72, 407–419.

Incera, S., Shah, A. P., McLennan, C. T., and Wetzel, M. T. (2017). Sentence context influences the subjective perception of foreign accents. *Acta Psychol.* 172, 71–76. doi: 10.1016/j.actpsy.2016.11.011

Jazwinski, A. H. (2007). *Stochastic Processes and Filtering Theory*. New York, NY: Courier Corporation.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Eng.* 82, 35–45.

Kapsouras, I., and Nikolaidis, N. (2014). Action recognition on motion capture data using a dynemes and forward differences representation. *J. Visual Commun. Image Represent.* 25, 1432–1445. doi: 10.1016/j.jvcir.2014.04.007

Ke, A. H., Epstein, S. D., Lewis, R., and Pires, A. (2017). The quantificational domain of DOU: an experimental study. *J. Psycholinguist. Res.* 47, 537–556. doi: 10.1007/s10936-017-9532-9

Kieslich, P. J., and Henninger, F. (2017). Mousetrap: an integrated, open-source mouse-tracking package. *Behav. Res. Methods* 49, 1652–1667. doi: 10.3758/s13428-017-0900-z

Koop, G. J. (2013). An assessment of the temporal dynamics of moral decisions. *Judg. Decis. Making* 8, 527–539.

Krpan, D. (2017). Behavioral priming 2.0: enter a dynamical systems perspective. *Front. Psychol.* 8:1204. doi: 10.3389/fpsyg.2017.01204

Kulkarni, J. E., and Paninski, L. (2008). State-space decoding of goal-directed movements. *IEEE Signal Process. Mag.* 25, 78–86. doi: 10.1109/MSP.2008.4408444

Lodewyckx, T., Tuerlinckx, F., Kuppens, P., Allen, N., and Sheeber, L. (2011). A hierarchical state space approach to affective dynamics. *J. Math. Psychol.* 55, 68–83. doi: 10.1016/j.jmp.2010.08.004

Luengo, D., and Martino, L. (2013). "Fully adaptive gaussian mixture metropolis-hastings algorithm," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE), 6148–6152.

Marin, J.-M., Mengersen, K., and Robert, C. P. (2005). Bayesian modelling and inference on mixtures of distributions. *Handb. Stat.* 25, 459–507. doi: 10.1016/S0169-7161(05)25016-2

Martino, L. (2018). A review of multiple try mcmc algorithms for signal processing. *Digital Signal Process.* 75, 134–152. doi: 10.1016/j.dsp.2018.01.004

McClintock, B. T., King, R., Thomas, L., Matthiopoulos, J., McConnell, B. J., and Morales, J. M. (2012). A general discrete-time modeling framework for animal movement using multistate random walks. *Ecol. Monogr.* 82, 335–349. doi: 10.1890/11-0326.1

Mendel, J. M. (1995). *Lessons in Estimation Theory for Signal Processing, Communications, and Control*. Upper Saddle River, NJ: Pearson Education.

Monaro, M., Gamberini, L., and Sartori, G. (2017). The detection of faked identity using unexpected questions and mouse dynamics. *PLoS ONE* 12:e0177851. doi: 10.1371/journal.pone.0177851

Mulder, K., and Klugkist, I. (2017). Bayesian estimation and hypothesis tests for a circular generalized linear model. *J. Math. Psychol.* 80, 4–14. doi: 10.1016/j.jmp.2017.07.001

Norris, D., and Kinoshita, S. (2008). Perception as evidence accumulation and bayesian inference: insights from masked priming. *J. Exp. Psychol. Gen.* 137, 434–455. doi: 10.1037/a0012799

Paninski, L., Ahmadian, Y., Ferreira, D. G., Koyama, S., Rahnama Rad, K., Vidne, M., et al. (2010). A new look at state-space models for neural data. *J. Comput. Neurosci.* 29, 107–126. doi: 10.1007/s10827-009-0179-x

Papesh, M. H., and Goldinger, S. D. (2012). Memory in motion: movement dynamics reveal memory strength. *Psychonom. Bull. Rev.* 19, 906–913. doi: 10.3758/s13423-012-0281-3

Quétard, B., Quinton, J. C., Mermillod, M., Barca, L., Pezzulo, G., Colomb, M., et al. (2016). Differential effects of visual uncertainty and contextual guidance on perceptual decisions: evidence from eye and mouse tracking in visual search. *J. Vis.* 16, 28–28. doi: 10.1167/16.11.28

Ramsay, J. O., and Silverman, B. W. (2007). *Applied Functional Data Analysis: Methods and Case Studies*. Springer.

Rastle, K., and Davis, M. H. (2008). Morphological decomposition based on the analysis of orthography. *Lang. Cogn. Process.* 23, 942–971. doi: 10.1080/01690960802069730

Särkkä, S. (2013). *Bayesian Filtering and Smoothing*, Vol. 3. New York, NY: Cambridge University Press.

Shampine, L. F. (2008). Vectorized adaptive quadrature in matlab. *J. Comput. Appl. Math.* 211, 131–140. doi: 10.1016/j.cam.2006.11.021

Shiffrin, R. M., Lee, M. D., Kim, W., and Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical bayesian methods. *Cogn. Sci.* 32, 1248–1284. doi: 10.1080/03640210802414826

Shumway, R. H., and Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the em algorithm. *J. Time Series Anal.* 3, 253–264.

Shumway, R. H., and Stoffer, D. S. (2006). *Time Series Analysis and Its Applications: With R Examples*. Springer Science & Business Media.

Smith, A. C., and Brown, E. N. (2003). Estimating a state-space model from point process observations. *Neural Comput.* 15, 965–991. doi: 10.1162/089976603765202622

Smith, P. L., and Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends Neurosci.* 27, 161–168. doi: 10.1016/j.tins.2004.01.006

Song, H., and Ferrer, E. (2009). State-space modeling of dynamic psychological processes via the kalman smoother algorithm: rationale, finite sample properties, and applications. *Struct. Equat. Model.* 16, 338–363. doi: 10.1080/10705510902751432

Spivey, M. J., and Dale, R. (2006). Continuous dynamics in real-time cognition. *Curr. Dir. Psychol. Sci.* 15, 207–211. doi: 10.1111/j.1467-8721.2006.00437.x

Spivey, M. J., Dale, R., Knoblich, G., and Grosjean, M. (2010). Do curved reaching movements emerge from competing perceptions? A reply to van der wel et al. (2009). *J. Exp. Psychol.* 36, 251–254. doi: 10.1037/a0017170

Stolier, R. M., and Freeman, J. B. (2017). A neural mechanism of social categorization. *J. Neurosci.* 37, 5711–5721. doi: 10.1523/JNEUROSCI.3334-16.2017

Tanawongsuwan, R., and Bobick, A. (2001). "Gait recognition from time-normalized joint-angle trajectories in the walking plane," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001*, Vol. 2.

Tanner, M. A. (1991). *Tools for Statistical Inference*, Vol. 3. New York, NY: Springer.

Van Der Wel, R. P., Eder, J. R., Mitchel, A. D., Walsh, M. M., and Rosenbaum, D. A. (2009). Trajectories emerging from discrete versus continuous processing models in phonological competitor tasks: a commentary on Spivey, Grosjean, and Knoblich (2005). *J. Exp. Psychol.* 35, 588–594. doi: 10.1037/0096-1523.35.2.588

Yang, M., and Chow, S.-M. (2010). Using state-space model with regime switching to represent the dynamics of facial electromyography (EMG) data. *Psychometrika* 75, 744–771. doi: 10.1007/s11336-010-9176-2

Yap, M. J., Balota, D. A., Tse, C.-S., and Besner, D. (2008). On the additive effects of stimulus quality and word frequency in lexical decision: evidence for opposing interactive influences revealed by RT distributional analyses. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 495–513. doi: 10.1037/0278-7393.34.3.495

Yu, B. M., Kemere, C., Santhanam, G., Afshar, A., Ryu, S. I., Meng, T. H., et al. (2007). Mixture of trajectory models for neural decoding of goal-directed movements. *J. Neurophysiol.* 97, 3763–3780. doi: 10.1152/jn.004 82.2006

Zgonnikov, A., Aleni, A., Piiroinen, P., O'Hora, D., and di Bernardo, M. (2017). Decision landscapes: visualizing mouse-tracking data. *Open Sci.* 4:170482. doi: 10.1098/rsos.170482

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 10 January 2020
doi: 10.3389/fpsyg.2019.02970

Machine Learning in Psychometrics and Psychological Research

Graziella Orrù[1], Merylin Monaro[2], Ciro Conversano[1], Angelo Gemignani[1] and Giuseppe Sartori[2]*

[1] Department of Surgical, Medical, Molecular and Critical Area Pathology, University of Pisa, Pisa, Italy, [2] Department of General Psychology, University of Padua, Padua, Italy

Recent controversies about the level of replicability of behavioral research analyzed using statistical inference have cast interest in developing more efficient techniques for analyzing the results of psychological experiments. Here we claim that complementing the analytical workflow of psychological experiments with Machine Learning-based analysis will both maximize accuracy and minimize replicability issues. As compared to statistical inference, ML analysis of experimental data is model agnostic and primarily focused on prediction rather than inference. We also highlight some potential pitfalls resulting from adoption of Machine Learning based experiment analysis. If not properly used it can lead to over-optimistic accuracy estimates similarly observed using statistical inference. Remedies to such pitfalls are also presented such and building model based on cross validation and the use of ensemble models. ML models are typically regarded as black boxes and we will discuss strategies aimed at rendering more transparent the predictions.

Keywords: machine learning, cross-validation, replicability, machine learning in psychological experiments, machine learning in psychometrics

OPEN ACCESS

Edited by:
Pietro Cipresso,
Italian Auxological Institute (IRCCS),
Italy

Reviewed by:
Raydonal Ospina,
Federal University of Pernambuco,
Brazil
Miguel Lozano,
University of Valencia, Spain

****Correspondence:***
Graziella Orrù
graziella.orru@unipi.it;
graziella.orru@gmail.com

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: *20 August 2019*
Accepted: *16 December 2019*
Published: *10 January 2020*

Citation:
Orrù G, Monaro M, Conversano C, Gemignani A and Sartori G (2020) Machine Learning in Psychometrics and Psychological Research. Front. Psychol. 10:2970. doi: 10.3389/fpsyg.2019.02970

INTRODUCTION

The use of Machine Learning (ML) in psychometrics has attracted media attention after the Cambridge Analytica affair which dominated headlines around the world after the election of President Trump. Originally, academics from the Psychometric Centre from the University of Cambridge United Kingdom, collected a huge number of social media data (on over 50.000 participants) in order to predict personality of Facebook (FB) profile owners on the basis of their FB behavior. This research yielded a highly influential publication (Kosinski et al., 2013) were the authors showed how FB-based behaviors (i.e., likes) could be used to identify private traits with high accuracy (Christianity vs. Islam AUC = 0.82; Democrats vs. Republican, AUC = 0.88). Widespread attention arose because these data were opaquely leaked from the academic researchers to Cambridge Analytica, the now-infamous firm that scraped Facebook psychometric test data to construct millions of psychographic profiles, which it then used to hyper-target voters with custom-made campaign ads in favor of the Candidate Donald Trump during the presidential race of 2016.

In short, Cambridge Analytica targeted "persuadable," voters whose psychographic profiles (mostly a Big Five profiling) suggested they were open to suggestion.

A less media-attracting example of the use of ML in psychological science is the field of Psychometric Credit Score. A Psychometric Credit Score is a predictive model based on a microcredit applicant psychological and behavioral profile which is a substitute of the FICO score used for banked applicants, which, in turn is mainly based on bureau data and credit cards historical records (e.g., Meier and Sprenger, 2012). Fintech mobile apps powered by machine learning psychometric evaluations are testing microcredit applicants (e.g., for estimating the personal risk of the applicant) and are granted access to the data of the applicant's smartphone which are fed into a machine learning model that extracts data relevant to the default prediction (e.g., number of phone calls during working hours is an indirect estimator of income, etc.). The psychological and behavioral data are used to estimate, using ML models, the default risk of the applicant and, for low risk applicants only, grant the loan asked for.

The above reported examples refer to the recent applications of ML and Deep Learning methods in psychological science that are emerging mainly outside the academic arena. However, the number of experiments reported in academic journals that use ML as analytical tools to complement statistical analysis is also increasing (Kosinski et al., 2013; Monaro et al., 2018; Pace et al., 2019). Machine learning has been successfully applied, for example, in the analysis of imaging data in order to classify psychiatric disorders (Orrù et al., 2012; Vieira et al., 2017), in genetics (Libbrecht and Noble, 2015; Navarin and Costa, 2017), in clinical medicine (Obermeyer and Emanuel, 2016), in forensic sciences (Pace et al., 2019) etc.

However, ML is not extensively used in the analysis of psychological experiments as compared to other fields (e.g., genetics). This seems particularly strange if we consider that mathematical modeling of cognitive/brain functioning had great advancements from psychology and neural network based cognitive modeling emerged as one of the main advancements in cognitive psychology (e.g., Seidenberg, 2005).

Experiments in psychological science has been traditionally analyzed with statistical inferential tools. However, recent controversies about the level of replicability in behavioral research of such analytical tools have cast interest in developing more efficient techniques for analyzing the results of psychological experiments (Pashler and Wagenmakers, 2012). ML has developed techniques that may control at least some forms of replicability, the replication of results with similar accuracy to unseen fresh new data.

The Theoretical Role of Psychological Science in the Emergence of Machine Learning and Deep Learning

Hebb (1949) pioneered the mathematical modeling of a neural network that is still at the base of model based on reinforcement learning. He proposed what has come to be known as Hebb's rule. He states, "*When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased.*"

Later, in 1958, Frank Rosenblatt, a Cornell psychologist (see Rosenblatt, 1962) in charge of The Perceptron Project designed what has been described as "*the first precisely specified, computationally oriented neural network*" (Anderson and Rosenfeld, 1988, p. 89).

Neural network modeling rebirth dated 1986 with the publication of David Rumelhart and Jay McClelland's influential two—volume textbook, *Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations, Volume 2: Psychological and biological models, commonly referred to as the PDP Volumes*. In 1987, Walter Schneider noted that the Parallel Distributed Processing (PDP) volumes were already the basis for many courses in connectionism and observed that they were likely to become classics (Schneider, 1987, p. 77). His prediction was borne out. A leading figure in the group was Geoffrey Hinton, a Canadian psychologist turned-data-scientist who contributed to the first papers of the PDP group (McClelland et al., 1987), Hinton is now regarded as a godfather of deep and is now chief scientist at Google.

Machine Learning in Analyzing the Results of Psychological Experiments

While psychology was in the front-end in theory building, is late in adopting ML as a tool for analyzing experimental results. In fact, psychological experiment results are largely analyzed by orthodox p-based statistical inference and more recently by effect size measures.

Here, we will not systematically review the recent advancement in modeling cognitive processes using ML/Deep Learning models (e.g., reinforcement learning) but rather focus on the benefits deriving from the more extensive use of ML methods in the analysis of results collected from psychological experiments as a complement to more traditional statistical inference techniques.

Here we claim that the use of ML could be a useful complement to inferential statistics and will help in achieving at least the following objectives:

- developing models which can generalize/replicate to fresh new data;
- developing models focused on prediction also at single subject level.

The Difference Between Statistics and Machine Learning

In the now classic paper, Breiman (2001) highlighted the difference between statistical modeling and ML. He stated that the classical orthodox statistical approach assumes that data are generated by a given stochastic data mode and the evaluation is more focused on the degree of fitness that the data have to the model. Statistical inference based on data modeling has been the standard *de facto* procedure in the analysis of scientific experiments since 1940.

Inference creates a mathematical model of the data-generation process to formalize understanding or test a hypothesis about how the system behaves. Statistical methods have a long-standing focus on inference, which is achieved through the creation and fitting of a project-specific probability model. The model allows us to compute a quantitative measure of confidence that a discovered relationship describes a 'true' effect that is unlikely to result from noise. Measures typically include *p*-values with a recent shift to effect size in order to contrast the improper use of *p*-based inferences that may lead to a lack of replicability (Ioannidis et al., 2011).

By contrast, ML approach treats the data as unknowns and is mainly focusing on predictive accuracy. Prediction aims at forecasting unobserved outcomes or future behavior. Prediction is also addressed in statistics but with models that are usually constrained by strong assumptions (e.g., linear regression and logistic regression). ML models are more focused on prediction and "model agnostic." It is a frequent observation that in most dataset analyzed with ML models similar predictions accuracies may be achieved using models that rely on very different assumptions (e.g., Support Vector Machine, Naive Bayes, Knn, Random Forest).

In ML models, prediction is achieved by using general-purpose learning algorithms to find patterns in often numerous and highly complex datasets.

ML methods are particularly helpful when one is dealing with datasets in which the number of input variables exceeds the number of subjects, as opposed to datasets where the number of subjects is greater than that of input variables.

ML makes minimal assumptions about the data-generating systems; they can be effective even when the data are gathered without a carefully controlled experimental design and in the presence of complicated non-linear interactions. However, despite convincing prediction results, the lack of an explicit model can make ML solutions difficult to directly relate to existing biological knowledge.

The boundary between statistical inference and ML is fuzzy and methods originally developed in statistics are included in the ML toolbox. For example, logistics among classifiers, linear regression among regression techniques, hierarchical clustering among clustering techniques and Principal components analysis among dimensionality reduction techniques are routinely included in all ML packages. Some of these models (e.g., logistics) usually compares favorably with more complex models (Zhang et al., 2019) with respect to accuracy.

Statistics requires us to choose a model that incorporates our knowledge of the system, and ML requires us to choose a predictive algorithm by relying on its empirical capabilities. Justification for an inference model typically rests on whether we feel it adequately captures the essence of the system. The choice of pattern-learning algorithms often depends on measures of past performance in similar scenarios. Inference and ML are complementary in pointing us to biologically meaningful conclusions.

The agnostic empirical approach of ML is best understood considering the Naive Bayes classifier. The Naive Bayes algorithm is an intuitive method that uses the probabilities of each feature (independent variable) predicts the class the individual case belongs to. It is referred to as "naive" because all features are regarded as independent, which is rarely the case in real life. Naive Bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class is independent of all other attributes. This is a strong and frequently false assumption but results in a fast and effective classification method. Despite the apparently unrealistic assumptions it has been shown the mathematical properties of the good performance of the classifier (Ng and Jordan, 2001). It has been shown that no matter how strong the dependencies among attributes are, Naive Bayes can still be optimal if the dependencies distribute evenly in classes, or if the dependencies cancel each other out (Zhang, 2004). Basically, Naive Bayes is finding the probability of given feature being associated with a label and assigning the label with the highest probability. Despite the assumption of independence the Naive Bayes classifier is usually performing well and is used in practice for a number of practical reasons (e.g., no need to handle inter-correlations, small computational time, performs well for categorical input data, needs less data with respect to other classifiers, e.g., logistics). The success of Naive Bayes classifier is an example of the empirical approach that is characterizing ML modeling. What counts is predictive efficiency rather than how well-prediction based on correct assumptions reliably approximate the data. We will see, in the simulation reported below, that Naive Bayes results among the best classifiers and among those that consistently generalizes across different datasets.

Machine Learning Models

ML models are typically distinguished in supervised models and unsupervised models. Supervised models are built from examples which are labeled. By contrast unsupervised models are developed using unlabeled examples and consists in grouping examples on the basis of their similarities (e.g., clustering, anomaly detectors, etc.) (Mohri et al., 2012).

Supervised models may be further distinguished in classifiers and regressors. Classifiers deal with classification problems when the output variable is a category (e.g., "disease" vs. "no disease"). Regressors address regression problems when the output variable is a real value (e.g., Reaction Time).

Some ML learning models deal only with classification problems (e.g., Naive Bayes) while others may be used both for classification and regression (e.g., Decision trees, Artificial neural Networks, Random Forest) and their use depends on the problem that is addressed.

Here, we will focus on supervised models used for classification among which we could list:

(1) **Decision Trees:** decision tree builds classification or regression models in the form of a tree structure. It utilizes an if-then rule set which is mutually exclusive and exhaustive for classification. The rules are learned sequentially using the training data. Each time a rule is learned, the tuples covered by the rules are removed. This process is continued on the training set until meeting

a termination condition. The tree is constructed in a top-down recursive divide-and-conquer manner. Simple decision trees have the advantage of transparency as the final user understands the prediction rules. However, complex decision tree models such as Random forest (e.g., Breiman, 2001) and Xgboost usually outperform the most simple decision trees.

(2) **Naive Bayes:** Naive Bayes (John and Langley, 1995) is a probabilistic classifier inspired by the Bayes theorem under a simple assumption which the attributes are conditionally independent. Even though the assumption is not valid in most cases since the attributes are dependent, surprisingly Naive Bayes performs impressively in a variety of datasets.

(3) **Artificial Neural Network (ANN)**: is a brain-inspired model with a set input/output units where each connection has a weight associated. ANNs were originally developed by psychologists and neurobiologists to develop and test computational analog of neurons. During the learning phase, the network learns by adjusting the weights (strength of the synapses of the virtual neuron) so as to be able to predict the correct class label of the input stimulus. ANN could be used both for classification and regression.

(4) **k-Nearest Neighbor**: is a lazy learning algorithm which stores all instances in a n-dimensional space. When an unknown new data must be classified, it analyses the closest k number of instances saved (nearest neighbors) and returns the most common class as the prediction. In the distance-weighted nearest neighbor algorithm, it weighs the contribution of each of the k neighbor's according to their distance using the giving greater weight to the closest neighbors (Aha et al., 1991). KNN could be used both for classification and regression.

(5) **Logistic Regression:** (Le Cessie and van Houwelingen, 1992) is a powerful statistical way of modeling a categorical outcome with one or more explanatory variables. It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution.

(6) **Ensemble Methods:** are learning algorithms that construct a set of classifiers and then classify new data points by taking a weighted vote of their individual predictions. The original ensemble method is Bayesian averaging, but more recent algorithms include error-correcting output coding, bagging, and boosting. Ensemble models, by combining different classifiers, usually perform better with a reduction of prediction variability when compared with their constituent classifiers. Ensemble methods usually outperform single classifiers as can be seen in Kaggle competition winners solutions. Ensemble methods usually are optimal solutions of the so called bias/variance trade-off. Usually Bias, the amount of systematic error in prediction, is related to the complexity of the model and highly complex models tend to have low bias but also overfit (e.g., Random Forest). By contrast, simple models, which make few assumptions, tend to underfit. Variance refers to the variability in the predictions, which is usually high in complex models and low in simple models.

There are two procedures that in some cases may enhance a classifier performance apart of ensembles models: feature selection and feature engineering and parameter tuning. Feature selection consists in selecting among the all features (independent variables) the most informative ones while feature engineering consists in deriving new features usually basing on domain knowledge and preliminary data analysis. In other words, feature engineering is about creating new input features from existing ones with the intention to boost the performance of ML models. In psychological test development, feature selection and engineering may be used to derive a subset of items (e.g., the original tests) that performs similarly to the full test and eventually enhance efficiency via developing combination of features.

Parameter tuning consists in selecting the optimal value for parameters of the model that are intended to be used. For example Knn, is a classification model with a single parameter which is the number of neighbors that are used to decide the category of which the new example belongs to. The winning class that is assigned to the new unlabeled case will result from computing the majority of neighbors. The dimension of the neighborhood (2, 3...10, 11) is a parameter that may be optimized and identified as the one that gives maximum performance. In some cases, such as in deep learning models of object detection, the number of parameters to be estimated is in the order of 100.000.

The Interpretability/Accuracy Trade-Off

Best performing models are usually hard to interpret giving rise to a clear interpretability/accuracy trade-off (Johansson et al., 2011). For example, Fernández-Delgado et al. (2014) evaluated the performances of 179 ML classifiers on 121 different datasets arriving to the conclusion that the best performer is Random Forest with support vector machine (SVM) notably second (no significant difference between the two). Additional investigations (Wainberg et al., 2016) re-analyzing the data claimed the Random Forest superiority was not significantly better than SVM and Neural Networks. However, for what counts here, Random Forest, as well as Neural Network and SVM are all hard to interpret. Simpler models, such as pruned decision rules (J48), Naive Bayes, Knn are easier to interpret but rarely result in having the best performance.

Some insight on the interpretability/accuracy trade-off may also come from inspecting the strategies used by Kaggle masters. Kaggle is a site where ML experts can compete in finding the best predictive model on a public dataset. The Netflix Prize was one of these competitions (prize $100.000). Best practices collected from such ML competitions indicate that winners systematically rely on the following strategies in deploying winning models: (i) feature engineering (finding new features usually combinations of the available ones), (ii) parameters tuning (finding the optimal parameters of the model that maximize performance), and (iii) ensemble learning (build a complex model which is a combination of more simple models). Ensemble learning performs better than the constituent classifiers but this reduces interpretability. An example is the difference in the interpretability of a single decision rule when contrasted with a random forest model on the same data. The single decision rule is transparent (e.g., *if $X > 3.5$ than class A else B*) while Random Forest (of decision rules) results in an uninterpretable random mixture of a high number (e.g., 100) decision rules that render opaque any understanding on the exact mechanism at the base of prediction.

In short, interpretable models usually are not the best performers and the best performers classifiers are usually not interpretable. This means that using ML models for analysis results of psychological experiments one could use hard-to-interpret ensemble models to have an estimate of the maximum accuracy possible while using easy-to-interpret decision rules for more confidence based evaluations.

Replicability of Results and Cross Validation

The recent focus on the lack of replicability in behavioral experiments is known with the term of replicability crisis. One source of potential problem leads back to the use of inferential statistics and its misunderstanding of *p*-values and underpowered experiments (Baker, 2016). Recent methodological discussions are related to procedures that guarantee replicable results (Browne, 2000). In summarizing their assessment of replicability Szucs and Ioannidis (2017) concluded that:

> "*Assuming a realistic range of prior probabilities for null hypotheses, false report probability is likely to exceed 50% for the whole literature. In light of our findings, the recently reported low replication success in psychology is realistic, and worse performance may be expected for cognitive neuroscience.*"

Replication of experimental results may be distinguished in exact and broad replication (Cumming, 2008). Exact replication refers to a replication using exactly the same procedure of the original experiment and is targeted by cross validation. The author (Cumming, 2008) proved, in a simulation study of 25 repetitions, that a result in the first experiment significant at $p < 0.05$ in the replications may vary from $p < 0.001$ to $p = 0.76$ (with a 10% chance of $p > 0.44$) showing that p is a very unreliable measure. To complicate the landscape, some researchers have also highlighted how failed replication are not immune from the same type of error that may be detected in the original studies (Bressan, 2019) and false negatives in replication studies have recently attracted attention (Bryan et al., 2019).

Similarly to analysis conducted with inferential statistics, ML workflow encounters the problem replication (Gardner et al., 2018; Gundersen and Kjensmo, 2018). In fact, it is easy to develop complex ML models (e.g., Random Forest) that on small datasets reach near perfect classification accuracies (McDermott et al., 2019). However, this accuracy does not replicate to fresh data which are not used to develop the model (holdout data). For this reason a *de facto* standard for handling this overfitting problem, that plagues not only ML models but also statistical models (e.g., logistics, linear regression) is cross validation.

Cross Validation (see **Figure 1**) is usually a very good procedure to measure how well a result may be replicable at least for what has been called exact replication (Cumming, 2008). Even if ideally it does not address reproducibility of the main finding when minor variations are introduced, exact replication refers to replication where all the conditions of the original experiment are maintained. As cross validation consists in evaluating models on a hold-out set of experimental examples, this set do not differ from the examples used for model development. While cross validation does not prevent the model to overfit, it still estimates the true performance.

In order to avoid overfitting, cross validation regards a compulsory step in ML analysis but its use is very limited in the analysis of psychological experiments. There are a number of different cross validation procedure but one which guarantees good result is the so called stratified 10-fold cross validation. In order to develop models able to generalize new data (unseen data) a good procedure envisages to: (1) remove the 20% of the data for validation; (2) run 10-fold cross validation on the remaining 80% with the aim to select optimal parameters; (3) train model with all 80% of the data with optimal parameters; (4) test the model on the 20% validation set. The result of step 4 will be the best approximation to exact replication of the experiment.

A special case of n-fold cross validation is the LOOCV (Cawley and Talbot, 2010) a method of choice in imaging studies with clinical samples (Orrù et al., 2012). In LOOCV, the statistical model is developed using only n-1 examples and tested on the remaining one exemplar. The procedure is repeated rotating systematically the left out case and the final out-of-sample classification error estimate is derived from the average error of the n-1 models.

When running a cross validation, special care is needed to control information leakage which is one of the reasons why cross validation goes wrong. For example, selecting a subset of predictors before cross validation is a form of leaking knowledge that reduces generalization.

Most psychometric investigations do not address the problem of generalization outside the sample used to develop the model. Clearly, avoiding cross validation yields inflates results, which are over optimistic and may not replicate when the model is applied to out-of-sample data. Similar results have been recently reported by Bokhari and Hubert (2018). The authors reanalyzed the results of the MacArthur Violence Risk Assessment Study using ML tree models and cross validation. Also Pace et al. (2019), in discussing the results of the B test (a test for detecting malingered cognitive symptoms), similarly observed that a decision rule developed on the whole dataset yielded a classification accuracy of whole dataset 88% but using LOOCV the expected accuracy drops to 66%.

Working Example: ML Analysis on Millon Clinical Multiaxial Inventory (MCMIIII)

The example below (**Table 1**) regards the psychometric identification of malingering (Sartori et al., 2016, 2017).

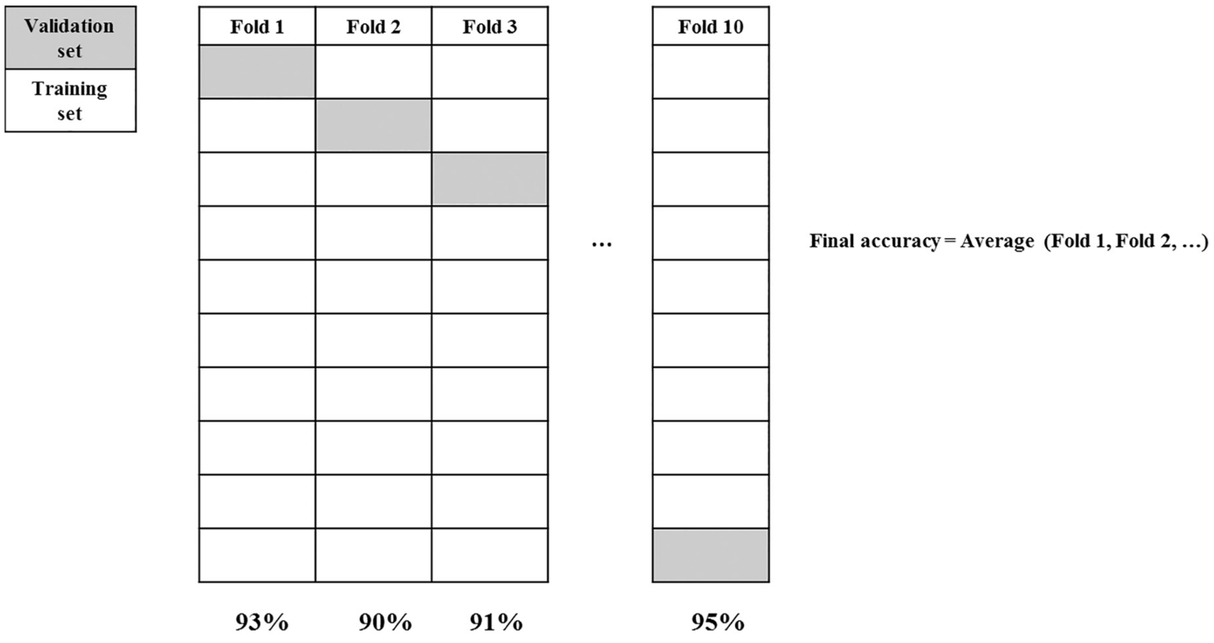


FIGURE 1 | 10-fold cross validation.

The dataset analyzed here consists in the raw scores on the personality questionnaire MCMI-III that was used to predict whether the test was collected in one of two settings. Both groups are low credibility groups, the first are fake good suspects (they have advantages from denying psychopathology) while the second are fake bad suspects (they have advantages from doctoring a get-out-of-jail psychopathology). One group was administered the test for a psychological assessment for reinstatement of driving license and child custody court assessment (n = 93) while the fake bad group included cases involved in a criminal trial who underwent a mental insanity assessment (n = 93). Input were a total of 27 MCMI-III scores, which were used to predict whether the test results were drawn from a Fake good setting or Fake bad, setting. To check the level of replicability, models were tested on 62 new cases extracted, as a first step of the procedure, from the original sample of 186 + 62 cases[1].

TABLE 1 | ML analysis conducted on 186 participants tested with the MCM III.

Classifier	Training set (n = 186)	Cross validation (n = 186)	Stratified holdout test set (n = 62)	Model overfitting training *minus* stratified holdout test accuracy
Naive Bayes	67%	65%	66%	1%
Logistic	75%	62%	58%	17%
SVM	74%	70%	67%	7%
Knn	79%	70%	64%	15%
OneR	70%	62%	67%	3%
CART	93%	62%	61%	32%
Random forest	100%	66%	64%	36%
Neural network	96%	66%	69%	27%
(Averaging)	81.6%	65.4%	65.3%	0.1%
	(12.7)	(3.33)	(3.37)	
Ensemble learner	80.6%	67.7%	69.4%	1.7%

Half of the participants belonged to the Fake-Good group and Half to the Fake-Bad group. A stratified holdout test set (n = 62) was used to evaluate the generalization/replicability of the predictions. Note how the 10-fold cross validation conducted on the training data (n = 186) is a good approximation of the performance on the holdout test set (n = 62). The ensemble model performed slightly better than the performance of the constituents ML models.

As seen above, if a model is developed on all the available data then the final accuracy will be an over optimistic estimate that is not confirmed when the model is tested on previously unseen data (out-of-sample dataset).

From the inspection of the above reported table it appears that:

- Developing the model on all the available training data leads to an accuracy which is not replicated on the test set (average; 81.6 vs. 65.3%).
- The 10-fold cross validation leads to accuracy estimates that correspond to that obtained on the test set. Exact replication on the test set show that the 10 fold cross validation does not lead to an overly optimistic estimate. In short, models developed with cross validation replicate well (see also Koul et al., 2018).
- There is no clear winner among the classifiers. Very simple classifiers (in terms of parameters that require estimation)

[1] All the analysis reported here were done in WEKA (Hall et al., 2009), one of the open sourced code-free ML tools available. Other no-code GUI-based software packages are: ORANGE, KNIME and for deep learning Uber's LUDWIG.

give comparable results to more complex models (compare Naive Bayes and Knn versus Random Forest and Multilayer Neural Network).

- The ensemble of many classifiers performs well on new data and therefore replicates well on fresh new data.
- Some very complex models with many parameters to estimate show extreme overfitting (Random Forest and Multiplayer Neural Network). For example, a Random Forest model developed on the training set yielded a perfect classification (100% accurate) while after a 10-fold cross validation accuracy drops to 66 and 64% on the stratified holdout test set. On the same data the figures for a Multilayer Neural network were 96% on the total of the sample while the result drops to 66% after a 10-fold cross validation which approximates well the 69% measured on the holdout test set. Cross validation is therefore approximating results in exact replication with high accuracy.
- Some very simple models (Naive Bayes) do not suffer much from overfitting when trained without cross validation.
- Also decision rules (usually developed in psychological test building for identifying test cut-off) when fine-tuned without cross validation may heavily overfit. Note that decision rules (e.g., OneR) are the method of choice in most neuropsychological and personality tests; they are simple, readily interpretable but they also need cross validation because they also suffer from overfitting and low replicability.

As regards to exact replicability, it has been noted that results, analyzed with statistical inferences techniques, when replicated show a reduced effect size. In short, an original experiment with an effect size of d = 0.8 when replicated shows an effect size d = 0.4. Repeated K-fold cross validation may derive a distribution of measures generalization/replication.

TABLE 2 | Different machine learning models trained using 10-fold cross validation.

CV on Fold	Tested on	Tested on	Tested on	Max% diff (average = 8.3%)
(a) Naive Bayes				
Fold 1 = 68%	Fold 2 = 73%	Fold 3 = 68%	Fold 4 = 66%	5%
Fold 2 = 69%	Fold 1 = 65%	Fold 3 = 66%	Fold 4 = 66%	4%
Fold 3 = 69%	Fold 1 = 63%	Fold 2 = 73%	Fold 4 = 66%	6%
Fold 4 = 65%	Fold 1 = 65%	Fold 2 = 66%	Fold 3 = 63%	2%
(b) SVM				
Fold 1 = 63%	Fold 2 = 70%	Fold 3 = 71%	Fold 4 = 69%	8%
Fold 2 = 69%	Fold 1 = 66%	Fold 3 = 66%	Fold 4 = 61%	8%
Fold 3 = 69%	Fold 1 = 70%	Fold 2 = 69%	Fold 4 = 61%	8%
Fold 4 = 65%	Fold 1 = 74%	Fold 2 = 67%	Fold 3 = 72%	9%
CV on fold	**Tested on**	**Tested on**	**Tested on**	**Max% diff (average = 9.5%)**
(c) Random forest				
Fold 1 = 62%	Fold 2 = 69%	Fold 3 = 67%	Fold 4 = 58%	7%
Fold 2 = 72%	Fold 1 = 66%	Fold 3 = 64%	Fold 4 = 67%	8%
Fold 3 = 71%	Fold 1 = 69%	Fold 2 = 67%	Fold 4 = 56%	15%
Fold 4 = 63%	Fold 1 = 66%	Fold 2 = 71%	Fold 3 = 64%	8%
CV on fold	**Tested on**	**Tested on**	**Tested on**	**Max% diff (average = 7%)**
(d) Ensemble				
Fold 1 = 65%	Fold 2 = 67%	Fold 3 = 69%	Fold 4 = 61%	5%
Fold 2 = 69%	Fold 1 = 64%	Fold 3 = 65%	Fold 4 = 63%	6%
Fold 3 = 68%	Fold 1 = 65%	Fold 2 = 74%	Fold 4 = 60%	8%
Fold 4 = 63%	Fold 1 = 71%	Fold 2 = 69%	Fold 3 = 72%	9%

Results are reported on testing a ML model on each of the four stratified folds (using 10-fold cross validation) and tested on each of the remaining 3. Results of three classifiers are reported as well as the results of an ensemble model built using all the classifiers included in ***Table 1****. The maximum error is reported as well as the average error%. CV, cross validation.*

Characteristics of the Dataset

High performance neural networks are trained with extremely large dataset. For example a deep neural network with 152 layers and trained on a Imagenet dataset (n = 1.2 mn images) has reduced to 3% the error in classifying images (He et al., 2016). It has been well-established that for a given problem, with large enough data, very different algorithms perform virtually the same.

However, in the analysis of psychological experiments typical number of data points is in the 100 range. Do ML classifiers trained on such small dataset maintain their performance?

In order to evaluate the capacity of ML models to replicate classification accuracies on small datasets, we ran a simulation using the dataset used for the simulations reported in **Table 1**. A total of 298 participants assessed in a low credibility setting (124 in the fake good group and 124 in the fake bad group) were administered the MCMI-III as a part of a forensic assessment. The whole dataset was split into four stratified subsets (folds). Each ML model was trained on one of these folds (using 10-fold cross validation) and tested on the remaining three. The results are reported in **Table 2**.

As shown in **Table 2** all the classifiers trained on a small dataset of 62 cases (32 per each of the two categories) perform well on each of the other test folds. Simple classifiers (e.g., Naive Bayes) perform slightly less erratically across holdout folds than more complex one (e.g., Random Forest). A good strategy in developing ML models that replicates well is to train simple classifiers or ensemble of classifiers rather than models with many parameters.

Balanced Versus Unbalanced Datasets and Priors

In all the examples reported above the number of cases for each class was equal. Unbalanced datasets are usually a problem for classifiers and usually performance of classifiers is generally poor on the minority class. For this reason a number of techniques have been developed in order to deal with unbalanced datasets.

Another problem often neglected is that the final accuracy is the result not only of the accuracy of the model but also depends on the prior probability of the class under investigation. For example, if the prior probability of the class is 10% and the accuracy of classifiers trained on a balanced dataset is 90% the actual probability that a case is correctly classified in the minority class is 50% (of the 18 classified 9/18 will be correct).

Comparing Statistical Inferences With Machine Learning Results

ML uses evaluation metrics mainly addressing accuracy in classification such as Accuracy, area under the curve (AUC), etc. By contrast, statistical metrics are different and more linked to inference (*p*-values) and more recently focusing on reporting effect sizes (e.g., Cohen's d etc.).

One problem that requires to be addressed when complementing statistical analysis with ML results is in the comparison between the metrics used in statistics (e.g., r, d, etc.) and the typical metrics used in ML (classification accuracy, F1, AUC).

Salgado (2018) addressed the problem of translating performance indicators from ML metrics and statistical metrics. It has been shown that the most used ML evaluation metrics can be mapped into effect size; for example, it has been shown that an AUC = 0.8 corresponds to a Cohen's $d = 1.19$. It is possible to transform the accuracy results obtained from ML models to more psychologically oriented effect size measures (Salgado, 2018). It is worth noting, that a Cohen's d of 0.8 is usually regarded as large but, when translated to classification accuracy among two categories, corresponds to an accuracy in classification of 71% due to an overlap between the two distributions of 69%. Using results from **Table 1** an out-of-sample accuracy of 65.3% resulting from the averaging of various classifiers corresponds to a Cohen's $d = 0.556$, usually regarded as a medium effect (Cohen, 1977). However, an accuracy of 65.3% in distinguishing fake good versus faked bad responders of MCM III is far from being of any practical utility when applying the test at single subject level (as in clinical usage of the test).

Model-Hacking in Machine Learning

One procedure which is believed to be at the origin of lack of replicability in reporting experimental results, analyzed with statistical inference, is the so called p-hacking (Nuzzo, 2014).

In ML analyses, there is a similar source of lack of replicability, which could be called model hacking. If many models are tested in order to report only the best model, we are in a condition similar to p- hacking. In the example reported in **Table 1**, using cross validation and reporting only the best performer among the classifiers, in this case SVM, would have produced an accuracy estimation in excess of 4.5% (SVM cross validation results = 70%; average of all cross validation results = 65.5%).

In order to avoid model hacking, one strategy is to verify that classification accuracy is not changing much among different classes of classifiers (see Monaro et al., 2018) as follows: if similar results are obtained by ML models relying on radically different assumptions, we may be relatively confident that the results are not dependent on such assumptions. Additionally, model stability may be addressed by combining different classifiers into an ensemble classifier that indeed reduces the variance in out-of-sample predictions and therefore gives more reliable predictions. Using ensembles instead of specific classifiers is a procedure that avoids model-hacking.

CONCLUSION

Academic psychologists have pioneered the contemporary ML/deep learning development (Hebb, 1949; Rumelhart et al., 1986) and cognitive theorists used connectionist modeling in the field of reading, semantics, attention (Seidenberg, 2005) and frequently anticipated the now much spoken about technology advancements in such fields such as Natural Language Processing (e.g., Word2vec and Lund and Burgess, 1996) and object recognition.

By contrast, ML/deep learning models used for cognitive theorizing have been rarely used in the analysis of psychological experiments and in psychometric test development (Mazza et al., 2019). Classification of brain images (both functional and structural) is a notable exception (Orrù et al., 2012; Vieira et al., 2017).

We have highlighted, in this paper, the reasons why ML should systematically complement statistical inferential analysis when reporting behavioral experiments. Advantages derived from using ML modeling in an analysis experimental results include the following:

- generalization/replication of results to unseen data is realistically estimated rather than optimistically inflated;
- n-fold cross validation guarantees replicable results also for small datasets (e.g., $n = 40$) which are typical in psychological experiments;
- practical and clearly understandable metrics (e.g., out-of-sample accuracy) are reported, rather than indirect inferential measures;
- personalized predictions at single subject level (specific single subjects estimations may be derived also when there are numerous predictors) and subjects which are classified erroneously may be individually analyzed;
- more realistic estimate about the utility of a diagnostic procedure.

Known potential pitfalls of ML data analysis that may obstacle a more extensive use of the ML methods are:

- model hacking. When only the single best performer model is reported rather than a variety of models with differing theoretical assumptions. Model hacking may lead to an overestimation of replicable results. A remedy against model hacking consists in reporting many ML models or ensemble models;
- lack of interpretability. Usually maximum accuracy in prediction is achieved with highly complex non-interpretable models such as XGboost, Random Forest and Neural Networks. This is probably the single most important problem in clinical applications where the clinician needs a set of workable rules to drive the diagnosis. To tamper the problem it may be useful to

report simple decision rules that may help in evaluating the cost of non-interpretability (accuracy achieved with simple interpretable models as compared to maximum accuracy achieved by complex less interpretable models). Interpretability is important in clinical setting where clinicians need simple and reliable decision rules (see Figure 3 in Mazza et al., 2019).

DATA AVAILABILITY STATEMENT

The datasets analyzed in this article are not publicly available. Requests to access the datasets should be directed to giuseppe.sartori@unipd.it.

AUTHOR CONTRIBUTIONS

GO devised the main research topic, and planned and carried out the ML analysis. GO, MM, and GS conceived the conceptual ideas and proof outline. GO, MM, CC, AG, and GS drafted the manuscript, revised the manuscript critically, and gave the final approval for the version to be published.

FUNDING

GO, MM, and GS were found within the BACKUP Project Horizon 2020.

REFERENCES

Aha, D., Kibler, D., and Albert, M. K. (1991). Instance-based learning algorithms. *Mach. Learn.* 6, 37–66. doi: 10.1007/bf00153759

Anderson, J., and Rosenfeld, E. (eds) (1988). *Neurocomputing: Foundations of Research*. Cambridge, MA: The MIT Press.

Baker, M. (2016). Is there a reproducibility crisis? *Nature* 533, 452–454.

Bokhari, E., and Hubert, L. (2018). The lack of cross-validation can lead to inflated results and spurious conclusions: a re-analysis of the macarthur violence risk assessment study. *J. Classif.* 35, 147–171. doi: 10.1007/s00357-018-9252-3

Breiman, L. (2001). Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat. Sci.* 16, 199–231. doi: 10.1214/ss/1009213726

Bressan, P. (2019). Confounds in "failed" replications. *Front. Psychol.* 10:1884. doi: 10.3389/fpsyg.2019.01884

Browne, M. W. (2000). Cross-validation methods. *J. Math. Psychol.* 44, 108–132. doi: 10.1006/jmps.1999.1279

Bryan, C. J., Yeager, D. S., and O'Brien, J. M. (2019). Replicator degrees of freedom allow publication of misleading failures to replicate. *Proc. Natl. Acad. Sci U.S.A.* doi: 10.1073/pnas.1910951116 [Epub ahead of print].

Cawley, G. C., and Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* 11, 2079–2107.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences*. Abingdon: Routledge.

Cumming, G. (2008). Replication and P intervals: P values predict the future only vaguely, but confidence intervals do much better. *Perspect. Psychol. Sci.* 3, 286–300. doi: 10.1111/j.1745-6924.2008.00079.x

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *J. Mach. Learn. Res.* 15, 3133–3181.

Gardner, J., Yang, Y., Baker, R., and Brooks, C. (2018). Enabling End-To-End machine learning replicability: a case study in educational data mining. *arXiv* [Preprint]. http://arxiv.org/abs/1806.05208 (accessed August, 2019).

Gundersen, O. E., and Kjensmo, S. (2018). "State of the art: reproducibility in artificial intelligence," in *Proceeding of the 32nd AAAI Conference on Artificial Intelligence*, Trondheim

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Peter Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explor.* 11, 10–18.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV.

Hebb, D. O. (1949). *The Organization of Behavior*. New York, NY: Wiley & Sons.

Ioannidis, J. P., Tarone, R., and McLaughlin, J. K. (2011). The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology* 24, 450–456. doi: 10.1097/EDE.0b013e31821b506e

Johansson, U., Sönströd, C., Norinder, U., and Boström, H. (2011). Trade-off between accuracy and interpretability for predictive in silico modeling. *Future Med. Chem.* 3, 647–663.

John, G. H., and Langley, P. (1995). "Estimating continuous distributions in bayesian classifiers," in *Proceeding of the 11th Conference on Uncertainty in Artificial Intelligence*, San Mateo.

Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl. Acad. Sci. U.S.A.* 110, 5802–5805. doi: 10.1073/pnas.1218772110

Koul, A., Becchio, C., and Cavallo, A. (2018). Cross-Validation Approaches for Replicability in Psychology. *Front. Psychol.* 9:1117. doi: 10.3389/fpsyg.2018.01117

Le Cessie, S., and van Houwelingen, J. C. (1992). Ridge Estimators in Logistic Regression. *Appl. Stat.* 41, 191–201.

Libbrecht, M. W., and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* 16, 321–332. doi: 10.1038/nrg3920

Lund, K., and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav. Res. Methods Instrum. Comput.* 28, 203–208. doi: 10.3758/BF03204766

Mazza, C., Monaro, M., Orrù, G., Burla, F., Colasanti, M., Ferracuti, S., et al. (2019). Introducing machine learning to detect personality faking-good in a male sample: a new model based on minnesota multiphasic personality inventory-2 restructured form scales and reaction times. *Front. Psychiatry* 10:389. doi: 10.3389/fpsyt.2019.00389

McClelland, J. L., Rumelhart, D. E., and PDP Research Group. (1987). *Parallel Distributed Processing*, Vol. 2. Cambridge, MA: MIT press.

McDermott, M. B. A., Wang, S., Marinsek, N., Ranganath, R., Ghassemi, M., Foschin, L., et al. (2019). Reproducibility in machine learning for health. *arXiv* [preprint]. https://arxiv.org/pdf/1907.01463.pdf (accessed August, 2019).

Meier, S., and Sprenger, C. D. (2012). Time discounting predicts creditworthiness. *Psychol. Sci.* 23, 56–58. doi: 10.1177/0956797611425931

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. Cambridge, MA: The MIT Press.

Monaro, M., Galante, C., Spolaor, R., Li, Q. Q., Gamberini, L., Conti, M., et al. (2018). Covert lie detection using keyboard dynamics. *Scientific Reports* 8:1976

Navarin, N., and Costa, F. (2017). An efficient graph kernel method for non-coding RNA functional prediction. *Bioinformatics* 33, 2642–2650. doi: 10.1093/bioinformatics/btx295

Ng, A. Y., and Jordan, M. I. (2001). On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. *Adv. Neural Inform. Process. Syst.* 14, 605–610.

Nuzzo, R. (2014). Scientific method: statistical errors. *Nature* 506, 150–152. doi: 10.1038/506150a

Obermeyer, Z., and Emanuel, E. J. (2016). Predicting the future: big data, machine learning, and clinical medicine. *N. Engl. J. Med.* 375, 1216–1219. doi: 10.1056/nejmp1606181

Orrù, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., and Mechelli, A. (2012). Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neurosci. Biobehav. Rev.* 36, 1140–1152. doi: 10.1016/j.neubiorev.2012.01.004

Pace, G., Orrù, G., Merylin, M., Francesca, G., Roberta, V., Boone, K. B., et al. (2019). Malingering detection of cognitive impairment with the B test is boosted using machine learning. *Front. Psychol.* 10:1650. doi: 10.3389/fpsyg.2019.01650

Pashler, H., and Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? *Perspect. Psychol. Sci.* 7, 528–530. doi: 10.1177/1745691612465253

Rosenblatt, F. (1962). *Principles of Neurodynamics; Perceptrons and the Theory of Brain Mechanisms*. Washington: Spartan Books.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Salgado, J. F. (2018). Transforming the area under the Normal Curve (AUC) into cohen's d, pearson's r pb, odds-ratio, and natural log Odds-Ratio: Two Conversion Tables. *Eur. J. Psychol. Appl. Leg. Context* 10, 35–47. doi: 10.5093/ejpalc2018a5

Sartori, G., Orrù, G., and Zangrossi, A. (2016). "Detection of malingering in personal injury and damage ascertainment," in *Personal Injury and Damage Ascertainment Under Civil Law* (Cham: Springer), 547–558.

Sartori, G., Zangrossi, A., Orrù, G., and Monaro, M. (2017). "Detection of malingering in psychic damage ascertainment," in *P5 Medicine and Justice*. ed. S. Ferrara (Berlin: Springer), 330–341. doi: 10.1007/978-3-319-67092-8_21

Schneider, W. (1987). Connectionism: is it a paradigm shift for psychology? *Behav. Res. Methods Instrum. Comput.* 19, 73–83. doi: 10.1007/s00221-016-4866-3

Seidenberg, M. S. (2005). Connectionist models of word reading. *Curr. Dir. Psychol. Sci.* 14, 238–242. doi: 10.1111/j.0963-7214.2005.00372.x

Szucs, D., and Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol.* 15:e2000797. doi: 10.1371/journal.pbio.2000797

Vieira, S., Pinaya, H., and Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. *Neurosci. Biobehav. Rev.* 74(Part A), 58–75. doi: 10.1016/j.neubiorev.2017.01.002

Wainberg, M., Alipanahi, B., and Frey, B. J. (2016). Are random forests truly the best classifiers? *J. Mach. Learn. Res.* 17, 3837–3841.

Zhang, H. (2004). "The optimality of naive bayes," in *Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference*, Michigan: FLAIRS.

Zhang, J. M., Harman, M., Ma, L., and Liu, Y. (2019). Machine learning testing: survey, landscapes and horizons. *arXiv* [Preprint]. arXiv:1906.10742

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

TECHNOLOGY AND CODE
published: 10 January 2020
doi: 10.3389/fpsyg.2019.02938

DscoreApp: A Shiny Web Application for the Computation of the Implicit Association Test D-Score

Ottavia M. Epifania, Pasquale Anselmi and Egidio Robusto*

Department of Philosophy, Sociology, Pedagogy, and Applied Psychology, University of Padova, Padua, Italy

Several options are available for computing the most common score for the Implicit Association Test, the so-called *D-score*. However, all these options come with some drawbacks, related to either the need for a license, for being tailored on a specific administration procedure, or for requiring a degree of familiarity with programming. By using the R **shiny** package, a user-friendly, interactive, and open source web application (DscoreApp) has been created for the *D-score* computation. This app provides different options for computing the *D-score* algorithms and for applying different cleaning criteria. Beyond making the *D-score* computation easier, DscoreApp offers the chance to have an immediate glimpse on the results and to see how they change according to different settings configurations. The resulting *D-score*s are immediately available and can be seen in easy-readable and interactive graphs, along with meaningful descriptive statistics. Graphical representations, data sets containing the *D-score*s, and other information on participants' performance are downloadable. In this work, the use of DscoreApp is illustrated on an empirical data set.

Keywords: implicit association test, implicit measures, shiny, web application, *D-score*, user-friendly, social cognition

OPEN ACCESS

Edited by:
Pietro Cipresso,
Italian Auxological Institute (IRCCS),
Italy

Reviewed by:
Antonio Calcagnì,
University of Padova, Italy
Lena Nadarevic,
Universität Mannheim, Germany

****Correspondence:***
Ottavia M. Epifania
marinaottavia.epifania@phd.unipd.it

Specialty section:
This article was submitted to Quantitative Psychology and Measurement,
a section of the journal Frontiers in Psychology

Received: *01 August 2019*
Accepted: *11 December 2019*
Published: *10 January 2020*

Citation:
Epifania OM, Anselmi P and Robusto E (2020) DscoreApp: A Shiny Web Application for the Computation of the Implicit Association Test D-Score. Front. Psychol. 10:2938. doi: 10.3389/fpsyg.2019.02938

1. INTRODUCTION

The Implicit Association Test (IAT; Greenwald et al., 1998) is one of the most common measures for assessing the strength of automatically activated associations between concepts. The resistance to self-presentation strategies (Egloff and Schmukle, 2002; Greenwald et al., 2009) and its ease of adaptation to different topics (Zogmaister and Castelli, 2006) make the IAT broadly used in studies on various issues, ranging from consumers behaviors (e.g., Karnal et al., 2016) and addiction behaviors (e.g., Chen et al., 2018) to self–esteem (e.g., Dentale et al., 2019) and personality traits (e.g., Steffens, 2004). Given its ability of overcoming self–presentation biases, the IAT finds many applications in social cognition studies, where it is employed for assessing implicit attitudes toward different social groups (e.g., Anselmi et al., 2015).

A convenient measure of the strength and direction of the implicit association assessed by the IAT is the *D-score* algorithm (Greenwald et al., 2003), for which different variations are available. The differences between each of the algorithms mainly concern the treatment for error and fast responses, while the core procedure for its computation is the same.

Despite many options are available for the *D-score* computation, like SPSS syntaxes, R packages, Inquisit scripts, they all come with some drawbacks. The use of SPSS syntaxes requires the SPSS license, programming skills are required for using R packages, and Inquisit scripts are tailored on Inquisit administration procedure. The aim of this study is to present an interactive Web

Application for the computation of the *D-score* able to combine an easy and intuitive User Interface (UI) with the computational power of R, while being completely Open Source.

2. THE IMPLICIT ASSOCIATION TEST—IAT

The IAT procedure (depicted in **Table 1**) is typically composed of seven different blocks, and is based on the speed and accuracy with which different type of stimuli (appearing sequentially at the center of the screen) are sorted in their reference categories (displayed at the top corners of the screen). Three blocks (Blocks 1, 2, and 5) are practice blocks, in which either object stimuli (e.g., images of *flowers* and *insects* in a Flowers-Insects IAT) or attribute stimuli (e.g., *Positive* and *Negative* words) are sorted in their reference categories. In the first associative condition (Blocks 3 and 4), flowers images and positive words are mapped with the same response key, while insects images and negative words are mapped with the opposite response key. In the second associative condition (Blocks 6 and 7), the labels for categorizing flowers and insects stimuli switch their positions on the screen. Thus, flowers images and negative words are mapped with the same response key, and insects images and positive words are mapped with the other response key. The categorization task is supposed to be easier in the condition consistent with respondents' automatically activated association (the so-called "compatible condition") than in the condition against their automatically activated association (the so-called "incompatible condition"). In a more general fashion, the two associative conditions can be arbitrarily identified as Mapping A (e.g., Blocks 3 and 4) and Mapping B (e.g., Blocks 6 and 7). The difference between respondents' performance in the two conditions results in the *IAT effect* that can be easily interpreted by means of the *D-score*.

The IAT administration procedure might include a feedback strategy, for which a red cross appears on the screen every time a stimulus is incorrectly categorized. Participants are then asked to correct their response to continue the experiment.

2.1. The *D-Score*

The *D-score* algorithms result from the combination of the various error correction and lower tail treatment strategies ("Error inflation" and "Lower tail treatment" in **Table 2**).

Grounding on the IAT administration procedure, the error correction may apply either a built-in or an *ex post* correction. In the former case (*D1* and *D2*), the response time considered for the *D-score* computation is the time at the first incorrect response increased by the time required to correct it. In the latter case (*D3*, *D4*, *D5*, and *D6*), the error response is replaced by the average response time of the block in which the error occurred, increased by a fixed penalty (i.e., either 600 ms or two times the standard deviation of the block response time). The *D-score* algorithms differ also according to the lower tail treatment, which concerns the decision to discard fast trials (< 400 ms) or not. Once the treatments for the error and fast responses have been applied according to the chosen algorithm, the *D-score* can be computed. Firstly, the *D-score*s for associative practice blocks (Equation 1) and associative test blocks (Equation 2) are computed:

$$D_{\text{practice}} = \frac{M_{\text{B6}} - M_{\text{B3}}}{sd_{\text{B6, B3}}}, \tag{1}$$

and

$$D_{\text{test}} = \frac{M_{\text{B7}} - M_{\text{B4}}}{sd_{\text{B7, B4}}}. \tag{2}$$

In both cases, the difference in the average response times between the two critical blocks is divided by the standard deviation computed on the pooled trials of both blocks. Once the *D-score*s for practice and test blocks are obtained, it is possible to compute the actual *D-score*:

$$D\text{-}score = \frac{D_{\text{practice}} + D_{\text{test}}}{2}. \tag{3}$$

The blocks order in Equations (1) and (2) is arbitrary, and can be reversed. The interpretation of the *D-score* clearly follows the order with which the subtraction between the blocks is computed. For instance, if the *D-score* of the Flowers-Insects IAT illustrated in **Table 1** is computed following the blocks order in Equation 1 ($M_{\text{B6}} - M_{\text{B3}}$) and Equation 2 ($M_{\text{B7}} - M_{\text{B4}}$), a positive score

TABLE 1 | IAT blocks and conditions, adapted from Greenwald et al. (2003).

Block	Function	Left key	Right key
B1	Practice	Flowers	Insects
B2	Practice	Good	Bad
B3	Practice Mapping A	Flowers + Good	Insects + Bad
B4	Test Mapping A	Flowers + Good	Insects + Bad
B5	Practice	Insects	Flowers
B6	Practice Mapping B	Insects + Good	Flowers + Bad
B7	Test Mapping B	Insects + Good	Flowers + Bad

The presentation order of the critical blocks B3 and B4 and the critical blocks B6 and B7 is counterbalanced across participants.

TABLE 2 | Overview of the *D-score* algorithms.

D-score	Error inflation	Lower tail treatment
D1	Built-in correction	No
D2	Built-in correction	Delete trials < 400 ms
D3	Mean (correct responses) + 2 sd	No
D4	Mean (correct responses) + 600 ms	No
D5	Mean (correct responses) + 2 sd	Delete trials < 400 ms
D6	Mean (correct responses) + 600 ms	Delete trials < 400 ms

For all the algorithms, trials with a latency >10,000 ms are discarded. Trials from Blocks 3, 4, 6, and 7 are used for computing the D-score. *Practice blocks (i.e., Blocks 1, 2, and 5) are discarded.*

would stand for a possible preference for flowers over insects (that is, faster responses would have been observed in B3 and B4 compared with B6 and B7). Vice versa, if the order of the blocks in Equation 1 and in Equation 2 is reversed ($M_{B3} - M_{B6}$ and $M_{B4} - M_{B7}$, respectively), a positive score would stand for a possible preference for insects over flowers.

Several options (illustrated in **Table 3**) are available for computing the *D-score*, namely Inquisit scripts, SPSS syntaxes, and `R` packages.

Inquisit scripts are probably the most straightforward way for obtaining the *D-score* since they compute it right after the IAT administration procedure and store the result along with other information on participants' performance (e.g., response time for each IAT trial, correct and incorrect responses). Nonetheless, these scripts work only when associated with the Inquisit administration procedure, and they can compute just one of the available *D-score* algorithms. Finally, Inquisit requires a license to be used.

SPSS syntaxes provides several information on participants' performance, and they are not tied to a specific administration software. Nonetheless, their use requires a certain degree of expertise with SPSS language, and SPSS license.

`R` provides the open-source alternative to both Inquisit scripts and SPSS syntaxes. Both **`IATanalytics`** and **`IATScore`** packages by Storage (2018a) and Storage (2018b) provide the users with just the function for computing the *D-score*. **`IATScore`** gives the chance to compute the score also for Brief-IAT (B-IAT; Sriram and Greenwald, 2009). Both **`IAT`** and **`IATScores`** provide functions for cleaning the original data set, for plotting the data, and for computing the different *D-score* algorithms. So far, only **`IATScores`** has built-in functions for computing IAT reliability (i.e., split–half and the test–retest IAT reliability).

Regardless of the specific `R` package one wants to use, the data preparation is not straightforward and easy. For some of the packages (e.g., **`IATanalytics`**), the columns identifying the variables for the computation of the *D-score* have to follow a specific order, otherwise the computation will fail. Also the coding of the variables might result counterintuitive: For example, in **`IAT`** package, error responses have to be coded as 1 and correct responses have to be coded as 0. Moreover, in both **`IATanalytics`** and **`IATScore`** it is possible to compute the *D-score* for just one participant at a time, and it is not well specified which *D-score* is computed. None of the above mentioned options provides the users with graphical representations of the *D-score*s.

TABLE 3 | Overview of the available options for computing the *D-score*.

	Open source	Programming skills	Multiple D-score	Plot	Reliability
SPSS syntaxes	No	A bit	Yes	No	No
Inquisit scripts	No	No	No	No	No
`IATanalytics`	Yes	Yes	Not clear	No	No
`IATScore`	Yes	Yes	Not clear	No	No
`IAT`	Yes	Yes	Yes	Yes	No
`IATScores`	Yes	Yes	Yes	Yes	Yes

R packages are reported in bold.

An interactive tool able to combine a user–friendly interface with the computational power of `R` and its open source philosophy could represent an optimal solution for the *D-score* computation, also for researchers with no experience with coding. Additionally, this tool might be of convenience for researchers more experienced in coding and data analysis that want to obtain a quick overview of IAT results.

In the next sections, the functioning of DscoreApp is illustrated through a practical example.

3. THE CHOCOLATE-IAT DATA SET

Data comes from the responses of 152 participants ($F = 63.82\%$, Age $= 24.03 \pm 2.82$) to a Dark-Milk Chocolate IAT. This IAT was developed for the assessment of dark and milk chocolate implicit preference. It followed the structure depicted in **Table 1**. The two critical conditions were made out of 60 trials each (i.e., 20 trials in each associative practice block and 40 trials in each associative test block). The associative condition in which Milk chocolate was associated with negative words and Dark chocolate was associated with positive words was identified as Mapping A (i.e., `.Milkbad` in **Figure 1**). Vice versa, the associative condition in which Dark chocolate was associated with negative words and Milk chocolate was associated with positive words was identified as Mapping B (i.e., `.Milkgood` in **Figure 1**). In case of an erroneous stimulus categorization, participants received no feedback. Results obtained from this data have been previously published in Epifania et al. (in press).

4. DSCOREAPP

DscoreApp was developed in `R` (R Core Team, 2018) by using **`shiny`** (Chang et al., 2018) and **`shinyjs`** (Attali, 2018) packages. DscoreApp can be retrieved at the URL: https://fisppa.psy.unipd.it/DscoreApp/, and its source code is available on GitHub. DscoreApp is platform independent, and is distributed under a MIT license. The UI is designed to be as clear and straightforward as possible, and the pop–up menus for the different functions are meant to make the use of the app more intuitive and interactive. The app is organized in different panels (i.e., "Input," "Read me first," "D-score results," and "Descriptive statistics") that will be presented in the next sections.

4.1. Read Me First Panel

The "Read Me First" panel includes important information regarding the app functioning. The interactive structure allows the users to jump directly to the instructions section they are interested in, making the information on the different app functions easily accessible. The **Download Template** button can be used to download a CSV template suggested for using the app. However, it is not strictly necessary to use the provided CSV template, or to specify the variables in the same order as in the template. As long as the uploaded file is in a CSV format (with comma set as separator) and contains the variables for

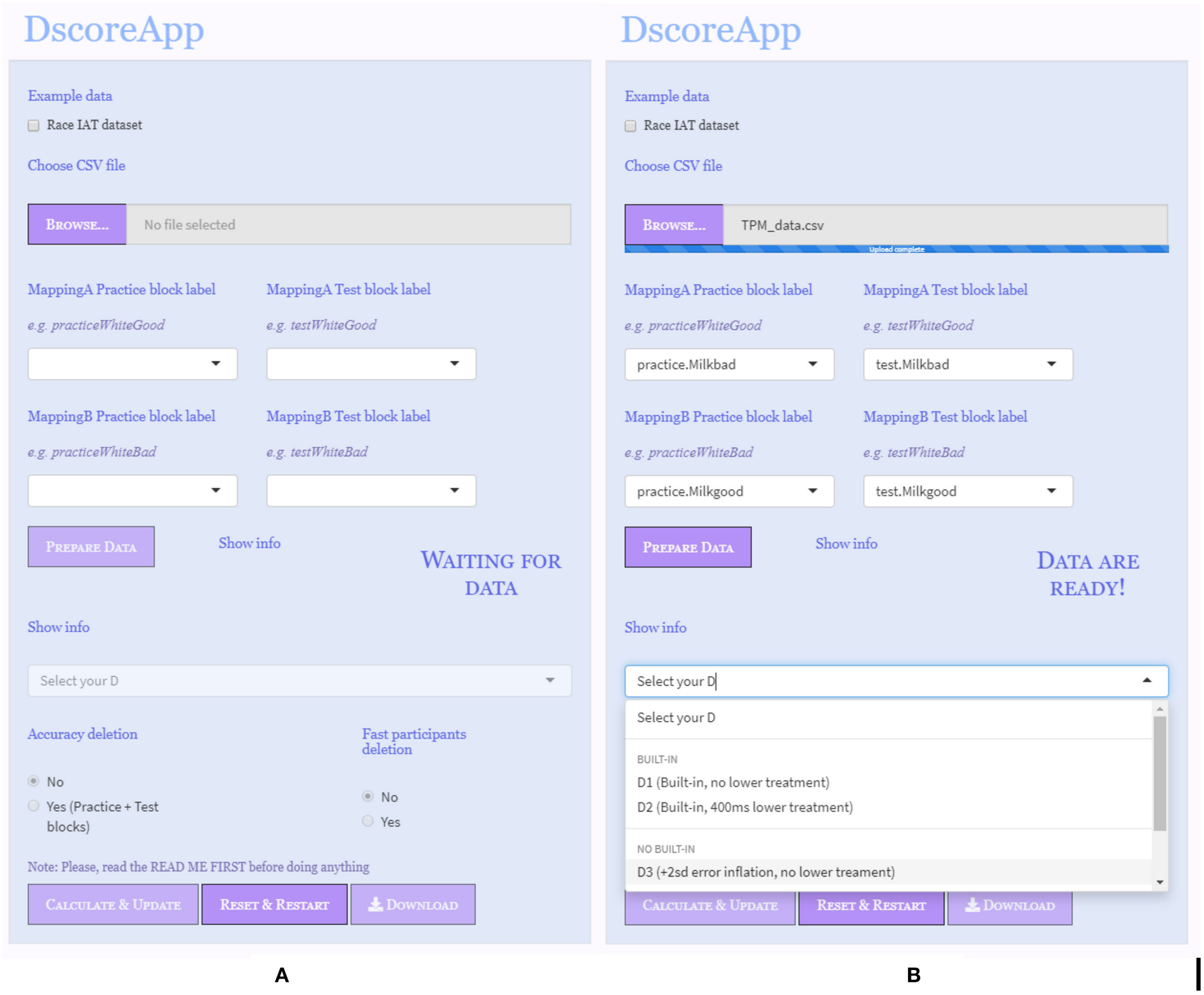


FIGURE 1 | Input panel. **(A)** Data correctly uploaded. **(B)** Data ready for computation.

the *D-score* computation with the same names as the variables in the CSV template, the app will work. Specifically, the data frame must contain a variable identifying participants' IDs (`participant`), the labels identifying the four critical blocks of the IAT (`block`), the latency of the responses expressed in milliseconds (`latency`), and the variable identifying the accuracy responses (`correct`).

The pure practice blocks (Blocks 1, 2, and 5) must be removed before using the app. If they are not removed, the app will throw an error. The `block` variable must be a character string that uniquely identify the four critical blocks of the IAT. This variable contains the information for distinguishing between the practice and test blocks of the two mapping conditions, such as "practiceMappingA," "praticeMappingB," "testMappingA," and "testMappingB." The specific name of each level is not important, as it is not important the order with which they have been presented to participants. In case the blocks labels are not unique, the app will throw an error. If the IAT administration procedure included a built-in correction, the `latency` variable must contain the already inflated response times. Otherwise, it must contain the raw response times. Finally, the `correct` variable must be a numeric variable with just two possible values, namely 0 identifying incorrect responses and 1 identifying correct responses. Usually, accuracy responses are automatically coded as 0 for incorrect and 1 for correct responses by the software for the IAT administration, unless otherwise specified by the users.

The "Read me first" panel also provides information about the different *D-score* algorithms, the blocks order for the *D-score* computation (i.e., MappingB − MappingA), and the downloadable file that can be retrieved at the end of the computation. Further details on the downloadable file are given in Section 4.5. The blocks order for the *D-score* computation can be changed by reversing the Mapping A and Mapping B labels (see section 4.2).

4.2. Input Panel

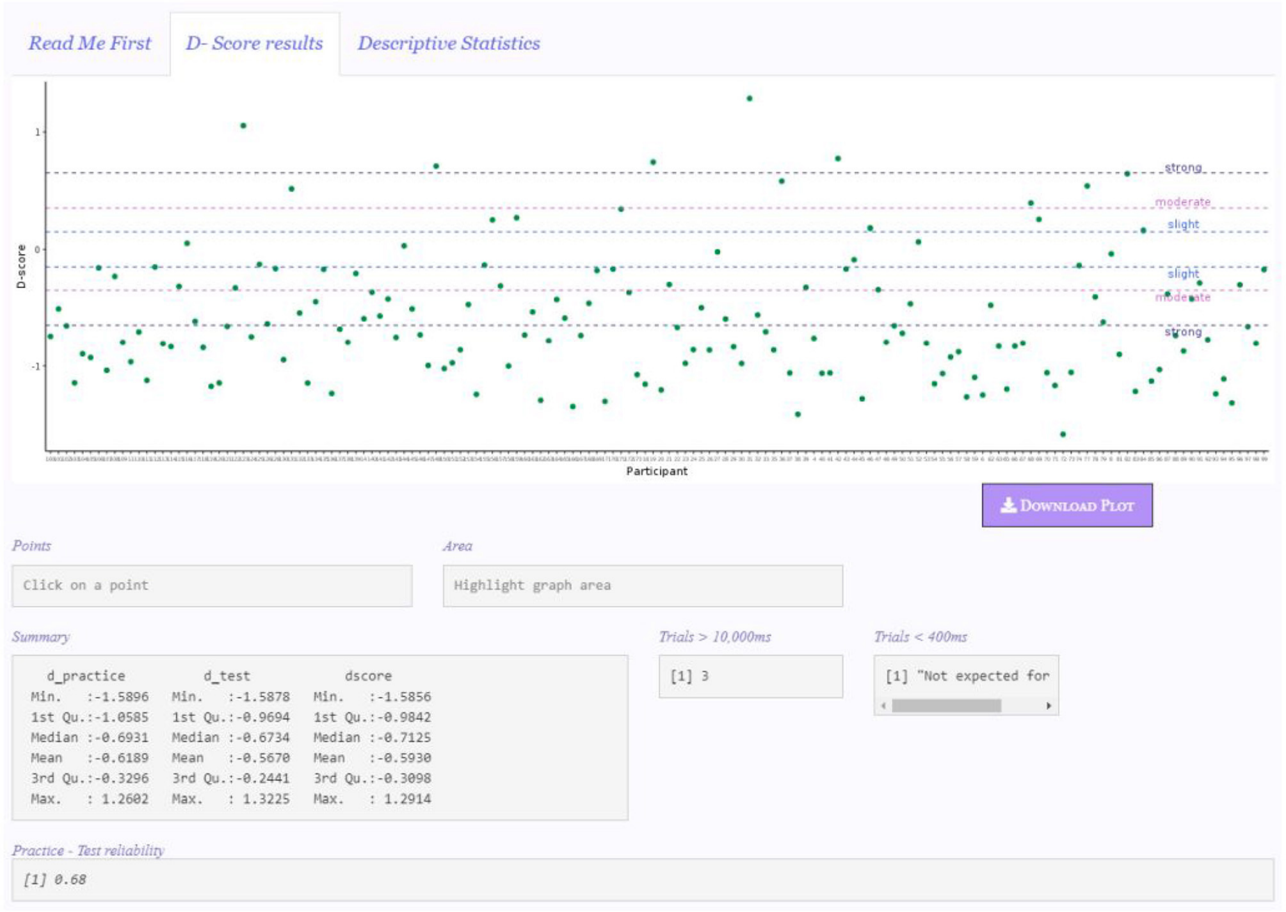
In the starting state of the app, none of the buttons are enabled, and the input drop-down menus for labeling the blocks are empty. The app comes with a toy data set that can be used to familiarize with the app functions, and that can be uploaded by clicking on the checkbox `Race IAT dataset`. Users can upload their own data by means of the **Browse** button.

Two different states of the "Input Panel" are illustrated in **Figure 1**.

Figure 1A depicts the app state when the data set has been correctly uploaded and read by the app. The name of the uploaded file and its extension appear right next to the **Browse** button. The labels of the four different blocks, as they are named in the data frame, are shown into their—alleged—positions (i.e., "MappingA practice block label," "MappingA test block label," "MappingB practice block label," "MappingB test block label" in **Figure 1A**). In case the uploaded data set has some problems, like it uses another column separator than the comma, the app will not be able to distinguish between the columns, and the drop-down menus for the assignment of the blocks labels will be empty. Users can redefine the labels for each block by clicking and selecting from the drop-down menus. To reverse the direction of the *D-score*, and hence the interpretation of its meaning, users can switch the labels for Mapping A and Mapping B. The **Prepare Data** button becomes active when the data are correctly uploaded and the labels for each level of the IAT blocks are defined. Once the **Prepare Data** button has been clicked and data are ready for the *D-score* computation, the alert message "Waiting for data" becomes "Data are ready," and the **Select your D** drop-down menu is enabled (**Figure 1B**). A brief description of the *D-score* algorithms is given next to each option.

The IAT administration procedure of the example data set did not include a built-in correction strategy, and hence a *D-score* algorithm with an *ex post* strategy for the error responses was chosen, specifically the *D3* one. Since the default direction for the *D-score* computation is (MappingB – MappingA), positive scores stand for faster response times in associating Milk chocolate with negative words and Dark chocolate with positive words.

The **Calculate & Update** button and the **Graphic display** options become active after a *D-score* has been selected, as well as the **Accuracy cleaning** option and the **Fast participants cleaning** option. The **Accuracy cleaning** option refers to the elimination of participants with an high percentage of incorrect responses in



Points

Click on a point

Area

Highlight graph area

Summary

```
  d_practice         d_test            dscore       
 Min.   :-1.5896   Min.   :-1.5878   Min.   :-1.5856  
 1st Qu.:-1.0585   1st Qu.:-0.9694   1st Qu.:-0.9842  
 Median :-0.6931   Median :-0.6734   Median :-0.7125  
 Mean   :-0.6189   Mean   :-0.5670   Mean   :-0.5930  
 3rd Qu.:-0.3296   3rd Qu.:-0.2441   3rd Qu.:-0.3098  
 Max.   : 1.2602   Max.   : 1.3225   Max.   : 1.2914  
```

Trials > 10,000ms

```
[1] 3
```

Trials < 400ms

```
[1] "Not expected for
```

Practice - Test reliability

```
[1] 0.68
```

FIGURE 2 | Results panel.

at least one of the two associative conditions, either Mapping A or Mapping B (Nosek et al., 2002). The default threshold is set at 25%, and participants with an error percentage exceeding this threshold are discarded. Users can modify the default threshold via the **Error percentage** option (active only when the "Yes" option of **Accuracy Cleaning** is selected). The **Fast participants cleaning** refers to the elimination of participants with more than 10% of trials with responses faster than 300 *ms* (Greenwald

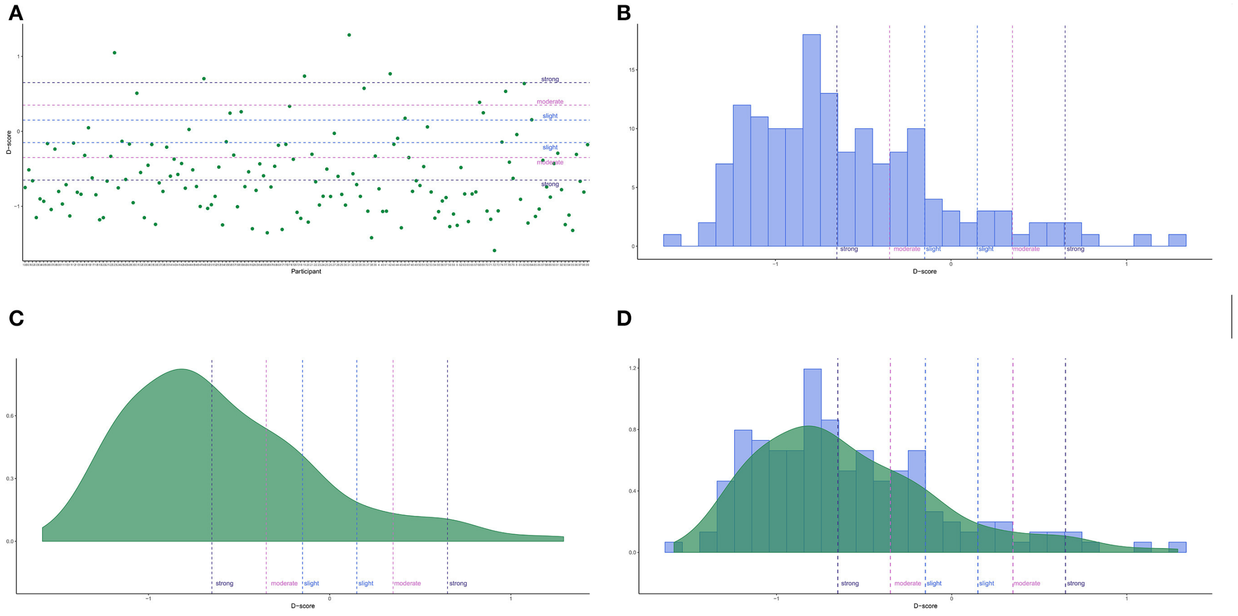
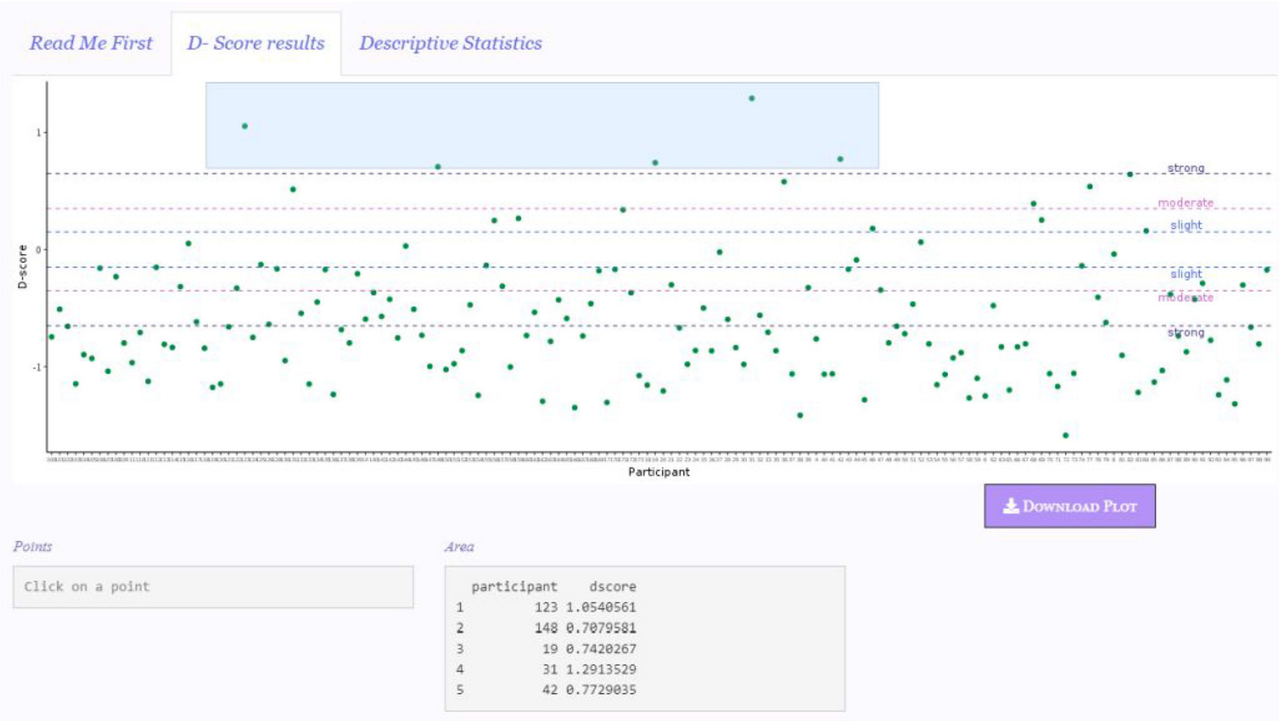


FIGURE 3 | Shiny App graphical representations. **(A)** Points (default). **(B)** Histogram. **(C)** Density. **(D)** Histogram and Density.



Points

```
Click on a point
```

Area

```
  participant    dscore
1         123 1.0540561
2         148 0.7079581
3          19 0.7420267
4          31 1.2913529
5          42 0.7729035
```

FIGURE 4 | Area highlighter for detecting participants' *D-score*.

et al., 2003). If one of these options is selected, the results of the participants meeting these elimination criteria are not displayed in the "D-Score results" panel, but their *D-score*s, and the information on their performance, will still be available in the downloadable file.

The **Download** button is enabled after the first *D-score* is computed.

4.3. D-Score Results Panel

When the **Calculate & Update** button is clicked, results appear in the "D-score results" panel (**Figure 2**). The **Calculate & Update** button must be clicked every time users want to make a settings change effective, otherwise the app will not be updated.

Despite not shown in **Figure 2**, the "Input Panel" remains visible on the left side, so that users can constantly check the specific configuration for the computation of the *D-score*.

The first object appearing in this panel is the graphical representation of the results, for which various options are available ("Points," "Histogram," "Density," "Histogram + Density," see section 4.3.1 for further details). The functioning of the `Points` and `Area` boxes is illustrated in section 4.3.1 as well. The default graph appearing is a points graph depicting each participant's *D-score*.

In the `Summary` box, the descriptive statistics (i.e., *Minimum*, *First Quartile*, *Median*, *Mean*, *Third quartile*, and *Maximum*) of *D-practice*, *D-test*, and *D-score* are presented. The `Trials` $> 10,000$ `ms` box reports the number of trials discarded because of a response time higher than 10,000 *ms*. If no trials meet this elimination criterion, the message "None" is displayed. When a *D-score* algorithm that eliminates trials faster than 400 ms (i.e., *D2*, *D5*, *D6*) is selected, the `Trials` $<$ `400 ms` box reports the number of discarded trials, otherwise the "Not expected for this D" message is shown, as in the example. Finally, the `Practice-Test reliability` box shows the IAT reliability computed as the correlation between D_{practice} and D_{test} across all participants (Gawronski et al., 2017).

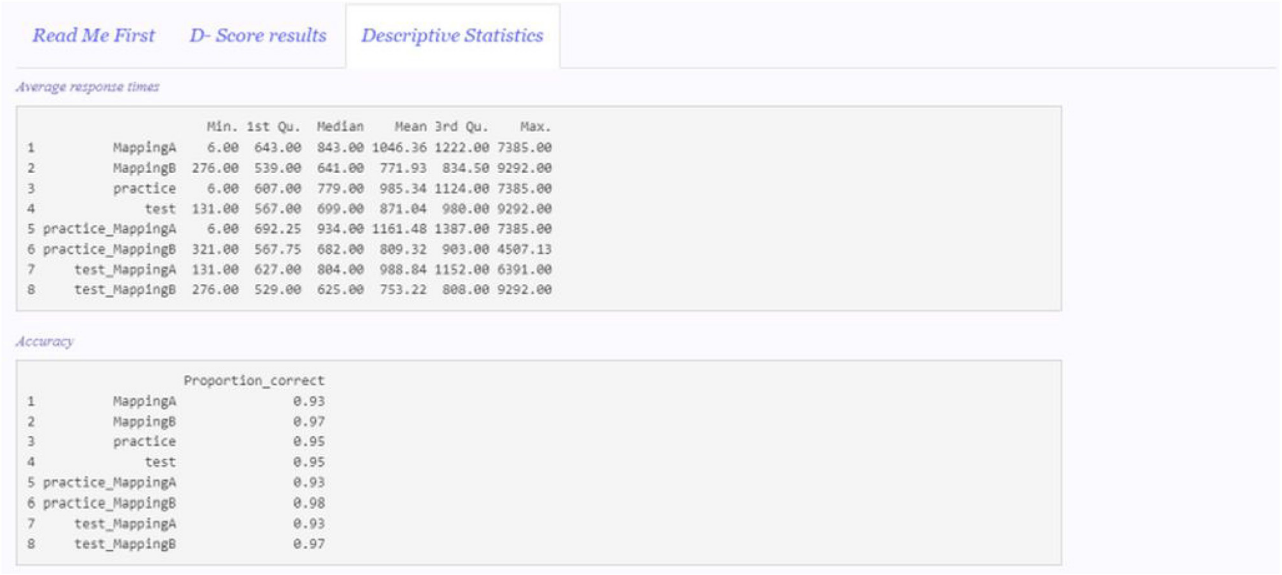
Figure 2 depicts the app appearance when the default settings are used (e.g., no participants are discarded, the plot of the *D-score* is the default representation plot). However, users are given the chance to customize the settings configuration for the *D-score* computation, and the display of the results, according to various criteria. For instance, if the **Accuracy cleaning** option is selected, the box `Accuracy deletion` would appear, reporting the number of participants with an error percentage higher than the selected threshold (if any). Likewise, if the **Fast participants cleaning** option is selected, the box `Participants` $<$ `300 ms` appears, reporting the number of participants with more than 10% of responses with latency faster than 300 ms (if any).

By looking at the graphical representation and the summary statistics of the results, it pops out that respondents' tended to have a preference (dislike) for Milk (Dark) chocolate, since they tended to be faster in Mapping B associative condition (i.e., the condition in which Milk chocolate was associated with positive words and Dark chocolate was associated with negative words). Moreover, the majority of the *D-score*s tended to have a strong effect (see section 4.3.1).

4.3.1. Graphic Display

DscoreApp provides the possibility to visually inspect the *D-score* results (**Figure 3**), both at the individual level (**Figure 3A**) and at the sample level (**Figures 3B–D**). The lines for interpreting the *D-score*s effect sizes are drawn at ±0.15 ("slight"), at ±0.35 ("moderate"), and at ±0.65 ("strong"), consistently with the guidelines in Project Implicit Website.

Users can customize the graphs to have a better inspection of the results. For instance, in the point graph participants



Average response times

```
                      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1          MappingA   6.00  643.00  843.00 1046.36 1222.00 7385.00
2          MappingB 276.00  539.00  641.00  771.93  834.50 9292.00
3          practice   6.00  607.00  779.00  985.34 1124.00 7385.00
4              test 131.00  567.00  699.00  871.04  980.00 9292.00
5 practice_MappingA   6.00  692.25  934.00 1161.48 1387.00 7385.00
6 practice_MappingB 321.00  567.75  682.00  809.32  903.00 4507.13
7     test_MappingA 131.00  627.00  804.00  988.84 1152.00 6391.00
8     test_MappingB 276.00  529.00  625.00  753.22  808.00 9292.00
```

Accuracy

```
                 Proportion_correct
1          MappingA            0.93
2          MappingB            0.97
3          practice            0.95
4              test            0.95
5 practice_MappingA            0.93
6 practice_MappingB            0.98
7     test_MappingA            0.93
8     test_MappingB            0.97
```

FIGURE 5 | Descriptive statistics panel.

order can be arranged by changing the options in the **Point graph** drop-down menu. The default representation ("None") follows the order participants had in the original data frame, while the "D-score: Increasing" and "D-score: Decreasing" options arrange participants by increasing or decreasing *D-score*, respectively. In the graphs including the histogram representation (**Figures 3B,D**), users can set the number of displayed bins by means of a slider, which appears only when either the "Histogram" or the "Histogram + Density" options are selected.

Graphical representation is a convenient way for inspecting the results, particularly for identifying extreme scores. However, it might be difficult to pinpoint a particular score in the graph, and then to link it to the corresponding participant in the data set. DscoreApp provides two useful and handy tools for linking specific points or area of the graphs to the corresponding participants and their *D-score*s. By clicking on a point in the points graph, the ID and the *D-score* of the participant corresponding to that point will appear in the `Point` box. By selecting an area in any of the graphs, the IDs and *D-score*s of the participants in the selected area will appear in the `Area` box.

All the graphs are downloadable by clicking on the **Download graph** button, which will be active only after the first graph is displayed. The default name of the graph will contain the type of graph and the specific *D-score* it shows. In the example in **Figure 2**, the default name will be "PointDefaultDscore3.pdf." All the graphs have a.pdf extension.

In the depicted example, five participants showed a *D-score* far from other participants' *D-score*s. By using the area highlighter, as illustrated in **Figure 4**, it is possible to immediately and conveniently identify the IDs of these participants (see `Area` box in the figure), and to check for any particular response pattern resulting in these scores in the original data set. Within these five participants, it is possible to note that there is also the participant obtaining the maximum *D-score* of the sample, namely Participant 31 (see `Summary` box in **Figure 2**).

4.4. Descriptive Statistics Panel

Figure 5 depicts the appearance of the "Descriptive statistics" panel.

The average response times and the proportion of correct responses in each of the mapping conditions and blocks of the IAT are reported. `MappingA` and `MappingB` include all the trials in both `practiceMappingA` and `testMappingA` and `practiceMappingB` and `testMappingB`, respectively. Practice blocks trials (`practiceMappingA` and `practiceMappingB`) compose `practice`, while test blocks trials (`testMappingA` and `testMappingB`) compose `test`. All the other categories (i.e., `practiceMappingA`, `practiceMappingB`,

TABLE 4 | Content of the Downloadable File.

Variable	Content
`participant`	Participants' IDs.
`n_trial`	Number of IAT trials (before data cleaning).
`slow10000`	Number of trials with latency > 10, 000 ms.
`num.300`	Number of trials with latency < 300 ms.
`num.400`	Number of trials with latency < 400 ms.
`mean.tot`	Average response time across all blocks.
`p_correct_block.practice.MappingA`	Proportion of correct responses in practice block of Mapping A.
`p_correct_block.practice.MappingB`	Proportion of correct responses in practice block of Mapping B.
`p_correct_block.test.MappingA`	Proportion of correct responses in test block of Mapping A.
`p_correct_block.test.MappingB`	Proportion of correct responses in test block of Mapping B.
`p_correct_bpool.practice`	Proportion of correct responses in practice blocks (`practiceMappingA` and `practiceMappingB`).
`p_correct_bpool.test`	Proportion of correct responses in test blocks (`testMappingA` and `testMappingB`).
`prop_correct_cond_MappingA`	Proportion of correct responses in Mapping A.
`prop_correct_cond_MappingB`	Proportion of correct responses in Mapping B.
`p_correct_tot`	Overall proportion of correct responses.
`d_practice.#`	*D-score* for the practice blocks.
`d_test.#`	*D-score* for the test blocks.
`dscore.#`	*D-score*.
`cond_ord`	Order of presentation of the two associative conditions (i.e., `MappingA_first` or `MappingB_first`).
`LegendMappingA`	Users' data set labels for Mapping A (e.g., `practiceMappingA_and_testMappingA`).
`LegendMappingB`	Users' data set labels for Mapping B (e.g., `practiceMappingB_and_testMappingB`).

testMappingA, and testMappingB) are composed by their respective number of trials in users' original data set.

The descriptive statistics are computed on the same data set on which the *D-score* is computed. For instance, if a *D-score* algorithm with the lower tail treatment is selected, the descriptive statistics are computed without considering the discarded trials. Likewise, if participants cleaning is selected, the descriptive statistics will not include the discarded participants.

4.5. Downloadable File

At the end of the computation, users can download a CSV file containing the last computed *D-score*. The default name of the file will contain the number of the selected *D-score* algorithm. The variables contained in the downloadable file are illustrated in **Table 4**.

The value in each column refers to the observed value for each participant. The # represents the number corresponding to the selected *D-score* algorithm.

In the depicted example, the default file name will be "ShinyAPPDscore3.csv."

5. FINAL REMARKS

The user-friendly and intuitive interface of DscoreApp makes its use straightforward, with no need for programming skills. Furthermore, the preparation of the data set for the analyses does not require any particular software or skill.

Beyond making the *D-score* computation easier, DscoreApp provides unique features that are not accessible with the available options for the *D-score* computation. First, DscoreApp provides the ability to immediately see the results and how they change in response to users' configurations. Additionally, since all the important information on participants performance and IAT functioning are available at the same time (e.g., *D-score*s, number of fast trials, IAT reliability), this app allows for grasping a complete overview of the functioning of the IAT. For instance, it allows for an immediate glimpse of how fast trials or inaccurate participants influence the results, and to identify critical aspects of the IAT that might deserve further investigation. Moreover, the preparation of the data set itself is particularly easy: Users will just have to eliminate the pure practice blocks and to rename the columns according to the instructions.

The downloadable file contains all the information that might be needed for further analysis on the IAT, or for plotting the results according to users' needs.

DscoreApp is constantly updated by the Authors, and new functions that are not present in this paper might be available in the future (e.g., other IAT reliability indexes). DscoreApp has been tested on several browsers (i.e., Google Chrome, Safari, Firefox, and Internet Explorer), and it has been found to have a reliable functioning. Problems encountered when using these browsers might be attributable to browsers security settings and/or poor internet connection.

DATA AVAILABILITY STATEMENT

The code used for the development of Dscore app can be found at OttaviaE/DscoreApp.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

OE was responsible for DscoreApp development and for the initial draft of the contribution. PA and ER revised the initial draft of the paper and provided useful comments for the final version.

REFERENCES

Anselmi, P., Voci, A., Vianello, M., and Robusto, E. (2015). Implicit and explicit sexual attitudes across genders and sexual orientations. *J. Bisexual.* 15, 40–56. doi: 10.1080/15299716.2014.986597

Attali, D. (2018). *shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds*. R package version 1.0.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2018). *shiny: Web Application Framework for R*. R package version 1.2.0.

Chen, L., Zhou, H., Gu, Y., Wang, S., Wang, J., Tian, L., et al. (2018). The neural correlates of implicit cognitive bias towards internet-related cues in internet addiction: an ERP study. *Front. Psychiatry* 9:421. doi: 10.3389/fpsyt.2018.00421

Dentale, F., Vecchione, M., Ghezzi, V., and Barbaranelli, C. (2019). Applying the latent state-trait analysis to decompose state, trait, and error components of the Self-Esteem Implicit Association Test. *Eur. J. Psychol. Assess.* 35, 78–85. doi: 10.1027/1015-5759/a000378

Egloff, B., and Schmukle, S. C. (2002). Predictive validity of an implicit association test for assessing anxiety. *J. Pers. Soc. Psychol.* 83:1441. doi: 10.1037//0022-3514.83.6.1441

Epifania, O. M., Anselmi, P., and Robusto, E. (in press). A fairer comparison between the implicit association test and the single category - implicit association test. *Test. Psychometrics Methodol. Appl. Psychol.*

Gawronski, B., Morrison, M., Phills, C. E., and Galdi, S. (2017). Temporal stability of implicit and explicit measures: a longitudinal analysis. *Pers. Soc. Psychol. Bull.* 43, 300–312. doi: 10.1177/0146167216684131

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: the implicit association test. *J. Pers. Soc. Psychol.* 74, 1464–1480. doi: 10.1037/0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., and Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *J. Pers. Soc. Psychol.* 85, 197–216. doi: 10.1037/0022-3514.85.2.197

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., and Banaji, M. R. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *J. Pers. Soc. Psychol.* 97:17. doi: 10.1037/a0015575

Karnal, N., Machiels, C. J., Orth, U. R., and Mai, R. (2016). Healthy by design, but only when in focus: communicating non-verbal health cues through symbolic meaning in packaging. *Food Qual. Preferen.* 52, 106–119. doi: 10.1016/j.foodqual.2016.04.004

Nosek, B. A., Banaji, M. R., and Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dyn.* 6, 101–115. doi: 10.1037/1089-2699.6.1.101

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Core Team.

Sriram, N., and Greenwald, A. G. (2009). The brief implicit association test. *Exp. Psychol.* 56, 283–294. doi: 10.1027/1618-3169.56.4.283

Steffens, M. C. (2004). Is the implicit association test immune to faking? *Exp. Psychol*. 51, 165–179. doi: 10.1027/1618-3169.51.3.165

Storage, D. (2018a). *IATanalytics: Compute Effect Sizes and Reliability for Implicit Association Test (IAT) Data*. R package version 0.1.1.

Storage, D. (2018b). *IATScore: Scoring Algorithm for the Implicit Association Test (IAT)*. R package version 0.1.1.

Zogmaister, C., and Castelli, L. (2006). La misurazione di costrutti impliciti attraverso l'Implicit Association Test. *Psicologia Sociale* 1, 65–94. doi: 10.1482/21502

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer AC declared a shared affiliation, with no collaboration, with the authors to the handling editor at the time of the review.

frontiers in Psychology

ORIGINAL RESEARCH
published: 14 January 2020
doi: 10.3389/fpsyg.2019.02893

Enhancing Statistical Inference in Psychological Research via Prospective and Retrospective Design Analysis

Gianmarco Altoè[1]*, *Giulia Bertoldo*[1], *Claudio Zandonella Callegher*[1], *Enrico Toffalini*[2], *Antonio Calcagnì*[1], *Livio Finos*[1] and *Massimiliano Pastore*[1]

[1] *Department of Developmental Psychology and Socialisation, University of Padova, Padova, Italy,* [2] *Department of General Psychology, University of Padova, Padova, Italy*

OPEN ACCESS

Edited by:
Pietro Cipresso,
Italian Auxological Institute (IRCCS),
Italy

Reviewed by:
Dirk Van Rooy,
Australian National University, Australia
Martha Michalkiewicz,
Heinrich Heine University of
Düsseldorf, Germany

***Correspondence:**
Gianmarco Altoè
gianmarco.altoe@unipd.it

Specialty section:
This article was submitted to
Quantitative Psychology and
Measurement,
a section of the journal
Frontiers in Psychology

Received: *22 September 2019*
Accepted: *06 December 2019*
Published: *14 January 2020*

Citation:
Altoè G, Bertoldo G,
Zandonella Callegher C, Toffalini E,
Calcagnì A, Finos L and Pastore M
(2020) Enhancing Statistical Inference
in Psychological Research via
Prospective and Retrospective Design
Analysis. Front. Psychol. 10:2893.
doi: 10.3389/fpsyg.2019.02893

In the past two decades, psychological science has experienced an unprecedented replicability crisis, which has uncovered several issues. Among others, the use and misuse of statistical inference plays a key role in this crisis. Indeed, statistical inference is too often viewed as an isolated procedure limited to the analysis of data that have already been collected. Instead, statistical reasoning is necessary both at the planning stage and when interpreting the results of a research project. Based on these considerations, we build on and further develop an idea proposed by Gelman and Carlin (2014) termed "prospective and retrospective design analysis." Rather than focusing only on the statistical significance of a result and on the classical control of type I and type II errors, a comprehensive design analysis involves reasoning about what can be considered a plausible effect size. Furthermore, it introduces two relevant inferential risks: the exaggeration ratio or Type *M* error (i.e., the predictable average overestimation of an effect that emerges as statistically significant) and the sign error or Type *S* error (i.e., the risk that a statistically significant effect is estimated in the wrong direction). Another important aspect of design analysis is that it can be usefully carried out both in the planning phase of a study and for the evaluation of studies that have already been conducted, thus increasing researchers' awareness during all phases of a research project. To illustrate the benefits of a design analysis to the widest possible audience, we use a familiar example in psychology where the researcher is interested in analyzing the differences between two independent groups considering Cohen's *d* as an effect size measure. We examine the case in which the plausible effect size is formalized as a single value, and we propose a method in which uncertainty concerning the magnitude of the effect is formalized via probability distributions. Through several examples and an application to a real case study, we show that, even though a design analysis requires significant effort, it has the potential to contribute to planning more robust and replicable studies. Finally, future developments in the Bayesian framework are discussed.

Keywords: prospective and retrospective design analysis, Type *M* and Type *S* errors, effect size, power, psychological research, statistical inference, statistical reasoning, R functions

> "*If statisticians agree on one thing, it is that scientific inference should not be made mechanically.*"
>
> Gigerenzer and Marewski (2015, p. 422)

> "***A**ccept uncertainty. Be **t**houghtful, **o**pen, and **m**odest. Remember 'ATOM'.*"
>
> Wasserstein et al. (2019, p. 2)

1. INTRODUCTION

In the past two decades, psychological science has experienced an unprecedented replicability crisis (Ioannidis, 2005; Pashler and Wagenmakers, 2012; Open Science Collaboration, 2015) that has uncovered a number of problematic issues, including the adoption of Questionable Research Practices (John et al., 2012) and Questionable Measurement Practices (Flake and Fried, 2019), the reliance on excessively small samples (Button et al., 2013), the misuse of statistical techniques (Pastore et al., 2019), and the consequent misleading interpretation and communication of research findings (Wasserstein et al., 2019).

Whereas some important reasons for the crisis are intrinsically related to psychology as a science (Chambers, 2019), leading to a renewed recommendation to rely on strong and well-formalized theories when planning a study, the use of statistical inference undoubtedly plays a key role. Specifically, the inferential approach most widely used in psychological research, namely Null Hypothesis Significance Testing (NHST), has been strongly criticized (Gigerenzer et al., 2004; Gelman, 2018; McShane et al., 2019). As a consequence, several alternative approaches have received increasing attention, such as the use of Bayes Factors for hypothesis testing and the use of both Frequentist and Bayesian methods to estimate the magnitude of the effect of interest with uncertainty (see Kruschke and Liddell, 2018, for a comprehensive historical review).

In the current paper, we focus on an upstream—but still neglected—issue that is unrelated to the approach chosen by the researcher, namely the need for statistical reasoning, i.e., "to reason about data, variation and chance" (Moore, 1998, p. 1253), during all phases of an empirical study. Our work was inspired by the famous statistician Ronald Fisher (1890–1962), who stated that, "To consult the statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of" (Fisher, 1938, p.17). Indeed, we argue that statistical inference is too often seen as an isolated procedure that is limited to the analysis of data that have already been collected. In particular, we emphasize the non-trivial importance of making statistical considerations at the onset of a research project. Furthermore, we stress that, although Fisher has ironically defined them as a "post-mortem examination," appropriate evaluations of published results can provide a relevant contribution to the progress of (psychological) science. The ultimate goal of this paper is to increase researchers' awareness by promoting active engagement when designing their research.

To achieve this goal, we build on and further develop an idea proposed by Gelman and Carlin (2014) called "prospective and retrospective design analysis," which is virtually absent in current research practice. Specifically, to illustrate the benefits of design analysis to the widest possible audience, we use a familiar example in psychology where the researcher is interested in analyzing the differences between two independent groups considering Cohen's d (Cohen, 1988) as an effect size measure.

In brief, the term *design analysis* has been proposed by Gelman and Carlin (2014) as a broader definition of power analysis—a concept that in the statistical literature traditionally indicates the determination of an appropriate sample size, at prespecified levels of Type I and Type II errors and a "plausible effects size" (Gigerenzer et al., 2004). Indeed, a comprehensive design analysis should also explicitly consider other two inferential risks: Type M error and Type S error. Type M error (where M stands for magnitude) is also known as exaggeration ratio and indicates how much a statistically significant effect is, on average, overestimated in comparison to a "plausible effect size." Type S error (where S stands for sign) indicates the risk that a statistically significant effect is estimated in the wrong direction. These two errors will be further discussed in the subsequent paragraphs with several examples. Notably, the estimation of these errors will require an effort from psychologists to introduce their expert knowledge and hypothesize what could be considered a "plausible effect size." As we will see later, a key aspect of design analysis is that it can be usefully carried out both in the planning phase of a study (i.e., prospective design analysis) and for the evaluation of studies that have already been conducted (i.e., retrospective design analysis).

Although the idea of a design analysis could be developed within different inferential statistical approaches (e.g., Frequentist and Bayesian), in this paper we will rely on the Neyman-Pearson (N-P) approach (Pearson and Neyman, 1928) as opposed to the widely used NHST. The rationale for this choice is that, in addition to other strengths, the N-P approach includes formalization of the *Null Hypothesis* (i.e., the absence of an effect) like NHST, but it also includes an explicit formalization of the *Alternative Hypothesis* (i.e., the magnitude of the expected effect). For a more comprehensive description of the difference between N-P and NHST approaches, we refer the reader to Gigerenzer et al. (2004).

In the next paragraphs, we will briefly review the main consequences of underpowered studies, discuss two relevant misconceptions concerning the interpretation of statistically significant results, and present a theoretical framework for design analysis, including some clarifications regarding the concept of "plausible effect size." In section 2, through familiar examples within psychological research, the benefits of prospective and retrospective design analysis will be highlighted. In section 3, we will propose a specific method that, by explicitly taking uncertainty issues into account, could further assist researchers in evaluating scientific findings. Subsequently, in section 4, a real case study will be presented and analyzed. Finally, in section 5, we will summarize the potentials, further developments, and limitations of our proposal.

To increase readability and ensure transparency of our work, we also include two **Appendices** as Supplementary Material:

- **Appendix A.** A detailed description concerning the computation and the interpretation of Cohen's *d*.
- **Appendix B.** A brief explanation of the *ad-hoc* R (R Core Team, 2018) functions used in the paper. Details on how to reproduce the presented examples and on how to use our R functions for other purposes are also provided. Furthermore, the source code of our functions, `PRDA.R`, is freely available at the Open Science Framework (OSF) at the link https://osf.io/j8gsf/files/.

1.1. The Consequences of Underpowered Studies in Psychology

In 1962, Cohen called attention to a problem affecting psychological research that is still very much alive today (Cohen, 1962). Researchers seemed to ignore the statistical power of their studies—which is not considered in NHST (Gigerenzer et al., 2004)—with severe consequences for the robustness of their research findings. In the N-P approach, the power of a statistical test is defined as the probability that the test has to reject the Null Hypothesis (H_0) when the Alternative Hypothesis (H_1) is true. One of the problems with underpowered studies is that the probability of finding an effect, if it actually exists, is low. More importantly, if a statistically significant result (i.e., "in general," when the observed p-value is <0.05 and consequently H_0 is rejected; see Wasserstein et al., 2019) is obtained in an underpowered study, the effect size associated with the observed p-value might be "too big to be true" (Button et al., 2013; Gelman and Carlin, 2014).

This inflation of the effect sizes can be seen when examining results of replication projects, which are usually planned to have higher power than the original studies. For example, the Open Science Collaboration (2015, pp. 4–5) reported that "Overall, original study effect sizes ($M = 0.403$, $SD = 0.188$) were reliably larger than replication effect sizes ($M = 0.197$, $SD = 0.257$)," and in the Social Science Replication Project (Camerer et al., 2018, p. 637), "the effect size of the replication was on average about 50% of the original effect size." These considerations contributed to the introduction in the literature of the term "decline effect," defined as "the notion that science routinely observes effect sizes decrease over repeated replications for reasons that are still not well-understood" (Schooler, 2014, p. 579).

Given that underpowered studies are widespread in psychology (Cohen, 1962; Sedlmeier and Gigerenzer, 1989; Maxwell, 2004), the shrinkage of effect sizes in replications could be partially explained by the fallacy of "what does not kill statistical significance makes it stronger" (Loken and Gelman, 2017) and by the trap of the "winner's curse" (Button et al., 2013).

1.2. The "What Does Not Kill Statistical Significance Makes It Stronger" Fallacy and the "Winner's Curse" Trap

When a statistically significant result is obtained in an underpowered study (e.g., power = 40%), in spite of the low probability of this event happening, the result might be seen as even more remarkable. In fact, the researcher might think, "If obtaining a statistically significant result is such a rare event, and in my experiment I obtained a statistically significant result, it must be a strong one." This is called the "what does not kill statistical significance makes it stronger" fallacy (Loken and Gelman, 2017). The reason why this is a fallacy lies in the fact that it is possible to obtain statistical significance due to the presence of many other factors that are different from the presence of a real effect. The researcher degrees of freedom, large measurement errors, and small sample sizes all contribute to the creation of noise in the data, thus inflating the perhaps true but small underlying effect. Then, if the procedure used to analyze those data is only focused on a threshold (like in NHST, with a conventional significance level of 0.05), the noise in the data allows it to pass this threshold.

In these situations, the apparent win in terms of obtaining a statistically significant result is actually a loss; "the lucky" scientist who makes a discovery is cursed by finding an inflated estimate of that effect (Button et al., 2013). This is called the "Winner's curse," and **Figure 1** shows an example of this. In this hypothetical situation, the researcher is interested in studying an effect that can plausibly be of small dimensions, e.g., Cohen's d of 0.20 (see **Appendix A**, for a detailed description of the calculation and interpretation of Cohen's d). If they decide to compare two groups on the outcome variable of interest, using 33 participants per group (and performing a two-tailed test), they will never be able to simultaneously reject H_0 and find an effect close to what it is plausible in that research field (i.e., 0.20). In fact, in this underpowered study (i.e., based on a d of 0.20, the actual power is only 13%) all the effects falling in the "rejection regions" are higher than 0.49 or smaller than −0.49, and 0.20 falls in the region where the decision rules state that you cannot reject H_0 under the NHST approach, and that you can accept H_0 under the N-P approach.

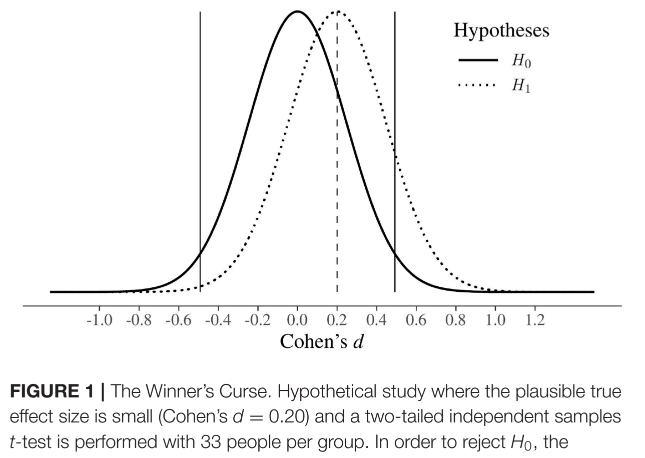


FIGURE 1 | The Winner's Curse. Hypothetical study where the plausible true effect size is small (Cohen's $d = 0.20$) and a two-tailed independent samples t-test is performed with 33 people per group. In order to reject H_0, the researcher has to overestimate the underlying true effect, which is indicated by the dashed vertical line. Note: the rejection regions of H_0, given a significance level of 0.05, lie outside the vertical black lines.

1.3. Beyond Power: The Design Analysis

As we saw in the previous example, relying solely on the statistical significance of a result can lead to completely misleading conclusions. Indeed, researchers should take into account other relevant information, such as the hypothesized "plausible effect size" and the consequent power of the study. Furthermore, to assist researchers with evaluating the results of a study in a more comprehensive way, Gelman and Carlin (2014) suggested that two other relevant types of errors should be considered in addition to the traditional Type I and Type II errors, namely Type *M* and Type *S* errors (see also Gelman and Tuerlinckx, 2000; Lu et al., 2019). Specifically, a Type *M* [magnitude] error or *exageration ratio* can be viewed as the expected average overestimation of an effect that emerges as statistically significant, whereas a Type *S* [sign] error can be viewed as the probability of obtaining a statistically significant result in the opposite direction with respect to the sign of the hypothesized plausible effect size.

Based on this consideration, Gelman and Carlin (2014) proposed the term "*design analysis*" to broadly identify the analysis of the properties of different studies, such as their statistical power as well as Type *M* and Type *S* errors. Moreover, as is shown in the next paragraph, in design analysis particular emphasis is given to the elicitation and formalization of what can be considered a *plausible effect size* (see also paragraph 1.4) for the study of interest. In this regard, it is important to make a clarification. Although Gelman and Carlin (2014) developed a design analysis relying on an unstandardized effect size measure (i.e., the difference between two means), we have, in this paper, adapted their method to deal with Cohen's *d*, a standardized measure of effect size that is more commonly used in psychology (see **Appendix A** for more details on the reasons that motivated this choice).

Given these premises, the steps to perform design analysis using Cohen's *d* as a measure of effect size can be summarized in three steps:

1. A plausible effect size for the study of interest needs to be identified. Rather than focusing on data at hand or on noisy estimates of a single pilot study, the formalization of a plausible effect size should be based on an extensive theoretical literature review and/or on meta-analyses. Moreover, specific tools (see for example Zondervan-Zwijnenburg et al., 2017; O'Hagan, 2019; Zandonella Callegher et al., 2019) that allow for the incorporation of expert knowledge can also be considered to increase the validity of the plausible effect size elicitation process[1].
2. Based on the experimental design of the study of interest (in our case, a comparison between two independent groups), a large number of simulations (i.e., 100,000) will be performed according to the identified plausible effect size. This procedure serves to provide information about what to expect if the experiment is replicated an infinite number of times and assuming that the pre-identified plausible effect is true.
3. Given a fixed level of Type I error (e.g., 0.05), power as well as type *M* and type *S* errors will be calculated. Specifically, power will be estimated as the ratio between the number of significant results obtained and the number of replicates (i.e., the higher the power, the higher the probability of detecting the plausible effect). A Type *M* error will be estimated as the ratio between the mean of the absolute values of the statistically significant replicated effect sizes and the plausible effect size. In this case, larger values indicate an expected large overestimation of the plausible effect size. Type *S* error will be the ratio between the number of significant results with opposite signs with regard to the plausible effect size and the total number of significant results. Put in other terms, a type *S* error estimates the probability of obtaining a significant result in the wrong direction.

Although the procedure may seem complex to implement, we have here https://osf.io/j8gsf/files/ (see also **Appendix B**) made available some easy-to-use R functions that allow others to perform different types of design analysis, even for less experienced users. The same functions will also be used in the examples and application presented in this paper.

To get a first idea of the benefits of design analysis, let us re-analyze the hypothetical study presented in **Figure 1**. Specifically, given a plausible effect size equal to $d = 0.20$ and a sample size of 33 participants per group, a design analysis will highlight the following information: power = 13%, Type *M* error = 3.11, and Type *S* error = 2%. Despite the low power, which shows that the study has only a 13% probability of detecting the plausible effect size, a type *M* error explicitly indicates that the expected overestimate of a result that will emerge as statistically significant is around three times the plausible effect. Furthermore, given a Type *S* error of 2%, there is also a non-negligible probability of obtaining a significant result in the wrong direction. Overall, the results of design analysis clearly tell the researcher that the study of interest could provide very poor support to both the existence and non-existence of a plausible effect size.

Another advantage of design analysis, which will be better explored in the following sections, is that it can be effectively used in the planning phase of a study, i.e., *prospective design analysis*, as well as in the evaluation of already obtained study results, i.e., *retrospective design analysis*. For example, in prospective design analysis, considerations concerning power as well as Type *M* and Type *S* errors could assist researchers in deciding the appropriate sample size for detecting the effect of interest (if it actually exists). In a retrospective design analysis, power as well as Type *M* and Type *S* errors (always calculated using the theoretically plausible effect size) can be used to obtain information about the extent to which the results of the study could be exaggerated and/or in the wrong direction. Most importantly, we believe that, engaging in a retrospective design analysis helps researchers to recognize the role of uncertainty and to make more reasonable statistical claims, especially in those cases at risk of falling in the aforementioned "Winner's Curse" trap.

[1] To obtain a more comprehensive picture of the inferential risks associated with their study, we suggest that researchers inspect different scenarios according to different plausible effect sizes and perform more than one design analysis (see for example our application to a real case study in section 4).

In conclusion, it is important to note that whatever the type of design analysis chosen (prospective or retrospective), the relationships between power, type M error, and type S error are the same. For illustrative purposes, these relationships are graphically displayed as a function of sample size in **Figure 2**. A medium-to-small effect of $d = 0.35$ (i.e., a reasonable average effect size for a psychological study in the absence of other relevant information, see also section 4) was considered as a plausible effect size, and Type I error was set at 0.05.

As expected, power increases as sample size increases. Moreover, type M and type S errors decrease as the size of the sample increases, with the latter showing a much steeper decrease.

From an applied perspective, issues with type M and S errors emerge with underpowered studies, which are very common in psychological research. Indeed, as can be seen in **Figure 2**, for a power of 40% (obtained with 48 participants per group), the type M error reaches the worrisome value of 1.58; for a power around 10% (i.e., with 10 participants per group), even a type S error becomes relevant (around 3%).

1.4. What Does "Plausible Effect Size" Mean?

> "*Thinking hard about effect sizes is important for any school of statistical inference [i.e., Frequentist or Bayesian], but sadly a process often neglected.*"
>
> Dienes (2008, p. 92)

The main and most difficult point rests on deciding what could be considered a "plausible effect size." Although this might seem complex, studies are usually not developed in a void. Hypotheses are derived from theories that, if appropriately formalized in statistical terms, will increase the validity of the inferential process. Furthermore, researchers are commonly interested in knowing the size and direction of effects; as shown above, this corresponds to control for a Type M [magnitude] error and a type S [sign] error.

From an epistemological perspective, Kruschke (2013) suggests an interesting distinction between *strong theories* and *weak theories*. Strong theories are those that try to make precise predictions and could be, in principle, more easily disconfirmed. For example, a strong theory could hypothesize a medium-sized positive correlation between two variables. In contrast, weak theories make broader predictions, such as the hypothesis that two variables are correlated without specifying the strength and direction of the correlation (Dienes, 2008). The former type allows many more research findings to disconfirm the hypothesis, whereas the latter type allows only the result of no correlation to disconfirm it. Specifically, following Karl Popper (1902–1994), it could be argued that theories explaining virtually everything and that are hard to disconfirm risk being out of the realm of science. Thus, scientific theories should provide at least a hint regarding the effect that is expected to be observed.

A challenging point is to establish the dimension of this effect. It might seem paradoxical that the researcher must provide an estimate of the effect size before running the experiment given that they will conduct the study with the precise aim of finding what that estimate is. However, strong theories should allow to make such predictions, and the way in which science accumulates should provide increasing precision to these predictions.

In practice, it might be undesirable to simply take the estimate found in a pilot study or from a single previous study published in the literature as the "plausible effect size." In fact, the plausible effect size refers to what could be approximately the true value of the parameter in the population, whereas the results of pilots or single studies (especially if underpowered) are noisy estimates of that parameter.

In line with Gelman and Carlin (2014), we suggest the use of information outside the data at hand, such as literature reviews and/or meta-analyses taking into account issues concerning publication bias (Borenstein et al., 2009). Moreover, as stated in the previous paragraph, promising procedures to elicit and formalize expert knowledge should also be considered. It is important to note that, whatever the procedures, all assumptions that will lead to the identification of a plausible effect size must be communicated in a transparent manner, thus increasing the information provided by a study and ensuring more reasonable statistical claims related to the obtained results, regardless of whether they are significant or not.

As we have seen, the identification of a plausible effect size (or a series of plausible effect sizes to explore different scenarios) requires significant effort from the researcher. Indeed, we believe that this kind of reasoning can make a substantial contribution to the planning of robust and replicable studies as well as to the efficient evaluation of obtained research findings.

To conclude, we leave the reader with a question: "All other conditions being equal, if you had to evaluate two studies of the same phenomenon, the first based on a formalization of the expected plausible effect sizes of interest that is as accurate as possible, and the second one in which the size of the effects of interest was not taken into account, the findings of which study would you believe the most?" (R. van de Schoot, personal communication).

2. PROSPECTIVE AND RETROSPECTIVE DESIGN ANALYSIS

To highlight the benefits of design analysis and to make familiar the concepts of Type M and Type S errors, we will start with a simple example that is well-known in psychological research, i.e., the comparison between the means of two independent groups[2].

In particular, the goal of our hypothetical case study was to evaluate the differences between two treatments that aim to improve a cognitive ability called Y. Both treatments have the same cost, but the first is innovative, whereas the second is traditional. To this end, the researchers recruited a sample of participants who were homogeneous with respect to pre-specified

[2] We remind the reader that **Appendix B** provides a brief explanation of the *ad-hoc* R functions used in the paper as well as details on how to reproduce the presented examples and on how to use our R functions for other purposes. The source code of our functions, `functions_PRDA.R`, is available at the link https://osf.io/j8gsf/files/.

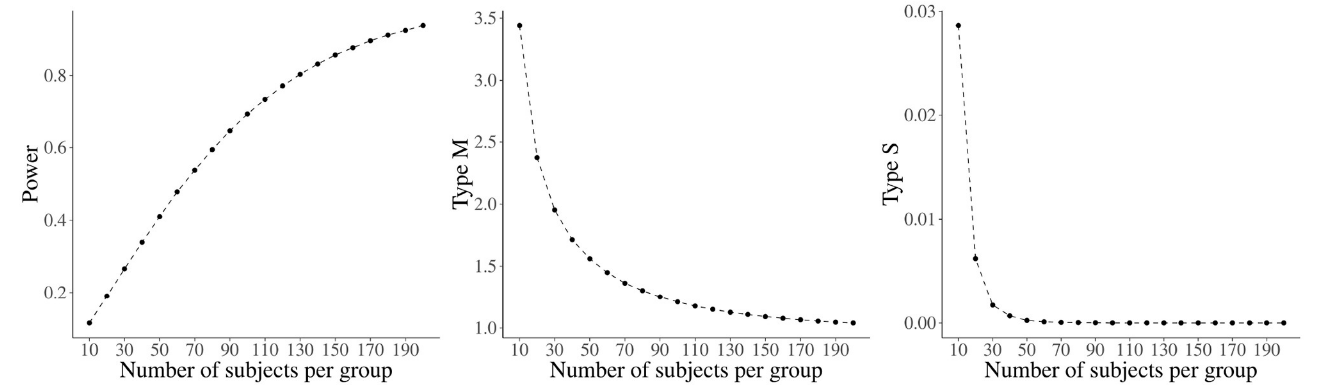


FIGURE 2 | Relationship between sample size and Power, Type *M*, and Type *S* for a Cohen's *d* of 0.35 in an independent samples *t*-test. Type I error is set at 0.05.

relevant study variables (i.e., age, IQ, etc.). Next, they randomly assigned each participant to one of the two conditions (i.e., innovative vs. traditional treatment). After the treatment phase was completed, the means of the two groups were compared.

2.1. Prospective Design Analysis

Before collecting data, the researchers planned the appropriate sample size to test their hypotheses, namely that there was a difference between the means of *G*1 (the group to which the innovative treatment was administered) and *G*2 (the group to which the traditional treatment was administered) vs. there was no difference.

After an extensive literature review concerning studies theoretically comparable to their own, the researchers decided that a first reasonable effect size for the difference between the innovative and the traditional treatment could be considered equal to a Cohen's *d* of 0.30 (see **Appendix A** for a detailed description of the calculation and interpretation of Cohen's *d*). Due to the possible presence of publication bias (Borenstein et al., 2009), which could lead to an overestimation of the effects of published studies, the researchers decided to be more conservative about the estimate of their plausible effect size. Thus, they decided to consider a Cohen's *d* of 0.25. Eventually, all researchers agreed that a Cohen's *d* of 0.25 could also represent a clinically relevant effect in order to support the greater efficacy of the innovative treatment.

Based on the above considerations, the researchers started to plan the sample size for their study. First, they fixed the Type I error at 0.05 and—based on commonly accepted suggestions from the psychological literature—fixed the power at 0.80. Furthermore, to explicitly evaluate the inferential risks connected to their choices, they calculated the associated Type *M* and Type *S* errors.

Using our `R` function `design_analysis`, they obtained the following results:

```
> design_analysis (d=0.25, power=0.80)
     d  power      n   typeS typeM
  0.25   0.80 252.00    0.00  1.13
```

Based on the results, to achieve a power of 0.80, a sample size of 252 for each group was needed (i.e., total sample size = 504). With this sample size, the risk of obtaining a statistically significant result in the wrong direction (Type *S* error) was practically 0, and the expected exaggeration ratio (Type *M* error) was 1.13. In other words, the expected overestimation related to effects that would emerge as statistically significant would be around 13% of the hypothesized plausible effect size.

Although satisfied in terms of expected type *S* and type *M* risks, the researchers were concerned about the economic feasibility of recruiting such a "large" number of subjects. After a long discussion, they decided to explore which inferential risks would result for a lower level of power, namely 60%[3].

Using the function `design_analysis`

```
> design_analysis (d=0.25, power=0.60)
     d  power      n   typeS typeM
  0.25   0.60 158.00    0.00  1.30
```

they discovered that: (1) the overall required sample size was considerably smaller (from 504 to 316 = 158 × 2), thus increasing the economic feasibility of the study; (2) the Type *S* error remained negligible (0%); and (3) the exaggeration ratio considerably increased (from 1.13 to 1.30); thus, an effect that will emerge as statistically significant will be on average 130% of the hypothesized plausible effect size.

The researchers had to make a decision. From a merely statistical point of view, the optimal choice would be to consider a power of 80% that is associated with a Type *M* error of 1.13 (i.e., mean overestimation of ∼10%) and a negligible Type *S* error close to zero. However, it is important to highlight that these values cannot be considered universal benchmarks. Indeed, other relevant aspects must be considered, such as the practical implications of an expected overestimation of the plausible effect size, the phase of the study (i.e., preliminary/exploratory, intermediate, or final/confirmatory), and feasibility constraints.

[3] Specifically, we agree with Gelman (2019) that an 80% level of power should not be used as an automatic routine, and that requirements of 80% power could encourage researchers to exaggerate their effect sizes when planning sample size.

Whatever the decision, the researchers must be aware of the inferential risks related to their choice. Moreover, when presenting the results, they must be transparent and clear in communicating such risks, thus highlighting the uncertainty associated with their conclusions.

2.2. Retrospective Design Analysis

To illustrate the usefulness of retrospective design analysis, we refer to the example presented in the previous paragraph. However, we introduce three new scenarios that can be considered as representative of what commonly occurs during the research process:

- **Scenario 1 (S1): Evaluating sample size based on a single published study**[4]
  Imagine that the researchers decide to plan their sample size based on a single published study in the phase of formalizing a plausible effect size, either because the published study presents relevant similarities with their own study or because there are no other published studies available.
  Question: What type of inferential risks can be associated with this decision?
  Issues: Using a single study as a reference point without considering other sources (e.g., theoretical framework, expert opinion, or a meta-analysis), especially when the study has a low sample size and/or the effect of interest is small, can lead to use an excessively optimistic estimate of the effect size when planning an appropriate sample size (Gelman and Carlin, 2014).
- **Scenario 2 (S2): Difficulty in recruiting the planned number of research participants**
  Imagine that, due to unforeseen difficulties (e.g., insufficient funding), the researchers are not able to recruit the pre-planned number of participants as defined based on prospective design analysis.
  Question: How do you evaluate the inferential risks associated with the new reduced sample size? How do you communicate the obtained results?
  Issues: Researchers are often tempted to evaluate the results of their study based on the observed effect size. This procedure, known as "*post-hoc* power analysis," has been strongly criticized, and many statistical papers explicity advise against its use (see for example, Goodman and Berlin, 1994; Gelman, 2019). Indeed, to evaluate the information provided by the obtained results, researchers should use the *a priori* plausible effect size, i.e., the one formalized before collecting their data.
- **Scenario 3 (S3): No prospective design analysis because the number of participants is constrained**
  Imagine the number of participants involved in the study have specific characteristics that make it impossible to yield a large sample size, or that the type of treatment is particularly expensive and cannot therefore be tested on a large sample. In this case, the only possibility is to recruit the largest possible number of participants.
  Question: What level of scientific quality can be provided by the results?
  Issues: Although study results can provide a useful contribution to the field, there are several associated inferential risks that the researchers need to communicate in a transparent and constructive way.

As we will see below, retrospective design analysis can be a useful tool to deal with the questions and the issues raised across all three scenarios.

For the sake of simplicity and without loss of generalizability, suppose that in each of the three scenarios the researchers obtained the same results (see **Table 1**).

At a first glance, the results indicated a statistically significant difference in favor of the innovative treatment (see **Table 1**), with a large effect size (i.e., $d = 0.90$). However, the 95% confidence interval for Cohen's d was extremely wide, suggesting that both medium-small (i.e., $d = 0.38$) and very large (i.e., $d = 1.43$) effects were consistent with the observed data.

A closer look indicated that the estimated effect size seemed too large when compared with the initial guess of the researchers (i.e., $d = 0.25$). Furthermore, an estimated d of 0.90 seemed, in general, implausibly large for a difference between two cognitive treatments (see also **Appendix A**). The latter interpretation seemed to be also supported by the fact that the hypothesized plausible effect size was not even included in the estimated confidence interval. Overall, in order to prevent the aforementioned "Winner's Curse" and "What Does Not Kill Statistical Significance Makes It Stronger" heuristics, results had to be evaluated and eventually communicated with caution and skepticism.

To obtain a clearer picture of the inferential risks associated with the observed results, we performed a retrospective design analysis using $d = 0.25$ as plausible effect size and 31 participants per group as sample size:

```
> design_analysis (n=31, d=0.25)
power typeS typeM
 0.16  0.01  2.59
```

As can be seen, the power was markedly low (i.e., only 16%) and the Type M error even suggested an expected overestimation around two and a half times the plausible effect size. Lastly, the Type S error, although small, indicated a 1% risk of obtaining a significant result in the wrong direction (i.e., the traditional treatment is better than the innovative treatment). Let's see how this information could be helpful to deal with the three presented scenarios.

TABLE 1 | Comparison of the cognitive skill Y between the two groups.

Group	n	M	SD	t (df)	p	Cohen's d (95% CI)
Innovative treatment	31	114	16	3.496 (60)	0.001	0.90 (0.38–1.43)
Traditional treatment	31	100	15			

[4] Even though, in this paper, we strongly recommend that one does not plan the sample size based on a single study, we propose this example to further emphasize the inferential risks associated with the information provided by a single underpowered study.

In S1, the researchers took a single noisy estimate as the plausible effect size from a study that found a "big" effect size (e.g., 0.90). The retrospective design analysis showed what happens if the plausible effect size is, in reality, much smaller (i.e., 0.25). Specifically, given the low power and the high level of Type *M* error, researchers should abandon the idea of planning their sample size based on a single published study. Furthermore, issues regarding the presence of Questionable Research Practices (John et al., 2012; Arrison, 2014) and Questionable Measurement Practices (Flake and Fried, 2019) in the considered published study must at least be explored. From an applied perspective, researchers should continue with a more comprehensive literature review and/or consider the opportunity of using an effect size elicitation procedure that is based on expert knowledge (Zondervan-Zwijnenburg et al., 2017; O'Hagan, 2019).

In S2, to check the robustness of their results, researchers might initially be tempted to conduct a power analysis based on their observed effect size ($d = 0.90$). Acting in this way, they would obtain a completely misleading *post-hoc* power of 94%. In contrast, the results of the retrospective design analysis based on the a-priori plausible effect size ($d = 0.25$) highlight the high level of inferential risks related to the observed results. From an applied perspective, researchers should be very skeptical about their observed results. A first option could be to replicate the study on an independent sample, perhaps asking for help from other colleagues in the field. In this case, the effort to recruit a larger sample could be well-justified based on the retrospective design analysis.

In S3, given the low power and the high level of Type *M* error, results should be presented as merely descriptive by clearly explaining the uncertainty that characterizes them. Researchers should first reflect on the possibility of introducing improvements to the study protocol (i.e., improving the reliability of the study variables). As a last option, if improvements are not considered feasible, the researchers might consider not continuing their study.

Despite its advantages, we need to emphasize that design analysis should not be used as an *automatic problem solver machine*: "Let's pull out an effect size …let me see the correct sample size for my experiment." In other words, to obtain reliable scientific conclusions there is no "free lunch." Rather, psychologists and statisticians have to work together, case by case, to obtain a reasonable effect size formalization and to evaluate the associated inferential risks. Furthermore, researchers are encouraged to explore different scenarios via a sensitivity analysis (see section 4) to better justify and optimize their choices.

3. INCORPORATING UNCERTAINTY CONCERNING EFFECT SIZE FORMALIZATION IN RETROSPECTIVE DESIGN ANALYSIS

As shown in the previous examples, a key point both in planning (i.e., prospective design analysis) and in evaluating (i.e., retrospective design analysis) a study is the formalization of a plausible effect size. Using a single value to summarize all external information and previous knowledge with respect to the study of interest can be considered an excessive simplification. Indeed, all uncertainty concerning the magnitude of the plausible effect size is not explicitly taken into consideration. In particular, the level of heterogeneity emerging from the examination of published results and/or from different opinions of the consulted experts, which can be poorly formalized. The aim of this paragraph is to propose a method that can assist researchers with dealing with these relevant issues. Specifically, we will focus on the evaluation of the results of a study (i.e., retroprospective design analysis).

Our method can be summarized in the three steps: (1) defining a lower and an upper bound within which the plausible effect size can reasonably vary; (2) formalizing an appropriate probability distribution that reflects how the effect size is expected to vary; and (3) conducting the associated analysis of power, Type *M* error, and Type *S* error.

To illustrate the procedure, we use the study presented in **Table 1** as a reference. Let us now hypothesize that, after a thorough evaluation of external sources, the researchers conclude that a plausible effect size could reasonably vary between 0.20 and 0.60 (instead of specifying a too simplistic single-point value). It should be noted that, from a methodological perspective, the specification of a "plausible interval" can be considered an efficient and informative starting point to elicit the researchers' beliefs (O'Hagan, 2019).

At this point, a first option could be to assume that, within the specified interval, all effect size values have the same probability of being true. This assumption can be easily formalized using a Uniform distribution, such as the one shown in **Figure 3** (left panel).

However, from an applied point of view it is rare for the researcher to expect that all values within the specified interval have the same plausibility. Indeed, in general conditions, it is more reasonable to believe that values around the center of the interval (i.e., 0.40 in our case) are more plausible, and that their plausibility gradually decreases as they move away from the center. This expectation can be directly formalized in statistical terms using the so-called "doubly truncated Normal distribution." On an intuitive level (for a more complete description see Burkardt, 2014), the doubly truncated Normal distribution can be seen as a Normal distribution whose values are forced to vary within a specific closed interval. In case of the formalization of the plausible effect size, we propose the use of doubly truncated Normal distribution with several parameters: a lower and an upper bound according to the pre-specified plausible interval, a mean fixed at the center of the interval, and a standard deviation that reflects the hypothesized uncertainty around the center. A standard deviation of $\frac{1}{10}$ the length of the chosen interval will produce a substantially Normal distribution. Higher values, like $\frac{1}{6}$ the length of the interval (see right panel of **Figure 3**) will lead to normal-like distributions with increased probability on the tails, thus reflecting greater uncertainty around the center.

Coming back to our example, suppose that the researchers want to evaluate the study of interest assuming a plausible interval for Cohen's d as the one represented in **Figure 3**.

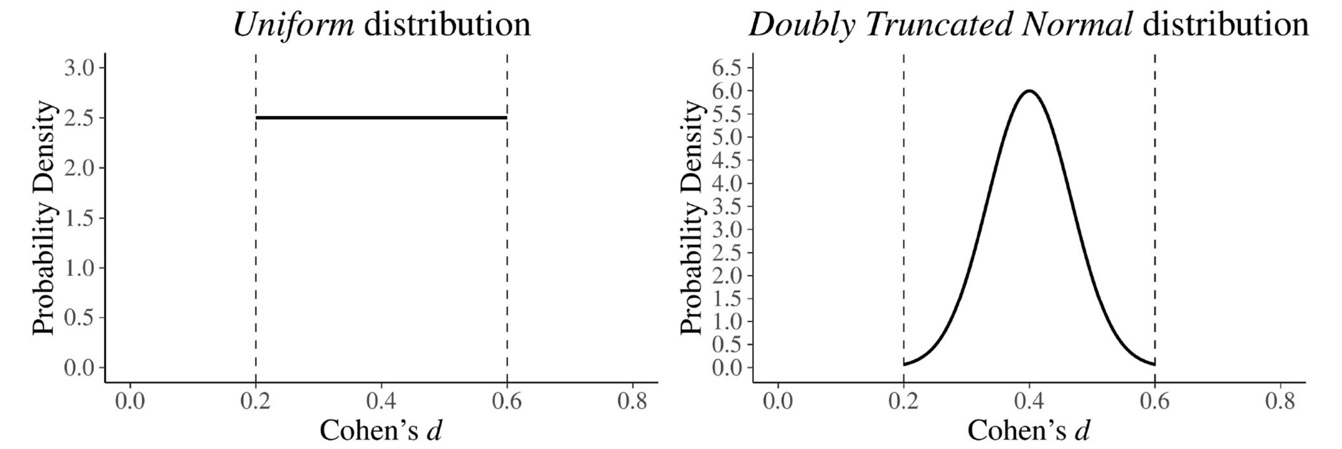


FIGURE 3 | Different ways to formalize a plausible interval for the effect size *d*. In the left panel, a Uniform distribution with lower bound = 0.20 and upper bound = 0.60 is used. In the right panel, a doubly truncated Normal distribution with lower bound = 0.20, upper bound = 0.60, mean = 0.40, and standard deviation equal to $\frac{1}{6}$ the length of the interval (i.e., $\frac{0.60-0.20}{6} = 0.067$) is used.

Using the *ad-hoc* function `design_est`[5] they will obtain this information :

```
> design_est(n1=31, n2=31, target_d_limits=
c(0.20,0.60), distribution="normal")
power typeS typeM
 0.35  0.00  1.73
```

To summarize, this information suggests that the results of the study of interest (see **Table 1**) should be taken very cautiously. Indeed, the expected power was low (35%), and the expected overestimation of the most plausible effect size (i.e., $d = 0.40$) was around 73%. Furthermore, it is important to note that the observed effect size of 0.90 fell abundantly outside the pre-specified plausible interval of 0.20–0.60, thus supporting the idea that the study of interest clearly overestimated the actual magnitude of the effect.

In general, when the observed effect size falls outside the pre-specified plausible interval, we can conclude that the observed study is not coherent with our theoretical expectations. On the other hand, we could also consider that our plausible interval may be unrealistic and/or poorly formalized. In these situations, researchers should be transparent and propose possible explanations that could be very helpful to the understanding of the phenomenon under study. Although this way of reasoning requires a notable effort, the information provided will lead to a more comprehensive inference than the one deriving from a simplistic dichotomous decision (i.e., "reject / do not reject") typical of the NHST approach. Indeed, in this approach the hypotheses are poorly formalized, and power, Type *M* error, and Type *S* error are not even considered.

4. AN ILLUSTRATIVE APPLICATION TO A CASE STUDY

To illustrate how design analysis could enhance inference in psychological research, we have considered a real case study. Specifically, we focused on Study 2 of the published paper "A functional basis for structure-seeking: Exposure to structure promotes willingness to engage in motivated action" (Kay et al., 2014).

The paper presented five studies arising from findings showing that human beings have a natural tendency to perceive structure in the surrounding world. Various social psychology theories propose plausible explanations that share a similar assumption that had never been tested before: that perceiving a structured world could increase people's willingness to make efforts and sacrifices toward their own goals. In Study 2, the authors decided to test this hypothesis by randomly assigning participants to two different conditions differing in the type of text they had to read. In the "random" condition, the text conveyed the idea that natural phenomena are unpredictable and random, whereas in the "structure" condition the phenomena were described as predictable and systematic. The outcome measure was the willingness to work toward a goal that each participant chose as their "most important." The expected result was that participants in the "structure" condition would report a higher score in the measure of goal-directed behavior than those in the "random" condition.

4.1. Prospective Design Analysis

As we saw in the previous paragraphs, before collecting data it is fundamental to plan an appropriate sample size via prospective design analysis. In this case, given the relative novelty of Study 2, was hard to identify a single plausible value for the size of the effect of interest. Rather, it seemed more reasonable to explore

[5] The idea behind this function is simple. First, we sample a large number (e.g., 100,000) of effect sizes *d* from the probability distribution associated with the plausible interval. Then, for each *d* we calculate power, type *M* error, and type *S* error based on the sample size of the two groups involved in the comparison, and we consider the center of the plausible interval as the most plausible effect size. In this way, a distribution for each of the three indices is finally obtained. In the output of the function, the means of the three distributions are presented as a summary value. For additional details, see **Appendix B**, which also shows (in section "design_est") how to obtain the expected distribution of power as well as Type *M* and Type *S* errors, given the plausible interval for *d*.

TABLE 2 | Sample size, Type *M* and Type *S* error by power and plausible effect size. Type I error is fixed at 0.05.

Power	Cohen's *d*	*n* (per sample)	Total *n*	Type *M* error	Type *S* error
0.80	0.20	392	784	1.13	0.00
	0.35	130	260		
	0.50	64	128		
0.60	0.20	244	488	1.30	0.00
	0.35	82	164		
	0.50	40	80		

different scenarios according to different plausible effect sizes and power levels. We started with a minimum d of 0.20, so that the study was planned to detect at least a "small" effect size. If the final results did not reach statistical significance, the researchers could conclude that it was unlikely that the true effect was equal to or >0.20, and they could eventually decide whether it would be worth it to replicate the study, perhaps by modifying their protocol.

As the most plausible effect size, we considered $d = 0.35$, which could be considered—at least in our opinion—a typical average level with which to test a hypothesis in psychological research in the absence of informative external sources (see for example the results reported in Open Science Collaboration, 2015)[6]. As extrema ratio, we included also a d of 0.5, which, in the words of Jacob Cohen, can be referred to as "differences that are large enough to be visible to the naked eye" (see Cohen 1988, p. 26 and **Appendix A**), and that, given the experiment under investigation, could be viewed as an extremely optimistic guess. Finally, to take issues concerning the feasibility of the study into account, we also considered two levels of power, namely 80 and 60%.

Overall, our "sensitivity" prospective design analysis (see **Table 2**) suggested that the sample size chosen by the authors ($n = 67$) was inadequate. Indeed, even in the least reasonable scenario ($d = 0.50$, power = 0.60), a minimum of 80 participants is required. Furthermore, is should be noted, that the associated Type M error was considerably high, i.e., 130%, signaling a high risk of overestimating the plausible effect.

A good compromise could be to consider the second scenario ($d = 0.35$, power = 0.80), which requires a total sample size of 260, guaranteeing optimal control of the Type M error. After conducting the study with this sample size, a significant result would lead to the acceptance of the researcher's hypothesis, while a non-significant result would indicate that, if an effect exists, the effect would presumably be <0.35. Whatever the result, the researchers could eventually present their findings in a transparent and informative way. In any case, the results could be used to improve scientific progress. As an example, other researchers could fruitfully use the observed results as a starting point for a replication study.

4.2. Retrospective Design Analysis

Let us now evaluate Study 2 from a retrospective point of view. Based on their results [$M_{\text{structure}} = 5.26$, $SD_{\text{structure}} = 0.88$, $M_{\text{random}} = 4.72$, $SD_{\text{random}} = 1.32$, $n_{\text{total}} = 67$; $t_{(65)} = 2.00$, $p = 0.05$, Cohen's $d = 0.50$][7], the authors concluded that "participants in the structure condition reported higher willingness to expend effort and make sacrifices to pursue their goal compared to participants in the random condition." Kay et al. (2014, p. 487), thus supporting their initial hypothesis.

To evaluate the inferential risks associated with this conclusion, we ran a sensitivity retrospective design analysis on the pre-identified plausible effect sizes (i.e., $d = 0.20$, $d = 0.35$, $d = 0.50$).

In line with the results that emerged from the prospective analysis, the retrospective design analysis indicated that the sample size used in Study 2 exhibited high inferential risks. In fact, both for a plausible effect of $d = 0.20$ (power = 0.13, type $M = 3.06$, type $S = 2\%$) and for a plausible effect of $d = 0.35$ (power = 0.29, type $M = 1.86$, type $S = 0\%$), the power was very low, and the Type M error reached worrying levels. For a d of 0.50 (chosen on the basis of plausible effects and not based on the results observed in Study 2), the Type M error was 1.40, indicating an expected overestimate of 40%. Furthermore, the power was 0.52, suggesting that if we replicated the study on a new sample with the same number of participants, the probability of obtaining a significant result would be around the chance level.

We also evaluated the results of Study 2 by performing a retrospective design analysis using the method presented in section 3. Specifically, we used a doubly truncated normal distribution centered at 0.35 (i.e., the most plausible effect size) with a plausible interval of 0.20–0.50. As could be expected, the results (i.e., power = 0.29, type $M = 1.86$, type $S = 0\%$) substantially confirmed what emerged from the sensitivity retrospective design analysis.

In summary, our retrospective design analysis indicated that, although statistically significant, the results of Study 2 were inadequate to support the authors' conclusions.

As mentioned at the beginning of this paragraph, Study 2 by Kay et al. (2014) was selected for illustrative purposes in a constructive perspective. For a more comprehensive picture, we invite interested readers to consult the "Many Labs 2 project" (Klein et al., 2018), which showed that with a large sample size ($n = 6506$) the original conclusion of Study 2 cannot be supported (i.e., $t(6498.63) = -0.94$, $p = 0.35$, $d = -0.02$,

[6] In the Open Science Collaboration (2015), the authors conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when possible. They found an average effect size of $r = 0.197$, i.e., $d = 0.41$. Given the heterogeneity of the 100 studies, we propose the use of a more conservative value to represent a typical average effect in psychology. Overall, it should be noted that all the pre-specified values of d, albeit plausible, are not based on a thorough theoretical revision and/or on the formalized knowledge of experts in the field. Indeed, an appropriate use of the latter two external sources would undoubtedly contribute to producing more reliable results, but a discussion of these strategies is beyond the scope of this paper.

[7] The authors reported only the total sample size ($n = 67$). Since participants were randomly assigned to each of the two experimental conditions, in the following we assumed, without loss of generalizability, that 34 participants were assigned to the "structure" condition and 33 to the "random" condition.

$95\%CI = [-0.07, 0.03]$, and neither can the subsequent response of the original authors (Laurin et al., 2018).

5. DISCUSSION AND CONCLUSIONS

In psychological research, statistical inference is often viewed as an isolated procedure that limits itself to the analysis of data that have already been collected. In this paper, we argue that statistical reasoning is necessary both at the planning stage and when interpreting the results of a research project. To illustrate this concept, we built on and further developed Gelman and Carlin's (2014) idea of "prospective and retrospective design analysis."

In line with recent recommendations (Cumming, 2014), design analysis involves in-depth reasoning on what could be considered as a plausible effect size within the study of interest. Specifically, rather than focusing on a single pilot or published study, we underlined the importance of using information outside the data at hand, such as extensive literature reviews and meta-analytic studies, taking issues related to publication bias into account. Furthermore, we introduced the potentials of elicitation of expert knowledge procedures (see for example Zondervan-Zwijnenburg et al., 2017; O'Hagan, 2019). Even though these procedures are still under-utilized in psychology, they could provide a relevant contribution to the formalization of research hypotheses.

Moving beyond the simplistic and often misleading distinction between significant and non-significant results, a design analysis allows researchers to quantify, consider, and explicitly communicate two relevant risks associated with their inference, namely exaggeration ratio (Type *M* error) and sign error (Type *S* error). As illustrated in the paper, the evaluation of these risks is particularly relevant in studies that investigate small effect sizes in the presence of high levels of intra- and inter-individual variability, with a limited sample size—a situation that is quite common in psychological research.

Another important aspect of design analysis is that it can be usefully carried out both in the planning phase of a study (i.e., prospective design analysis) and to evaluate studies that have already been conducted (i.e., retrospective design analysis), reminding researchers that the process of statistical inference should start before data collection and does not end when the results are obtained. In addition, design analysis contributes to have a more comprehensive and informative picture of the research findings through the exploration of different scenarios and according to different plausible formalizations of the effect of interests.

To familiarize the reader with the concept of design analysis, we included several examples as well as an application to a real case study. Furthermore, in addition to the classic formalization of the effect size with a single value, we proposed an innovative method to formalize uncertainty and previous knowledge concerning the magnitude of the effect via probability distributions within a Frequentist framework. Although not directly presented in the paper, it is important to note that this method could also be efficiently used to explore different scenarios according to different plausible probability distributions.

Finally, to allow researchers to use all the illustrated methods with their own data, we also provided two easy-to-use R functions (see also **Appendix B**), which are available at the Open Science Framework (OSF) at the link https://osf.io/j8gsf/files/.

For the sake of simplicity, in this paper we limited our consideration to Cohen's *d* as an effect size measure within a Frequentist approach. However, the concept of design analysis could be extended to more complex cases and to other statistical approaches. For example, our R functions could be directly adapted to other effect size measures, such as Hedges' *g*, Odds Ratio, η^2, and R^2. Moreover, concerning the proposed method to formalize uncertainty and prior knowledge, other probability distributions beyond those proposed in this paper (i.e., the uniform and the doubly truncated normal) could be easily added. This was one of the main reasons behind the choice to use resampling methods to estimate power as well as Type *M* and Type *S* errors instead of using an analytical approach.

Also, it is important to note that our considerations regarding design analysis could be fruitfully extended to the increasingly used Bayesian methods. Indeed, our proposed method to formalize uncertainty via probability distributions finds its natural extension in the concept of Bayesian prior. Specifically, design analysis could be useful to evaluate the properties and highlight the inferential risks (such as Type *M* and Type *S* errors) associated with the use of Bayes Factors and parameter estimation with credible Bayesian intervals.

In summary, even though a design analysis requires significant effort, we believe that it has the potential to contribute to planning more robust studies and promoting better interpretation of research findings. More generally, design analysis and its associated way of reasoning helps researchers to keep in mind the inspiring quote presented at the beginning of this paper regarding the use of statistical inference: "Remember ATOM."

DATA AVAILABILITY STATEMENT

All R scripts used to reproduce the examples presented in the paper are reported in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

GA conceived the original idea and drafted the paper. GB, CZ, and ET contributed to the development of the original idea and drafted sections of the manuscript. MP and GA wrote the R functions. GA, MP, and CZ took care of the statistical analysis and of the graphical representations. LF and AC provided the critical and useful feedback. All authors contributed to the manuscript revision, read, and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02893/full#supplementary-material

REFERENCES

Arrison, T. S. (2014). *Responsible Science: Ensuring the Integrity of the Research Process*. Technical report. Washington, DC: National Academy of Sciences.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: John Wiley & Sons.

Burkardt, J. (2014). *The Truncated Normal Distribution*. Tallahassee, FL: Department of Scientific Computing Website, Florida State University.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14, 365–376. doi: 10.1038/nrn3475

Camerer, C., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., et al. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nat. Hum. Behav.* 2, 637–644. doi: 10.1038/s41562-018-0399-z

Chambers, C. (2019). *The Seven Deadly Sins of Psychology: A Manifesto for Reforming the Culture of Scientific Practice*. Princeton, NJ: Princeton University Press.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *J. Abnorm. Soc. Psychol.* 65, 145–153. doi: 10.1037/h0045186

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Lawrence Erlbaum Associates.

Cumming, G. (2014). The new statistics: why and how. *Psychol. Sci.* 25, 7–29. doi: 10.1177/0956797613504966

Dienes, Z. (2008). *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. New York, NY: Macmillan International Higher Education.

Fisher, R. A. (1938). Presidential address, first Indian statistical congress. *Sankhyā* 4, 14–17.

Flake, J. K., and Fried, E. I. (2019). Measurement schmeasurement: questionable measurement practices and how to avoid them. *PsyArXiv [Preprint]* (2019). Available online at: https://psyarxiv.com/hs7wm (accessed August 5, 2019).

Gelman, A. (2018). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Pers. Soc. Psychol. Bull.* 44, 16–23. doi: 10.1177/0146167217729162

Gelman, A. (2019). Don't calculate *post-hoc* power using observed estimate of effect size. *Ann. Surg.* 269, e9–e10. doi: 10.1097/SLA.0000000000002908

Gelman, A., and Carlin, J. (2014). Beyond power calculations: assessing type s (sign) and type m (magnitude) errors. *Perspect. Psychol. Sci.* 9, 641–651. doi: 10.1177/1745691614551642

Gelman, A., and Tuerlinckx, F. (2000). Type S error rates for classical and bayesian single and multiple comparison procedures. *Comput. Stat.* 15, 373–390. doi: 10.1007/s001800000040

Gigerenzer, G., Krauss, S., and Vitouch, O. (2004). "The null ritual: what you always wanted to know about significance testing but were afraid to ask," in *The Sage Handbook of Quantitative Methodology for the Social Sciences*, ed D. Kaplan (Thousand Oaks, CA: Sage), 391–408.

Gigerenzer, G., and Marewski, J. N. (2015). Surrogate science: the idol of a universal method for scientific inference. *J. Manag.* 41, 421–440. doi: 10.1177/0149206314547522

Goodman, S. N., and Berlin, J. A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann. Intern. Med.* 121, 200–206. doi: 10.7326/0003-4819-121-3-199408010-00008

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med.* 2:e124. doi: 10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* 23, 524–532. doi: 10.1177/0956797611430953

Kay, A. C., Laurin, K., Fitzsimons, G. M., and Landau, M. J. (2014). A functional basis for structure-seeking: exposure to structure promotes willingness to engage in motivated action. *J. Exp. Psychol. Gen.* 143, 486–491. doi: 10.1037/a0034462

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B. Jr., Alper, S., et al. (2018). Many labs 2: investigating variation in replicability across samples and settings. *Adv. Methods Pract. Psychol. Sci.* 1, 443–490. doi: 10.1177/2515245918810225

Kruschke, J. (2013). Bayesian estimation supersedes the *t* test. *J. Exp. Psychol. Gen.* 142, 573–603. doi: 10.1037/a0029146

Kruschke, J. K., and Liddell, T. M. (2018). The bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychon. Bull. Rev.* 25, 178–206. doi: 10.3758/s13423-016-1221-4

Laurin, K., Kay, A. C., and Landau, M. J. (2018). Structure and goal pursuit: individual and cultural differences. *Adv. Methods Pract. Psychol. Sci.* 1, 491–494. doi: 10.1177/2515245918797130

Loken, E., and Gelman, A. (2017). Measurement error and the replication crisis. *Science* 355, 584–585. doi: 10.1126/science.aal3618

Lu, J., Qiu, Y., and Deng, A. (2019). A note on type s/m errors in hypothesis testing. *Br. J. Math. Stat. Psychol.* 72, 1–17. doi: 10.1111/bmsp.12132

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychol. Methods* 9, 147–163. doi: 10.1037/1082-989X.9.2.147

McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2019). Abandon statistical significance. *Am. Stat.* 73, 235–245. doi: 10.1080/00031305.2018.1527253

Mersmann, O., Trautmann, H., Steuer, D., and Bornkamp, B. (2018). truncnorm: truncated normal distribution, R package version 1.0-8. Available online at: https://cran.r-project.org/package=truncnorm

Moore, D. S. (1998). Statistics among the liberal arts. *J. Am. Stat. Assoc.* 93, 1253–1259. doi: 10.1080/01621459.1998.10473786

O'Hagan, A. (2019). Expert knowledge elicitation: subjective but scientific. *Am. Stat.* 73, 69–81. doi: 10.1080/00031305.2018.1518265

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349:aac4716. doi: 10.1126/science.aac4716

Pashler, H., and Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? *Perspect. Psychol. Sci.* 7, 528–530. doi: 10.1177/1745691612465253

Pastore, M., Lionetti, F., Calcagnì, A., and Altoè, G. (2019). La potenza è nulla senza controllo—Power is nothing without control. *Giorn. Ital. Psicol.* 46, 359–378. doi: 10.1421/93796

Pearson, J., and Neyman, E. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: part I. *Biometr. A* 20A, 175–240. doi: 10.1093/biomet/20A.1-2.175

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ruscio, J. (2008). A probability-based measure of effect size: robustness to base rates and other factors. *Psychol. Methods* 13, 19–30. doi: 10.1037/1082-989X.13.1.19

Schooler, J. (2014). Turning the lens of science on itself: verbal overshadowing, replication, and metascience. *Perspect. Psychol. Sci.* 9, 579–584. doi: 10.1177/1745691614547878

Sedlmeier, P., and Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychol. Bull.* 105, 309–316. doi: 10.1037//0033-2909.105.2.309

Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019). Moving to a world beyond "p<0.05". *Am. Stat.* 73, 1–19. doi: 10.1080/00031305.2019.1583913

Zandonella Callegher, C., Toffalini, E., and Altoè, G. (2019). Eliciting effect size - Shiny App (version 687 v1.0.0). Available online at: https://zenodo.org/record/2564852#.Xfz2out7nwc

Zondervan-Zwijnenburg, M., van de Schoot-Hubeek, W., Lek, K., Hoijtink, H., and van de Schoot, R. (2017). Application and evaluation of an expert judgment elicitation procedure for correlations. *Front. Psychol.* 8:90. doi: 10.3389/fpsyg.2017.00090

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

BRIEF RESEARCH REPORT
published: 22 January 2020
doi: 10.3389/fpsyg.2019.03090

An Item-Level Analysis for Detecting Faking on Personality Tests: Appropriateness of Ideal Point Item Response Theory Models

Jie Liu[1,2] *and Jinfu Zhang*[2]*

[1] *School of Mathematics and Statistics, Southwest University, Chongqing, China,* [2] *Faculty of Psychology, Southwest University, Chongqing, China*

How to detect faking on personality measures has been investigated using various methods and procedures. As previous findings are mixed and rarely based on ideal point item response theory models, additional research is needed for further exploration. This study modeled the responses of personality tests using ideal point method across instructed faking and honest responding conditions. A sample of undergraduate students participated the within-subjects measures to examine how the item location parameter derived from the generalized graded unfolding model changed, and how individuals' perception about items changed when faked. The mean test scores of faking group was positively correlated to the magnitude of within-subjects score change. The item-level analysis revealed both conscientiousness items (18.8%) and neuroticism items (50.0%) appeared significant shifts on item parameters, suggesting that response pattern changed from honest to faking conditions. The direction of the change appeared both in positive and negative way, demonstrating that faking could increase or decrease personality factor scores. The results indicated that the changes of perceptions on items could be operated by faking, offering some support for the ideal point model to be an adequate measure for detecting faking. However, the findings of diagnostic accuracy analysis also implied that the appropriateness of ideal point models for detecting faking should be under consideration, also be used with caution. Implications, further research directions, and limitations are discussed.

Keywords: item parameter, ideal point model, faking detection, item response theory, personality tests, appropriate measurement

Edited by:
Jason C. Immekus,
University of Louisville, United States

Reviewed by:
Avi Allalouf,
National Institute for Testing and Evaluation (NITE), Israel
Merylin Monaro,
University of Padua, Italy
Cornelius J. König,
Saarland University, Germany

****Correspondence:***
Jinfu Zhang
zhangjf@swu.edu.cn

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *10 September 2019*
Accepted: *31 December 2019*
Published: *22 January 2020*

Citation:
Liu J and Zhang J (2020) An Item-Level Analysis for Detecting Faking on Personality Tests: Appropriateness of Ideal Point Item Response Theory Models. Front. Psychol. 10:3090. doi: 10.3389/fpsyg.2019.03090

INTRODUCTION

For many years, faking on personality measures has been perceived as a response distortion or intentional dissimulation. From theoretical perspective, it is well known that the measurement validity of the tests would be significantly affected due to faking, which can negatively impact the quality of the potential personality measures (Topping and O'Gorman, 1997; Stark et al., 2001; Pauls and Crost, 2004, 2005; Holden, 2008; Komar et al., 2008; Buehl et al., 2019). In practical contexts, the typical case is that the candidates who want to improve their chance to be accepted to a job are more likely to fake, even if without any help, they still try to find a way to bring the answers closer to the expectations of the organizations. However, the decision is therefore effected when substantial proportions of applicants would be incorrectly admitted as increasing the likelihood that an organization would hire the fakers (Rosse et al., 1998; Donovan et al., 2014; Niessen et al., 2017). Additionally, even non-real-life-applicants under experimental conditions also can fake when instructed to do so (Thumin and Barclay, 1993; Dalen et al., 2001; Mueller-Hanson et al., 2003; Nguyen et al., 2005; Griffith et al., 2007; Day and Carroll, 2008; Berry and Sackett, 2009; Buehl et al., 2019). Thus, there has been a considerable research interest focused on detecting faking using various methods and procedures.

Many methodologies and techniques have been developed for detecting response distortion over the years, for example, machine learning models, reaction times, regression analysis, etc. (Dunn et al., 1972; Sellbom and Bagby, 2010; Jiménez Gómez et al., 2013; Monaro et al., 2018; Roma et al., 2018; Mazza et al., 2019). Still, there is a concern about the perceptions and interpretations of the change on items due to intentional dissimulation. From an item-level perspective, the changing-item paradigm (Zickar and Robie, 1999) posits that not the standing on the latent trait changes when individuals fake, but the item locations on the continuum that change. In other words, when response distortion occurs, the individuals' level of the latent trait is fixed without the impact of faking, but the items will be positioned a higher or lower standing on the latent continuum than what is actually possessed. In this case, when the difference of item locations between faking situation and honest situation is captured (i.e., assessed at the item level), the fakability would be identified.

The research following the changing-item paradigm has often employed differential item functioning (DIF) techniques to address changes over items. As item response theory (IRT) provides a formal statistical model for the relationship between the item response and the latent characteristic, IRT-based DIF is deservedly appropriate for modeling the change of item locations over different responding conditions (Zickar and Robie, 1999; Stark et al., 2001). To describe how people respond to personality measures, the ideal point response process assumes that individuals will have a higher probability to endorse an item that is closer to their "true" latent levels (Roberts, 1996; Roberts and Laughlin, 1996). Specifically, an item response function (IRF) is shown in **Figure 1** (Stark et al., 2006). For example, on a measure of conscientiousness (i.e., θ), the agreement probability (i.e., vertical axis) on a statement will be the highest when the item locates nearest the true level of conscientiousness (i.e., horizontal axis). When the distance between conscientiousness level and item location increases, an individual will less likely endorse the item. The generalized graded unfolding model (GGUM) is used as the ideal point model in past years (Roberts and Laughlin, 1996; Roberts et al., 2000). There has already been many previous research that identified advantages of the GGUM in working with personality and attitude data, including the use of understanding faking (Stark et al., 2006; Chernyshenko et al., 2007; Weekers and Meijer, 2008; Tay et al., 2009; Carter and Dalal, 2010; O'Brien and LaHuis, 2011; Speer et al., 2016).

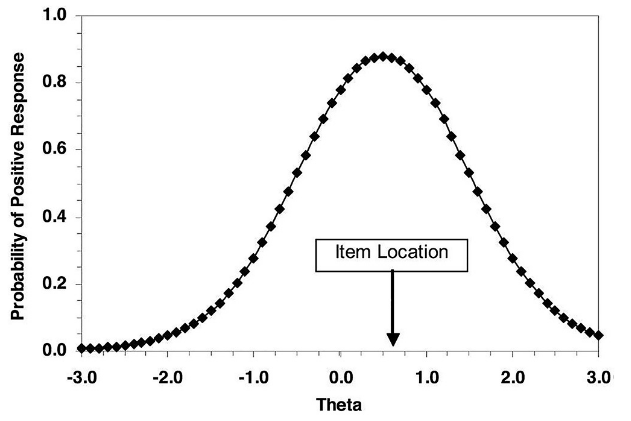


FIGURE 1 | Example of item response function for an ideal point response process.

In this study, we performed an item-level analysis to investigate the valence of ideal point IRT models that focus on how perceptions of personality items change when individuals are responding honestly or faking. The within-subjects design was employed to form the comparison groups, under which participants completed both conscientiousness and neuroticism scales. In summary, it can be expected that there is an overall tendency to response distortion that is reflected in different conditions of responding. The hypothesis concerns that different groups of subjects differ in their pattern of selecting options regarding to instructed faking and honestly responding sessions. It is hypothesized that not only the change of test scores can be significantly identified with faking condition, but also the item locations would shift with a dishonest response pattern and the shifts can be examined. Finally, whether the GGUM is adequate for detecting faking needs to be under consideration with caution.

METHODS

Participants

Respondents consisted of 568 undergraduate students from four Chinese colleges. They volunteered for the study and received extra credit in exchange for their participation. Approximately 78.4% of the participants were female, the average age was 19.84 years (SD = 1.11 years), and non-psychology students. In total, 499 valid cases remained in conscientiousness factor, 547 remained in neuroticism factor. The subjects were excluded from data analysis for two reasons: (a) only one or two response options were selected for all the items (i.e., straight-column answers), and (b) pairwise deleted the data that without an identifying number.

Design

The response instructions were the within-subjects factor in both experimental sessions. At Time 1, about half of the sample was randomly assigned to respond to the questionnaires honestly, while the other half was assigned to complete the questionnaires with fake instructions. At Time 2, respondents received the opposite set of instructions.

Procedure

The study was approved by the Institutional Review Board of the Southwest University of China. All participants provided written informed consent after being fully informed of the research procedure.

The questionnaires were administered in paper-and-pencil version in classrooms. The instructions for the honest condition were as follows:

> *Please complete this personality inventory as honestly as you can. There are no good or bad answers to the items. It is very important that you respond to this survey by describing yourself as you really are and not as you want to be or as you want others to see you.*

The instructions for the faking-good condition were as follows:

> *Imagine that you are applying for a job you really want. Please complete this personality inventory to increase your chances of being hired. To try to give a good impression to the organization, you should present yourselves as the candidates think the organization would like, regardless of your truthful opinions.*

After a retest interval of 3 weeks, the second session was the same as the first one except that participants received the other set of response instructions.

Measures

The International Personality Item Pool (IPIP) is a public-domain measure of the Five-factor model of personality. The IPIP conscientiousness and neuroticism factors are two core personality characteristics that more likely susceptive related to faking (Topping and O'Gorman, 1997; McFarland and Ryan, 2000; Mueller-Hanson et al., 2006; Komar et al., 2008). In this study, the two factors were measured by 20 items from the IPIP, respectively (40 total items). Thus the Conscientiousness Scale and Neuroticism Scale were constructed for measuring the extent to which each item described the respondent on a five-point rating scale ranging from 0 (very inaccurate) to 4 (very accurate). Each scale consists of 10 items that are reverse-coded, and higher composite scores indicate higher levels of traits. The forward–backward procedure was applied to translate the scales from English to Chinese. Participants completed the final Chinese version of the two scales.

Analyses

Firstly, to examine the veracity of the unidimensional data assumption, a parallel analysis and the matrix of polychoric correlations were performed separately for each response condition on conscientiousness and neuroticism factors. Then, the chi-square test (Drasgow et al., 1995), with the MODFIT program (Stark, 2001) was employed separately for each response condition on both personality factors to examine the fit of the GGUM to the data.

Secondly, the GGUM2004 program (Roberts et al., 2006) was used to obtain the item and person parameters derived from the marginal maximum likelihood estimation method and the expected *a posteriori* estimation method, respectively. Then the GGUMLINK program (Roberts and Huang, 2003) was performed for equating the parameter estimates by transforming the metric of the fake condition group to the same metric of the honest condition group.

Finally, to examine the impact of response distortion on each item, a statistical comparison based on (Scherbaum et al., 2013)' study was conducted between the GGUM parameter estimates obtained separately under honest and faking conditions. Then we used receiver operating characteristic (ROC) curves analyses to evaluate the diagnostic accuracy of model estimates in detecting faking-induced change[1].

RESULTS

Descriptive Statistics

Descriptive statistics of the row scores of two personality scales in each condition are presented in **Table 1**. The amount of faking refers as within-subjects change in row scores between two experimental sessions. The intraclass correlation coefficient of the 3-week test–retest was 0.74 (0.70–0.79) for the conscientiousness scale and 0.75 (0.70–0.79) for the neuroticism scale. Under the fake response condition, we observed significant higher scores on conscientiousness ($t(498) = 5.85$, $p < 0.05$, $d = 0.24$), and significant lower scores on neuroticism ($t(546) = -3.36$, $p < 0.05$, $d = -0.13$), compared to the honest response condition, indicating that the faking manipulation was effective. The order effects of response instructions was not statistically significant for conscientiousness ($t(497) = 0.04$, $p > 0.05$, $d = 0.04$), or neuroticism ($t(545) = 0.72$, $p > 0.05$, $d = 0.06$).

Correlation Between Faking Scores and Score Changes

According to the results of the correlation matrix (see **Supplementary Table 1** in **Supplementary Material**), scores of personality factors in faking condition were significantly correlated with the magnitude of score change from the faking to honest context, but with moderate correlation coefficients. For conscientiousness, $r = 0.50$ (0.43–0.56, $p < 0.05$), and for neuroticism, $r = 0.46$ (0.41–0.52, $p < 0.05$). This finding suggests that the overall tendency of the change for score elevation is consistent with the test scores related to faking condition, supporting the hypothesis regarding the tendency.

Test of GGUM Assumptions and Model Fit

One of the assumptions of GGUM is to model data that obtained in unidimensionality personality tests (Roberts et al., 2000). The results of parallel analysis and polychoric correlation coefficients demonstrated that both the conscientiousness and neuroticism data met this assumption. As presented in **Table 2**, the results of GGUM model fit were reasonably good, except for several items. Hence these four items ("Am always prepared"; "Get chores

[1] We would like to thank the reviewers for raising this suggestion.

TABLE 1 | Descriptive statistics and reliability of study measures.

	Honest			Faking			Amount of faking			*t*	*d*
	M	**SD**	**α**	**M**	**SD**	**α**	**M**	**SD**	**95%CI**		
C	2.33	0.48	0.85	2.45	0.51	0.87	0.12	0.44	0.08–0.16	5.85***	0.24
N	1.71	0.49	0.83	1.65	0.50	0.83	−0.06	0.44	−0.10–0.03	−3.36**	−0.13

*C = conscientiousness; N = neuroticism; amount of faking = change in scores calculated as fake response scores minus honest response scores of pairwise data; α = Cronbach's α coefficient; 95% CI = 95% confidence interval; t = result of t-test comparing mean scores in faking and honest response conditions; d = Cohen's d, computed using the standard formula of the difference between the means of faking and honest scores divided by the pooled standard deviation. **p < 0.01, ***p < 0.001.*

TABLE 2 | Model fit results of GGUM by scales and conditions.

Measure	**Honest**			**Faking**		
	Number of items with χ^2/df < 3	***M***	**SD**	**Number of items with χ^2/df < 3**	***M***	**SD**
C	17	0.06	0.03	18	0.17	0.15
N	20	0.09	0.13	19	0.07	0.09

C = conscientiousness; N = neuroticism.

done right away"; "Do just enough wore to get by"; "Do things according to a plan") in the Conscientiousness scale under both two conditions were pair-wised excluded from the subsequent analyses for the reliable veracity of model assumptions, as well as a neuroticism item ("Feel comfortable with myself") under the faking condition, although most IRT estimation procedures are generally tolerant of slight to moderate violations of the unidimensionality assumption (Hulin et al., 1983).

Model Parameter Estimates and Shifts in Item Parameter

The item location parameters (i.e., δ) were estimated from GGUM to indicate the location of each item on the latent trait continuum. All of the δ values were positive, as the negatively worded items were recoded and rescored in the positive direction. A test was conducted to identify the differences between the location parameters from the two response groups in order to estimate the shifts. As the differences between item parameters from an IRT model can be considered an effect size (Steinberg and Thissen, 2006), the effect size indicator (i.e., *d*) in this case was the one-to-one difference of the δ (**Table 3**).

From the table, nearly 20% of conscientiousness items and over 50% of neuroticism items demonstrated statistically significant shifts in the item location parameter. These significant changes occurred in opposite directions in the two personality factors. As the δ is also helpful to index a respondent's θ level above or below the item location, and the distance between the location of the person and the item, with regard to positive shifts, individuals who were actually at lower levels of this trait tended to select higher order options and appeared as if they were really located on the positive side of the latent trait continuum. Correspondingly, the implication for negative shifts indicated that individuals with high levels of this factor were not likely to select a higher order option and appear as if they were lower on the trait than they really were. These findings supported the hypothesis that the item location could be changed due to response pattern changed and the changes could be modeled using an ideal point IRT model.

ROC Analyses for Diagnostic Accuracy

Receiver operating characteristic analyses evaluated the shifts of item location parameter for detecting faking-good items versus honest items (see **Supplementary Table 2** in **Supplementary Material**). The area under the curve (AUC) of ROC were 0.74 (SE = 0.12) and 0.64 (SE = 0.13) for conscientiousness factor and neuroticism factor, respectively. Although these AUCs indicated moderate diagnostic accuracy, they are evaluated without statistical significance ($p > 0.05$), suggesting that the effectiveness of the item parameter shifts for examining the faking-induced change of item response pattern was not powerful enough.

DISCUSSION

The current study used an ideal point IRT model to identify dishonest responses at the item level. We found that the magnitude of score change was positively correlated to the test scores of motived faking group. Parts of the item location parameters derived from the GGUM showed statistically significant shifts across honest and faking conditions in the within-subjects' response pattern, which indicates that, to some extent, the shifts of item parameters play the role as indicators of faking. Moreover, the accuracy of the indicators was moderately weak for evidencing the appropriateness of ideal point IRT models that used for detect faking.

It was noteworthy that the deltas significantly differed in two response conditions for some items. This demonstrates that operating the response instructions could lead to changes of item positions on the latent trait continuum, and the ideal point IRT model might provide some insight into how faking impacts individuals' perception of personality items. Specifically, almost all conscientiousness items experienced positive shifts. In this

TABLE 3 | Item parameters for conditions and shifts for each item.

Measure	Item	δ		*t*	*d*
		Honest	Faking		
C	Waste my time	4.61	6.50	0.19	1.89
C	Pay attention to details	4.66	5.78	0.13	1.12
C	Find it difficult to get down to work	4.06	2.69	−0.22	−1.37
C	Carry out my plans	3.79	4.73	0.30	0.94
C	Do not see things through	4.37	4.98	0.08	0.61
C	Make plans and stick to them	3.71	5.19	0.66	1.48
C	Shirk my duties	3.81	4.36	0.11	0.55
C	Complete tasks successfully	4.30	5.35	0.15	1.05
C	Mess things up	2.52	4.13	5.03***	1.61
C	Leave things unfinished	3.86	2.78	−0.22	−1.08
C	Am exacting in my work	2.70	4.28	2.87**	1.58
C	Don't put my mind on the task at hand	2.67	5.21	7.26***	2.54
C	Finish what I start	4.06	4.35	0.05	0.29
C	Make a mess of things	3.46	4.58	0.33	1.12
C	Follow through with my plans	3.96	3.94	0.00	−0.02
C	Need a push to get started	4.37	5.29	0.12	0.92
N	Often feel blue	0.98	0.44	9.00***	−0.54
N	Seldom feel blue	0.82	0.59	−2.56*	−0.23
N	Dislike myself	0.68	0.37	−4.43***	−0.31
N	Am often down in the dumps	1.05	2.55	25.00***	1.50
N	Rarely get irritated	0.21	0.47	2.60**	0.26
N	Have frequent mood swings	0.93	1.30	3.70***	0.37
N	Am not easily bothered by things	0.49	0.63	1.40	0.14
N	Panic easily	0.88	0.86	−0.33	−0.02
N	Am very pleased with myself	0.47	0.34	−1.44	−0.13
N	Am filled with doubts about things	1.31	0.66	−3.82***	−0.65
N	Am relaxed most of the time	0.65	0.30	−3.50***	−0.35
N	Feel threatened easily	1.05	0.99	−0.60	−0.06
N	Seldom get mad	0.41	0.49	1.00	0.08
N	Get stressed out easily	1.01	1.14	0.68	0.13
N	Am not easily frustrated	0.56	0.18	−4.22***	−0.38
N	Fear for the worst	1.19	0.92	−1.08	−0.27
N	Remain calm under pressure	0.47	4.06	29.92***	3.59
N	Worry about things	1.11	1.07	−0.33	−0.04
N	Rarely lose my composure	0.29	0.08	−1.75	−0.21

*C = conscientiousness; N = neuroticism; δ = item location parameter; t = test statistic of the difference between the δ parameters under faking and honest conditions divided by the standard error of the parameter estimates; d = the indictor of effect size, calculated as faking δ values minus honest δ values of pairwise data. *p < 0.05, **p < 0.01, ***p < 0.001.*

case, individuals with lower levels of the personality characteristic were likely to endorse higher-order options and appear to be higher on the factor than they really were. All the items with significant shifts on the conscientiousness factor showed the same pattern. On the other hand, however, not all the significant neuroticism items followed the same pattern in the direction of the shifts (i.e., negative shifts). The significant reverse shifts demonstrate that the response patterns are complex and sensitive to the characteristic assessed by an item even if such characteristic is not seen as a desirable behavior in the faking condition.

We also found that the magnitude of the shifts was large for many conscientiousness items, whereas it was universally small for neuroticism items. Given that the one-to-one difference of deltas is regarded as an effect size, these values can demonstrate how far apart the item parameters are on the distribution of standardized latent trait. It could be the case that neuroticism is generally not seen as a desirable characteristic and therefore there might not be a uniform perception about these items when respondents fake, so that the direction of distortion varied to generate smaller value of effect size. In addition, the items might show fake in both sides of directions (i.e., positive or negative), which results in counteractions between possible shifts thus less significant shifts in item parameter, and negative impact on accuracy of the IRT-based procedure.

Implications

Ideal point IRT models (e.g., the GGUM used here) provide an effective means to extend the research on response distortion at the item level. These procedures could quantitatively model the impact of response behavior on personality items and therefore detect the change of response patterns under different response conditions. Positive shifts suggested that the item location on the continuum was higher in the faking condition, whereas negative values indicated that the δ parameter was lower in the faking condition. These findings are consist with the hypothesis that concerning different groups of subjects differ in their pattern of selecting options with respect to different experimental sessions. Not only the change of test scores is significantly identified with instructed faking, but also the item locations shift with a dishonest response pattern and consequently the shifts are examined via an IRT model.

Given that the diagnostic accuracy had appeared unexpected results, the valence of IRT item-analysis might be considered with the issues of appropriateness for ideal point models. It is suggested that if responders compare their self-perception to a certain threshold rather than to the statement's location, when responding to items, ideal point models should not be used (Brown and Maydeu-Olivares, 2010). Second, focus on the precision of item estimates, it is inherently more difficult to recover true item parameters for ideal point models with the normal probability density function model, if comparing with that for dominance models which derive item estimates with the normal ogive model (Brown and Maydeu-Olivares, 2010). Considering GGUM's mathematical complexity for estimation difficulties, some studies related to detect faking used other methods, for example, techniques based on reaction times, and scored invalidity scales (Sellbom and Bagby, 2010; Monaro et al., 2018; Roma et al., 2018; Mazza et al., 2019), generally obtained superior accurate outcomes. Finally, practically speaking, the use of ideal point models seems not to result in any improvement for predictive validity, if comparing with dominance models (Zhang et al., 2019). Hence there are still some issues with ideal point models when used for modeling faking response data.

The results of the present study also point to some areas for further research. Firstly, we need to better understand the various direction of the parameter shifts on personality factors. Although the shifts showed a pattern similar to that found in previous research, there is no readily unambiguous explanation for the opposite direction to that being hypothesized. Then, as (Ferrando and Anguiano-Carrasco, 2013) noted, the effectiveness of mixed procedures is higher than that of previous single procedure. The research on faking could benefit from traditional IRT models combined with other recent model-based approaches such as multilevel IRT analysis or mixture IRT models as a starting point.

Limitations

One potential limitation of this study is the insufficient proportion of double-barreled items and vague quantifiers. If only extreme items are used, dominance and ideal point models will more likely yield a similar fit with nearly monotonical IRFs of personality items (Drasgow et al., 2010). In this case, intermediate statements should be used more frequently for larger effect sizes thereby allowing the researchers to accurately identify an item's position on the latent continuum underlying faking.

We see an additional limitation regarding the measures of consequent outcomes for the validity of studies under simulated applicant-situations. Generally, these following criterion measures on scales or work performance in real-life context will more accurately predict or estimate the number or percent of the "benefited" items and responders due to faking behavior. It may well be that it provides an available way to examine the internal accuracy and external generalizability.

Conclusion

Taken together, we find that the test scores in faking condition corresponded with the amount of faking, moreover, the ideal point IRT models in some cases to be an adequate measure for detecting faking at the item level. The shifts of item location parameters offer direct support for the change of individuals' response pattern due to motivated faking. However, the diagnostic accuracy of the detection is not such ideal so that the usage of ideal point models should be approached with caution. On the whole, this study presents a possible useful method that is worth further investigation.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

ETHICS STATEMENT

This studies involving human participants were reviewed and approved by the Institutional Review Board of the Southwest University of China. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

Both authors contributed to the conception, design of the study, revised the manuscript, read, and approved the submitted version. JZ organized the experimental sessions and led the data collection. JL performed the data analysis and wrote the original draft of the manuscript.

ACKNOWLEDGMENTS

The authors thank Yanqin Li, Wei Chen, Chenghai Yan, and Yongping Zhao for their assistance with data collection, Stephen Stark for providing support with the computer software, and Wenjiu Du for help with modeling processing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.03090/full#supplementary-material

REFERENCES

Berry, C. M., and Sackett, P. R. (2009). Faking in personnel selection: tradeoffs in performance versus fairness resulting from two cut-score strategies. *Person. Psychol.* 62, 833–863. doi: 10.1111/j.1744-6570.2009.01159.x

Brown, A., and Maydeu-Olivares, A. (2010). Issues that should not be overlooked in the dominance versus ideal point controversy. *Indus. Org. Psychol.* 3, 489–493. doi: 10.1111/j.1754-9434.2010.01277.x

Buehl, A.-K., Melchers, K. G., Macan, T., and Kühnel, J. (2019). Tell me sweet little lies: how does faking in interviews affect interview scores and interview validity? *J. Bus. Psychol.* 34, 107–124. doi: 10.1007/s10869-018-9531-3

Carter, N. T., and Dalal, D. K. (2010). An ideal point account of the JDI Work satisfaction scale. *Pers. Individ. Differ.* 49, 743–748. doi: 10.1016/j.paid.2010.06.019

Chernyshenko, O. S., Stark, S., Drasgow, F., and Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: toward increasing the flexibility of personality measures. *Psychol. Assess.* 19, 88–106. doi: 10.1037/1040-3590.19.1.88

Dalen, L. H., Stanton, N. A., and Roberts, A. D. (2001). Faking personality questionnaires in personnel selection. *J. Manag. Dev.* 20, 729–742. doi: 10.1108/02621710110401428

Day, A. L., and Carroll, S. A. (2008). Faking emotional intelligence (EI): comparing response distortion on ability and trait-based EI measures. *J. Org. Behav.* 29, 761–784. doi: 10.1002/job.485

Donovan, J. J., Dwight, S. A., and Schneider, D. (2014). The impact of applicant faking on selection measures, hiring decisions, and employee performance. *J. Bus. Psychol.* 29, 479–493. doi: 10.1007/s10869-013-9318-5

Drasgow, F., Chernyshenko, O. S., and Stark, S. (2010). 75 years after Likert: thurstone was right! *Indus. Org. Psychol.* 3, 465–476. doi: 10.1111/j.1754-9434.2010.01273.x

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., and Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Appl. Psychol. Measure.* 19, 143–165. doi: 10.1177/014662169501900203

Dunn, T. G., Lushene, R. E., and O'Neil, H. F. (1972). Complete automation of the MMPI and a study of its response latencies. *J. Consult. Clin. Psychol.* 39, 381–387. doi: 10.1037/h0033855

Ferrando, P. J., and Anguiano-Carrasco, C. (2013). A structural model–based optimal person-fit procedure for identifying faking. *Educ. Psychol. Measure.* 73, 173–190. doi: 10.1177/0013164412460049

Griffith, R. L., Chmielowski, T., and Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Person. Rev.* 36, 341–355. doi: 10.1108/00483480710731310

Holden, R. R. (2008). Underestimating the effects of faking on the validity of self-report personality scales. *Pers. Individ. Diff.* 44, 311–321. doi: 10.1016/j.paid.2007.08.012

Hulin, C. L., Drasgow, F., and Parsons, C. K. (1983). *Item Response Theory: Application to Psychological Measurement*. Homewood, IL: Dorsey Press.

Jiménez Gómez, F., Sánchez Crespo, G., and Ampudia Rueda, A. (2013). Is there a social desirability scale in the MMPI-2-RF? *Clín. Salud* 24, 161–168. doi: 10.5093/cl2013a17

Komar, S., Brown, D. J., Komar, J. A., and Robie, C. (2008). Faking and the validity of conscientiousness: a Monte Carlo investigation. *J. Appl. Psychol.* 93, 140–154. doi: 10.1037/0021-9010.93.1.140

Mazza, C., Monaro, M., Orrù, G., Burla, F., Colasanti, M., Ferracuti, S., et al. (2019). Introducing machine learning to detect personality faking-good in a male sample: a new model based on Minnesota multiphasic personality inventory-2 restructured form scales and reaction times. *Front. Psychiatry* 10:389. doi: 10.3389/fpsyt.2019.00389

McFarland, L. A., and Ryan, A. M. (2000). Variance in faking across noncognitive measures. *J. Appl. Psychol.* 85, 812–821. doi: 10.1037/0021-9010.85.5.812

Monaro, M., Toncini, A., Ferracuti, S., Tessari, G., Vaccaro, M. G., De Fazio, P., et al. (2018). The detection of malingering: a new tool to identify made-up depression. *Front. Psychiatry* 9:249. doi: 10.3389/fpsyt.2018.00249

Mueller-Hanson, R., Heggestad, E. D., and Thornton, G. C. (2003). Faking and selection: considering the use of personality from select-in and select-out perspectives. *J. Appl. Psychol.* 88, 348–355. doi: 10.1037/0021-9010.88.2.348

Mueller-Hanson, R. A., Heggestad, E. D., and Thornton, G. C. (2006). Individual differences in impression management: an exploration of the psychological processes underlying faking. *Psychol. Sci.* 48, 288–312.

Nguyen, N. T., Biderman, M. D., and McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment test. *Int. J. Select. Assess.* 13, 250–260. doi: 10.1111/j.1468-2389.2005.00322.x

Niessen, A. S. M., Meijer, R. R., and Tendeiro, J. N. (2017). Measuring non-cognitive predictors in high-stakes contexts: the effect of self-presentation on self-report instruments used in admission to higher education. *Pers. Individ. Diff.* 106, 183–189. doi: 10.1016/j.paid.2016.11.014

O'Brien, E., and LaHuis, D. M. (2011). Do applicants and incumbents respond to personality items similarly? A comparison of dominance and ideal point response models. *Int. J. Select. Assess.* 19, 109–118. doi: 10.1111/j.1468-2389.2011.00539.x

Pauls, C. A., and Crost, N. W. (2004). Effects of faking on self-deception and impression management scales. *Pers. Individ. Diff.* 37, 1137–1151. doi: 10.1016/j.paid.2003.11.018

Pauls, C. A., and Crost, N. W. (2005). Effects of different instructional sets on the construct validity of the NEO-PI-R. *Pers. Individ. Diff.* 39, 297–308. doi: 10.1016/j.paid.2005.01.003

Roberts, J. S. (1996). *Item Response Theory Approaches to Attitude Measurement*. Dissertation, University of South Carolina, Columbia, SC.

Roberts, J. S., Donoghue, J. R., and Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Appl. Psychol. Measure.* 24, 3–32. doi: 10.1177/01466216000241001

Roberts, J. S., Fang, H. R., Cui, W. W., and Wang, Y. J. (2006). GGUM2004: a windows-based program to estimate parameters in the generalized graded unfolding model. *Appl. Psychol. Measure.* 30, 64–65. doi: 10.1177/0146621605280141

Roberts, J. S., and Huang, C. W. (2003). GGUMLINK: a computer program to link parameter estimates of the generalized graded unfolding model from item response theory. *Behav. Res. Methods Instrument. Comput.* 35, 525–536. doi: 10.3758/bf03195532

Roberts, J. S., and Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Appl. Psychol. Measure.* 20, 231–255. doi: 10.1177/014662169602000305

Roma, P., Verrocchio, M. C., Mazza, C., Marchetti, D., Burla, F., Cinti, M. E., et al. (2018). Could time detect a faking-good attitude? a study with the MMPI-2-RF. *Front. Psychol.* 9:1064. doi: 10.3389/fpsyg.2018.01064

Rosse, J. G., Stecher, M. D., and Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *J. Appl. Psychol.* 83, 634–644. doi: 10.1037/0021-9010.83.4.634

Scherbaum, C. A., Sabet, J., Kern, M. J., and Agnello, P. (2013). Examining faking on personality inventories using unfolding item response theory models. *J. Pers. Assess.* 95, 207–216. doi: 10.1080/00223891.2012.725439

Sellbom, M., and Bagby, R. M. (2010). Detection of overreported psychopathology with the MMPI-2 RF form validity scale. *Psychol. Assess.* 22, 757–767. doi: 10.1037/a0020825

Speer, A. B., Robie, C., and Christiansen, N. D. (2016). Effects of item type and estimation method on the accuracy of estimated personality trait scores: polytomous item response theory models versus summated scoring. *Pers. Individ. Diff.* 102, 41–45. doi: 10.1016/j.paid.2016.06.058

Stark, S. (2001). *MODFIT: A Computer Program for Model-Data Fit. [Software]*. Available at: http://cehs.unl.edu/edpsych/software-urls-and-other-interesting-sites/ (accessed June 14, 2017).

Stark, S., Chernyshenko, O. S., Chan, K., Lee, W., and Drasgow, F. (2001). Effects of the testing situation on item responding: cause for concern. *J. Appl. Psychol.* 86, 943–953. doi: 10.1037/0021-9010.86.5.943

Stark, S., Chernyshenko, O. S., Drasgow, F., and Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: should ideal point methods be considered for scale development and scoring? *J. Appl. Psychol.* 91, 25–39. doi: 10.1037/0021-9010.91.1.25

Steinberg, L., and Thissen, D. (2006). Using effect sizes for research reporting: examples using item response theory to analyze differential item functioning. *Psychol. Methods* 11, 402–415. doi: 10.1037/1082-989X.11.4.402

Tay, L., Drasgow, F., Rounds, J., and Williams, B. A. (2009). Fitting measurement models to vocational interest data: are dominance models ideal? *J. Appl. Psychol.* 94, 1287–1304. doi: 10.1037/a0015899

Thumin, F. J., and Barclay, A. G. (1993). Faking behavior and gender differences on a new personality research instrument. *Consult. Psychol. J.* 45, 11–22. doi: 10.1037/1061-4087.45.4.11

Topping, G. D., and O'Gorman, J. G. (1997). Effects of faking set on validity of the NEO-FFI. *Pers. Individ. Diff.* 23, 117–124. doi: 10.1016/S0191-8869(97)00006-8

Weekers, A. M., and Meijer, R. R. (2008). Scaling response processes on personality items using unfolding and dominance models. *Eur. J. Psychol. Assess.* 24, 65–77. doi: 10.1027/1015-5759.24.1.65

Zhang, B., Cao, M., Tay, L., Luo, J., and Drasgow, F. (2019). Examining the item response process to personality measures in high-stakes situations: issues of measurement validity and predictive validity. *Person. Psychol.* 1–28. doi: 10.1111/peps.12353 (in press).

Zickar, M. J., and Robie, C. (1999). Modeling faking good on personality items: an item-level analysis. *J. Appl. Psychol.* 84, 551–563. doi: 10.1037/0021-9010.84.4.551

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 22 January 2020
doi: 10.3389/fpsyg.2019.03036

Validation of Subjective Well-Being Measures Using Item Response Theory

Ali Al Nima[1,2]*, *Kevin M. Cloninger*[1,3], *Björn N. Persson*[1,4], *Sverker Sikström*[5] and *Danilo Garcia*[1,2,6]*

[1] Blekinge Center of Competence, Region Blekinge, Karlskrona, Sweden, [2] Department of Psychology, University of Gothenburg, Gothenburg, Sweden, [3] Anthropedia Foundation, St. Louis, MO, United States, [4] Department of Psychology, University of Turku, Turku, Finland, [5] Department of Psychology, Lund University, Lund, Sweden, [6] Department of Behavioral Science and Learning, Linköping University, Linköping, Sweden

OPEN ACCESS

Edited by:
Pietro Cipresso,
Italian Institute for Auxology (IRCCS),
Italy

Reviewed by:
Ioannis Tsaousis,
University of Crete, Greece
Andrea Bonanomi,
Catholic University of the Sacred
Heart, Italy

***Correspondence:**
Ali Al Nima
alinor_1979@yahoo.co.uk
Danilo Garcia
danilo.garcia@icloud.com

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: 24 August 2019
Accepted: 23 December 2019
Published: 22 January 2020

Citation:
Nima AA, Cloninger KM,
Persson BN, Sikström S and Garcia D
(2020) Validation of Subjective
Well-Being Measures Using Item
Response Theory.
Front. Psychol. 10:3036.
doi: 10.3389/fpsyg.2019.03036

Background: Subjective well-being refers to the extent to which a person believes or feels that her life is going well. It is considered as one of the best available proxies for a broader, more canonical form of well-being. For over 30 years, one important distinction in the conceptualization of subjective well-being is the contrast between more affective evaluations of biological emotional reactions and more cognitive evaluations of one's life in relation to a psychologically self-imposed ideal. More recently, researchers have suggested the addition of harmony in life, comprising behavioral evaluations of how one is doing in a social context. Since measures used to assess subjective well-being are self-reports, often validated only using Classical Test Theory, our aim was to focus on the psychometric properties of the measures using Item Response Theory.

Method: A total of 1000 participants responded to the Positive Affect Negative Affect Schedule. At random, half of the participants answered to the Satisfaction with Life Scale or to the Harmony in life Scale. First, we evaluate and provide enough evidence of unidimensionality for each scale. Next, we conducted graded response models to validate the psychometric properties of the subjective well-being scales.

Results: All scales showed varied frequency item distribution, high discrimination values (*Alphas*), and had different difficulty parameters (*Beta*) on each response options. For example, we identified items that respondents found difficult to endorse at the highest/lowest point of the scales (e.g., "Proud" for positive affect; item 5, "If I could live my life over, I would change almost nothing," for life satisfaction; and item 3, "I am in harmony," for harmony in life). In addition, all scales could cover a good portion of the range of subjective well-being (*Theta*): −2.50 to 2.30 for positive affect, −1.00 to 3.50 for negative affect, −2.40 to 2.50 for life satisfaction, and −2.40 to 2.50 for harmony in life. Importantly, for all scales, there were weak reliability for respondents with extreme latent scores of subjective well-being.

Conclusion: The affective component, especially low levels of negative affect, were less accurately measured, while both the cognitive and social component were covered to an equal degree. There was less reliability for respondents with extreme latent scores of subjective well-being. Thus, to improve reliability at the level of the scale, at the item level and at the level of the response scale for each item, we point out specific items that need to be modified or added. Moreover, the data presented here can be used as normative data for each of the subjective well-being constructs.

Keywords: Harmony in Life Scale, item response theory, Positive Affect Negative Affect Schedule, Satisfaction with Life Scale, subjective well-being

INTRODUCTION

Subjective well-being refers to the extent to which a person believes or feels that his or her life is going well and is considered as one of the best available proxies for a broader, more canonical form of well-being (Diener et al., 2018). This line of research has led to important contributions with regard to physical, psychological, and social health (e.g., Cloninger, 2004; Eid and Larsen, 2008; Lyubomirsky, 2008; Diener et al., 2009; Kjell et al., 2013), thus, making subjective well-being a popular and interesting construct (OECD, 2013). For over 30 years, subjective well-being has been conceptualized as comprising affective and cognitive evaluations of one's life (Diener, 1984; Diener et al., 2018). The affective component is conceptualized as affective evaluations of the emotions people experience in their daily lives, emotions such as, sadness, fear, anger, joy, etc. (cf. Watson et al., 1988). The cognitive component, on the other hand, is conceptualized as the way people evaluate their life as a whole in relation to a self-imposed ideal (Diener et al., 1985). Hence, one important distinction in the conceptualization of subjective well-being is the contrast between more affective evaluations that are obtained when asking about a person's typical emotional experience and more cognitive, judgment-focused evaluations like life satisfaction (Diener et al., 2018).

Despite some debates regarding the best way to conceptualize and measure the affective component of subjective well-being (e.g., how frequent or how intensive positive and negative emotions are experienced, whether it is best to use experience sampling methods or recollections of experienced emotions), most researchers agree that the frequency of emotions, rather than how intensive emotions are experienced, is a better measure of the affective component (Diener et al., 2018). For instance, people who experience high levels of well-being experience intensive positive emotions very rarely (only 2.6% of the time); instead they feel contented or mildly happy very frequently (Diener and Diener, 1996; Diener and Seligman, 2002; Garcia and Erlandsson, 2011). Judgments of life satisfaction, on the other hand, have been the undisputed way to conceptualize the cognitive component of subjective well-being. More recently, however, researchers have suggested harmony in life as a complement or supplement to life satisfaction (Kjell et al., 2016; Kjell, 2018). Nevertheless, in contrast to the focus on a psychologically self-imposed ideal involved in evaluations of life satisfaction, harmony is the sense of balance and flexibility that an individual experience in relation to the world around her (Li, 2008a,b). Moreover, harmony is distinctive from life satisfaction, not only by means of relations to different constructs or psychometric properties of measures (i.e., the Satisfaction with Life Scale vs. the Harmony in Life Scale), but also through how people pursue harmony in their life (Kjell et al., 2016; Garcia et al., 2020b). Indeed, when people are asked to describe how they pursue harmony, the most frequent words they use are: *peace*, *balance*, *unity*, *agreement*, *calm*, *mediation*, *cooperation*, *tolerant*, *nature*, *forgiveness*, etc. (Kjell et al., 2016). In contrast, when asked to describe how they pursue life satisfaction, the most frequent words are: *job*, *money*, *achievement*, *education*, *success*, *wealth*, *house*, *gratification*, etc. (Kjell et al., 2016). Thus, conceptually, harmony is different from life satisfaction, not because it is a different cognitive component, but because the concept comprises behaviors and notions of a person being in balance, in agreement, or striving for equilibrium or unity with the world around her (Garcia et al., 2020b).

In sum, life satisfaction comprises cognitive evaluations of one's life in relation to a psychologically self-imposed ideal (Diener et al., 1985), harmony comprises behavioral evaluations of how one is doing in a social context, and positive and negative affect comprises affective evaluations of biological emotional reactions. This is in line with the definition of health by the World Health Organization [WHO] (1946), in which health pertains not merely to the absence of disease or infirmity, but also to a state of physical, mental, and social well-being (see also Cloninger, 2004; VanderWeele, 2017). What is even more, it also corresponds to the biopsychosocial model, which is a scientific model that refers to a dynamic and complex interaction of physiological, psychological, and social factors that can both result in and contribute to health (Engel, 1977, 1980; Cloninger, 2004). Thus, we argue that the three subjective well-being components together are extremely important for our understanding of a complete biopsychosocial (cf. affect-cognition-behavior) model of subjective well-being (Garcia et al., 2020b). In this context, because most measures used to assess subjective well-being are self-reports, the cornerstone of research on a tentatively biopsychosocial model of subjective well-being should be to focus on the psychometric properties of the measures (Pavot, 2018). At a general level, the existing self-report measures exhibit strong psychometric properties including unidimensionality, high internal consistency, moderately strong test-retest reliability, and theoretically meaningful patterns of associations with other constructs and criteria (for reviews see Diener et al., 2009; Diener et al., 2013; for criticism regarding well-being measures see Brown et al., 2018). A clear majority of these analyses have implemented Classical Test Theory (CTT), which is a useful theory for understanding latent traits. To the best of our knowledge, there is little debate about the quality of these subjective well-being measures when researchers use these traditional methods (Diener et al., 2018; for criticism regarding well-being measures see Brown et al., 2018). However, evaluations of psychometric information of items and scales using CTT is dependent on the number of items and on the sample's size and other features, so any changes of these features can strongly affect both item and the total psychometric properties of the scale (Oishi, 2007). For instance, more precise estimates of reliability coefficients and their confidence intervals are obtained in large sample sizes of at least 400 respondents (Charter, 1999), which is no so common when these measures have been tested (Leue and Lange, 2011). Moreover, using CTT researchers can only report a single value to represent the reliability of the scale that is under investigation. This is problematic because by using this type of analysis, researchers implicitly assume that the standard error of measurement is equal across all points in the continuum of the concept being measured (Oishi, 2007). Therefore, this type of analyses does not provide sufficient information at different points along the trait continuum (e.g., ranging from extremely satisfied with life to extremely unsatisfied with life). In other words, CTT does not yield detailed feedback about which items provide the most reliable information across range of true scores (Oishi, 2007). Instead, CTT considers a summated scale as a measure of the latent trait although it is created without any justification from the sum of item scores.

Indeed, as suggested by others, many of the advantages of modern methods (e.g., Item Response Theory, IRT) have been ignored when subjective well-being measures have been validated (Oishi, 2007). IRT is as relatively modern psychometric technique that overcomes some of these limitations. One of IRT's biggest advantages is that we can determine how suitable items are to measure the latent traits, so it can increase reliable information and validity of the scale as a whole. The error and the reliable information obtained using IRT vary from one item to another and throughout the trait continuum of the scale, sometimes widely for one part of the scale compared with other parts (Oishi, 2007). In short, the aim of the present study is to apply IRT to evaluate existing well-validated measures[1] that might constitute a tentative biopsychosocial model of subjective well-being (i.e., Positive Affect Negative Affect Schedule, Satisfaction with Life Scale, and Harmony in Life Scale). Next, we briefly present research regarding the psychometric properties of each of the measures.

The Positive Affect Negative Affect Schedule

The Positive Affect Negative Affect Schedule was developed by Watson et al. (1988) as an attempt to provide better measures of positive and negative affect than contemporary measures at that time. These scales have been used in several studies to assess the affective or biological component of subjective well-being. Watson and colleagues started by selecting 60 adjectives representing affect from the factor analyses conducted by Zevon and Tellegen (1982). The selection criterion was that the adjectives were strongly correlated to one corresponding affect dimension but exhibited a weak correlation to the other. Throughout meticulous multiple rounds of selection and preliminary analyses, Watson et al. (1988) ended up with 10 items for each of the scales (see also Watson and Clark, 1994). That is, a total of 20 items consisting of 10 adjectives that measure positive affect (i.e., "Interested," "Enthusiastic," "Proud," "Alert," "Inspired," "Determined," "Attentive," "Active," "Excited," and "Strong") and 10 adjectives that measure negative affect ("Distressed," "Upset," "Guilty," "Afraid," "Hostile," "Irritable," "Ashamed," "Nervous," "Jittery," and "Scared") with a 5-point Likert (1 = *not at all*, 5 = *very much*). Watson et al. (1988) suggested that the orthogonal rotation of the factors is the best representation of positive and negative affect's latent structure because of the opposing pleasant-unpleasant relationship in the factor loadings. The scales have shown high internal consistency in different studies — *Cronbach's alphas* raging between 0.83 to 0.90 for positive affect and between 0.85 to 0.93 for negative affect (see Watson and Clark, 1994; Leue and Lange, 2011).

Nevertheless, researchers have reported a two-factor model with positive affect and negative affect as uncorrelated factors and correlated factors (e.g., Kercher, 1992; Krohne et al., 1996; Crocker, 1997; Mackinnon et al., 1999; Terraciano et al., 2003; Crawford and Henry, 2004) and also subfactors of positive affect and negative affect as uncorrelated and correlated first-order factors (e.g., Mehrabian, 1997; Killgore, 2000; Gaudreau et al., 2006). Moreover, validation studies (see Crawford and Henry, 2004) using structural equation modeling suggest that best-fitting models are achieved by specifying correlations between error in items closely related to each other in meaning: Distressed-Upset, Guilty-Ashamed, Scared-Afraid, Nervous-Jittery, Hostile-Irritable, Interested-Alert-Attentive, Excited-Enthusiastic-Inspired, Proud-Determined, and Strong-Active. Hence, these covariances suggest the possibility of item reduction without serious repercussions on the content domain or internal consistency reliability of the positive and negative affect scales (Thompson, 2007, 2017). Finally, despite a robust and impressive body of research, only a few studies have conducted IRT analyses to validate the Positive Affect Negative Affect Schedule (e.g., Pires et al., 2013 who showed, in a Brazilian sample, that the item Alert was the one with highest difficulty[2] and worst fit statistics). Thus, IRT analyses are an important endeavor for the development of accurate and effective operationalization of the affective component of subjective well-being.

The Satisfaction With Life Scale

The Satisfaction with Life Scale was originally developed by Diener et al. (1985) as a brief assessment of an individual's general sense of satisfaction with her life (see also Pavot and Diener, 1993, 2008). It has been used in thousands of studies to assess the cognitive or psychological component of subjective well-being. Diener et al. (1985) developed the scale by first generating a pool of 48 items intended to reflect life satisfaction and well-being. Using factor analysis, they identified 10 items with high

[1] There are different well-validated scales that can be used to measure each component of subjective well-being, for a compilation of the most common, the reader is advised to see Lopez and Snyder (2004).

[2] Throughout the manuscript the term "difficult" or "difficulty" refers to "endorsement rate" or "probability of endorsement."

loadings (0.60 or above) on a common factor interpreted as global evaluations of a person's life. After eliminating items with redundancies, Diener et al. further reduced the number of items to five (i.e., "In most ways my life is close to my ideal," "The conditions of my life are excellent," "I am satisfied with my life," "So far I have gotten the important things I want in life," and "If I could live my life over, I would change almost nothing") with a 7-point Likert response scale (1 = *strongly disagree* to 7 = *strongly agree*).

The scale has high internal consistency as indicated by *Cronbach's alphas* raging between 0.79 and 0.89 in some studies (Pavot and Diener, 1993), 0.87 (Adler and Fagley, 2005) and 0.86 (Steger et al., 2006) in other studies (for a meta-analysis see Vassar, 2008). Moreover, in the original article (Diener et al., 1985), the researchers showed that a principal-axis factor analysis on the Satisfaction with Life Scale resulted in a single factor solution, in which the single factor accounted for 66% of the variance of the scale. Despite the fact that the single factor solution has been replicated in several studies, the fifth item of the scale ("If I could live my life over, I would change almost nothing") often shows lower factor loadings and item-total correlations than the first four items of the scale (e.g., Senécal et al., 2000). Pavot and Diener (2008) suggested that, because this specific item strongly implies a summary evaluation over past years, responses to it may involve a different cognitive recollection than the responses to the other items of the scale that imply a focus on the present (e.g., "The conditions of my life are excellent") or a temporal summation (e.g., "In most ways my life is close to my ideal"). One way or the other, both CTT and the few studies using IRT methodology (e.g., Oishi, 2006) indicate that the fifth item of the Satisfaction with Life Scale is somewhat distinct from the other four items (Pavot and Diener, 2008). Since this item is highly correlated with the other four, however, it is not costume nor necessary or recommended to drop it from the measure (Pavot and Diener, 2008).

The few studies using IRT (Vittersø et al., 2005; Oishi, 2006) suggest that, in some cases, comparisons based on raw scores of the Satisfaction with Life Scale may be misleading. In one study, for example, although initial analyses showed no mean difference between Greenlanders and Norwegians, when IRT was applied, it was revealed that some Greenlanders were more prone to random responding, and to use extreme response categories. After controlling for these tendencies, Norwegians showed higher life satisfaction than Greenlanders, with the exception of a specific latent class of Greenlanders, who were in turn more satisfied than the Norwegian sample (Vittersø et al., 2005).

The Harmony in Life Scale

The Harmony in Life Scale was developed by Kjell et al. (2016) who suggested that focusing solely on life satisfaction as the cognitive component of subjective well-being is problematic since individuals think about their life in various ways (cf. Delle Fave et al., 2011). Based on a literature review of global contexts, such as, lifestyle, surroundings, conditions, environment, society and the world, Kjell et al. (2016) generated 29 items that included essential key concepts such as harmony, being attuned, fitting in, acceptance, adaptation, adjustment, and peace of mind. These items were evaluated by 5 experts within psychological research who were presented with a review of the aims and theories underlying the scale and asked to rate each item based on relevance (cf. Davis, 1992). Based on these evaluations the final numbers of items amounted to 15. The 15 items were randomly presented, with the same instructions and Likert Scale as the Satisfaction with Life Scale, to 476 respondents. Kjell et al. (2016) used an exploratory factor analysis based on maximum likelihood and promax rotation to explore the factor structure of the scale. The analysis revealed a clear single factor model with the total eigenvalue of 9.40 explaining 62.64%, while the factor loadings for the 15 items ranged from 0.56 to 0.86. The researchers eliminated redundant items and chose five items (i.e., "My lifestyle allows me to be in harmony," "Most aspects of my life are in balance," "I am in harmony," "I accept the various conditions of my life," and "I fit well with my surroundings") that they found relevant to their theoretical framework and with factor loadings ranging from 0.73 to 0.86 (see also Singh et al., 2016 for factor loadings ranging from 0.75 to 0.90) and a *Cronbach's alpha* of 0.90 (see also Garcia et al., 2014 for a *Cronbach's alpha* of 0.91, Kjell et al., 2019 for *Cronbach's alphas* between 0.89 and 0.95, and Singh et al., 2016 for *Cronbach's alphas* between 0.83 and 0.87).

In a second study in the same article (Time 1 n_1 = 787 and Time 2 n_2 = 545), Kjell et al. (2016) showed that the Harmony in Life Scale had good test-retest reliability (r = 0.77) and that it correlated as expected to other well-being related scales, such as, the Satisfaction with Life Scale (r = 0.76) and the Subjective Happiness Scale (r = 0.67). Interestingly, CTT analyses showed that despite a strong correlation between life satisfaction and harmony in life, the two-factor models, rather than single factor models, were considerable better at both Time 1 [$\chi^2(34) = 191.70$, $p < 0.001$; $CFI = 0.97$; $RMSEA = 0.08$] and Time 2 [$\chi^2(34) = 120.72$, $p < 0.001$; $CFI = 0.98$; $RMSEA = 0.07$]. Moreover, to the best of our knowledge, the Harmony in Life Scale has only been used in three published articles besides the original study (i.e., Garcia et al., 2014; Singh et al., 2016; Kjell et al., 2019) and no study has used IRT as a method for psychometric testing.

Item Response Theory and the Present Study

IRT is a family of psychometric methods for analysis of items, item responses as well as whole scale properties. The basic premise of IRT is that the probability of a response is a function of an underlying trait, continuum (latent dimension) or ability that is denoted by Theta (θ). Theta represents a person's true latent trait (e.g., subjective well-being), which has been standardized to follow standard normal distribution with a range from −3.00 to 3.00, with 0.00 representing the average score (Baker, 2001). The primary goal of using IRT is to validate and modify existing scales that measure how much of a latent trait one person has, in this case positive affect, negative affect, life satisfaction, and harmony. For example, IRT can be applied to investigate which items that haven't enough reliable information about the construct and which parts of that construct that the items don't measure. IRT analyses can also differentiate items' properties (e.g., discrimination and difficulty) among individuals across a much wider range of the construct at hand. If the analyses show that there is such a problem with some items, the researcher can remove/modify those items or add new items that help to measure these parts of the construct, thus, providing information that can differentiate people across a much greater range of the latent trait and increases the validity of the whole scale (Oishi, 2007). Also, IRT analyses might help clinicians to understand patients' behavior regarding a difficult or easy item, which might be helpful for intervention as well as for normative data (Pires et al., 2013).

The items of the scales used to measure subjective well-being (i.e., Positive Affect Negative Affect Schedule, the Satisfaction with Life Scale, and the Harmony in Life Scale) are ordinal and scored on Likert scales, so the appropriate IRT model for them is a graded response model (GRM). In GRM each item has its own estimated difficulty scores or threshold parameter (i.e., Beta, β) that represents the underlying latent trait for each response for each person. More specifically, Beta represents the level of the underlying trait at which the next response option has 50% chance of being endorsed. Moreover, each item in GRM has also its discrimination parameter (i.e., Alpha, α) which reflects how well the items discriminate between different levels of the latent trait. Moreover, Alpha is used to reflect how strongly an item is related with this latent trait, so it can be considered roughly equivalent to factor loadings used in CTT. The discrimination parameter values can be from $-\infty$ to $+\infty$, but values are typically at about 0 to $+2.50$. Here, item discrimination values of 0.01–0.34 are considered very low; 0.34–0.64 low; 0.65–1.34 moderate; 1.35–1.69 high; and 1.70 and above very high (Baker, 2001). It is usually recommended to delete the items with negative value, because this might suggest that something is wrong with the item since it indicates that the probability of a correct response decreases while the ability increases (Baker, 2001).

In order to use IRT models, there are some basic assumptions regarding unidimensionality, local independence, monotonicity (shape of curve) and differential item functioning (DIF). Unidimensionality states that the set of items in the questionnaire/test are expected to load on only one latent factor to explain the item response patterns. This is tested using factor analysis. Local independence means that the latent trait score explains most of the variance of participants' responses to the items in the scale. This is tested by verifying that the residuals for each item is not significantly correlated to the residuals of any other item in the scale. Monotonicity refers to item characteristics that reflect the true relationship between the person's latent trait score and the participant's actual response to the item. In other words, IRT models assume that the levels of the person's latent trait increase, as a monotonical function, as the probability to choosing the answer in each item that represents the participants actual level of the trait increases. DIF is applied to investigate so that the differences regarding the responses to each item does not vary across different groups (e.g., men and women).

Again, more sophisticated statistical techniques based on IRT (e.g., techniques described above that address the properties of the whole scale, items, and item responses at the population and subpopulation level) seem to present a promising way forward for the measurement of subjective well-being (Oishi, 2007; OECD, 2013). Our aim was to investigate, using IRT methods, the psychometric properties of the two instruments that are commonly used to measure the affective (or biological) and cognitive (or psychological) components of subjective well-being (i.e., the Positive Affect Negative Affect Schedule and the Satisfaction with Life Scale) along a new measure, tentatively suggested to measure the behavioral (or social) component (i.e., the Harmony in Life Scale). These measures are not only the most common when measuring the different components, but as reviewed in the introduction, they have good psychometric properties and are unidimensional in nature as analyzed using CTT in past research. Unidimensionality, is by the way, an important assumption for IRT analyses. To the best of our knowledge, this is the first study to examine these three subjective well-being instruments in the same study using IRT.

MATERIALS AND METHODS

Ethics Statement

Ethics approval was not required at the time the research was conducted as per national regulations. The consent of the participants was obtained by virtue of survey completion after they were provided with all relevant information about the research (e.g., anonymity).

Participants and Data Collection Procedure

The participants were recruited through Amazon's Mechanical Turk[3,4]. All participants originated from the United States and spoke English as their first language. Participants were informed that the survey was voluntary, anonymous, that they could terminate the survey at any time and that those who accepted would receive $0.50 as compensation for their participation. We added two control questions to the survey, to control for automatic responses (e.g., "This is a control question, please answer "either agree or disagree"). The final sample, after taking away those who responded erroneously to one or both of the control questions (*n* = 100, 9.09% of all respondents) consisted of 1000 participants (404 males and 596 females), including two who did not report their age (age *mean* for 998 participants = 34.22, *SD* = 12.73, range from 18 to 74). All 1000 participants responded to the Positive Affect Negative Affect Schedule. However, since the instructions, the format, and response scale of the Satisfaction with Life Scale and the Harmony in life Scale are exactly the same, participants were randomly presented with the Satisfaction with Life Scale (age *mean* for 498 participants = 34.08, *SD* = 12.55, range from 18 to 74; male = 217 and female = 283) or the Harmony in Life Scale among the participants (age *mean* for 500

[3] Amazon's Mechanical Turk MTurk allows data collectors to recruit participants (i.e., workers) online for completing different tasks for money (for a review on the validity of this method for data collection see among others: Buhrmester et al., 2011; Rand, 2012).

[4] http://www.mturk.com/mturk/welcome

participants = 34.36, *SD* = 12.92, range from 18 to 73; male = 187 and female = 313). This was done in order to avoid any likeness between the scales to influence participants' responses.

Measures

The Positive Affect Negative Affect Schedule (Watson et al., 1988) measures a person's experience of positive and negative affect. The respondents are asked to estimate and rate to which extent they have felt 10 positive (e.g., "Attentive") and 10 negative (e.g., "Hostile") feelings and moods during the last week on a five-point scale (1 = *very slightly or not at all*, 5 = *extremely*).

The Satisfaction with Life Scale (Diener et al., 1985) measures individuals' global cognitive judgments of their life as a whole in relation to a self-imposed ideal using five items (e.g., "In most ways my life is close to my ideal") and a seven-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*).

The Harmony in Life Scale (Kjell et al., 2016) assess a person's global sense of harmony in life and consists of five statements (e.g., "My lifestyle allows me to be in harmony") for which respondents are asked to indicate degree of agreement on a seven-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*).

Statistical Procedure

We used the following software to analyze the data: STATA version 14, R, SPSS version 24, and AMOS version 24. First, we used exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) to replicate past evidence showing that the correlation among items in each measure is explained by only a single latent trait (i.e., showing unidimensional factor structures). The lack of unidimensionality, for instance, might lead to biased results regarding IRT parameter estimates[5]. For each of the subjective well-being measures, EFA showed that the scree plot of eigenvalues suggested a single latent factor. The first eigenvalues of each scale (3.56 for life satisfaction, 3.74 for harmony in life, 5.08 for positive affect, and 1.05 for negative affect) were much greater than the others, which were less than 1.06. The ratio of the first to the second eigenvalue was greater than 5.00. Hence, for all scales there is evidence of unidimensionality (cf. Sattelmayer et al., 2017). Item loadings ranged from 0.63 to 0.80 for positive affect, 0.63 to 0.80 for negative affect, 0.74 to 0.90 for life satisfaction, and 0.79 to 0.91 for harmony in life.

The basic single factor CFA model for positive affect showed that the *chi-square* value was significant (χ^2 = 443.59, *df* = 35, $p < 0.001$), the *goodness of fit index* was 0.91, the *incremental fit index* was 0.91, and the *Root Mean Square Error of Approximation* fit statistic was slightly outside the acceptable rang 0.108 (for more details see **Supplementary Figure S1**). After one modification, a path between the error measurement for Alert-Attentive, the *chi-square* value was lower, but still significant (χ^2 = 307.55, *df* = 34, $p < 0.001$). Nevertheless, after this modification, all other fit indexes were acceptable (the *goodness of fit index* was 0.94, the *incremental fit index* was 0.94, and the *Root Mean Square Error of Approximation* fit statistic that was 0.09). All factor loadings were significant at $p < 001$ (**Supplementary Figures S1, S2**).

The basic single factor CFA model for negative affect showed that the *chi-square* value was significant (χ^2 = 1055.38, *df* = 35, $p < 0.001$). Fit indexes were slightly outside the traditional acceptable range: the *goodness of fit index* was 0.80, the *incremental fit index* was 0.82, and the *Root Mean Square Error of Approximation* fit statistic that was 0.17 (for more details see **Supplementary Figure S3**). After three modifications, paths between the error measurements for Guilty-Ashamed, Hostile-Irritable, and Afraid-Scared, the *chi-square* value was lower but still significant (χ^2 = 438.53, *df* = 32, $p < 0.001$). Nevertheless, after these modifications, all other fit indexes were acceptable (the *goodness of fit index* was 0.91, the *incremental fit index* was 0.93, and the *Root Mean Square Error of Approximation* fit statistic that was 0.11). All factor loadings were significant at $p < 001$ (for more details see **Supplementary Figures S3, S4**).

The basic single factor CFA model for life satisfaction fitted well (**Supplementary Figure S5**). The results showed that the *chi-square* value was not significant (χ^2 = 10.14, *df* = 5, $p = 0.07$), the *goodness of fit index* was 0.99, the *incremental fit index* was 1.00, and the *Root Mean Square Error of Approximation* fit statistic that was 0.04. Thus, indicating that the model fit was acceptable (cf. Bollen, 1989; Browne and Cudeck, 1993). All factor loadings were significant at $p < 001$.

The basic single factor CFA model for harmony in life fitted also well (**Supplementary Figure S6**). The results showed that the *chi-square* value was significant (χ^2 = 31.68, *df* = 5, $p < 0.001$). The *goodness of fit index* was 0.98, the *incremental fit index* was 0.99, and the *Root Mean Square Error of Approximation* fit statistic that was 0.10. That is, all indexes indicated that the model fit was acceptable. All factor loadings were significant at $p < 001$.

Previous research suggests that fit indexes that are slightly outside the traditional acceptable range can be considered as sufficiently unidimensional for further IRT analysis (Cook et al., 2009; Stepp et al., 2012). In addition, although significant for some of the models, the *chi-square* statistic is heavily influenced by sample size (Kline, 2010), with larger samples leading to a larger value and therefore, a larger likelihood of being significant. Thus, given the results of the scree plot of eigenvalues, eigenvalues, ratios, item loadings and the results of the CFA, we considered that our results provide sufficient evidence of unidimensionality of single latent trait for each one of these four main measures of a biopsychosocial model of subjective well-being.

Regarding local independence, our analyses showed that, for all scales, the residuals (i.e., differences between the individuals' observed scores and their respective predicted scores) of almost each paired correlation were significant. That is, most of the items can be considered as locally dependent and that our data had a tendency for multidimensionality. See **Supplementary Tables S2a,b** for the details. Result regarding Monotonicity indicated that the response function of the probability of getting correct response of each item of each scale increased when the person's latent trait level increased. See **Supplementary Table S3** and **Supplementary Figure S7** for the details. The result exhibited uniform Differential Item Function (DIF) for each item

[5] Some researchers, however, confirm that IRT analyses are reasonably robust to violations to unidimensional factor structure assumptions (Ip, 2010).

in SWLS across gender. This indicated that the ability of a person to answer does not change due to gender characteristics. See **Supplementary Figure S8** for the details.

We tested the item fit statistic using the Orlando–Thissen–Bjorner *item fit S-χ2 statistic* to determine absolute fit of the model to each item. Regarding S-χ2 statistic, a value that is not significant indicates that the model adequately fits an item. The result indicated that 25 items were adequately fit, while four items were statistically significant at $p < 0.05$ and one item at $p < 0.01$. The S-χ2 statistic is sensitive and influenced by sample size, test length and multiple comparisons, with larger samples, small test length and multiple comparisons leading to a larger value and therefore, a larger likelihood of being significant (*Type I error*). In other words, these five valid items were falsely identified as mis-fitting when in fact the model fitted the data/items, so the root mean square error of approximation (RMSEA) was used but it was based on the S-χ2 statistic (RMSEA S-χ2). Traditional cut-offs for RMSEA tend to be RMSEA ≤ 0.08 to determine absolute fit of the model to each item. The result exhibited that the largest value of RMSEA S-χ2 was 0.03, so this result indicated an adequate item-level model-data fit. Nevertheless, we applied the Benjamini–Hochberg criterion for *p*-value adjustment (Benjamini and Hochberg, 1995). Three items ("Scared," "My lifestyle allows me to be in harmony," and "I fit in well with my surroundings") were still significant after correction (see **Supplementary Table S4**). We checked these items' information, difficulty, and discrimination parameter in order to decide whether they needed to be excluded from the analyses. Since these three items provided with reliable information, discrimination and difficulty, along good properties overall (see for example analyses regarding monotonicity), we decided to keep them. For example, the item "Scared," was still significant after correction, but this item had good reliable information, high discrimination parameter 3.49 and difficulty parameters between 0.26 and 1.94, which are even better values that some of the items that were not significant after correction. See **Supplementary Table S4** for the details.

Comparisons Among GRM, RSM and PCM

In order to determine the most appropriate IRT model to our data, we compared the model we chose, GRM, with both *Rating Scale Model* (RSM), which is for ordinal responses to items that share the same rating scale structure, and *Partial Credit Model* (PCM), which is for ordinal responses to item that have its own rating scale structure. We used three fit indices to evaluate model fit: *Log-likelihood*, *Bayesian information criterion* (BIC) and *Akaike information criterion* (AIC). The result showed that GRM was preferable. See **Supplementary Table S1** for the details.

RESULTS

IRT Analyses of the Positive Affect Negative Affect Schedule

Positive Affect

We found that the frequency distributions for each of the items in the positive affect scale were different (see **Table 1**), for example, for the item "Determined" 20.80% of the participants reported the highest levels (5 = *extremely*) compared with the item "Enthusiastic" for which only 10.30% of the participants reported the highest levels (5 = *extremely*). The item "Enthusiastic" was more difficult, explained through the proportion of participants choosing the highest point of the scale, than the item "Determined." This is important, if the items vary in their difficulty, the correlations among items would be small. Moreover, in this analysis each item gets its own discrimination/slope (Alpha) and own 'location' parameter (Beta); the differences between categories around that location are not equal across items (see **Table 2** and **Figure 1**). Regarding item discrimination, all items had high discrimination values (Alphas from 1.37 to 2.65) and demonstrated a steeper slope, which indicates that the items can differentiate well between persons with high and low levels of the latent score of positive affect (see **Table 2** and **Figure 1**). Regarding the estimated threshold/difficulty parameter (Beta) for the positive affect scale were between -2.54 and 1.65 (see **Table 2**). The item "Alert" had the highest estimated difficulty parameter on response 5 (β = 1.65) and the item "Interested" had the lowest estimated difficulty parameter on response 1 (β = −2.54). To understand the difficulty parameter, let's exemplify with the first item, "Interested." A respondent with −2.54 in positive affect has a 50% chance of answering 1 (*very slightly or not at all*), versus greater or equal chance of answering 2 (i.e., responses 2, 3, 4, or 5). A respondent with −1.36 in positive affect has a 50% chance of answering 1 or 2, rather than greater or equal chance of answering 3 (i.e., responses 3, 4, or 5). A person with 1.33 in positive affect has a 50% chance of picking response 5 (*extremely*), rather than less or equal chance of answering 4 (i.e., responses 1, 2, 3, or 4).

Furthermore, the differences between categories around difficulty parameters (Beta) are not equal across items. That is, for each item a response of, for example, 5 (*extremely*) was treated differently: β = 1.65 for item "Alert" while it was 1.15 for item "Determined." Moreover, the differences in difficulty varied within each item (i.e., distances between responses for each item). For example, for the item "Interested" (see **Table 2**), the difference between ≥2 and ≥3 is −2.54 – (−1.36) = −1.18, while the difference between ≥3 and ≥4 is −1.36 – (−0.12) = −1.24. Thus, participants' total score of positive affect will differ from totals scores using CTT, where differences are treated as equal and added without further justification (for more details see **Table 2** and **Figure 1**).

The graph regarding category characteristic curves (**Figure 2**) gives information about the relationship between the level of the participants' positive affect (i.e., the latent trait) and the probability of responding to specific points in the scale for each item, respectively. The graphs show the location where the next category becomes more likely (not 50%), that is, the points where the adjacent categories cross represent transitions from one response point to the next. For example, for the item "Interested," participants with positive affect (latent trait) levels below −2.46 are more likely to respond 1 (*very slightly or not at all*) while the participants with positive affect levels between −2.46 and −1.38 are most likely to respond 2, and so on. Moreover, the probability

TABLE 1 | The frequency distributions of the positive affect scale of the Positive Affect Negative Affect Schedule (*N* = 1000).

Item	Points in the Likert Scale				
	1	2	3	4	5
Interested					
Frequency	32	113	309	396	150
Percent	3.20	11.30	30.90	39.60	15.00
Cumulating	3.20	14.50	45.40	85.00	100.00
Enthusiastic					
Frequency	115	183	300	299	103
Percent	11.50	18.30	30.00	29.90	10.30
Cumulating	11.50	29.80	59.80	89.70	100.00
Proud					
Frequency	199	205	263	209	124
Percent	19.90	20.50	26.30	20.90	12.40
Cumulating	19.90	40.40	66.70	87.60	100.00
Alert					
Frequency	79	152	273	347	149
Percent	7.90	15.20	27.30	34.70	14.90
Cumulating	7.90	23.10	50.40	85.10	100.00
Inspired					
Frequency	175	212	269	227	117
Percent	17.50	21.20	26.90	22.70	11.70
Cumulating	17.50	38.70	65.60	88.30	100.00
Determined					
Frequency	71	125	244	352	208
Percent	7.10	12.50	24.40	35.20	20.80
Cumulating	7.10	19.60	44.00	79.20	100.00
Attentive					
Frequency	55	101	301	373	170
Percent	5.50	10.10	30.10	37.30	17.00
Cumulating	5.50	15.60	45.70	83.00	100.00
Active					
Frequency	119	198	328	233	122
Percent	11.90	19.80	32.80	23.30	12.20
Cumulating	11.90	31.70	64.50	87.80	100.00
Excited					
Frequency	169	243	290	188	110
Percent	16.90	24.30	29.00	18.80	11.00
Cumulating	16.90	41.20	70.20	89.00	100.00
Strong					
Frequency	154	214	281	231	120
Percent	15.40	21.40	28.10	23.10	12.00
Cumulating	15.40	36.80	64.90	88.00	100.00

of option 1 and 5 for this item are about equal and very high (For more details see **Figure 2**).

We also investigated the item information function (see **Figure 3A**) for each item to see how much information each item provides as estimated by their location on the continuum (i.e., difficulty parameter) for the latent factor of positive affect and to investigate what level of the continuum each item has most or least information or reliability. In other words, the item information function reflects the properties of each item in terms of both its difficulty (Beta) and discrimination (Alpha) index. Moreover, this analysis helped us to evaluate where additional items would be useful to develop the scale. For instance, the items “Enthusiastic” and “Excited” had the highest discrimination estimates and seem to provide more information than the remaining items, while the items “Alert” and “Attentive” provide lesser information. In general, the items cover the distribution of the true range of positive affect (Theta, θ) from low (−2.50) up to high (2.30). Moreover, we show that we get reliable information at $\theta = 0$ (vertical red line in **Figure 3A**) at about 1.90 from the item “Enthusiastic,” at about 1.30 from the item “Excited,” at about 1.20 from the item “Proud,” at about 1.10 from the item “Interested,” at about 1.05 from item “Strong,” and so on.

TABLE 2 | Item response analysis of the positive affect scale in the Positive Affect Negative Affect Schedule (N = 1000).

Item	Coef.	*SE*	*Z*	*P*	95% CI	
Interested						
Discrimination	2.03	0.12	17.58	0.00	1.81	2.26
Difficulty						
≥2	−2.54	0.14	−18.54	0.00	−2.81	−2.28
≥3	−1.36	0.08	−17.62	0.00	−1.51	−1.21
≥4	−0.12	0.05	−2.40	0.02	−0.22	−0.02
= 5	1.33	0.08	17.39	0.00	1.18	1.48
Enthusiastic						
Discrimination	2.65	0.15	17.56	0.00	2.35	2.94
Difficulty						
≥2	−1.45	0.07	−19.72	0.00	−1.59	−1.30
≥3	−0.61	0.05	−11.89	0.00	−0.71	−0.51
≥4	0.31	0.05	6.72	0.00	0.22	0.41
= 5	1.48	0.07	19.81	0.00	1.33	1.63
Proud						
Discrimination	2.00	0.11	17.43	0.00	1.77	2.22
Difficulty						
≥2	−1.09	0.07	−15.52	0.00	−1.23	−0.95
≥3	−0.28	0.05	−5.39	0.00	−0.38	−0.18
≥4	0.58	0.06	10.52	0.00	0.47	0.69
= 5	1.50	0.08	17.78	0.00	1.33	1.66
Alert						
Discrimination	1.37	0.09	15.83	0.00	1.20	1.54
Difficulty						
≥2	−2.31	0.14	−16.06	0.00	−2.59	−2.03
≥3	−1.17	0.09	−13.63	0.00	−1.34	−1.00
≥4	0.00	0.06	0.05	0.96	−0.12	0.12
=5	1.65	0.11	15.20	0.00	1.43	1.86
Inspired						
Discrimination	1.81	0.11	17.20	0.00	1.60	2.02
Difficulty						
≥2	−1.29	0.08	−16.14	0.00	−1.44	−1.13
≥3	−0.38	0.06	−6.79	0.00	−0.49	−0.27
≥4	0.56	0.06	9.69	0.00	0.44	0.67
=5	1.62	0.09	17.40	0.00	1.44	1.80
Determined						
Discrimination	1.71	0.10	16.90	0.00	1.51	1.91
Difficulty						
≥2	−2.10	0.12	−17.57	0.00	−2.34	−1.87
≥3	−1.15	0.08	−15.00	0.00	−1.31	−1.00
≥4	−0.16	0.05	−3.00	0.00	−0.27	−0.06
=5	1.15	0.08	15.06	0.00	1.00	1.30
Attentive						
Discrimination	1.58	0.10	16.35	0.00	1.39	1.77
Difficulty						
≥2	−2.41	0.14	−16.92	0.00	−2.69	−2.14
≥3	−1.44	0.09	−15.67	0.00	−1.62	−1.26
≥4	−0.13	0.06	−2.26	0.02	−0.24	−0.02
=5	1.38	0.09	15.46	0.00	1.21	1.56
Active						
Discrimination	1.78	0.10	17.29	0.00	1.57	1.98
Difficulty						
≥2	−1.63	0.09	−17.27	0.00	−1.82	−1.45

(Continued)

TABLE 2 | Continued

Item	Coef.	*SE*	*Z*	*P*	95% CI	
≥3	−0.62	0.06	−10.36	0.00	−0.74	−0.51
≥4	0.51	0.06	8.81	0.00	0.39	0.62
=5	1.59	0.09	17.18	0.00	1.41	1.78
Excited						
Discrimination	2.11	0.12	17.61	0.00	1.87	2.34
Difficulty						
≥2	−1.22	0.07	−16.82	0.00	−1.37	−1.08
≥3	−0.28	0.05	−5.57	0.00	−0.38	−0.18
≥4	0.67	0.06	11.98	0.00	0.56	0.78
=5	1.58	0.08	18.66	0.00	1.41	1.74
Strong						
Discrimination	1.87	0.11	17.27	0.00	1.65	2.08
Difficulty						
≥2	−1.37	0.08	−16.69	0.00	−1.53	−1.21
≥3	−0.44	0.06	−7.86	0.00	−0.54	−0.33
≥4	0.52	0.06	9.24	0.00	0.41	0.63
=5	1.58	0.09	17.58	0.00	1.40	1.75

Moreover, the 10 items together provide a lot of information to measure positive affect among participants that vary within range −2.50 up to about 2.30 (Theta) of the level of the scale of positive affect (see **Figure 3B**, test information function and the standard error, that is, measurement error). This means that the positive affect scale has good reliability and small standard error in this range. The test information highest level is located at −0.50 (Theta), thus indicating that this score has the smallest standard error and provides the most information of the scale. However, there is almost no reliable information below -3.50 and above 3.50 (Theta) and the standard error increases quickly for both smaller and larger Theta values. The reliability for different levels of positive affect are shown in **Table 3**. These results showed that the scale's reliability is very strong (between 0.88 to 0.91) at $\theta = -2.00$, $\theta = -1.00$, $\theta = 0.00$, $\theta = 1.00$, and $\theta = 2.00$, that reliability is good (0.75) at $\theta = -3.00$, but that reliability is week (0.64) at $\theta = 3.00$.

Figure 3C shows the test characteristic curve for the whole scale, which indicates the expected score against the latent trait (i.e., positive affect) as a sum of the probabilities. Since the positive affect scale of the Positive Affect Negative Affect Schedule has 10 items with a five-point Likert scale (1 = *very slightly or not at all*, 5 = e*xtremely*), the expected scores are between 10 and 50. Our results showed that the expected score for participants that have positive affect at level of −1.96 (Theta) and below, is 15.50 or less. That is, these participants are most likely to choose the answer coded 1 or 2 on most items. With critical values (−1.96 and 1.96) coding to the standard normal distribution we can expect 95% of randomly selected participants have a score between 15.50 and 46.50 (see **Figure 3C**).

Negative Affect

We found that the frequency distributions for each of the items in the negative affect scale varied (see **Table 4**). For example, for the item "Distressed," 7.20% of participants report a high negative affect (5 = *extremely*) compared with the item "Hostile" for which only 1.60% of participants report high negative affect (5 = *extremely*). In other words, the item "Hostile" differ in its difficulty compared with the item "Distressed" that has less difficulty (for more details see **Table 5**). Regarding item discrimination, all items had high discrimination values (Alphas from 1.53 to 3.49) and had a steeper slope (see **Table 5** and **Figure 4**). Thus, indicating that that the items can differentiate well between persons with high and low levels of the latent score of negative affect. The difficulty parameters estimations (Beta) for the negative affect scale are between −0.70 and 3.14 (see **Table 5**). The item "Hostile" has the highest estimated difficulty parameter on response 5 ($\beta = 3.14$) and the item "Irritable" has the lowest estimated difficulty parameter on response 1 ($\beta = -0.70$). Our results also showed that the differences between categories around difficulty parameters are not equal across the negative affect scale items. For example, 5 (*extremely*) was 3.14 for the item "Hostile," while it was 1.71 for the item "Distressed." Moreover, the differences in difficulty varied within each item (i.e., distances between responses for each item). For example, for the item "Distressed," the difference between ≥ 2 and ≥3 is −0.69 – (0.44) = −0.15, while the difference between ≥3 and ≥4 is 0.44 – (1.03) = 0.59. Thus, participants' total score of negative affect will differ from totals scores using CTT, where differences are treated as equal and added without further justification (for more details see **Table 5** and **Figure 4**).

Figure 5, the category characteristic curves, shows the transitions from one category to the next. For example, for the item "Distressed," participants with negative affect (i.e., latent trait) levels below −0.65 are most likely to respond 1 (*very slightly or not at all*), while the participants with negative affect levels between 0.62 and 0.98 are most likely to respond 2, and so on. Moreover, the probability of responding 1 and 5 for this item are equal and very high (see **Figure 5** for more details).

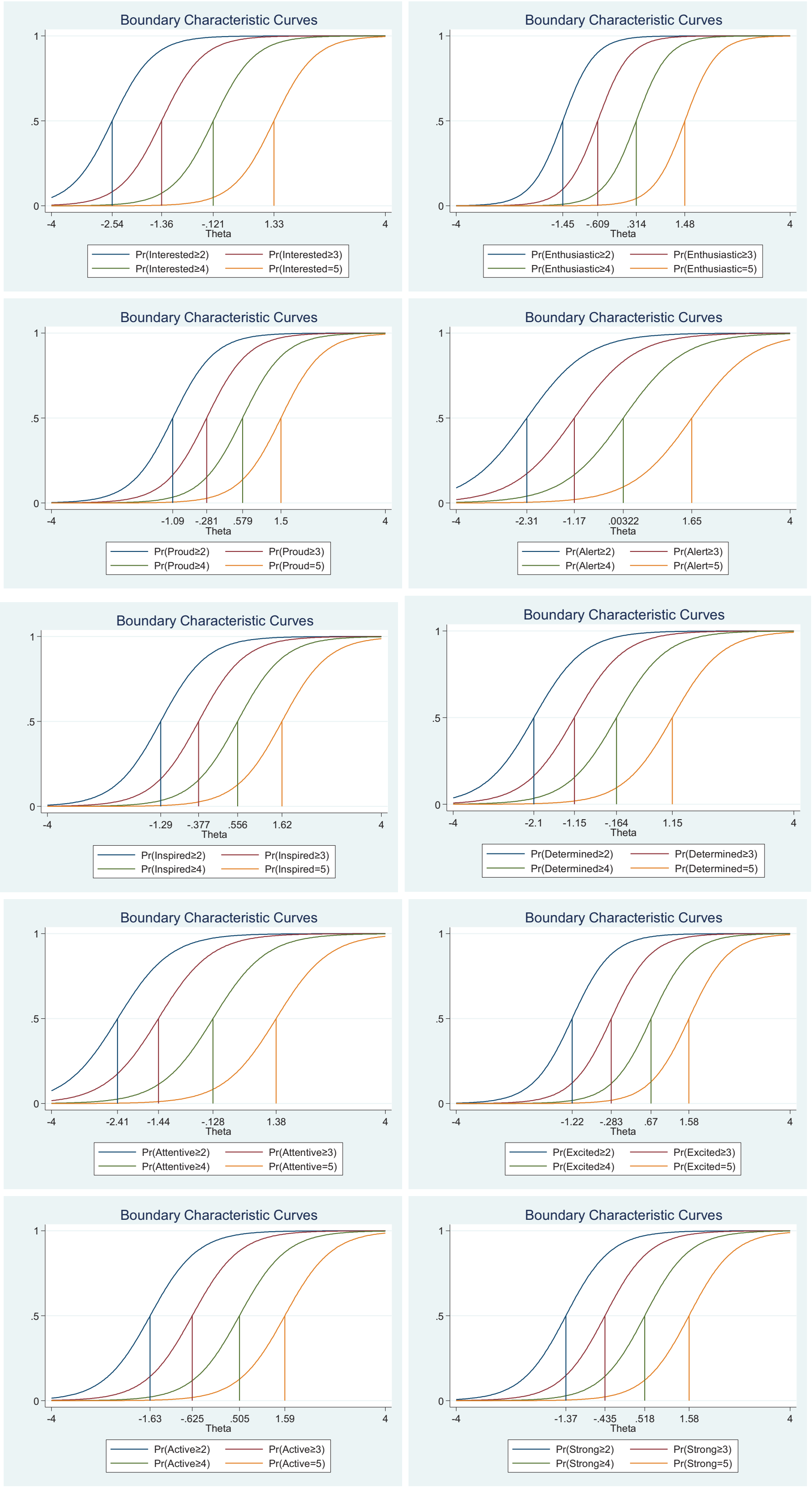


FIGURE 1 | Boundary characteristic curves for each item of the positive affect scale of the Positive Affect Negative Affect Schedule (*N* = 1000).

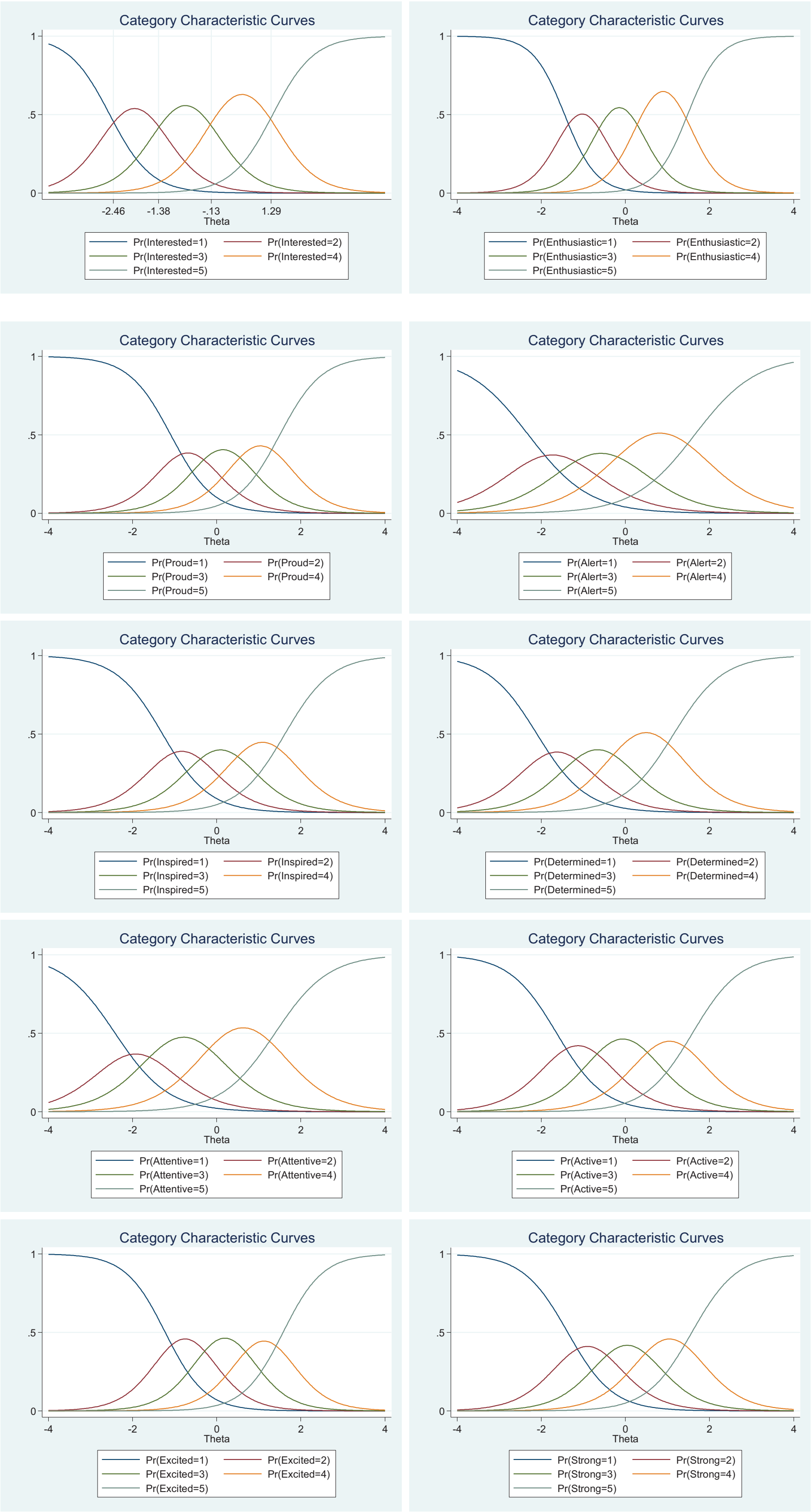


FIGURE 2 | Category characteristic curves for the items in the positive affect scale of the Positive Affect Negative Affect Schedule (*N* = 1000).

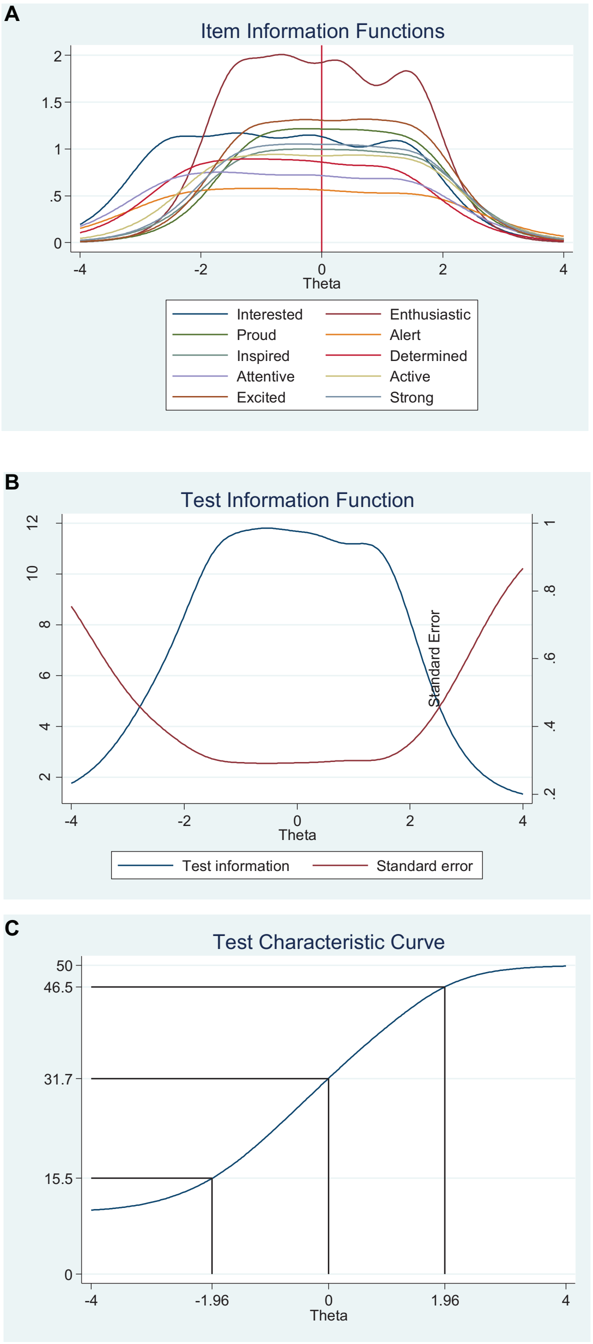


FIGURE 3 | Items information function graphs for graded response and with vertical line at θ = 0 **(A)** and information and standard error graph for graded response **(B)** and test characteristic curve **(C)** for the whole positive affect scale of the Positive Affect Negative Affect Schedule (*N* = 1000).

The item information function analyses indicate that the items "Scared" and "Afraid" have the two highest discrimination estimates and provide more information than the remaining items, while the items "Jittery" and "Hostile" provided the lesser information (see **Figure 6A**). Moreover, we show that we get reliable information at θ = 0 (vertical red line in **Figure 6A**) at about 2.60 from the item "Scared," at about 1.80 from the item "Afraid," at about 1.75 from the item "Distressed," at about 1.70 from the items "Nervous" and "Irritable," and so on. Moreover, the ten items together provide a lot of information to measure negative affect among participants that vary within range −1.00 up to about 3.00 (Theta) of the level of the scale of negative affect (see **Figure 6B**, test information function and the standard error, that is, measurement error). This means that the negative affect scale of the Positive Affect Negative Affect Schedule has good reliability and small standard error in this range. The test information highest level is located at 1.80 (Theta), thus indicating that this score has the smallest standard error and provides the most information of the negative affect scale. However, there is almost no reliable information about below −2.00 and about above 4.00 (Theta) and the standard error increases quickly for both smaller and larger Theta values. The reliability for different levels of negative affect are shown in **Table 3**. These results showed that the scale's reliability is very strong at θ = −1.00, θ = 0.00, θ = 1.00, θ = 2.00, and θ = 3.00 (between 0.84 to 0.95), but that reliability is weak (0.46) at θ = −2.00 and very week (0.10) at θ = −3.00.

Figure 6C shows the test characteristic curve for the whole scale, which indicates the expected score against the latent trait of negative affect as a sum of the probabilities. Since the negative affect scale of the Positive Affect Negative Affect Schedule has 10 items with a five-point Likert scale (1 = *very slightly or not at all*, 5 = *extremely*), the expected scores are between 10 and 50. Our results showed that the expected score for participants that have negative affect at level of −1.96 (Theta) and below, is 10.30 or less. That is, these participants are most likely to choose the answer coded 1 on all items. With critical values (−1.96 and 1.96) coding to the standard normal distribution we can expect 95% of randomly selected participants have a score between 15.50 and 46.50 (see **Figure 3C**). With critical values (−1.96 and 1.96) coding to the standard normal distribution we can expect 95% of randomly selected participants have expected score between 10.30 and 39.20 (see **Figure 6C**).

IRT Analyses of the Satisfaction With Life Scale

Again, as for the positive and negative affect scales, the frequency distributions for each of the items in the Satisfaction with Life Scale varied (see **Table 6**). Thus, suggesting that some items differ in difficulty compared to other items in the scale. For example, for item 4 ("So far I have gotten the important things I want in life"), 12.40% of the participants reported high satisfaction with life (7 = *strongly agree*), while only 7% of the participants report 7 when answering item 1 ("In most ways my life is close to my ideal"). Moreover, all items had very high discrimination values (from 1.74 to 4.50) and a steeper slope, which indicates that the items can differentiate well between persons with high and low levels of the latent score of satisfaction with life (see **Table 7** and **Figure 7**). In addition, the difficulty parameters estimations for the Satisfaction with Life Scale are between −1.69 and 1.76. Here, Item 5 ("If I could live my life over, I would change

TABLE 3 | Reliability of the fitted graded response IRT model of the positive and negative affect scales of the Positive Affect Negative Affect Schedule (*N* = 1000).

Theta	Positive Affect			Negative Affect		
	Test Information Function	Test Information Function-*SE*	Reliability IRT GRM	Test Information Function	Test Information Function-*SE*	Reliability IRT GRM
−3.00	4.00	0.50	0.75	1.11	0.95	0.10
−2.00	8.37	0.35	0.88	1.86	0.73	0.46
−1.00	11.66	0.29	0.91	6.33	0.40	0.84
0.00	11.68	0.29	0.91	14.96	0.26	0.93
1.00	11.19	0.30	0.91	18.89	0.23	0.95
2.00	8.17	0.35	0.88	17.69	0.24	0.94
3.00	2.80	0.60	0.64	6.80	0.38	0.85

Cronbach's alpha for the whole scale using CTT were 0.90 for the positive affect scale and 0.91 for negative affect scale.

almost nothing") has the highest estimated difficulty parameter on response 7 (1.76) and item 4 ("So far, I have gotten the important things I want in life") has the lowest estimated difficulty parameter on response 1 (−1.67). Our results showed also that the differences between categories around difficulty parameters are not equal across items. This means that for item 3 ("I am satisfied with my life"), for example, a response of 7 (*strongly agree*) was 1.28, while it was 1.76 for item 5 ("If I could live my life over, I would change almost nothing"). Moreover, the differences in difficulty varied within each item (i.e., distances between responses for each item). Thus, participants' total score of life satisfaction will differ from totals scores using CTT, where differences are treated as equal and added without further justification. For example, for item 1 ("In most ways my life is close to my ideal"), the difference between ≥2 and ≥3 is −1.25 – (−0.73) = −0.52, while the difference between ≥3 and ≥4 is −0.73 – (−0.35) = −0.38 (for more details see **Table 7** and **Figure 7**).

Figure 8, the category characteristic curves, shows the transitions from one category to the next. For example, for item 1 ("In most ways my life is close to my ideal"), participants with satisfaction with life (latent trait) levels below -1.18 are most likely to respond 1 (*strongly disagree*), while participants with satisfaction with life levels between 1.18 and −0.66 are most likely to respond 2, and so on. Moreover, the probability of option 1 and 7 for this item are equal and very high (see **Figure 8** for all the details).

The item information function analyses, **Figure 9A**, showed that items 1 ("In most ways my life is close to my ideal") and item 3 ("I am satisfied with my life") have the two highest discrimination estimates and provide more information than the remaining items, while item 5 ("If I could live my life over, I would change almost nothing") provides lesser information. In general, the results suggest that a lot of information of the true range of life satisfaction is covered between low (Theta = −2.00) up to high (Theta = 2.00) values. Moreover, we show that we get reliable information at θ = 0.00 at about 5.80 from item 1 ("In most ways my life is close to my ideal"), at about 3.30 from item 2 ("The conditions of my life are excellent"), at about 4.30 from item 3 ("I am satisfied with my life"), at about 1.80 from item 4 ("So far, I have gotten the important things I want in life") and at about 1.20 from item 5 ("If I could live my life over, I would change almost nothing") (see **Figure 9B**, test information function and the standard error, that is, measurement error). This means that the Satisfaction with Life Scale has good reliability and small standard error in this range. The test information highest is located at about −0.30 (Theta), thus indicating that this score has the smallest standard error and provides the most information of the scale. However, there is almost no reliable information about below −2.40 and about above 2.50 (Theta) and the standard error increases quickly for both smaller and larger Theta values. The reliability for different levels of life satisfaction are shown in **Table 8**. These results showed that the scale's reliability is very strong at θ = −2.00, θ = −1.00, θ = 0.00, θ = 1.00, and θ = 2.00, but that reliability is weak at θ = −3.00 and θ = 3.00. Since the Satisfaction with Life Scale has five items with a seven-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*), the expected scores are between 5 and 35. Our results showed that the expected score for participants that have life satisfaction at the level −1.96 (Theta) and below, is 6.35 or less. That is, these participants are most likely to choose the answer coded 1 on all or most items. With critical values (−1.96 and 1.96) coding to the standard normal distribution we can expect 95% of randomly selected participants to have a score between 6.35 and 33.6 (see **Figure 9C**).

IRT Analyses of the Harmony in Life Scale

As for the other subjective well-being measures, the frequency distributions for each of the items in the Harmony in Life Scale varied (see **Table 9**). Hence, suggesting that some items differ in difficulty compared to other items in the scale. For example, while 12.20% of the participants reported harmony in life (7 = *strongly agree*) for item 4 ("I accept the various conditions of my life"), only 5.20% of the participants reported high harmony in life (7 = *strongly agree*) for item 3 ("I am in harmony"). Moreover, all items had very high discrimination values (from 2.05 to 5.23) and a steeper slope, which indicates that the items can differentiate well between persons with high and low levels of the latent score of harmony in life (see **Table 10** and **Figure 10**). Furthermore, the difficulty parameters estimations for the Harmony in Life scale are between −2.09 and 1.64. Here, Item 3 ("I am in harmony")

TABLE 4 | The frequency distributions of the negative affect scale of the Positive Affect Negative Affect Schedule (N = 1000).

Item	Points of Likert scale				
	1	2	3	4	5
Distressed					
Frequency	275	365	169	119	72
Percent	27.50	36.50	16.90	11.90	7.20
Cumulating	27.50	64.00	80.90	92.80	100.00
Upset					
Frequency	328	338	169	110	55
Percent	32.80	33.80	16.90	11.00	5.50
Cumulating	32.80	66.60	83.50	94.50	100.00
Guilty					
Frequency	647	222	64	46	21
Percent	64.70	22.20	6.40	4.60	2.10
Cumulating	64.70	86.90	93.30	97.90	100.00
Afraid					
Frequency	574	244	84	64	34
Percent	57.40	24.40	8.40	6.40	3.40
Cumulating	57.40	81.80	90.20	96.60	100.00
Hostile					
Frequency	611	230	97	46	16
Percent	61.10	23.00	9.70	4.60	1.60
Cumulating	61.10	84.10	93.80	98.40	100.00
Irritable					
Frequency	297	353	187	106	57
Percent	29.70	35.30	18.70	10.60	5.70
Cumulating	29.70	65.00	83.70	94.30	100.00
Ashamed					
Frequency	661	205	69	47	18
Percent	66.10	20.50	6.90	4.70	1.80
Cumulating	66.10	86.60	93.50	98.20	100.00
Nervous					
Frequency	405	301	150	92	52
Percent	40.50	30.10	15.00	9.20	5.20
Cumulating	40.50	70.60	85.60	94.80	100.00
Jittery					
Frequency	573	257	81	63	26
Percent	57.30	25.70	8.10	6.30	2.60
Cumulating	57.30	83.00	91.10	97.40	100.00
Scared					
Frequency	585	264	63	51	37
Percent	58.50	26.40	6.30	5.10	3.70
Cumulating	58.50	84.90	91.20	96.30	100.00

has the highest estimated difficulty parameter on response 7 (1.64) and item 5 ("I fit in well with my surroundings") has the lowest estimated difficulty parameter on response 1 (−2.09). Our result also showed that the differences between categories around difficulty parameters are not equal across items. This means that for item 3 ("I am in harmony"), for example, a response of 7 (*strongly agree*) was 1.64, while it was 1.49 for item 4 ("I accept the various conditions of my life"). Moreover, the differences in difficulty varied within each item (i.e., distances between responses for each item). Thus, participants' total score of harmony in life will differ from totals scores using CTT, where differences are treated as equal and added without further justification. For example, for item 1 ("Most aspects of my life are in balance"), the difference between ≥2 and ≥3 is −1.62 – (−1.00) = −0.62, while the difference between ≥3 and ≥4 is −1.00− (−0.58) = −0.42 (see **Table 10** and **Figure 7**).

The analyses of the category characteristic curves showed that, for example, for item 1 ("My lifestyle allows me to be in harmony"), participants with harmony in life (latent trait) levels below −1.60 are most likely to respond 1 (*strongly disagree*), while participants with harmony in life levels between −1.60 and −0.95 are most likely to respond 2, and so on. Moreover, the probability

TABLE 5 | Item response analysis of the negative affect scale in the Positive Affect Negative Affect Schedule (*N* = 1000).

	Coef.	*SE*	*Z*	*P*	95% CI	
Distressed						
Discrimination	2.66	0.15	17.57	0.00	2.36	2.96
Difficulty						
≥2	−0.69	0.05	−12.71	0.00	−0.80	−0.58
≥3	0.44	0.05	9.37	0.00	0.35	0.53
≥4	1.03	0.06	17.62	0.00	0.91	1.14
5.00	1.71	0.08	20.44	0.00	1.54	1.87
Upset						
Discrimination	2.47	0.14	17.37	0.00	2.19	2.75
Difficulty						
≥2	−0.52	0.05	−9.77	0.00	−0.62	−0.41
≥3	0.55	0.05	10.99	0.00	0.45	0.64
≥4	1.18	0.06	18.43	0.00	1.06	1.31
5.00	1.92	0.10	19.87	0.00	1.73	2.11
Guilty						
Discrimination	2.05	0.14	14.57	0.00	1.78	2.33
Difficulty						
≥2	0.49	0.05	9.35	0.00	0.39	0.60
≥3	1.42	0.08	17.34	0.00	1.26	1.58
≥4	1.92	0.11	17.75	0.00	1.70	2.13
5.00	2.67	0.17	16.15	0.00	2.35	3.00
Afraid						
Discrimination	3.28	0.22	14.84	0.00	2.85	3.71
Difficulty						
≥2	0.24	0.04	5.44	0.00	0.15	0.32
≥3	1.00	0.05	18.38	0.00	0.89	1.11
≥4	1.43	0.07	20.99	0.00	1.30	1.57
5.00	2.03	0.10	20.72	0.00	1.84	2.22
Hostile						
Discrimination	1.70	0.12	14.34	0.00	1.46	1.93
Difficulty						
≥2	0.41	0.06	7.15	0.00	0.29	0.52
≥3	1.41	0.09	15.88	0.00	1.23	1.58
≥4	2.19	0.13	16.36	0.00	1.93	2.45
5.00	3.14	0.22	14.23	0.00	2.70	3.57
Irritable						
Discrimination	1.89	0.11	16.95	0.00	1.67	2.11
Difficulty						
≥2	−0.70	0.06	−11.23	0.00	−0.82	−0.58
≥3	0.53	0.06	9.67	0.00	0.43	0.64
≥4	1.32	0.08	17.05	0.00	1.16	1.47
5.00	2.12	0.12	17.94	0.00	1.89	2.35
Ashamed						
Discrimination	2.29	0.16	14.77	0.00	1.99	2.60
Difficulty						
≥2	0.52	0.05	10.32	0.00	0.43	0.62
≥3	1.36	0.07	18.27	0.00	1.22	1.51
≥4	1.88	0.10	18.89	0.00	1.68	2.07
5.00	2.66	0.16	16.78	0.00	2.35	2.97
Nervous						
Discrimination	2.47	0.15	17.01	0.00	2.19	2.76
Difficulty						
≥2	−0.27	0.05	−5.41	0.00	−0.36	−0.17

(Continued)

TABLE 5 | Continued

	Coef.	*SE*	*Z*	*P*	95% CI	
≥3	0.66	0.05	12.80	0.00	0.56	0.76
≥4	1.29	0.07	18.98	0.00	1.16	1.42
5.00	1.97	0.10	19.86	0.00	1.77	2.16
Jittery						
Discrimination	1.53	0.11	14.21	0.00	1.32	1.74
Difficulty						
≥2	0.27	0.06	4.63	0.00	0.16	0.39
≥3	1.39	0.09	14.89	0.00	1.21	1.58
≥4	2.01	0.13	15.75	0.00	1.76	2.26
5.00	3.01	0.21	14.64	0.00	2.61	3.42
Scared						
Discrimination	3.49	0.24	14.34	0.00	3.01	3.97
Difficulty						
≥2	0.26	0.04	6.15	0.00	0.18	0.35
≥3	1.14	0.06	19.95	0.00	1.03	1.25
≥4	1.49	0.07	21.72	0.00	1.36	1.63
5.00	1.94	0.09	21.06	0.00	1.76	2.12

of option 1 and 7 for this specific item are equal and very high (see **Figure 11** for more details).

The item information function analyses, **Figure 12A**, showed that items 2 ("Most aspects of my life are in balance") and item 3 ("I am in harmony") have the two highest discrimination estimates and provide more information than the remaining items, while items 4 ("I accept the various conditions of my life") and 5 ("I fit in well with my surroundings") provide lesser information. In general, the results suggest that a lot of information of the true range of harmony in life is covered between low (θ = −2.00) up to high (θ = 2.00) values. For instance, we showed that we get reliable information at θ = 0.00 at about 7.20 from item 2 ("Most aspects of my life are in balance"), at about 7.00 from item 3 ("I am in harmony"), at about 4.80 from item 1 ("My lifestyle allows me to be in harmony") and at about 1.50 from both item 4 ("I accept the various conditions of my life") and 5 ("I fit in well with my surroundings") (see **Figure 12B**, test information function and the standard error, that is, measurement error). This means that the Harmony in Life Scale has good reliability and small standard error in this range. The test information highest is located at about −0.30 (Theta), hence indicating that this score has the smallest standard error and it provides the most information of the scale. However, there is almost no reliable information about below −2.40 and about above 2.50 (Theta) and the standard error increases quickly for both smaller and larger Theta values. The reliability for different levels of harmony in life are shown in **Table 8**. These results showed that the scales reliability is very strong at θ = −2.00, θ = −1.00, θ = 0.00, θ = 1.00, and θ = 2.00 (between 0.87 and 0.96), but weak (0.50) at θ = −3.00 and very week (0.32) at θ = 3.00.

The Harmony in Life Scale has five items with a seven-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*), so the expected scores range from 5 to 35. Our results showed that the expected score for participants that have harmony in life at the level −1.96 (Theta) and below is 7.44 and less. Hence, these participants are most likely to choose the answer coded 1 on most items. With critical values (−1.96 and 1.96) coding to the standard normal distribution, we can expect 95% of randomly selected participants have a score between 7.44 and 33.9 (see **Figure 12**).

Convergent and Discriminant Validity

Finally, in order to test *convergent* and *discriminant validity* we investigated the Pearson correlations between the different scales. The Satisfaction with Life Scale ($r = 0.30$; $p < 0.001$) and Harmony in Life Scale ($r = 0.46$; $p < 0.001$) were positively and significantly correlated with the positive affect scale. Conversely, the Satisfaction with Life Scale ($r = -0.30$; $p < 0.001$) and Harmony in Life Scale ($r = -0.38$; $p < 0.001$) were negatively and significantly correlated with the negative affect scale. Moreover, positive and negative were negatively and significantly correlated with each other ($r = -0.25$; $p < 0.001$). Hence, there is sufficient convergent and discriminant validity.

DISCUSSION

Since measures used to assess subjective well-being are self-reports, often validated only using CTT methodology, our aim was to focus on the psychometric properties of three subjective well-being measures using IRT methods. More specifically, we used GRM to validate and suggest psychometric modifications to the Positive Affect Negative Affect Schedule, the Satisfaction with Life, and the Harmony in Life Scale. We argued that health is biopsychosocial and suggested that these three scales operationalize a biopsychosocial model of subjective well-being (cf. affect-cognition-behavior). Since past research shows that each scale has a unidimensional structure, our first step here was to validate each scale at the item level.

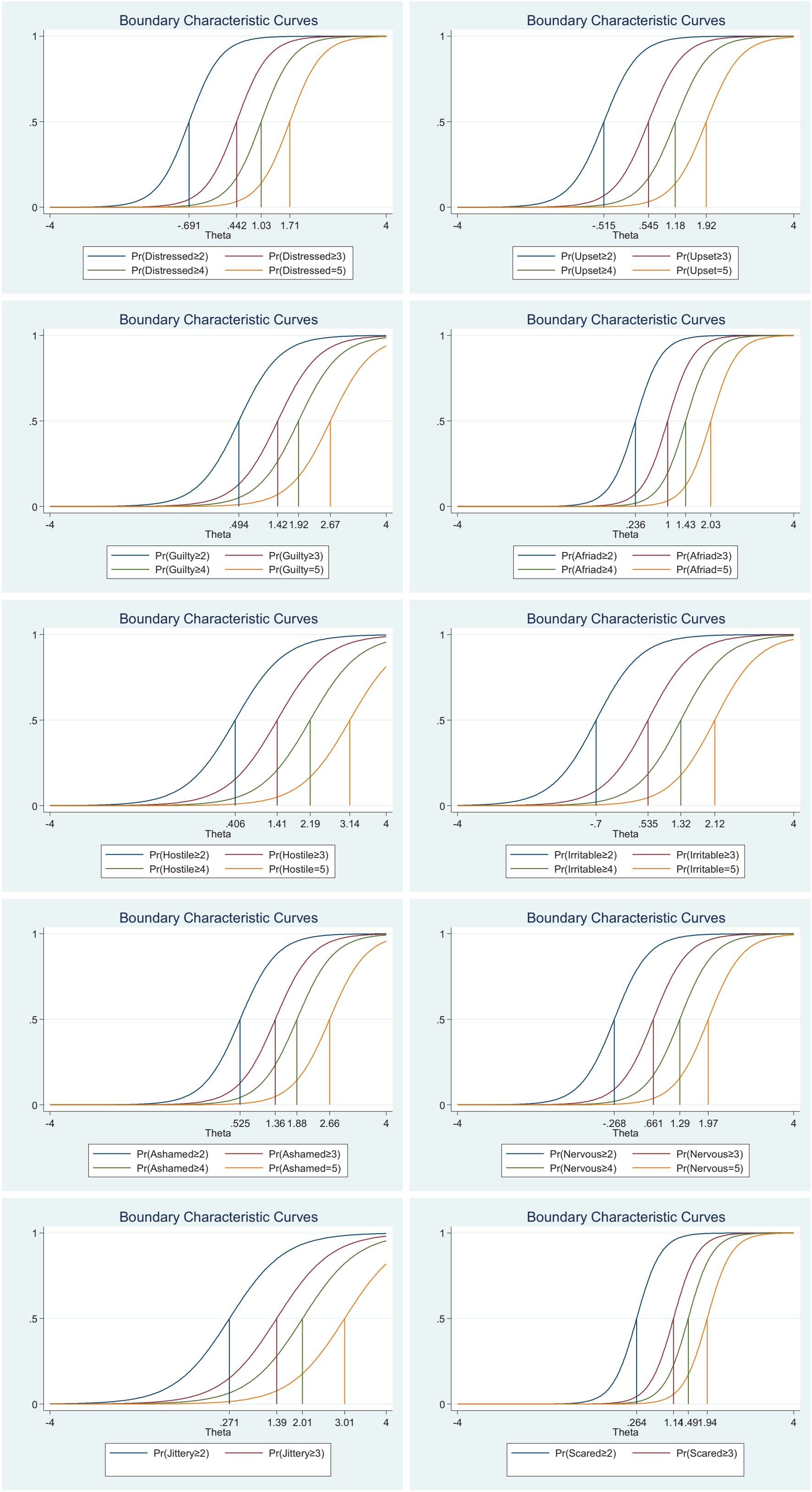


FIGURE 4 | Boundary characteristic curves for each item of the negative affect scale of the Positive Affect Negative Affect Schedule (*N* = 1000).

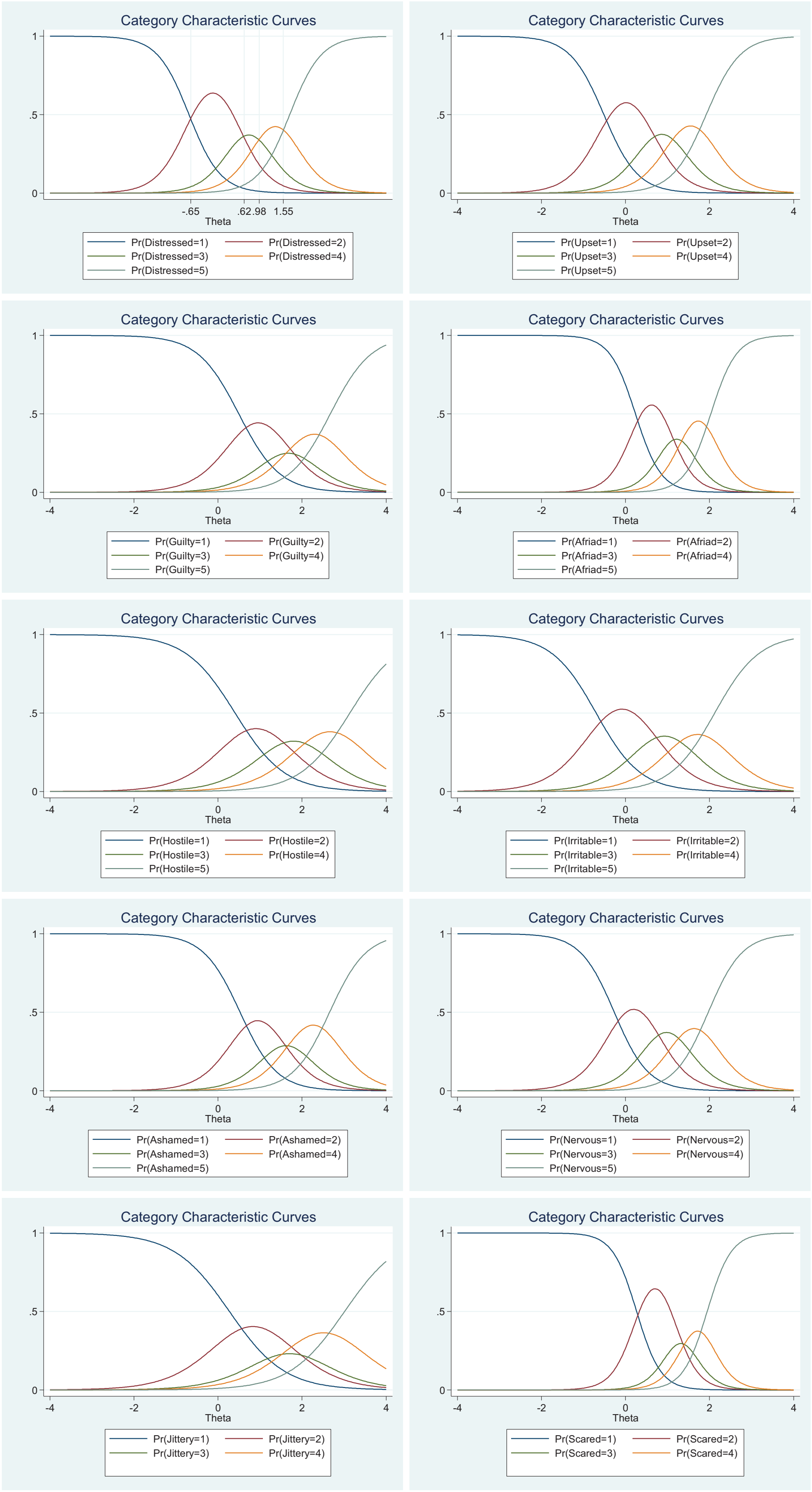


FIGURE 5 | Category characteristic curves for the items in the negative affect scale of the Positive Affect Negative Affect Schedule (*N* = 1000).

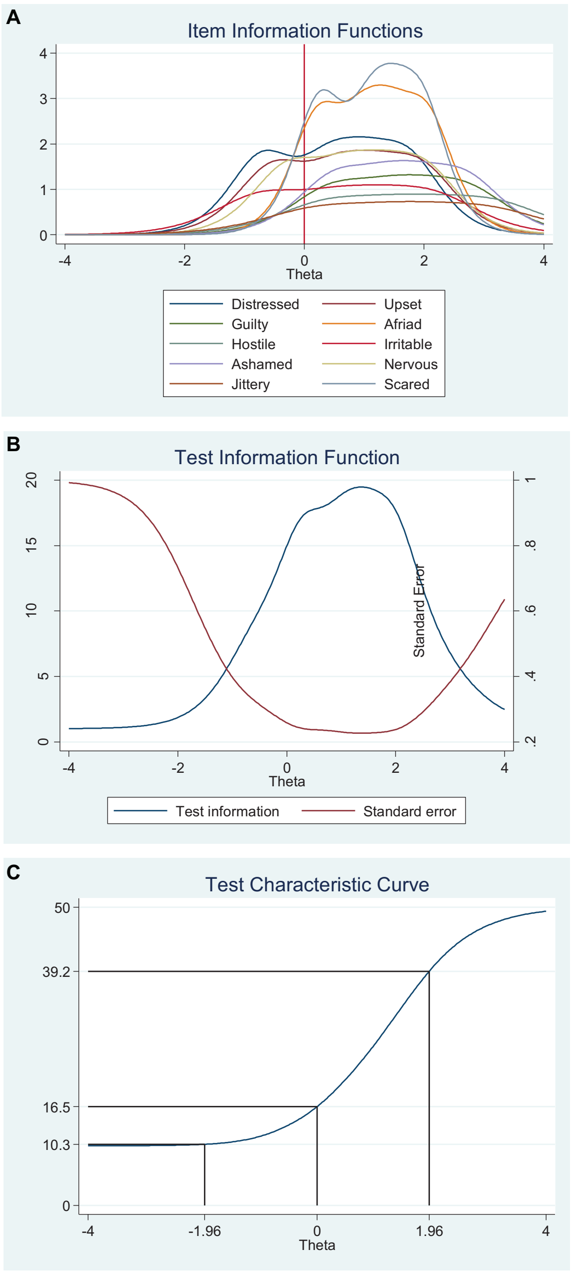


FIGURE 6 | Items information function graphs for graded response with vertical line at θ = 0 **(A)** and information and standard error graph for graded response **(B)** and test characteristic curve **(C)** for the whole negative affect scale of the Positive Affect Negative Affect Schedule (*N* = 1000).

The Affective or Biological Component: Positive Affect Negative Affect Schedule

The results showed that, despite having a varied frequency distribution, all items measuring positive and negative affect had high discrimination values (Alphas from 1.37 to 2.65 for positive affect and 1.53 to 3.49 for negative affect). In other words, indicating that all items in the scales can differentiate well between persons with high and low levels of positive and negative affect. Moreover, certain items had different difficulty parameter (Beta) for each specific response option. For example, participants were relatively less prone to choose the highest point in the Likert scale (5 = *Extremely*) when evaluating to which extent they have felt *alert* and *hostile* and more prone to choose this response when evaluating to which extent they have felt *determined* and *distressed*. In addition, participants were relatively more prone to choose the lowest point in the Likert scale (1 = *Very slightly or not at all*) when evaluating to which extent they have felt *proud* and *ashamed* and less prone to choose this response when evaluating to which extent they have felt *interested* and *irritable*. In this context, validation studies using CTT (e.g., Crawford and Henry, 2004) suggest that best-fitting models are achieved by specifying correlations between error in items closely related to each other in meaning, for example, Interested-Alert-Attentive, Proud-Determined, Excited-Enthusiastic-Inspired, Distressed-Upset, Guilty-Ashamed, Scared-Afraid, Nervous-Jittery, Hostile-Irritable. Therefore, researchers have suggested that these covariances, that form constellations of items, indicate the possibility of item reduction without serious repercussions on the content domain or internal consistency reliability of the scales (e.g., Thompson, 2007, 2017). For instance, the CFA analysis conducted in our study to replicate the unidimensionality of the scales showed similar covariance between errors regarding Alert-Attentive and even more for the negative affect scale. Nevertheless, our IRT results suggest that choosing which item to delete is more complex than just looking at the covariances between items closely related in meaning. For instance, for the constellation Proud-Determined, "Determined" was here shown to cover the highest levels of the Likert scale and "Proud" to be able to cover the lowest levels and for the constellation Guilty-Ashamed, we need to consider that, "Guilty" covers the lowest, while "Distressed" from the constellation Distressed-Upset covers the highest levels of the Likert scale. So, deleting any of these two items has repercussions for which item should be kept from other item constellations, since the scale will need an item that covers for lower/higher values. In other words, in contrast to what is implied by CTT models, the deletion of any of these items will have repercussions on the psychometric properties of the scale.

Furthermore, the items "Enthusiastic," "Excited," "Proud," "Interested," "Strong," "Scared," "Afraid," "Distressed," "Irritable," and "Nervous" provided satisfactory information values and seem useful to differentiate well between respondents. More specifically, the items "Enthusiastic," "Excited," "Scared," and "Afraid" had two of the highest discrimination estimates (Alpha) and provided more information than all the remaining items, while the items "Alert," "Attentive," "Jittery," and "Hostile" provided lesser information. Moreover, the test's highest amount of information was located within positive affect levels from −2.50 up to about 2.30 and within negative affect levels from −1.00 up to about 3.50 (Theta). However, even if some items, like "Alert" and "Attentive," had good discrimination values (Alpha), the information value was low. Hence, suggesting again that the item "Alert" can be removed, or even better, replaced with an equally good discriminating item that better covers lower values

TABLE 6 | The frequency distributions of the items in the Satisfaction with Life Scale (*N* = 500).

Item	Points of Likert scale						
	1	2	3	4	5	6	7
In most ways my life is close to my ideal							
Frequency	61	63	61	52	128	100	35
Percent	12.20	12.60	12.20	10.40	25.60	20.00	7.00
Cumulating	12.20	24.80	37.00	47.40	73.00	93.00	100.00
The conditions of my life are excellent							
Frequency	45	47	68	61	115	125	39
Percent	9.00	9.40	13.60	12.20	23.00	25.00	7.80
Cumulating	9.00	18.40	32.00	44.20	67.20	92.20	100.00
I am satisfied with my life							
Frequency	58	42	54	43	108	137	58
Percent	11.60	8.40	10.80	8.60	21.60	27.40	11.60
Cumulating	11.60	20.00	30.80	39.40	61.00	88.40	100.00
So far I have gotten the important things I want in life							
Frequency	45	44	70	50	95	134	62
Percent	9.00	8.80	14.00	10.00	19.00	26.80	12.40
Cumulating	9.00	17.80	31.80	41.80	60.80	87.60	100.00
If I could live my life over, I would change almost nothing							
Frequency	77	85	82	50	84	70	52
Percent	15.40	17.00	16.40	10.00	16.80	14.00	10.40
Cumulating	15.40	32.40	48.80	58.80	75.60	89.60	100.00

of the scale and provides more information for the whole ideal range (Theta −3.00 to +3.00). Last but not the least, reliability was relatively week for responses were Theta is at or above 3.00 for positive affect and at and below −2.00 for negative affect, suggesting that the standard error increases quickly for higher values of positive and negative affect. Hence, choosing deletion or addition of items that cover the ideal range of affect (Theta −3.00 to +3.00) needs to consider items that complement each other in their difficulty and discrimination levels. In general, in addition to what is implied by CTT models, the information provided in our study should be useful for further development of the scales of the Positive Affect Negative Affect Schedule.

The Cognitive or Psychological Component: The Satisfaction With Life Scale

As for the results of the affective component measure, all items of the Satisfaction with Life Scale had a varied frequency distribution and can differentiate well between persons with high and low levels of the latent score of life satisfaction (Alphas from 1.74 to 4.50). Moreover, certain items had different difficulty parameter (Beta) for each specific response option. For example, participants were relatively less prone to choose the highest point in the Likert scale (7 = *Extremely agree*) when evaluating the statement in item 5 ("If I could live my life over, I would change almost nothing") and more prone to choose this response when evaluating the statement in item 3 ("I am satisfied with my life"). In this context, studies using CTT methods suggest that the fifth item of the scale shows often lower factor loadings and item-total correlations than the first four items of the scale (e.g., Senécal et al., 2000; see also our CFA analysis for this scale, which replicate these results in the **Supplementary Material**). We agree with Pavot and Diener (2008) who suggested that, because this specific item strongly implies a summary evaluation over past years, responses to it might involve a different cognitive recollection than the responses to items that imply a focus on, for example, a temporal summation (e.g., Item 3: "I am satisfied with my life"). Moreover, as in our study, the few studies using IRT methodology indicate that the fifth item is somewhat distinct from the other four items of the scale, something that makes comparisons based on raw scores in certain populations misleading (e.g., Vittersø et al., 2005; Oishi, 2006). In addition, participants were relatively more prone to choose the lowest point in the Likert scale (1 = *Extremely disagree*) when evaluating item 1 ("In most ways my life is close to my ideal"), and less prone to choose this response when evaluating item 4 ("So far I have gotten the important things I want in life"). We interpret this as participants not seeing "get the important things in my life" as equal to being close to their own self-imposed ideal, which per definition is how life satisfaction has been conceptualized (Diener et al., 1985; Pavot and Diener, 1993, 2008). Thus, suggesting that responses to these items will have repercussions on the psychometric properties of the Satisfaction with Life Scale and to comparisons between groups based on raw scores of the scale (cf. Oishi, 2006). In this line, CTT methods suggest that a life satisfaction score of 20 represents the neutral point on the scale, while a scores between 5 and 9 indicates that the respondent is extremely dissatisfied with life, scores from 15 to 19 are interpreted as falling in the slightly dissatisfied range, scores between 21 and 25 represent slightly satisfied, and scores

TABLE 7 | Item response analysis of the Satisfaction with Life Scale (*N* = 500).

	Coef.	*SE*	*Z*	*P*	95% CI	
In most ways my life is close to my ideal						
Discrimination	4.50	0.38	11.82	0.00	3.75	5.24
Difficulty						
≥2	−1.25	0.08	−15.38	0.00	−1.41	−1.09
≥3	−0.73	0.07	−11.21	0.00	−0.86	−0.60
≥4	−0.35	0.06	−6.00	0.00	−0.47	−0.24
≥5	−0.06	0.06	−1.07	0.29	−0.17	0.05
≥6	0.65	0.06	10.20	0.00	0.53	0.78
7	1.57	0.10	16.37	0.00	1.38	1.76
The conditions of my life are excellent						
Discrimination	3.25	0.24	13.66	0.00	2.78	3.72
Difficulty						
≥2	−1.53	0.10	−15.38	0.00	−1.72	−1.33
≥3	−1.01	0.08	−13.04	0.00	−1.16	−0.86
≥4	−0.53	0.07	−8.08	0.00	−0.65	−0.40
≥5	−0.17	0.06	−2.85	0.00	−0.29	−0.05
≥6	0.49	0.07	7.46	0.00	0.36	0.61
7	1.58	0.10	15.30	0.00	1.38	1.78
I am satisfied with my life						
Discrimination	3.93	0.31	12.70	0.00	3.33	4.54
Difficulty						
≥2	−1.32	0.09	−15.44	0.00	−1.49	−1.15
≥3	−0.92	0.07	−12.77	0.00	−1.06	−0.78
≥4	−0.52	0.06	−8.30	0.00	−0.64	−0.40
≥5	−0.27	0.06	−4.48	0.00	−0.38	−0.15
≥6	0.31	0.06	5.14	0.00	0.19	0.43
7	1.28	0.08	15.07	0.00	1.11	1.45
So far I have gotten the important things I want in life						
Discrimination	2.30	0.17	13.58	0.00	1.97	2.63
Difficulty						
≥2	−1.67	0.12	−14.00	0.00	−1.91	−1.44
≥3	−1.12	0.09	−12.31	0.00	−1.30	−0.94
≥4	−0.56	0.07	−7.61	0.00	−0.70	−0.41
≥5	−0.23	0.07	−3.37	0.00	−0.36	−0.10
≥6	0.37	0.07	5.25	0.00	0.23	0.51
7	1.45	0.11	13.58	0.00	1.24	1.66
If I could live my life over, I would change almost nothing						
Discrimination	1.74	0.14	12.79	0.00	1.47	2.01
Difficulty						
≥2	−1.42	0.12	−11.79	0.00	−1.65	−1.18
≥3	−0.61	0.09	−7.17	0.00	−0.78	−0.44
≥4	−0.04	0.08	−0.54	0.59	−0.19	0.11
≥5	0.30	0.08	3.89	0.00	0.15	0.45
≥6	0.96	0.10	9.94	0.00	0.77	1.15
7	1.76	0.14	12.50	0.00	1.48	2.04

between 31 and 35 indicate that the respondent is extremely satisfied with life (Pavot and Diener, 2008). In contrast, our IRT analysis suggest a score of 22.30 as the neutral point of the scale and that 95% of the participants are within scores 6.35–33.60. Thus, IRT might be useful to create normative data for this scale and the others.

In general terms, however, item 1 ("In most ways my life is close to my ideal"), item 2 ("The conditions of my life are excellent"), item 3 ("I am satisfied with my life"), and item 4 ("So far I have gotten the important things I want in life") provided satisfactory information values and could differentiate well between respondents. Specifically, item 1 and 3 have the highest discrimination estimates (Alphas) and provide more information than the remaining items. The test's highest amount of information was located within life satisfaction levels from −2.00 up to about 2.00 (Theta). Additionally, although item 5

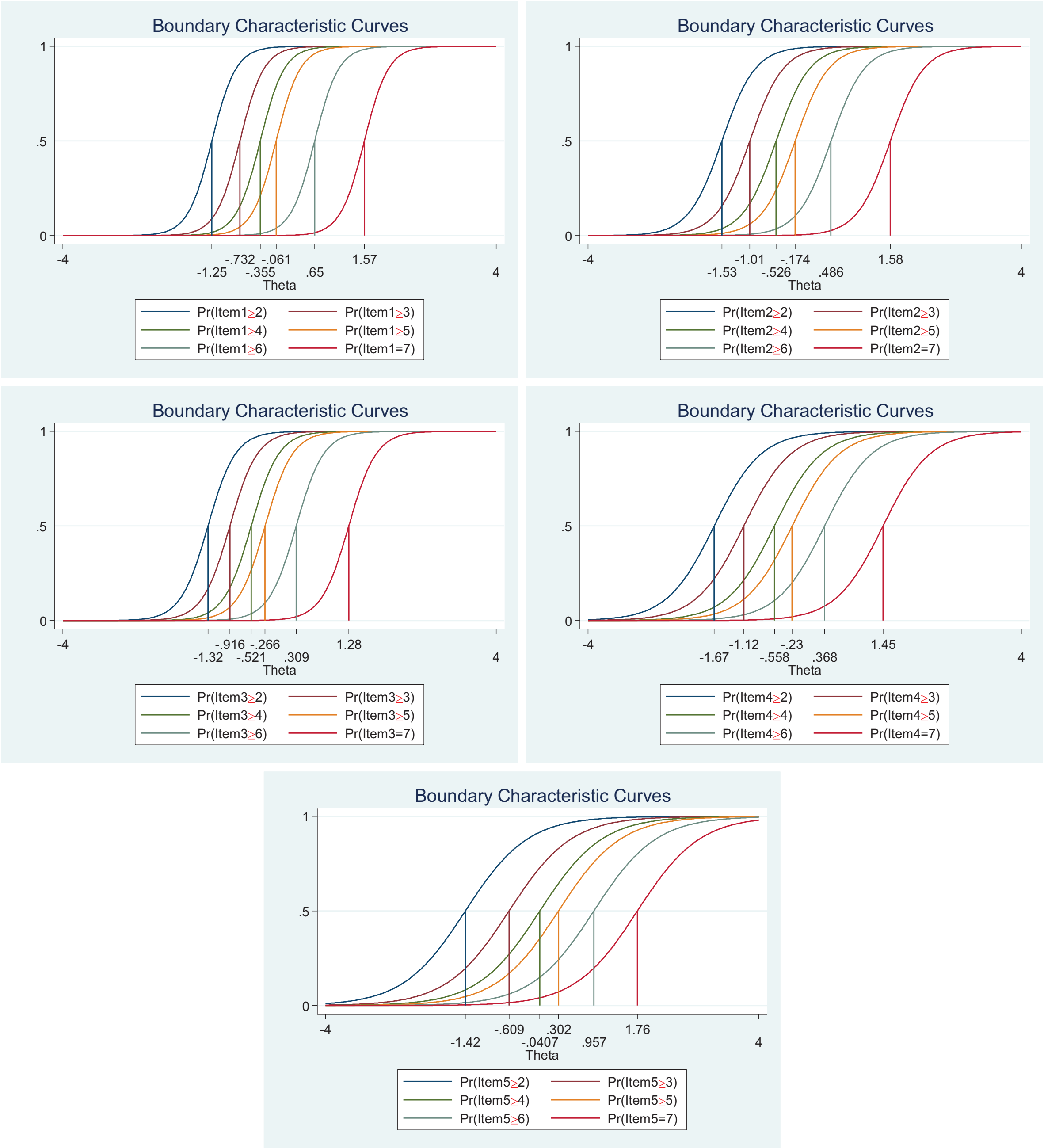


FIGURE 7 | Boundary characteristic curves for each item of the Satisfaction with Life Scale (*N* = 500). Item 1: "In most ways my life is close to my ideal"; Item 2: "The conditions of my life are excellent"; Item 3: "I am satisfied with my life"; Item 4: "So far, I have gotten the important things I want in life"; and Item 5: "If I could live my life over, I would change almost nothing."

had very high discrimination values (Alpha), it provided low information. Hence, reinforcing that item 5 should be removed or modified to develop the psychometric properties of the scale and that there is no reliable information for Theta values at and about below −2.40 and at and about above 2.50. In these specific location coefficients, the standard error increases quickly, thus, the scale's reliability is very weak. The information provided in our study should be useful for further development of the Satisfaction with Life Scale in order to cover the ideal range of the scale (Theta −3.00 to +3.00).

The Behavioral or Social Component: Harmony in Life Scale

As for the results of the other subjective well-being measures, the items of the Harmony in Life Scale showed varied frequency

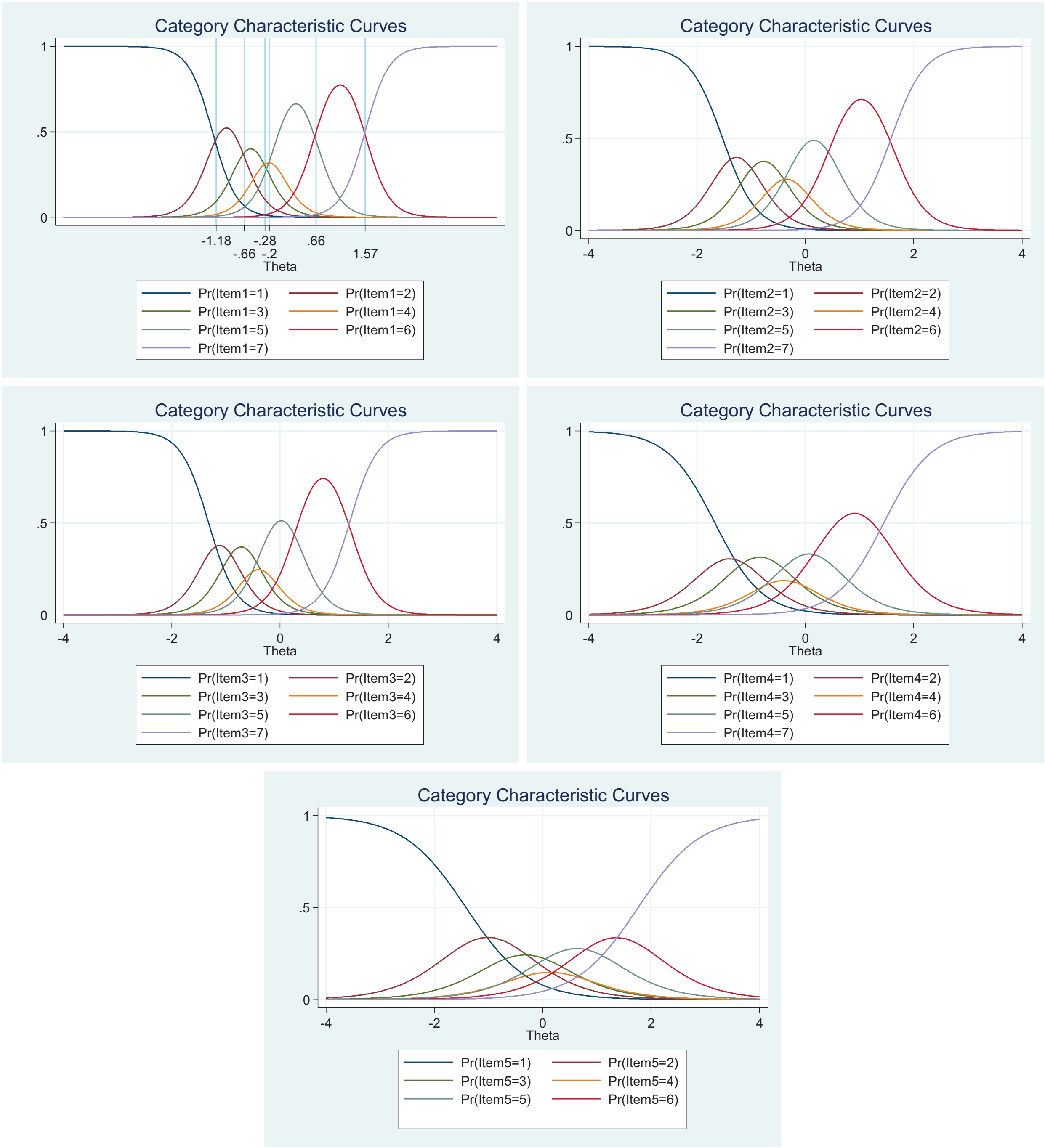


FIGURE 8 | Category characteristic curves for each item of the Satisfaction with Life Scale (*N* = 500). Item 1: "In most ways my life is close to my ideal"; Item 2: "The conditions of my life are excellent"; Item 3: "I am satisfied with my life"; Item 4: "So far, I have gotten the important things I want in life"; and Item 5: "If I could live my life over, I would change almost nothing."

distribution, high discrimination values (Alphas from 2.05 to 5.23) and had different difficulty parameters (Beta) on each specific response option. Here, participants were relatively less prone to choose the highest point in the Likert scale (7 = *Extremely agree*) when evaluating the statement in item 3 ("I am in harmony") and more prone to choose this response when evaluating the statement in item 4 ("I accept the various conditions of my life"). Moreover, participants were relatively more prone to choose the lowest point in the Likert scale (1 = *Extremely disagree*) when evaluating the statement in

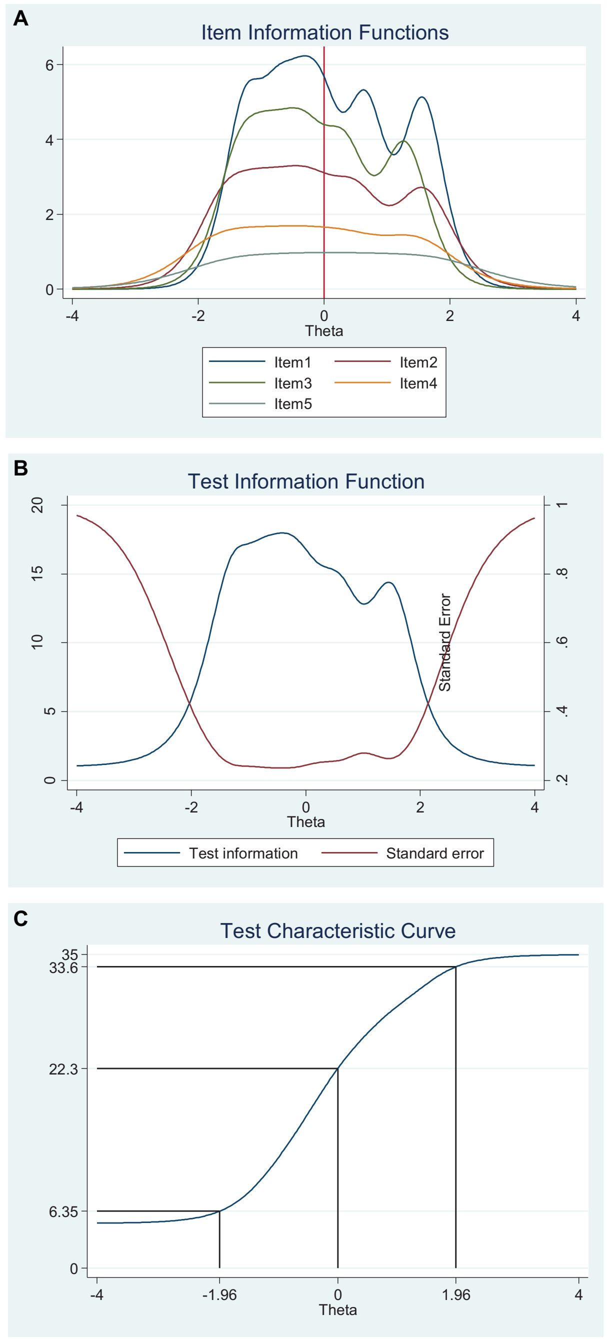


FIGURE 9 | Items information function graphs for graded response with vertical line at θ = 0 **(A)** and information and standard error graph for graded response **(B)** and test characteristic curve **(C)** for the whole Satisfaction with Life Scale (*N* = 500). Note: Item 1: "In most ways my life is close to my ideal"; Item 2: "The conditions of my life are excellent"; Item 3: "I am satisfied with my life"; Item 4: "So far, I have gotten the important things I want in life"; and Item 5: "If I could live my life over, I would change almost nothing."

item 3 ("I am in harmony") and less prone to choose this response when evaluating the statement in item 5 ("I fit in well with my surroundings"). In addition, items 2 ("Most aspects of my life are in balance") and 3 ("I am in harmony") have the highest discrimination estimates (Alpha) and provide more information than the remaining items. These two items together with item 1 ("My lifestyle allows me to be in harmony") provide satisfactory information values, thus, they differentiate well between respondents with high and low levels in harmony in life. Although beyond the scope of our study, we argue that these results reinforce our suggestion about seeing harmony in life as the behavioral or social component of subjective well-being. All relevant items suggest evaluations of behaviors (e.g., "My lifestyle. . .") and evaluations of social interactions between the self and the world around (e.g., ". . .in balance").

In addition, although item 4 ("I accept the various conditions of my life") and 5 ("I fit in well with my surroundings") had very high discrimination values (Alphas), the information that these items cover is low. With regard to item 4, the statement is probably more related to the concept of self-acceptance, rather than harmony *per se*. Self-acceptance has been conceptualized as one sub-trait in the personality trait of Self-directedness (Cloninger, 2004). In other words, even if self-acceptance has been identified as an important trait that promotes well-being, it is a personality trait rather than a construct of subjective well-being. With regard to item 5, perhaps the word "surroundings" is too narrow or confuses the respondents. In other words, "surroundings" might be misinterpreted only as the physical environment or adjacent area, which stands in contrast to both the concept of harmony as the sense of balance and flexibility that an individual experience in relation to the *world* around her (Li, 2008a,b) and the way people describe *how* they pursue harmony—that is, using words that describe more than just adjacent areas, such as, *nature*; in contrast to words people use to describe *how* they pursue life satisfaction, such as, *job* and *house* (see Kjell et al., 2016), which might be what some respondents interpret as their "surroundings." A tentative modification, for example, could be to change the statement in item 5 to "I fit in well with the world around me (e.g., nature)."

Last but not the least, the test's highest amount of information was located within Theta values from −2.00 up to about 2.00 and the scale has almost no reliable information for Theta values at and below −2.40 and at and about above 2.50. At these values, reliability is week and the standard error increases quickly. Hence, as for the other measures, our results are useful for further development of the Harmony in Life Scale in order to cover the ideal range of the scale (Theta −3.00 to +3.00).

Strengths and Limitations of the Present Study

IRT methodology is different from CTT in several important ways (see Hambleton and Swaminathan, 1985; Embretson and Reise, 2000 for details). One of the most significant differences is that in CTT the standard error of measurement is assumed to apply to the whole sample, while in IRT it varies depending on the latent trait score. Using IRT allowed us to consider additional sources of error, such as a person's latent score and person-by-item interaction (Oishi, 2007). In contrast, CTT indices such as *Cronbach's Alpha* do not provide information

TABLE 8 | Reliability of the fitted graded response IRT model of the Satisfaction with Life Scale (*N* = 500) and the Harmony in Life Scale (*N* = 500).

Theta	Satisfaction with Life Scale			Harmony in Life Scale		
	Test Information Function	Test Information Function-*SE*	Reliability IRT GRM	Test Information Function	Test Information Function-*SE*	Reliability IRT GRM
−3.00	1.51	0.81	0.34	2.02	0.70	0.50
−2.00	5.93	0.41	0.83	7.94	0.35	0.87
−1.00	17.21	0.24	0.94	22.88	0.21	0.96
0.00	16.80	0.24	0.94	22.21	0.21	0.95
1.00	12.82	0.28	0.92	11.87	0.29	0.92
2.00	7.43	0.37	0.87	9.68	0.32	0.90
3.00	1.58	0.80	0.37	1.48	0.82	0.32

Cronbach's alpha for the whole scale using CTT were 0.90 for the Satisfaction with Life Scale and 0.92 for the Harmony in Life Scale.

TABLE 9 | The frequency distributions of the items in the Harmony in Life Scale (*N* = 500).

Item	Points of Likert scale						
	1	2	3	4	5	6	7
My lifestyle allows me to be in harmony							
Frequency	35	54	55	71	120	131	34
Percent	7.00	10.80	11.00	14.20	24.00	26.20	6.80
Cumulating	7.00	17.80	28.80	43.00	67.00	93.20	100.00
Most aspects of my life are in balance							
Frequency	44	56	71	46	109	142	32
Percent	8.80	11.20	14.20	9.20	21.80	28.40	6.40
Cumulating	8.80	20.00	34.20	43.40	65.20	93.60	100.00
I am in harmony							
Frequency	53	58	64	55	126	118	26
Percent	10.60	11.60	12.80	11.00	25.20	23.60	5.20
Cumulating	10.60	22.20	35.00	46.00	71.20	94.80	100.00
I accept the various conditions of my life							
Frequency	32	32	33	40	145	157	61
Percent	6.40	6.40	6.60	8.00	29.00	31.40	12.20
Cumulating	6.40	12.80	19.40	27.40	56.40	87.80	100.00
I fit in well with my surroundings							
Frequency	28	27	44	63	118	168	52
Percent	5.60	5.40	8.80	12.60	23.60	33.60	10.40
Cumulating	5.60	11.00	19.80	32.40	56.00	89.60	100.00

whether some items measured some individuals' evaluations of their subjective well-being better than others (Oishi, 2007). As showed here, the first take home message is that there was less reliability for respondents with extreme latent scores of the different components of subjective well-being. Thus, we have suggested the need of modification or addition of specific items in order to improve reliability at the level of the scale, at the item level and at the level of the response scale for each item. This, however, is complex since our results imply that we need to consider both difficulty and discrimination scores and not only covariances between items as suggested by CTT methods. Importantly, in CTT, if two respondents answered the same number of items with the highest/lowest point in the scale, they will get the same total score even if they answered different items as high/low. In contrast, in IRT, the person who answered high to the most "difficult" items (i.e., the items less frequently answered as high) would receive a higher total score than the person who answered high to less difficult items. In addition, since IRT parameters are not sample dependent as in CTT, the score computed in IRT can be compared across different test forms and samples (Oishi, 2007). Hence, the data presented here can be used as normative data for each of the subjective well-being constructs.

Nevertheless, IRT methodology does not address the issue of response style or social desirability (cf. Oishi, 2007). For instance, item difficulty parameters might be influenced by response tendencies such as a mid-point use or extreme scale use (Oishi, 2007; see Chen et al., 1995, for cultural differences in response tendencies). Also, social desirability for specific items might be different across individuals depending on their culture or personal goals and values. For instance, items that we identified as more difficult (e.g., "Proud" in the Positive Affect Negative Affect

TABLE 10 | Item response analysis of the Harmony in Life Scale (*N* = 500).

	Coef.	*SE*	*Z*	*P*	95% CI	
My lifestyle allows me to be in harmony						
Discrimination	4.05	0.30	13.58	0.00	3.47	4.64
Difficulty						
≥2	−1.62	0.10	−16.40	0.00	−1.82	−1.43
≥3	−1.00	0.07	−13.77	0.00	−1.15	−0.86
≥4	−0.58	0.06	−9.26	0.00	−0.71	−0.46
≥5	−0.17	0.06	−2.88	0.00	−0.28	−0.05
≥6	0.48	0.06	7.88	0.00	0.36	0.61
7	1.56	0.10	15.80	0.00	1.36	1.75
Most aspects of my life are in balance						
Discrimination	5.23	0.44	11.82	0.00	4.37	6.10
Difficulty						
≥2	−1.43	0.09	−16.70	0.00	−1.59	−1.26
≥3	−0.88	0.07	−13.23	0.00	−1.01	−0.75
≥4	−0.40	0.06	−6.94	0.00	−0.52	−0.29
≥5	−0.13	0.06	−2.31	0.02	−0.24	−0.02
≥6	0.44	0.06	7.59	0.00	0.33	0.56
7	1.52	0.09	16.31	0.00	1.34	1.70
I am in harmony						
Discrimination	5.08	0.42	12.05	0.00	4.25	5.91
Difficulty						
≥2	−1.33	0.08	−16.33	0.00	−1.49	−1.17
≥3	−0.83	0.07	−12.70	0.00	−0.96	−0.71
≥4	−0.40	0.06	−6.80	0.00	−0.51	−0.28
≥5	−0.09	0.06	−1.59	0.11	−0.20	0.02
≥6	0.58	0.06	9.58	0.00	0.46	0.70
7	1.64	0.10	16.15	0.00	1.44	1.84
I accept the various conditions of my life						
Discrimination	2.05	0.15	13.48	0.00	1.76	2.35
Difficulty						
≥2	−2.03	0.15	−13.86	0.00	−2.32	−1.75
≥3	−1.46	0.11	−13.10	0.00	−1.68	−1.24
≥4	−1.08	0.09	−11.53	0.00	−1.27	−0.90
≥5	−0.74	0.08	−9.04	0.00	−0.90	−0.58
≥6	0.23	0.07	3.20	0.00	0.09	0.37
7	1.49	0.12	12.88	0.00	1.26	1.71
I fit in well with my surroundings						
Discrimination	2.06	0.15	13.59	0.00	1.76	2.36
Difficulty						
≥2	−2.09	0.15	−13.59	0.00	−2.39	−1.79
≥3	−1.55	0.12	−13.20	0.00	−1.78	−1.32
≥4	−1.06	0.09	−11.53	0.00	−1.25	−0.88
≥5	−0.58	0.08	−7.55	0.00	−0.73	−0.43
≥6	0.20	0.07	2.86	0.00	0.06	0.34
7	1.62	0.12	13.29	0.00	1.38	1.86

Scale; item 5, “If I could live my life over, I would change almost nothing,” in the Satisfaction with Life Scale; and item 3, “I am in harmony,” in the Harmony in Life scale) might be seen as socially undesirable to endorse at the highest point of the scales among individuals who value modesty (cf. Oishi, 2007; see Kitayama and Markus, 2000, for cross-cultural studies on happiness). Hence, since we cannot account if our IRT results have been affected by response tendencies and social desirability, our suggestions for modifications should be interpreted as guidelines rather than rules (Oishi, 2007).

Finally, the basic 1-factor CFA model used in this study showed that some fit indexes were slightly outside the traditional acceptable range. The high values of REMSEA, for example, may suggest that the high large residuals in these models could be

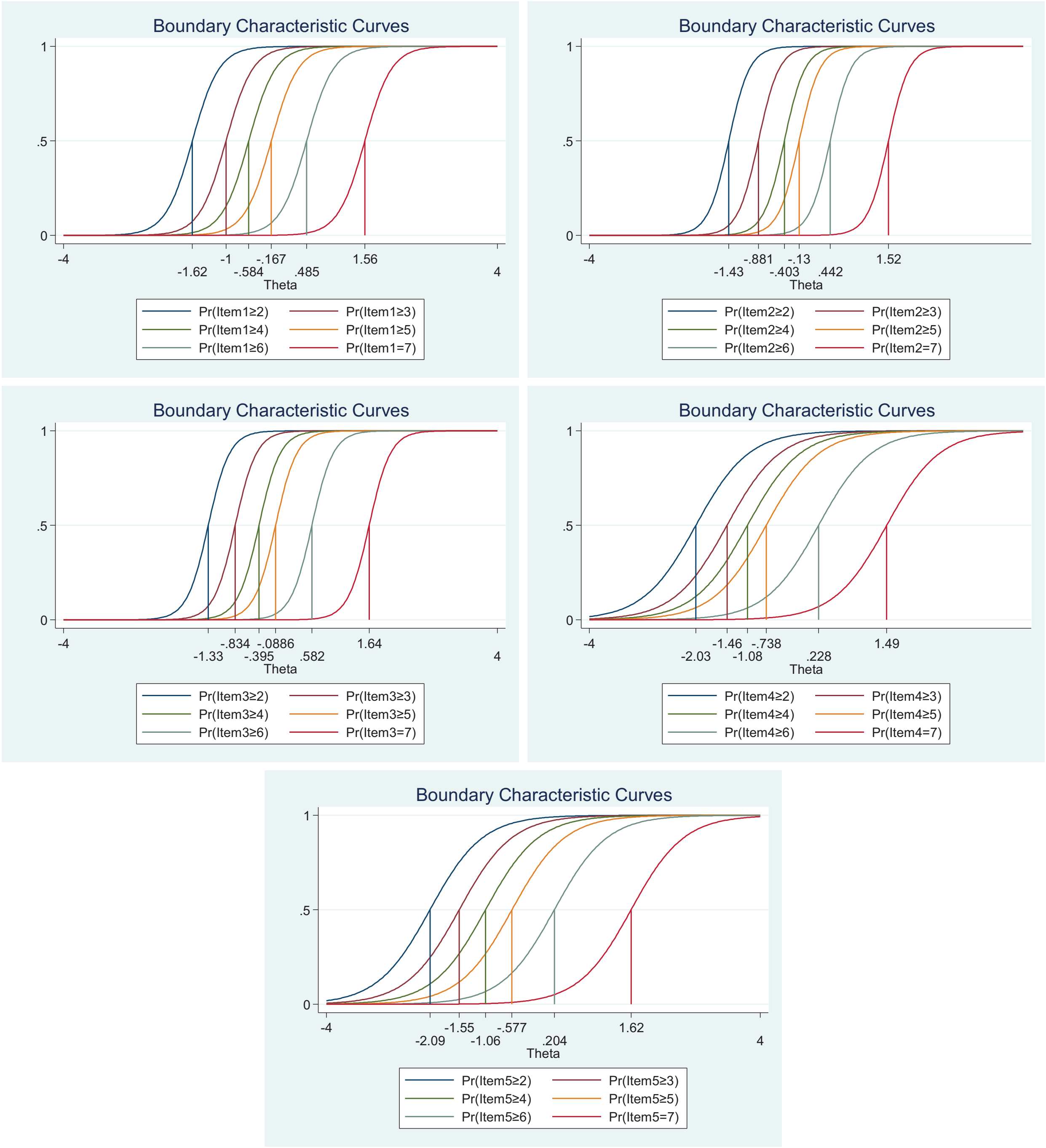


FIGURE 10 | Boundary characteristic curves for each item of the Harmony in Life Scale (*N* = 500). Item 1: "My lifestyle allows me to be in harmony"; Item 2: "Most aspects of my life are in balance"; Item 3: "I am in harmony"; Item 4: "I accept the various conditions of my life"; and Item 5: "I fit in well with my surroundings."

caused by latent multidimensional structure in the data, so this did not allow us to strongly confirm the unidimensionality of our data and cast doubts concerning the remaining dimensionality. Indeed, the result regarding local independence showed that the residuals were mostly significantly correlated, thus indicating also that the data had tendency for multidimensionality. We recommend that further research should apply both *Bifactor analysis* and *multidimensional item response theory* (MIRT) to investigate any multidimensionality regarding these measures. Tentatively, this multidimensionality, we argue, is related to our assumption of a general factor for subjective well-being (i.e., the biopsychosocial model of subjective well-being).

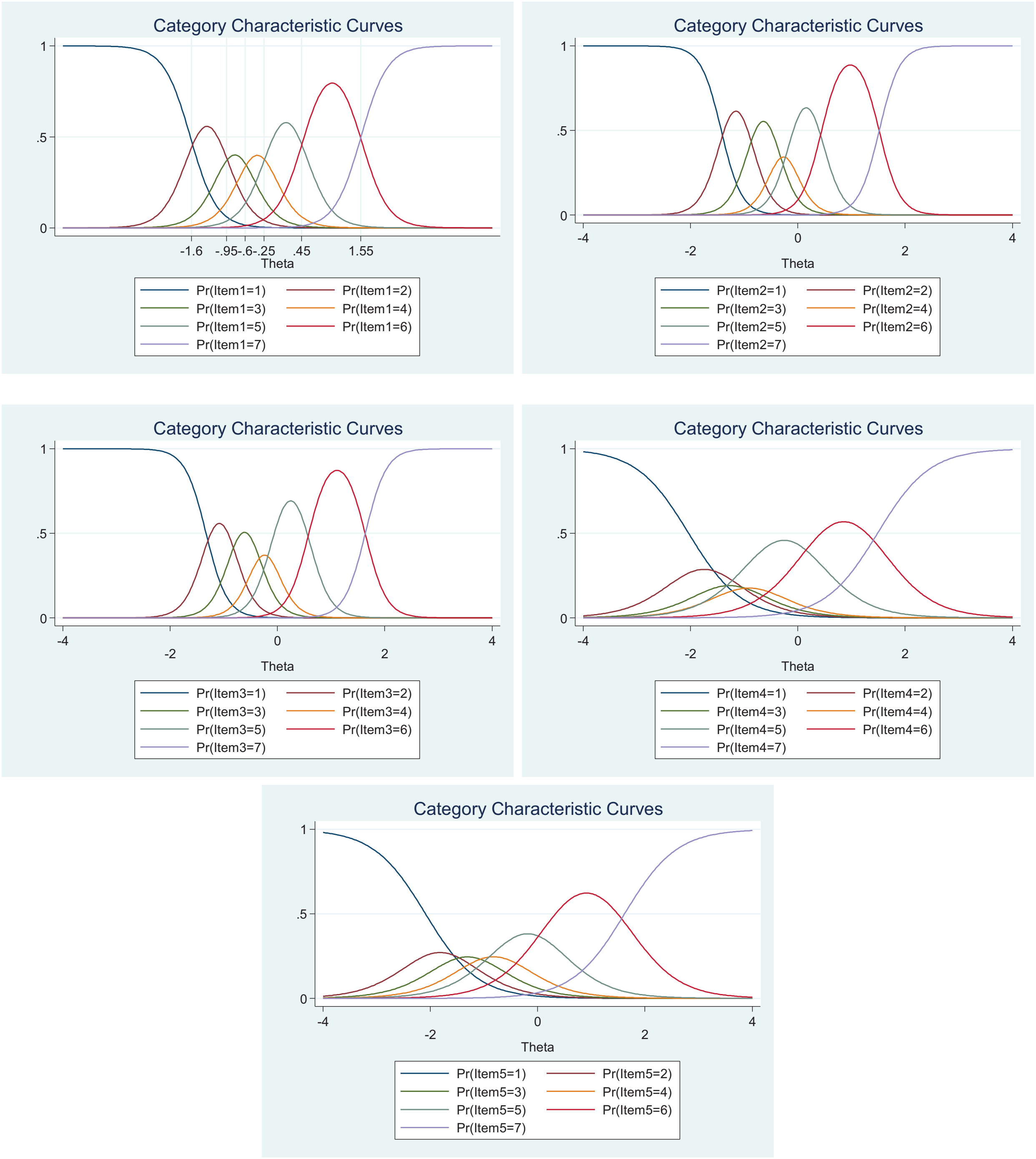


FIGURE 11 | Category characteristic curves for each item of the Harmony in Life Scale (*N* = 500). Item 1: "My lifestyle allows me to be in harmony"; Item 2: "Most aspects of my life are in balance"; Item 3: "I am in harmony"; Item 4: "I accept the various conditions of my life"; and Item 5: "I fit in well with my surroundings."

CONCLUSION AND FINAL REMARKS

In sum, all subjective well-being measures showed varied frequency distribution, high discrimination values (Alphas), and had different difficulty parameters (Beta) on each response options. For example, we identified items that respondents found difficult to endorse at the highest and lowest points of the scale. In addition, while all scales could cover a good portion of the latent trait of subjective well-being, there was less reliability for respondents with scores at the extremes

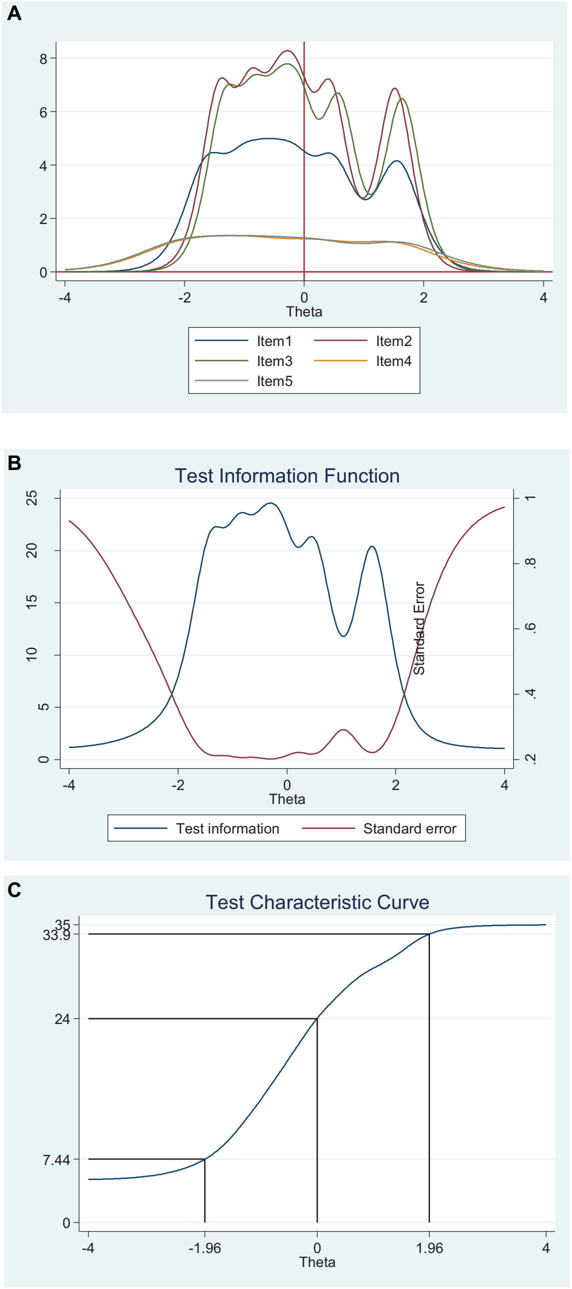


FIGURE 12 | Items information function graphs for graded response with vertical line at θ = 0 **(A)** and information and standard error graph for graded response **(B)** and test characteristic curve **(C)** for the whole Satisfaction with Life Scale (N = 500). Item 1: "My lifestyle allows me to be in harmony"; Item 2: "Most aspects of my life are in balance"; Item 3: "I am in harmony"; Item 4: "I accept the various conditions of my life"; and Item 5: "I fit in well with my surroundings."

of the scales. The affective component seems to be less accurately measured, especially the negative affect scale; while the measures for both the cognitive and social components seem to cover equal range of each latent construct. Although, the scales can be modified by deletion/addition of items that have less/more difficulty to cover the ideal range of subjective well-being, in contrast to what is implied by only focusing on CTT models, the deletion/addition of items needs to consider the additional sources of error we found here. We suggest the replication of our results and the use of other methods or a combination of methods before modifications are implemented. For instance, in recent studies our research team has used artificial intelligence to use words and narratives in relation to the measurement of health (Kjell et al., 2019), subjective well-being (Garcia and Sikström, 2013), happiness (Garcia et al., 2016; Garcia et al., 2020b), and personality (Garcia and Sikström, 2014, 2019; Garcia et al., 2015; Garcia et al., 2020a,c). In one study, the scales used here seem to be related to both different and similar words people use to describe what they *relate* to the concept of happiness and what *makes* them happy (Garcia et al., 2020b). These advanced and innovative techniques can probably be applied to validate items and constructs using peoples own narratives—a method we tentatively call Quantitative Semantics Test Theory, QuSTT. Together with CTT, IRT and qualitative methods, QuSTT might contribute to more rigorous systematic process for item deletion/addition (Sikström and Garcia, 2020). Indeed, many researchers have accurately pointed out the need for improvement in the conceptualization and measurement of well-being using good qualitative, intuitive and quantitative methodology, and consideration and implementation of past research (for critical positive psychology see Brown et al., 2018).

Here, we have argued (see also Garcia et al., 2020b) that these three scales operationalize a biopsychosocial model of subjective well-being (cf. affect-cognition-behavior). We only apply the logic of health being physical, mental, and social to the concept of subjective well-being (cf. World Health Organization [WHO], 1946; Engel, 1980; Cloninger, 2004). Since past research suggests that the proposed scales measuring these constructs are unidimensional, our first step was to validate each scale at the item level. Nevertheless, we need to acknowledge that a holistic view of the human being consists of body, mind and psyche, hence, also spiritual or existential components need to be adapted and tested for a more robust and accurate conceptualization of subjective well-being (Ryff, 1989; cf. Cloninger, 2004; Vaillant, 2008; VanderWeele, 2017; MacDonald, 2018). How this is done, is important because without good measurement to discern the actual concept of subjective well-being, without understanding that it is in itself a complex system (cf. Cloninger, 2004), and without considering how people express their well-being and past relevant research beyond a specific field (e.g., the biopsychosocial model of health), we risk ending up with "quick and dirty measures" that lack a comprehensive theory (cf. Wong and Roy, 2018) and suffer of "jingle-jangle" fallacy[6] (cf. Block, 1995).

> *"Let no one ignorant of geometry enter"*
> *Plato*

[6] Jingle refers to two constructs with equivalent labels that really reflect different phenomena, whereas jangle refers to when one construct is given multiple names (Kelley, 1927; Block, 1995).

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

ETHICS STATEMENT

Ethics approval was not required at the time the research was conducted as per national regulations. The consent of the participants was obtained by virtue of survey completion after they were provided with all relevant information about the research (e.g., anonymity).

AUTHOR CONTRIBUTIONS

AN and DG conceived, designed, and performed the experiments, analyzed the data, wrote the manuscript, prepared the figures and/or tables, and reviewed drafts of the manuscript. KC, BP, and SS reviewed drafts of the manuscript.

FUNDING

The development of this article was funded by a grant from the Swedish Research Council (Dnr. 2015-01229).

ACKNOWLEDGMENTS

We would like to express our gratitude to Professor Magnus Lindwall and Professor Jan Johansson Hanse for their most valuable comments on the first draft of the manuscript. We would like to thank the editor and the reviewers for their comments and suggestions, which substantially improved our article. We would also like to thank Emmy, Sophia, Linnéa, and Embla for their help during the correction of the proofs.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.03036/full#supplementary-material

REFERENCES

Adler, M. G., and Fagley, N. S. (2005). Appreciation: Individual differences in finding value and meaning as a unique predictor of subjective well-being. *J. Pers.* 73, 79–114. doi: 10.1111/j.1467-6494.2004.00305.x

Baker, F. B. (2001). *The Basics of Item Response Theory*. College Park: ERIC Clearinghouse on Assessment and Evaluation.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychol. Bull.* 117, 187–215. doi: 10.1037//0033-2909.117.2.187

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York, NY: John Wiley & Sons, Inc.

Brown, N. J. L., Lomas, T., and Eiroa-Orosa, J. (2018). *The Routledge International Handbook of Critical Positive Psychology*. New York, NY: Routledge.

Browne, M. W., and Cudeck, R. (1993). "Alternative ways of assessing model fit," in *Testing Structural Equation Models*, eds K. A. Bollen, and J. S. Long, (Newbury Park, CA: Sage), 136–162.

Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's mechanical Turk: a new source of inexpensive, yet high-quality, data? *Perspect. Psychol. Sci.* 6, 3–5. doi: 10.1177/1745691610393980

Charter, R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *J. Clin. Exp. Neuropsychol.* 21, 559–566. doi: 10.1076/jcen.21.4.559.889

Chen, C., Lee, S. Y., and Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychol. Sci.* 6, 170–175. doi: 10.1111/j.1467-9280.1995.tb00327.x

Cloninger, C. R. (2004). *Feeling Good: The Science of Well-Being*. Oxford: Oxford University Press.

Cook, K. F., Kallen, M. A., and Amtmann, D. (2009). Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Q. Life Res.* 18, 447–460. doi: 10.1007/s11136-009-9464-4

Crawford, J. R., and Henry, J. D. (2004). The positive and negative affect schedule (PANAS): construct validity, measurement properties and normative data in a large non-clinical sample. *Br. J. Clin. Psychol.* 43, 245–265. doi: 10.1348/0144665031752934

Crocker, P. R. E. (1997). A confirmatory factor analysis of the positive affect negative affect schedule (PANAS) with a youth sport sample. *J. Sport Exer. Psychol.* 19, 91–97. doi: 10.1123/jsep.19.1.91

Davis, L. L. (1992). Instrument review: getting the most from a panel of experts. *Appl. Nurs. Res.* 5, 194–197. doi: 10.1016/s0897-1897(05)80008-4

Delle Fave, A., Brdar, I., Freire, T., Vella-Brodrick, D., and Wissing, M. P. (2011). The eudaimonic and hedonic components of happiness: qualitative and quantitative findings. *Soc. Indic. Res.* 100, 185–207. doi: 10.1007/s11205-010-9632-5

Diener, E. (1984). Subjective well-being. *Psychol. Bull.* 95, 542–575.

Diener, E., and Diener, C. (1996). Most people are happy. *Psychol. Sci.* 7, 181–185. doi: 10.1111/j.1467-9280.1996.tb00354.x

Diener, E., Inglehart, R., and Tay, L. (2013). Theory and validity of life satisfaction scales. *Soc. Indic. Res.* 112, 497–527. doi: 10.1007/s11205-012-0076-y

Diener, E., Lucas, R., Helliwell, J. F., Helliwell, J., and Schimmack, U. (2009). *Well-Being for Public Policy*. Oxford: Oxford University Press.

Diener, E., Lucas, R. E., and Oishi, S. (2018). Advances and open questions in the science of subjective well-being. *Collabra Psychol.* 4:15. doi: 10.1525/collabra.115

Diener, E., and Seligman, M. E. (2002). Very happy people. *Psychol. Sci.* 13, 81–84. doi: 10.1111/1467-9280.00415

Diener, E. D., Emmons, R. A., Larsen, R. J., and Griffin, S. (1985). The satisfaction with life scale. *J. Pers. Assess.* 49, 71–75.

Eid, M., and Larsen, R. J. (eds) (2008). *The Science Of Subjective Well-Being*. New York, NY: Guilford Press.

Embretson, S. E., and Reise, S. P. (2000). *Multivariate Applications Books Series. Item response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Engel, G. L. (1977). The need for a new medical model: a challenge for biomedicine. *Science* 196, 129–136. doi: 10.1126/science.847460

Engel, G. L. (1980). The clinical application of the biopsychosocial model. *Am. J. Psychiatr.* 137, 535–544. doi: 10.1176/ajp.137.5.535

Garcia, D., Anckarsäter, H., Kjell, O. N. E., Archer, T., Rosenberg, P., Cloninger, C. R., et al. (2015). Agentic, communal, and spiritual traits are related to the semantic representation of written narratives of positive and negative life events. *Psychol. Well Being Theor. Res. Pract.* 5, 1–20.

Garcia, D., and Erlandsson, A. (2011). The relationship between personality and subjective well-being: different association patterns when measuring the affective component in frequency and intensity. *J. Happiness Stud.* 12, 1023–1034. doi: 10.1007/s10902-010-9242-6

Garcia, D., Kjell, O. N. E., and Sikström, S. (2016). "A collective picture of what makes people happy: words representing social relationships, not money or material things, are recurrent with the word 'happiness' in online newspapers," in *The Psychology of Social Networking. Identity and Relationships in Online Communities*, eds G. Riva, B. K. Wiederhold, and P. Cipresso, (Cambridge: Academic press), 2.

Garcia, D., Nima, A. A., and Kjell, O. N. E. (2014). The affective profiles, psychological well-being, and harmony: environmental mastery and self-acceptance predict the sense of a harmonious life. *PeerJ* 2:e259. doi: 10.7717/peerj.259

Garcia, D., and Sikström, S. (2013). Quantifying the semantic representations in adolescents' memories of positive and negative life events. *J. Happiness Stud.* 14, 1309–1323. doi: 10.1007/s10902-012-9385-8

Garcia, D., and Sikström, S. (2014). The dark side of Facebook – dark triad of personality predicts semantic representation of status updates. *Pers. Indiv. Dif.* 67, 92–94. doi: 10.1016/j.paid.2013.10.001

Garcia, D., and Sikström, S. (2019). "The ten words personality inventory (10WPI)," in *Encyclopedia of Personality and Individual Differences*, eds V. Zeigler-Hill, and T. Shackelford, (Cham: Springer), 1–6. doi: 10.1007/978-3-319-28099-8_2314-1

Garcia, D., Cloninger, K. M., Sikström, S., Anckarsäter, H., and Cloninger, C. R. (2020a). "A ternary model of personality: temperament, character, and identity," in *Statistical Semantics: Methods and Applications*, eds S. Sikström, and D. Garcia, (Cham: Springer).

Garcia, D., Nima, A. A., Kjell, O. N. E., Granjard, A., and Sikström, S. (2020b). "The (Mis)measurement of happiness: words we associate to happiness (semantic memory) and narratives of what makes us happy (Episodic Memory)," in *Statistical Semantics: Methods and Applications*, eds S. Sikström, and D. Garcia, (Cham: Springer).

Garcia, D., Rosenberg, P., and Sikström, S. (2020c). "Dark identity: distinction between malevolent character traits through self-descriptive language," in *Statistical Semantics: Methods and Applications*, eds S. Sikström, and D. Garcia, (Cham: Springer).

Gaudreau, P., Sanchez, X., and Blondin, J.-P. (2006). Positive and negative affective states in a performance-related setting: testing the factorial structure of the PANAS across two samples of French-Canadian participants. *Eur. J. Psychol. Assess.* 22, 240–249. doi: 10.1027/1015-5759.22.4.240

Hambleton, R. K., and Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston: Kluwer-Nijhoff.

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *Br. J. Math. Stat. Psychol.* 63, 395–416. doi: 10.1348/000711009X466835

Kelley, T. L. (1927). *Interpretation of Educational Measurements*. New York, NY: World Book Co.

Kercher, K. (1992). Assessing subjective well-being in the old-old: The PANAS as a measure of orthogonal dimensions of positive and negative affect. *Res. Aging* 14, 131–168. doi: 10.1177/0164027592142001

Killgore, W. D. S. (2000). Evidence for a third factor on the positive and negative affect schedule in a college student sample. *Percept. Mot. Skills* 90, 147–152. doi: 10.2466/pms.2000.90.1.147

Kitayama, S., and Markus, H. R. (2000). "The pursuit of happiness and the realization of sympathy: Cultural patterns of self, social relations, and well-being," in *Cultural and Subjective Well-Being* eds E. Diener, and E. M. Suh, (Cambridge, MA: MIT press).

Kjell, O. N., Kjell, K., Garcia, D., and Sikström, S. (2019). Semantic measures. *Psychol. Methods* 24, 92–115. doi: 10.1037/met0000191

Kjell, O. N., Nima, A. A., Sikström, S., Archer, T., and Garcia, D. (2013). Iranian and Swedish adolescents: differences in personality traits and well-being. *PeerJ* 1:e197. doi: 10.7717/peerj.197

Kjell, O. N. E. (2018). *Conceptualizing and Measuring Well-Being Using Statistical Semantics and Numerical Rating Scales*. Doctoral Thesis, Lund University, Lund.

Kjell, O. N. E., Daukantaitë, D., Hefferon, K., and Sikström, S. (2016). The harmony in life scale complements the satisfaction with life scale: expanding the conceptualization of the cognitive component of subjective well-being. *Soc. Indic. Res.* 126, 893–919. doi: 10.1007/s11205-015-0903-z

Kline, R. B. (2010). *Principles and Practice of Structural Equation Modeling*, 3rd Edn, New York, NY: Guilford Press.

Krohne, H. W., Egloff, B., Kohlmann, C. W., and Tausch, A. (1996). Investigations with a German version of the positive and negative affect schedule (PANAS). *Diagnostica* 42, 139–156.

Leue, A., and Lange, S. (2011). Reliability generalization: an examination of the positive affect and negative affect schedule. *Assessment* 18, 487–501. doi: 10.1177/1073191110374917

Li, C. (2008a). The ideal of harmony in ancient Chinese and greek philosophy. *Dao* 7, 81–98. doi: 10.1007/s11712-008-9043-3

Li, C. (2008b). The philosophy of harmony in classical confucianism. *Philos. Compass* 3, 423–435. doi: 10.1111/j.1747-9991.2008.00141.x

Lyubomirsky, S. (2008). *The How of Happiness: A Scientific Approach to Getting the Life You Want*. City of Westminster: Penguin.

MacDonald, D. A. (2018). "Taking a closer look at well-being as a scientific construct. Delineating its conceptual nature and boundaries in relation to spirituality and existential functioning," in *The Routledge International Handbook of Critical Positive Psychology*, eds N. J. L. Brown, T. Lomas, and J. Eiroa-Orosa, (New York, NY: Routledge), 26–52. doi: 10.4324/9781315659794-5

Mackinnon, A., Jorm, A. F., Christensen, H., Korten, A. E., Jacomb, P. A., and Rodgers, B. (1999). A short form of the positive and negative affect schedule: evaluation of factorial validity and invariance across demographic variables in a community sample. *Pers. Individ. Diff.* 27, 405–416. doi: 10.1016/s0191-8869(98)00251-7

Mehrabian, A. (1997). Comparison of the PAD and PANAS as models for describing emotions and for differentiating anxiety from depression. *J. Psychopathol. Behav. Assess.* 19, 331–357. doi: 10.1007/bf02229025

OECD (2013). *OECD Guidelines on Measuring Subjective Well-Being*. Paris: OECD Publishing, doi: 10.1787/9789264191655-en

Oishi, S. (2006). The concept of life satisfaction across cultures: an IRT analysis. *J. Res. Pers.* 40, 411–423. doi: 10.1016/j.jrp.2005.02.002

Oishi, S. (2007). "The application of structural equation modeling and item response theory to cross-cultural positive psychology research," in *Series in Positive Psychology. Oxford Handbook of Methods In Positive Psychology*, eds A. D. Ong, and M. H. M. van Dulmen, (New York, NY: Oxford University Press), 126–138.

Pavot, W. (2018). "The cornerstone of research on subjective well-being: valid assessment methodology," in *Handbook of Well-Being. Noba Scholar Handbook Series: Subjective Well-Being*, eds E. Diener, S. Oishi, and L. Tay, (Salt Lake City, UT: DEF Publishers).

Pavot, W., and Diener, E. (1993). Review of the satisfaction with life scale. *Psychol. Assess.* 5, 164–172.

Pavot, W., and Diener, E. (2008). The satisfaction with life scale and the emerging construct of life satisfaction. *J. Posit. Psychol.* 3, 137–152. doi: 10.1080/17439760701756946

Pires, P., Filgueiras, A., Ribas, R., and Santana, C. (2013). Positive and negative affect schedule: psychometric properties for the Brazilian Portuguese version. *Span. J. Psychol.* 16:E58.

Rand, D. G. (2012). The promise of mechanical turk: how online labor markets can help theorists run behavioral experiments. *J. Theor. Biol.* 299, 172–179. doi: 10.1016/j.jtbi.2011.03.004

Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *J. Pers. Soc. Psychol.* 57, 1069–1081. doi: 10.1037//0022-3514.57.6.1069

Sattelmayer, M., Hilfiker, R., Luomajoki, H., and Elsig, S. (2017). Use of Rasch analysis to investigate structural validity of a set of movement control tests for the neck. *Musculoskelet. Sci. Pract.* 27, 131–136. doi: 10.1016/j.math.2016.07.006

Senécal, C., Nouwen, A., and White, D. (2000). Motivation and dietary self-care in adults with diabetes: are self-efficacy and autonomous self-regulation complementary or competing constructs? *Health Psychol.* 19, 452–457. doi: 10.1037/0278-6133.19.5.452

Sikström, S., and Garcia, D. (2020). *Statistical Semantics: Methods and Applications*. Cham: Springer.

Singh, K., Mitra, S., and Khanna, P. (2016). Psychometric properties of hindi version of peace of mind, harmony in life and sat-chit-ananda scales. *Indian J. Clin. Psychol.* 43, 58–64.

Steger, M. F., Frazier, P., Oishi, S., and Kaler, M. (2006). The meaning in life questionnaire: Assessing the presence of and search for meaning in life. *J. Counsel. Psychol.* 53, 80–93. doi: 10.1080/00223891.2013.765882

Stepp, S. D., Yu, L., Miller, J. D., Hallquist, M. N., Trull, T. J., and Pilkonis, P. A. (2012). Integrating competing dimensional models of personality: linking the SNAP, TCI, and NEO using item response theory. *Pers. Disord.* 3:107. doi: 10.1037/a0025905

Terraciano, A., McCrae, R. R., and Costa, P. T. Jr. (2003). Factorial and construct validity of the italian positive and negative affect schedule (PANAS). *Eur. J. Psycholo. Assess.* 19, 131–141. doi: 10.1027//1015-5759.19.2.131

Thompson, E. R. (2007). Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). *J. Cross Cult. Psychol.* 38, 227–242. doi: 10.1177/0022022106297301

Thompson, E. R. (2017). "Positive and negative affect schedule (PANAS)," in *Encyclopedia of Personality and Individual Differences*, eds V. Zeigler-Hill, and T. K. Shackelford, (Berlin: Springer).

Vaillant, G. E. (2008). *Spiritual Evolution*. New York, NY: Doubleday.

VanderWeele, T. J. (2017). On the promotion of human flourishing. *Proc. Natl. Acad. Sci. U.S.A.* 114, 8148–8156.

Vassar, M. (2008). A note on the score reliability for the satisfaction with life scale: an RG study. *Soc. Indic. Res.* 86, 47–57. doi: 10.1007/s11205-007-9113-7

Vittersø, J., Biswas-Diener, R., and Diener, E. (2005). The divergent meanings of life satisfaction: Item response modeling of the satisfaction with life scale in Greenland and Norway. *Soc. Indic. Res.* 74, 327–348. doi: 10.1007/s11205-004-4644-7

Watson, D., and Clark, L. A. (1994). *The PANAS-X: Manual for the Positive Affect and Negative Affect Schedule–Expanded Form*. Iowa City: University of Iowa.

Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *J. Pers. Soc. Psychol.* 54, 1063–1070. doi: 10.1037//0022-3514.54.6.1063

Wong, P. T. P., and Roy, S. (2018). "Critique of positive psychology and positive interventions," in *The Routledge International Handbook of Critical Positive Psychology*, eds N. J. L. Brown, T. Lomas, and J. Eiroa-Orosa, (New York, NY: Routledge), 142–160. doi: 10.4324/9781315659794-12

World Health Organization [WHO] (1946). *Definition of Health. In: Preamble to the Constitution of the World Health Organization*, Vol. 2. Geneva: World Health Organization.

Zevon, M. A., and Tellegen, A. (1982). The structure of mood change: an idiographic/nomothetic analysis. *J. Pers. Soc. Psychol.* 43, 111–122. doi: 10.1037/0022-3514.43.1.111

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 05 February 2020
doi: 10.3389/fpsyg.2020.00004

Validity of Social Cognition Measures in the Clinical Services for Autism Spectrum Disorder

Maria Chiara Pino[1,2], Francesco Masedu[1], Roberto Vagnetti[1], Margherita Attanasio[1], Chiara Di Giovanni[2], Marco Valenti[1,2] and Monica Mazza[1,2]*

[1] Department of Applied Clinical Sciences and Biotechnology, University of L'Aquila, L'Aquila, Italy, [2] Regional Centre for Autism, Abruzzo Region Health System, L'Aquila, Italy

The current study evaluated three social cognition (SC) tests for their clinical utility in aiding autism diagnosis. To do so, we compared the performance of 86 children with autism spectrum disorder (ASD) and 68 typically developing (TD) children, all aged from 4 to 10 years old, on three SC tasks [the Social Information Processing Interview (SIPI), the Comic Strip Task (CST), and the children's version of the Eyes Task] and calculated threshold scores that best differentiated the two groups. While difficulties in these abilities appear to represent the "central core" of ASD, services have largely ignored SC tests when supporting autism diagnoses. Therefore, this study attempted to validate and evaluate the diagnostic potential of these three tasks for children with ASD. To investigate the accuracy of these SC tests, we used the receiver operating characteristic (ROC) curve. As expected, the ASD group performed worse than the TD group on the SIPI and CST, but contrary to our prediction, the groups did not significantly differ on the Eyes Task. Specifically, the overall area under the curve (AUC) for the SIPI was 0.87, with a sensitivity of 73.5% and a specificity of 83.9% at the best cutoff point (score range 0–36; best cutoff = 31). The overall AUC for the CST was 0.75, with a sensitivity of 71.1% and a specificity of 77.0% at the best cutoff point (score range 0–15; best cutoff = 11). The overall AUC for the Eyes Task was 0.51, with a sensitivity of 50.3% and a specificity of 40.2% at the best cutoff point (score range 0–54; best cutoff = 45). In conclusion, the results showed that the SIPI test has good predictive power for classifying children with ASD. It should provide substantial supplementary clinical information and help to consolidate diagnostic procedures based on standard tools. Moreover, the results of the study have substantial implications for clinical practice: the better the knowledge of SC functioning in children with ASD, the more effective the intervention program for rehabilitation.

Keywords: autism spectrum disorder, diagnostic process, receiver operating characteristic (ROC) curve, social cognition, Theory of Mind, clinical utility

Edited by:
Pietro Cipresso,
Italian Auxological Institute (IRCCS),
Italy

Reviewed by:
David Vagni,
Italian National Research Council, Italy
Beatrice G. Kuhlmann,
Universität Mannheim, Germany

****Correspondence:***
Maria Chiara Pino
mariachiara.pino@univaq.it;
mariachiara.pino@graduate.univaq.it

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *29 October 2019*
Accepted: *03 January 2020*
Published: *05 February 2020*

Citation:
Pino MC, Masedu F, Vagnetti R, Attanasio M, Di Giovanni C, Valenti M and Mazza M (2020) Validity of Social Cognition Measures in the Clinical Services for Autism Spectrum Disorder. Front. Psychol. 11:4. doi: 10.3389/fpsyg.2020.00004

Abbreviations: ADOS-2, Autism Diagnostic Observation Schedule–Second Edition; ASD, autism spectrum disorder; AUC, area under the curve; CST, Comic Strip Task; DSM-5, *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*; ROC, receiver operating characteristic; SC, social cognition; SIPI, Social Information Processing Interview; TD, typically developing; ToM, Theory of Mind; TROG-2, Test for Reception of Grammar–Second Edition; VMA, verbal mental age.

INTRODUCTION

Autism spectrum disorder (ASD) is a neurodevelopment condition characterized by deficits in two domains: (1) social communication and social interaction and (2) restricted, repetitive patterns of behaviors, interests, or activities (American Psychiatric Association, 2013). A large body of research supports the hypothesis that difficulties in social interaction and communication can be explained by a deficit in social cognition (SC) abilities (Happé and Frith, 2014; Lai et al., 2014; Mazza et al., 2017; Pino et al., 2017). SC is a set of cognitive abilities involved in the processing and interpretation of the social world (Mazza et al., 2010; Bishop-Fitzpatrick et al., 2017; Pino et al., 2017). A main component of SC is the Theory of Mind (ToM), namely the ability to understand the mental and emotional states of other people (Mazza et al., 2014); it affects the development of social behavior from birth. A crucial development of ToM occurs around 3–4 years of age, when children acquire false belief attributions and realize that mental states, such as beliefs or the intentions of other people, may not be true (Mazza et al., 2017). Thus, ToM deficits are related to social communication and social interaction criteria of the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (DSM-5).

Several studies (Shamay-Tsoory and Aharon-Peretz, 2007; Shamay-Tsoory et al., 2010; Mazza et al., 2014; Baron-Cohen et al., 2015; Pino et al., 2017) suggest that ToM is not a unitary construct; rather, it involves two distinct components: cognitive and affective. Specifically, the cognitive component of ToM includes the ability to understand what other people are thinking and make inferences about their beliefs, intentions, and motivations. The affective ToM component is the ability to understand what other people are feeling in a specific emotional context and comprehend their emotions. Understanding another person's cognitive or affective state is a crucial ability for development and production of adequate social behaviors (Krebs and Russell, 1981; Hoffman, 1984; Batson, 1987; Mazza et al., 2017).

According Happé and Frith (2014), social behavior develops around 5 years of age, when children are able to differentiate their own internal states form those of others (Mazza et al., 2017). Children with ASD show difficulties in understanding other people's mental state and their perspectives, and this deficit might compromise social behavior development (Frith and Happé, 1994; Happé, 1994; Frith and Frith, 2003; Jones et al., 2010; Mazza et al., 2014; Ziv et al., 2014).

The ToM hypothesis of ASD was first introduced by Baron-Cohen et al. (1985) three decades ago; it demonstrates difficulties for children with ASD in passing false belief tasks. Recent studies suggest that adults with ASD have difficulties in implicit mentalization tasks (measured by spontaneous looking patterns), despite the fact that they can pass classic explicit mentalizing tasks (direct questions about others mental states; Jones et al., 2018). Differentiation between the theoretical ToM components is crucial for future research in ASD (Altschuler et al., 2018).

Some mentalizing tests, such as the Eyes Test (Baron-Cohen et al., 1997, 2001), require emotion recognition to infer mental states (Jones et al., 2018). This test should reflect the mentalizing process and the ability to understand other's mental states, such as emotions, thoughts, desires, beliefs, and goals (Peterson and Slaughter, 2009; Franco et al., 2014). Children and adults with ASD present lower performance on the Eyes Test (Baron-Cohen et al., 2001; Franco et al., 2014). Specifically, individuals with ASD have difficulties in processing information from the faces of others, such as facial expression and eye gaze, which play a significant role in SC (Hadjikhani et al., 2004; Pellicano et al., 2007; Ramachandran et al., 2010; Franco et al., 2014).

Deficits of social interaction in individuals with ASD are not related to general intellectual functioning. Rather, they are specific to the SC competences (Baron-Cohen et al., 1997; Ziv et al., 2014; Mazza et al., 2017). Ziv and Sorongon (2011), following Crick and Dodge's (1994) social information processing model, suggested that many mental steps occur before individuals implement a behavioral response to social cues, such as the encoding of social cues, interpretation of the cues, clarification of goals, generation of a behavioral response, response construction, response decision, and realization of the behavior response (Crick and Dodge, 1994; Ziv and Sorongon, 2011; Ziv et al., 2014; Mazza et al., 2017). According to this model, these internal processes include the ability to understand thoughts, intentions, and feelings of others (ToM) and select the adequate social responses (Crick and Dodge, 1994). Subsequently, Ziv et al. (2014) showed deficits in social information processing abilities in preschool children with ASD using the Social Information Processing Interview (SIPI), an instrument that allows one to evaluate social behavior and the pattern of social information processing based on Crick and Dodge's (1994) model. Ziv et al. (2014) demonstrated that children with ASD had a specific difficulty in social information processing; the ToM deficits were related to inadequate social behavior and poor social communication skills (Lerner et al., 2011; Ziv et al., 2014; Mazza et al., 2017). According to Mazza et al. (2017), social behavior is a consequence of how children process social cues. Considering that severe difficulties in social interaction are a defining feature of individuals with autism (Fletcher-Watson et al., 2014; Mazza et al., 2017), the SC assessment in ASD individuals, including psychometric evaluation of commonly used SC tasks, might help clinicians collect additional information and plan the best treatment in ASD research (National Advisory Mental Health, 2016; Morrison et al., 2019).

In ASD research, the SC construct is widely investigated, but it is rarely considered in the clinical practice due to a lack of well-validated tests with established psychometric data, as highlighted by Morrison et al. (2019). In contrast, the use of an SC test in ASD services might improve the diagnostic process and be exceedingly useful for prognoses and creating specific rehabilitation treatments for different age groups. Thus, the aim of the present study was to evaluate three SC tests for their clinical utility in aiding autism diagnosis. We compared performance by ASD and typically developing (TD) children on three SC tasks. Specifically, we chose to use the SIPI (Ziv and Sorongon, 2011; Ziv et al., 2014) for evaluation of social information process abilities, the Comic Strip Task (CST, Cornish et al., 2010; Sivaratnam et al., 2012) to assess the ToM sub-components (beliefs, emotions and intentions), and the children's version of the Eyes Task to evaluate the ability to understand and infer mental and emotional states regardless of the child's language level. For each test, we calculated threshold scores that best differentiated the two groups using the receiver operating characteristic (ROC) curve.

MATERIALS AND METHODS

Participants

One hundred-fifty-four children participated in this study: 86 children with ASD (75 males and 11 females, from 4 to 10 years old, recruited by the Reference Regional Centre for Autism in L'Aquila in the Abruzzo Region, Italy) and 68 TD children (60 males and 8 females, from 4 to 10 years old). The TD children were recruited from a nursery (for 4- to 5-year-old children) and a primary school (for 6- to 10-year-old children) located in L'Aquila. We chose to match the two groups by verbal mental age (VMA), as assessed by the Test for Reception of Grammar (TROG-2; Bishop, 2003). Differences between the two groups emerged for chronological age, where ASD children (mean = 7.64 years, SD = 1.53) were older than TD children [mean = 6.62 years, SD = 1.79; $t(152) = 3.81$, $p < 0.001$] but did not differ in VMA [ASD: mean = 6.96 years, SD = 2.35; TD: mean = 7.52 years, SD = 2.47; $t(152) = 1.43$, $p = 0.15$]. The exclusion criterion was intellectual disability; the participants had an IQ > 80.

The ASD sample comprised children who came for a first-time diagnosis as well as those who came for a second evaluation. All previously diagnosed ASD children received special education through a support teacher. They also followed therapies provided by the National Health System: speech therapy, psychomotor intervention, and Applied Behavioral Analysis.

The clinical process for ASD diagnosis commences with an experienced neuropsychiatrist who observes the child and interviews caregivers. Thereafter, an experienced psychologist performs the Autism Diagnostic Observation Schedule–Second Edition (ADOS-2; Lord et al., 2012). Finally, they consult with one another to make the ASD diagnosis according to the DSM-5 (American Psychiatric Association, 2013) criteria and ADOS-2 outcomes. Clinicians directly involved in the clinical practice participated in the study. ASD participants were level 1, according to DSM-5 criteria (American Psychiatric Association, 2013): most of them showed a delayed language development. ADOS-2 comparison scores of our sample ranged from low to moderate autism-related symptoms. None of the participants had comorbidities with other disorders. All the children were native Italian speakers.

Procedure

This study was performed in accordance with the recommendations of the Ethics Committee of Local Health Unit 1 (ASL1-Avezzano, Sulmona, L'Aquila), Abruzzo Region, L'Aquila, Italy. The Ethics Committee approved the protocol (number 186061/17) prior to the recruitment of participants, according to the principles established by the Declaration of Helsinki. Informed consent from the child and her or his parents was obtained before participation. Children with ASD were tested at the Reference Regional Centre for Autism, Abruzzo Region Health System, L'Aquila, Italy, whereas TD children were tested in their nurseries or schools. All children were tested individually by an expert psychologist in a quiet room according to the principles established by the Declaration of Helsinki.

VMA Measure

According to recent literature (Pino et al., 2017), children with ASD show a delay in developing SC abilities based on chronological age, whereas VMA seems to be a good predictor of ToM abilities (Happé, 1994; Pino et al., 2017). Moreover, social difficulty does not appear to be based on the general IQ level, whereas VMA appears to be a more promising associated measure (Pino et al., 2017, 2018).

The literature suggests that children with ASD can use verbal strategies to support their reasoning during ToM tasks (Happé, 1995; Durrleman et al., 2019). Grammatical skills are particularly important during mentalizing (Fisher et al., 2005; de Villiers, 2007; Milligan et al., 2007). For these reasons, we chose to match two groups based on VMA, as assessed with the TROG-2 (Bishop, 2003), a standardized measure of receptive language that allows one to evaluate the ability to understand verbal language. The TROG-2 evaluates the comprehension of grammatical structures and contrasts grammatical indicated by suffixed, functional words, and order word. The test examines 20 syntactic constructions, each of which is examined with a block of four items. Participants select the picture–out of four presented choices–that corresponds to the sentence read by examiner. Standard and age-equivalent scores are made by the total number of blocks passed.

SC Measures

SIPI

The SIPI (Ziv and Sorongon, 2011; Ziv et al., 2014) is a 20-min structured interview based on a storybook-easel that depicts a series of vignettes in which a protagonist is either rejected by two other peers or provoked by another peer. Each type of vignette is combined with each type of peer intent to generate four stories: (1) a non-hostile peer-entry rejection story, (2) an ambiguous peer-entry rejection story, (3) an accidental provocation story, and (4) an ambiguous provocation story. According to Ziv et al. (2014), the scores correspond to four of the five mental steps of social information-processing proposed by Crick and Dodge's (1994) model: (1) encoding, (2) interpretation of cues, (3) response construction, and (4) response evaluation.

An example of a SIPI story is the following: Michael is watching the other children playing. Michael walks up to the other children and asks them: "Can I play with you?" The child says: "Sorry. The teacher said only two can play in the block area" (for details, see Ziv et al., 2014).

The Encoding component evaluates the level of detail that the child recalls across the four stories. Thus, the examiner asks the child: "Tell me what happened in the story, from the beginning to the end." A code of 0 is given to children who recall no correct details from the stories and a code of 1 to children who correctly recall all the details in all the stories. An overall score is then calculated (ranging from 0 to 4).

The Interpretation component evaluates hostile attribution to others' behavior (the question is: "Do you think the other children who didn't let Michael play are mean or not mean?"). The answers are coded with 0 or 1, and an overall score (0–4) is then calculated, with higher scores representing higher levels of hostile attribution bias. Scores for this component are inversely encoded compared with the other SIPI components; that is, a higher score indicates a major tendency to consider the behavior of other children as hostile.

The Response Generation score is derived from the child's responses to the open-ended question: "Pretend that you ask your friends if you can play with them and they say that only two can play in the block area. What would you do?" For each story, the examiner encodes the response as competent or non-competent and assigns a code of 1 if the child's response is classified as competent and of 0 if the answer is classified as non-competent. An overall score (from 0 to 8) is then calculated.

The Response Evaluation items examine the way in which the child assesses the behavior of other people as being right or wrong. This score is obtained by combining the 36 response evaluation questions (4 stories × 3 presented responses × 3 questions per presented response). The three response variables for these steps are: (1) a competent response (e.g., Michael could say, "Then can I play next?"); (2) an aggressive response (e.g., Michael could kick apart the blocks and say to the other children, "If I can't play, then you can't play either!"); and (3) an avoidant or inappropriate response (e.g., Michael could cry and say, "It's not fair"; Pino et al., 2018). The total number of non-competent responses (aggressive and avoidant responses) are subtracted from the total number of competent responses and adjusted for negative scores in order to obtain a score (from 0 to 36).

For the purpose of this study, we also calculated a total score. In our analysis, we did not include the Encoding subscale because one item showed poor psychometric properties (Ziv and Sorongon, 2011). Instead, we used the three main SIPI scores as reported by Ziv and Sorongon (2011): Interpretation, Response Generation, and Response Evaluation. A higher score on the Interpretation subscale (range 0–4) represents hostile attribution. Therefore, we first converted this scale into a non-hostile attribution scale (called Positive Interpretation) by calculating its complementary scale using the following formula: 4 − the number of hostile responses. Next, we summed the Positive Interpretation, Response Generation, and Response Evaluation scores to obtain a total SIPI score.

We decided to use the SIPI because it can evaluate the social cue processing that is closely related to the ability to understand and recognize the intentions, beliefs, and emotions of other people (ToM). According to Mazza et al. (2017), if a child has difficulties in processing social cues within a context, she or he will show difficulties in the ability to evaluate whether other people's social behavior is right or wrong and she or he will respond inadequately in social situations. This phenomenon will impair social relations with others. The test is coded by considering different aspects of the social information process, including the hostile style of attribution and the generation of socially competent, avoidant, or hostile responses. This factor represents an added value in the diagnostic evaluation; in fact, during the assessment, some behavioral problems may arise that should be considered for future intervention or evaluation. Indeed, Ziv and Sorongon (2011) demonstrated that preschoolers with aggressive tendencies evaluate aggressive responses as better ones. However, future research should deepen this aspect in the clinical setting for details see **Supplementary Material**.

CST

The CST (Cornish et al., 2010; Sivaratnam et al., 2012) is a 21-item measure that was developed to assess three aspects of ToM: understanding Beliefs, Intentions, and Emotions. There are five items in each component, and each comprises a five-picture comic strip that illustrates everyday social scenarios involving interpersonal interactions that are familiar to young children. Each component has a maximum score of 5, with a total test score range of 0–15 (higher scores correspond to better ToM). We used the CST because it does not require verbal abilities, a factor that allows one to measure ToM deficits *per se*. Moreover, the CST is suitable for a wide swath of the ASD population; it was designed for 4- to 8-year-old children, but it can be used in both younger and older children (Philpott et al., 2013). We also suppose that the use of comics might attract the attention of children, and the formal administration is very brief (10–15 min; Sivaratnam et al., 2012).

Eyes Task–Children's Version

The Eyes Task (Franco et al., 2014) consists of a series of black and white photos of children's eyes; they portray either mental states or primary emotions. The expressions selected as primary emotions were happy and surprised (positive/neutral valence) and sad and angry (negative valence), while excited and thinking (positive/neutral valence) and worried and shy (negative valence) were selected to represent mental states (for further details, see Franco et al., 2014; Pino et al., 2017). A total of 56 images are presented to the child; each represents one of the stimuli described above with two possible responses. If the child responds correctly, the item is coded as 1; otherwise, it is coded as 0. A total score is then calculated by adding the correct responses to the primary emotions and mental states. Total scores range from 0 to 56 (with higher scores indicating better ToM performance). We used the version by Franco et al. (2014) because stimuli are derived from naturalistic pictures of children rather than posed adults like the version of Baron-Cohen et al. (2001). Moreover, the Eyes Task (Franco et al., 2014) requires fewer cognitive demands because it shows one eye picture with two possible responses. This design is suitable even for low-functioning autism. Score calculations for each test are shown in **Table 1**.

Data Analysis

Descriptive Analysis

Demographic parameters and total scores for the SIPI, the CST, and the Eyes Task were recorded for both groups (ASD and TD).

TABLE 1 | Score construction.

	Score Construct	Count (#)	Range
SIPI	**Interpretation**		
	$Score_{I-}$ (Negative interpretation)	#(Hostile responses)	0–4
	$Score_{I+}$ (Positive interpretation)*	4-#(Hostile responses)	0–4
	Response generation		
	$Score_{G}$	#(Competent responses) + [4-#(Non-competent responses)]	0–8
	Response evaluation		
	$Score_{E}$	#(Competent responses) + [24-#(Non-competent responses)]	0–36
	Total SIPI Score		
	$Score_{SIPI}$	$Score_{I+} + Score_{G} + Score_{E}$	0–48
Comic Strip Task	**Intention**		
	$Score_{i}$	#(Correct responses)	0–5
	Beliefs		
	$Score_{b}$	#(Correct responses)	0–5
	Emotions		
	$Score_{e}$	#(Correct responses)	0–5
	Total CST Score		
	$Score_{CST}$	$Score_{i} + Score_{b} + Score_{e}$	0–15
Eyes Task	**Primary emotions**		
	$Score_{P}$	#(Correct responses)	0–28
	Mental states		
	$Score_{M}$	#(Correct responses)	0–28
	Total Eyes Task Score		
	$Score_{ET}$	$Score_{P} + Score_{M}$	0–56

**This score is used to calculate the total SIPI score.*

Reliability and Internal Consistency

We assessed the internal consistency and reliability, in relation to the overall measure, for each ToM measure (the SIPI, the CST and the Eyes Task) using Cronbach's α.

ROC Analysis

The overall goal of the ROC analysis was to estimate the cutoff points for the ToM measures that could distinguish between the two groups. ROC analysis is used to assess the diagnostic properties of tests, specifically, to assess the way in which various measures generally discriminate between categories of subjects. In order to do this, a cutoff point must be established. Based on the cutoff point, we can determine whether a person with a certain score belongs to one category or another (e.g., normal/non-clinical or clinical group). ROC analysis can also be used when comparing the diagnostic performance of two or more tests (Westin, 2001).

In a ROC curve, the true-positive rate (sensitivity) is plotted as a function of the false-positive rate (100 - specificity) for various cutoff points. The obtained area under the curve (AUC) signifies how well a parameter distinguishes between two groups. In order to establish a diagnostic threshold and corresponding test sensitivity and specificity, we established the cutoff as the value where the highest percentage of true positives was correctly classified as positive and true negatives was correctly classified as negative (Cleves, 1999). In our study, ROC curve analysis was performed to evaluate the accuracy of the total score of ToM measures (the SIPI, CST, and children's version of the Eyes Task) in discriminating between ASD and TD children, using ADOS-2 and DSM-5 criteria as the *gold standard*. The analysis was performed using STATA version 14 statistical software (StataCorp, 2015).

Optimizing Diagnostic Performance

To improve diagnostic performance, we constructed a test based on a linear combination of the SIPI, CST, and Eyes Task scores. A multivariate logistic regression was performed to obtain the respective logit scores. The logit model allowed us to assess the marginal diagnostic advantage of the SIPI, CST, and Eyes Task and test their statistical significance. Their marginal diagnostic gain can be viewed in terms of the AUC of the ROC curve of the new logit score.

RESULTS

Descriptive Analysis

Compared with TD children, children with ASD scored significantly lower on the SIPI [$t(152) = 9.19$, $p < 0.001$] and the CST [$t(152) = 5.59$, $p < 0.001$], but they recorded similar scores on the Eyes Task [$t(152) = 0.43$, $p = 0.66$]. The results are shown in **Table 2**.

Internal Consistency Results

The results for the CST demonstrated high internal consistency (Cronbach's α = 0.80), the results for SIPI demonstrated good internal consistency (Cronbach's α = 0.76), and the results for Eyes Task demonstrated high internal consistency (Cronbach's α = 0.80).

ROC Analysis

For the SIPI, the overall AUC_{SIPI} was 0.87 (*SE* = 0.02). The optimal cutoff value was 31 (correctly classified = 79.3%), which corresponded to a sensitivity of 73.5% and a specificity of 83.9%. For the CST, the overall AUC_{CST} was 0.75 (*SE* = 0.03), and the optimal cutoff value was 11 (correctly classified = 71.1%). This value corresponded to a sensitivity of 63.0% and a specificity of 77.0%. For the ET, the overall AUC_{ET} was 0.51 (*SE* = 0.04). The optimal cutoff value was 45 (correctly classified = 50.3); this value corresponded to a sensitivity of 63.24% and a specificity of 40.2%. The analysis revealed a significant difference between AUC measures ($\chi^2 = 60.9$, $p < 0.001$). The results are reported in **Table 3**, and the ROC curves are displayed in **Figure 1**.

Diagnostic Performance Optimization

The logistic model showed that the SIPI (β = 0.26, *SE* = 0.04, $z = 5.23$, $p < 0.001$) and Eyes Task (β = -0.10, *SE* = 0.03, $z = -3.12$, $p < 0.001$) were statistically significant diagnostic predictors, while the CST (β = 0.18, *SE* = 0.11, $z = 1.66$, $p = 0.09$) was not. When merging the two tests into one new test (hereafter referred

TABLE 2 | Between-group differences for demographic data, clinical information, and social cognition measures.

	ASD (*N* = 86) Mean (SD)	TD (*N* = 68) Mean (SD)	*t* (df = 152)	*p*
Chronological age (in years)	7.64 (1.53)	6.62 (1.79)	3.81	<0.001*
Verbal mental age (in years)	6.96 (2.35)	7.52 (2.47)	1.43	0.15
ADOS-Social communication and social interaction	8.34 (3.50)	–	–	–
ADOS-Repetitive and stereotyped behaviors	1.26 (1.12)	–	–	–
ADOS total scores	9.78 (3.62)	–	–	–
Social cognition measures (total score)				
SIPI	22.3 (9.22)	34.3 (6.13)	9.19	<0.001*
CST	9.01 (2.48)	11.2 (2.22)	5.59	<0.001*
Eyes Task	44.3 (8.01)	43.6 (12.0)	0.43	0.66

**Significant difference for $p < 0.05$.*

TABLE 3 | ToM measures' AUCs and cut-offs with respective sensitivity and specificity.

Social cognition measures	AUC*	SE	95% CI	Cutoff	Sensitivity(%)	Specificity (%)	Correctly Classified (%)
SIPI	0.87	0.02	0.81–0.92	31	73.5	83.9	79.4
CST	0.75	0.03	0.67–0.82	11	63.1	77.0	71.1
Eyes Task	0.51	0.04	0.42–0.60	45	63.2	40.2	50.3

**Comparison between AUC show a significant difference ($\chi^2 = 60.9$, $p < 0.001$).*

to as SIPI-ET), we observed an improvement in overall diagnostic performance (AUC$_{SIPI-ET}$ = 0.89, *SE* = 0.02). However, there was no statistically significant difference between AUC$_{SIPI-ET}$ and AUC$_{SIPI}$ (χ^2 = 2.39, p = 0.12), a finding that indicates that there was no statistically significant improvement. **Figure 2** shows AUC$_{SIPI}$ versus AUC$_{SIPI-ET}$.

DISCUSSION

This study highlights the utility of including a SC battery of tests to improve the quality and quantity of information collected during procedures for diagnosing ASD. According to Lai et al. (2014), social difficulties in children with autism have been reported since 1985, when it was first highlighted by Baron-Cohen and collaborators. This impaired ability is believed to play a central role in the social communication and interaction deficits (the first diagnostic criterion in DSM-5; American Psychiatric Association, 2013) of ASD individuals. In fact, this criterion requests clinicians to evaluate abilities as "reduced sharing of interest, emotion or affect" (criterion A1/DSM-5; American Psychiatric Association, 2013); "deficits in social–emotional reciprocity" (criterion A1/DSM-5; American Psychiatric Association, 2013), "deficits in non-verbal communicative behaviors used for social interaction" (criterion A2/DSM-5; American Psychiatric Association, 2013), and "deficits in developing, maintaining, and understanding relationships" (criterion A3/DSM-5; American Psychiatric Association, 2013).

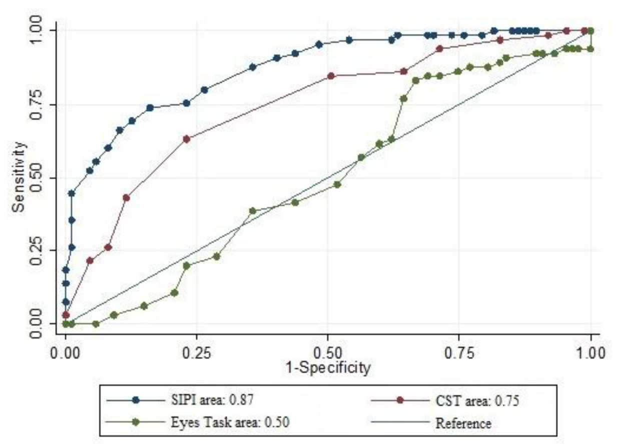


FIGURE 1 | Comparison of ROC curves of SC measures (SIPI, CST, and Eyes Task) with relative AUCs.

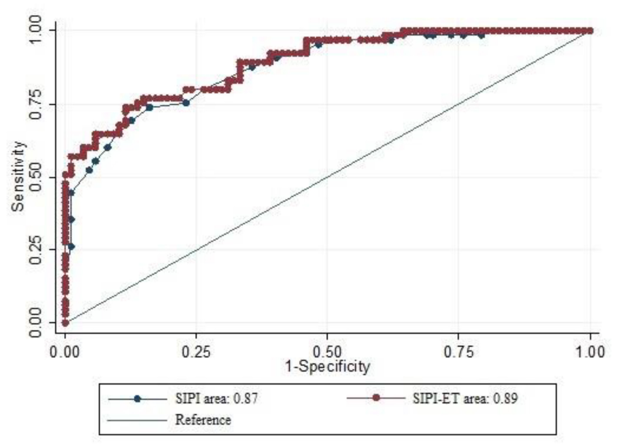


FIGURE 2 | Comparison of ROC curves of SIPI and SIPI-ET with relative AUCs.

All of these competences are part of the complex cognitive construct of SC. Despite the significant role exerted by SC components, such as ToM, in ASD diagnoses, assessment of these competences is neglected in Italian clinical services. Indeed, the use of ToM tests is limited to the research field. For this reason, we evaluated the accuracy of SC measures–using an ROC curve–to discriminate ASD from TD children in a small Italian sample. Additionally, we determined the best cutoff point for the three SC measures used: the SIPI, CST, and Eyes Task.

The results of the ROC analysis showed that the SIPI had good predictive power in terms of accurately classifying children with ASD. On the other hand, the CST showed moderate predictive power, while the Eyes Task showed no ability to correctly distinguish between ASD and TD.

Regarding the Eyes Task, Franco et al. (2014) found that ASD were less accurate compared to TD children, but based on our results, the difference between the groups would not allow us to characterize the ASD individuals during the diagnostic process. In fact, ASD children around 5–6 years old can recognize simple emotional and mental states (i.e., happy, sad, angry, and worried). Thus, there were no distinguishing characteristics in their performance compared to their TD peers. For the CST, the original authors administered the test to 4- to 8-year-old children with high functioning ASD (Sivaratnam et al., 2012). They performed significantly worse compared to controls on the overall two-subscale CST (belief- and intention-understanding). There were no group differences in the emotion understanding subscale performance (Cornish et al., 2010; Sivaratnam et al., 2012). In our study, unlike the authors of CST, we matched subjects by VMA. This method reduced differences in mentalizing ability due to delayed development based on chronological age. Additionally, the participants in our research presented a wider age range compared to Sivaratnam et al. (2012). On the basis of these results, the SIPI represents a useful instrument to support the ASD diagnosis. Specifically, the SIPI assesses the ability to correctly interpret the presented social scenarios (interpretation), “put oneself in another’s shoes” (response generation), and determine whether other people’s social behaviors are right or wrong (response evaluation).

Our results regarding the SIPI are consistent with a previous study that demonstrated differences between ASD and TD children on this task (Pino et al., 2018). Additionally, Mazza et al. (2017) showed that mentalizing ability plays a key role in the development of social abilities, and the lack of ToM competences in children with ASD impairs their competent social behavior (Mazza et al., 2017). Thus, these components are closely related and improved mentalizing ability might also enhance social behavior.

Collection of the data examined in this study should allow clinicians to plan a treatment focused on social abilities to improve the relationship with other people and avoid isolation and the emergence of other clinical symptomatology, such as depression or anxiety disorder. Furthermore, the systematic use of SC measures in clinical evaluations might help monitor improvements related to treatment and therapy.

In conclusion, we think that the data provided in this study are valuable because they emphasize the utility of incorporating SC measures into diagnostic processes in ASD clinical practice. In particular, the SIPI showed valid accuracy in distinguishing between ASD and TD children. These findings indicate that this test can be implemented into the diagnostic procedure. Additionally, the data provided by our work suggest the cutoff points for each of the examined SC tests (**Table 3**); these data should allow examiners to use these tests with normative values.

We are aware that the present study has some limitations. (1) Our two samples differed in chronological age. However, we stress that the development of SC competencies, particularly mentalizing ability, is related more to mental rather than to chronological age (Pino et al., 2018). (2) This study is also limited by the small Italian sample size; future studies are needed to demonstrate the generalizability of our results. (3) Performance would also need to be compared to other clinical conditions to determine whether these tasks adequately discriminate autism from competing diagnoses. Given that other clinical conditions also present with impairments in social performance, it is necessary to investigate the utility of these tasks for selectively aiding an ASD diagnosis.

DATA AVAILABILITY STATEMENT

The datasets generated during the current study are not publicly available because the data were obtained in the course of mental health care. However, they are available from the corresponding author on reasonable request.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee approved the protocol (number 186061/17) prior to the recruitment of participants, according to the principles established by the Declaration of Helsinki. Written informed consent to participate in this study was provided by the participants’ legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

MM and MP designed the research. RV, MA, and CD collected the data. FM, MP, and MV analyzed the data. All authors contributed to writing the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00004/full#supplementary-material

REFERENCES

Altschuler, M., Sideridis, G., Kala, S., Warshawsky, M., Gilbert, R., Carroll, D., et al. (2018). Measuring individual differences in cognitive, affective, and spontaneous theory of mind among school-aged children with autism spectrum disorder. *J. Autism. Dev. Disord.* 48, 3945–3957. doi: 10.1007/s10803-018-3663-1

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. Virginia: American Psychiatric Association.

Baron-Cohen, S., Bowen, D. C., Holt, R. J., Allison, C., Auyeung, B., Lombardo, M. V., et al. (2015). The "reading the mind in the eyes" test: complete absence of typical sex difference in˜ 400 men and women with autism. *PloS One* 10:e0136521. doi: 10.1371/journal.pone.0136521

Baron-Cohen, S., Leslie, A. M., and Frith, U. (1985). Does the autistic child have a 'theory of mind'? *Cognition* 21, 37–46. doi: 10.1016/0010-0277(85)90022-8

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., and Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *J. Child Psychol. Psychiatr. Allied Disciplines* 42, 241–251. doi: 10.1037/xlm0000745

Baron−Cohen, S., Jolliffe, T., Mortimore, C., and Robertson, M. (1997). Another advanced test of theory of mind: evidence from very high functioning adults with autism or Asperger syndrome. *J. Child Psychol. Psychiatry* 38, 813–822. doi: 10.1111/j.1469-7610.1997.tb01599.x

Batson, C. D. (1987). "Prosocial motivation: Is it ever truly altruistic?," in *Advances in Experimental Social Psychology*, Vol. 20, (Cambridge, MA: Academic Press.), 65–122. doi: 10.1016/s0065-2601(08)60412-8

Bishop, D. (2003). *The Test for Reception of Grammar (TROG2)—Version 2*. London: The Psychological Corporation.

Bishop-Fitzpatrick, L., Mazefsky, C. A., Eack, S. M., and Minshew, N. J. (2017). Correlates of social functioning in autism spectrum disorder: the role of social cognition. *Res. Autis. Spectr. Disord.* 35, 25–34. doi: 10.1016/j.rasd.2016.11.013

Cleves, M. A. (1999). "sg120: Receiver Operating Characteristic (ROC) analysis," in *Stata Technical Bulletin* 52, 19–33. In *Stata Technical Bulletin Reprints* 9, 212–229. College Station, TX: Stata Press.

Cornish, K., Rinehart, N., Gray, K., and Howlin, P. (2010). *Comic Strip Task*. Melbourne: Monash University Developmental Neuroscience.

Crick, N. R., and Dodge, K. A. (1994). A review and reformulation of social information: processing mechanisms in children's social adjustment. *Psychol. Bull.* 115, 74–101. doi: 10.1093/deafed/enw030

de Villiers, J. (2007). The interface of language and theory of mind. *Lingua* 117, 1858–1878. doi: 10.1016/j.lingua.2006.11.006

Durrleman, S., Burnel, M., De Villiers, J. G., Thommen, E., and Delage, H. (2019). The impact of grammar on mentalizing: a training study including children with autism spectrum disorder and developmental language disorder. *Front. Psychol.* 10:2478. doi: 10.3389/fpsyg.2019.02478

Fisher, N., Happé, F., and Dunn, J. (2005). The relationship between vocabulary, grammar, and false belief task performance in children with autistic spectrum disorders and children with moderate learning difficulties. *J. Child Psychol. Psychiatry* 46, 409–419. doi: 10.1111/j.1469-7610.2004.00371.x

Fletcher−Watson, S., McConnell, F., Manola, E., and McConachie, H. (2014). Interventions based on the Theory of Mind cognitive model for autism spectrum disorder (ASD). *Cochrane Database Syst. Rev.* 21, CD008785. doi: 10.1002/14651858.CD008785

Franco, F., Itakura, S., Pomorska, K., Abramowski, A., Nikaido, K., and Dimitriou, D. (2014). Can children with autism read emotions from the eyes? The eyes test revisited. *Re. Dev. Disabil.* 35, 1015–1026. doi: 10.1016/j.ridd.2014.01.037

Frith, U., and Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 358, 459–473.

Frith, U., and Happé, F. (1994). Autism: beyond "theory of mind". *Cognition* 50, 115–132. doi: 10.1016/0010-0277(94)90024-8

Hadjikhani, N., Joseph, R. M., Snyder, J., Chabris, C. F., Clark, J., Steele, S., et al. (2004). Activation of the fusiform gyrus when individuals with autism spectrum disorder view faces. *Neuroimage* 22, 1141–1150. doi: 10.1016/j.neuroimage.2004.03.025

Happé, F., and Frith, U. (2014). Annual research review: towards a developmental neuroscience of atypical social cognition. *J. Child Psychol. Psychiatry* 3, 553–577. doi: 10.1111/jcpp.12162

Happé, F. G. (1994). An advanced test of theory of mind: understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *J. Autism Dev. Disord.* 24, 129–154. doi: 10.1007/BF02172093

Happé, F. G. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Dev.* 66, 843–855.

Hoffman, M. L. (1984). "Interaction of affect and cognition in empathy," in *Emotions, Cognition, and Behavior*, Chap. Cambridge, eds C. E. Izard, J. Kagan, and R. B. Zajonc, (Cambridge University Press), 103–131.

Jones, A. P., Happé, F. G., Gilbert, F., Burnett, S., and Viding, E. (2010). Feeling, caring, knowing: different types of empathy deficit in boys with psychopathic tendencies and autism spectrum disorder. *J. f Child Psychol. Psychiatry* 51, 1188–1197. doi: 10.1111/j.1469-7610.2010.02280.x

Jones, C. R., Simonoff, E., Baird, G., Pickles, A., Marsden, A. J., Tregay, J., et al. (2018). The association between theory of mind, executive function, and the symptoms of autism spectrum disorder. *Autism Res.* 11, 95–109. doi: 10.1002/aur.1873

Krebs, D., and Russell, C. (1981). "Role-taking and altruism: when you put yourself in the shoes of another, will they take you to their owner's aid?," in *Altruism and Helping Behavior: Social, Personality, and Developmental Perspectives*, eds J. P. Rushton, and R. M. Sorrentino (New Jersey: Lawrence Erlbaum), 137–165.

Lai, M. C., Lombardo, M. V., and Baron-Cohen, S. (2014). Autism. *Lancet* 383, 896–910. doi: 10.1016/S0140-6736(13)61539-1

Lerner, M. D., Hutchins, T. L., and Prelock, P. A. (2011). Brief report: preliminary evaluation of the theory of mind inventory and its relationship to measures of social skills. *J. Autism Dev. Disord.* 41, 512–517. doi: 10.1007/s10803-010-1066-z

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., and Bishop, S. L. (2012). *Autism Diagnostic Observation Schedule (ADOS 2): Manual,* 2nd Edn. Los Angeles: CA: Western Psychological Services.

Mazza, M., Lucci, G., Pacitti, F., Pino, M. C., Mariano, M., Casacchia, M., et al. (2010). Could schizophrenic subjects improve their social cognition abilities only with observation and imitation of social situations? *Neuropsychol. Rehabil.* 20, 675–703. doi: 10.1080/09602011.2010.486284

Mazza, M., Mariano, M., Peretti, S., Masedu, F., Pino, M. C., and Valenti, M. (2017). The role of theory of mind on social information processing in children with autism spectrum disorders: a mediation analysis. *J. Autism Dev. Disord.* 47, 1369–1379. doi: 10.1007/s10803-017-3069-5

Mazza, M., Pino, M. C., Mariano, M., Tempesta, D., Ferrara, M., De Berardis, D., et al. (2014). Affective and cognitive empathy in adolescents with autisms spectrum disorder. *Front. Hum. Neurosci.* 7:791. doi: 10.3389/fnhum.2014.00791

Milligan, K., Astington, J. W., and Dack, L. A. (2007).Language and theory of mind: meta−analysis of the relation between language ability and false−belief understanding. *Child Dev.* 78, 622–646. doi: 10.1111/j.1467-8624.2007

Morrison, K. E., Pinkham, A. E., Kelsven, S., Ludwig, K., Penn, D. L., and Sasson, N. J. (2019). Psychometric evaluation of social cognitive measures for adults with autism. *Autism Res.* 12, 766–778. doi: 10.1002/aur.2084

National Advisory Mental Health (2016). *Council Workgroup on Tasks and Measures for Research Domain Criteria. Behavioral Assessment Methods for RDoC Constructs*. Bethesda, MA: National Advisory Mental Health.

Pellicano, E., Jeffery, L., Burr, D., and Rhodes, G. (2007). Abnormal adaptive face-coding mechanisms in children with autism spectrum disorder. *Curr. Biol.* 17, 1508–1512. doi: 10.1016/j.cub.2007.07.065

Peterson, C. C., and Slaughter, V. (2009). Theory of mind (ToM) in children with autism or typical development: links between eye-reading and false belief understanding. *Re. Autism Spectr. Disord.* 3, 462–473. doi: 10.1016/j.rasd.2008.09.007

Philpott, A. L., Rinehart, N. J., Gray, K. M., Howlin, P., and Cornish, K. (2013). Understanding of mental states in later childhood: an investigation of theory of mind in autism spectrum disorder and typical development with a novel task. *Int. J Dev. Disabil.* 59, 108–117. doi: 10.1179/2047387713y.0000000015

Pino, M. C., Mariano, M., Peretti, S., D'Amico, S., Masedu, F., Valenti, M., et al. (2018). When do children with autism develop adequate social behaviour? Cross-sectional analysis of developmental trajectories. *Eur. J. Dev. Psychol.* 17, 71–78. doi: 10.1080/17405629.2018.1537876

Pino, M. C., Mazza, M., Mariano, M., Peretti, S., Dimitriou, D., Masedu, F., et al. (2017). Simple mind reading abilities predict complex theory of mind: developmental delay in autism spectrum disorders. *J. Autism Dev. Disord.* 47, 2743–2743. doi: 10.1007/s10803-017-3194-1

Ramachandran, R., Mitchell, P., and Ropar, D. (2010). Recognizing faces based on inferred traits in autism spectrum disorders. *Autism* 14, 605–618. doi: 10.1177/1362361310372777

Shamay-Tsoory, S. G., and Aharon-Peretz, J. (2007). Dissociable prefrontal networks for cognitive and affective theory of mind: a lesion study. *Neuropsychologia* 45, 3054–3067. doi: 10.1016/j.neuropsychologia.2007.05.021

Shamay-Tsoory, S. G., Harari, H., Aharon-Peretz, J., and Levkovitz, Y. (2010). The role of the orbitofrontal cortex in affective theory of mind deficits in criminal offenders with psychopathic tendencies. *Cortex* 46, 668–677. doi: 10.1016/j.cortex.2009.04.008

Sivaratnam, C. S., Cornish, K., Gray, K. M., Howlin, P., and Rinehart, N. J. (2012). Brief report: Assessment of the social-emotional profile in children with autism spectrum disorders using a novel comic strip task. *J. Autism Dev. Disord.* 42, 2505–2512. doi: 10.1007/s10803-012-1498-8

StataCorp, (2015). *Stata Statistical Software: Release 14*. College Station, TX: StataCorp LP. doi: 10.1007/s10803-012-1498-8

Westin, L. (2001). *Receiver Operating Characteristic (ROC) Analysis: Evaluating Discriminance Effects Among Decision Support systems. Technical Report*. Sweden: Umea University.

Ziv, Y., Hadad, B. S., Khateeb, Y., and Terkel-Dawer, R. (2014). Social information processing in preschool children diagnosed with autism spectrum disorder. *J. Autism Dev. Disord.* 44, 846–859. doi: 10.1007/s10803-013-

Ziv, Y., and Sorongon, A. (2011). Social information processing in preschool children: relations to sociodemographic risk and problem behaviour. *J. Exp. Child Psychol.* 3, 412–429. doi: 10.1016/j.jecp.2011.02.009

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

CASE REPORT
published: 07 February 2020
doi: 10.3389/fpsyg.2020.00123

Virtual Reality for the Assessment of Everyday Cognitive Functions in Older Adults: An Evaluation of the Virtual Reality Action Test and Two Interaction Devices in a 91-Year-Old Woman

Andrea Chirico[1]*, *Tania Giovannetti*[2], *Pietro Neroni*[3,4], *Stephanie Simone*[2], *Luigi Gallo*[3], *Federica Galli*[1], *Francesco Giancamilli*[1], *Marco Predazzi*[5], *Fabio Lucidi*[1], *Giuseppe De Pietro*[3] *and Antonio Giordano*[6,7]

[1] Department of Psychology of Developmental and Socialization Processes, Sapienza University of Rome, Rome, Italy, [2] Psychology Department, Temple University, Philadelphia, PA, United States, [3] Institute for High Performance Computing and Networking, National Research Council, Naples, Italy, [4] Department of Engineering, Parthenope University of Naples, Naples, Italy, [5] Fondazione Il Melo Onlus, Gallarate, Italy, [6] Sbarro Institute for Cancer Research and Molecular Medicine, Center for Biotechnology, College of Science and Technology, Temple University, Philadelphia, PA, United States, [7] Department of Medical Biotechnologies, University of Siena, Siena, Italy

OPEN ACCESS

Edited by:
Pietro Cipresso,
Italian Auxological Institute (IRCCS),
Italy

Reviewed by:
Silvia Serino,
Lausanne University Hospital (CHUV),
Switzerland
Giulia Corno,
Université du Québec en Outaouais,
Canada

***Correspondence:**
Andrea Chirico
andrea.chirico@uniroma1.it

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: 06 September 2019
Accepted: 16 January 2020
Published: 07 February 2020

Citation:
Chirico A, Giovannetti T, Neroni P, Simone S, Gallo L, Galli F, Giancamilli F, Predazzi M, Lucidi F, De Pietro G and Giordano A (2020) Virtual Reality for the Assessment of Everyday Cognitive Functions in Older Adults: An Evaluation of the Virtual Reality Action Test and Two Interaction Devices in a 91-Year-Old Woman. Front. Psychol. 11:123. doi: 10.3389/fpsyg.2020.00123

Performance-based functional tests for the evaluation of daily living activities demonstrate strong psychometric properties and solve many of the limitations associated with self- and informant-report questionnaires. Virtual reality (VR) technology, which has gained interest as an effective medium for administering interventions in the context of healthcare, has the potential to minimize the time-demands associated with the administration and scoring of performance-based assessments. To date, efforts to develop VR systems for assessment of everyday function in older adults generally have relied on non-immersive systems. The aim of the present study was to evaluate the feasibility of an immersive VR environment for the assessment of everyday function in older adults. We present a detailed case report of an elderly woman who performed an everyday activity in an immersive VR context (Virtual Reality Action Test) with two different types of interaction devices (controller vs. sensor). VR performance was compared to performance of the same task with real objects outside of the VR system (Real Action Test). Comparisons were made on several dimensions, including (1) quality of task performance (e.g., order of task steps, errors, use and speed of hand movements); (2) subjective impression (e.g., attitudes), and (3) physiological markers of stress. Subjective impressions of performance with the different controllers also were compared for presence, cybersickness, and usability. Results showed that the participant was capable of using controllers and sensors to manipulate objects in a purposeful and goal-directed manner in the immersive VR paradigm. She performed the everyday task similarly across all conditions. She reported no cybersickness and even indicated that interactions in the VR environment were pleasant and relaxing. Thus, immersive VR is a feasible approach for function assessment even with older adults who might have very limited computer experience, no prior VR exposure, average educational experiences, and mild cognitive difficulties. Because of inherent limitations of single case reports (e.g., unknown generalizability, potential practice effects, etc.), group studies are needed to establish the full psychometric properties of the Virtual Reality Action Test.

Keywords: activities of daily living, everyday action, virtual reality, cognitive aging, psychometric assessment

INTRODUCTION

Performance-based tests, that evaluate the ability to perform everyday tasks in the laboratory/clinic, solve many of the limitations associated with the use of self- and informant-report questionnaires of everyday functioning in people with cognitive impairment (see Giovannetti et al., 2013 for a review). Performance-based, functional tests are objective, standardized, allow a systematic comparison between individuals and provide detailed information on behaviors during the natural performance of activities. The validity of performance-based measures is supported by studies showing expected differences between clinical groups and controls (Giovannetti et al., 2002, 2008a, 2018; Schwartz et al., 2002; Allain et al., 2014; Gold et al., 2015; Rycroft et al., 2018), significant (though modest) relations with cognitive tests (Giovannetti et al., 2002, 2008a, 2018; Schwartz et al., 2002; Kessler et al., 2007; Allain et al., 2014; Rycroft et al., 2018), and informant and clinician reports of functioning (Giovannetti et al., 2002, 2008b; Schwartz et al., 2002; Allain et al., 2014). Detailed analyses of errors and error-types afforded by performance-based tests of everyday function also have promoted theoretical frameworks to better characterize the breakdown of everyday function due to cognitive impairment (see Schwartz, 2006; Giovannetti et al., 2013; for a review). Despite their validity, objectivity and potential for characterization of functional difficulties, performance-based measures have not been widely adopted in clinics or research studies, because generally they require an extraordinary effort to administer and score, especially when used to assess minor difficulties.

Virtual reality (VR) technology has recently gained interest as an effective medium for administering different interventions in the context of healthcare (Cipresso and Serino, 2014; Chirico et al., 2016; Indovina et al., 2018). Several observational studies and a small number of controlled studies have found VR to be effective for a variety of health issues (Cipresso et al., 2016). VR also has been proposed to improve clinical assessments, as automated VR systems could dramatically reduce the time required for administration and scoring traditional performance-based assessments without sacrificing ecological validity. To date, efforts to develop VR systems for assessment of function in older adults have mostly relied on non-immersive systems (Cipresso et al., 2014). In 2014, Allain et al. (2014) reported results from the Virtual Kitchen (VK), a non-immersive activity that required participants to use a mouse to select and move target objects and avoid distractor objects on a computer screen to prepare a cup of coffee. In 2019 Giovannetti et al. (2019) reported preliminary data from a modified VK, called the Virtual Kitchen Challenge (VKC), which included complex tasks to enable assessment of participants with mild cognitive difficulties and requires participants to use a touch screen interface instead of a mouse. Automated scores from the VKC were significantly associated with scores from the same tasks performed with real objects in a real kitchen.

Immersive VR systems also have been proposed to assess everyday function, as they have the advantage of creating a sense of realism or “presence” in the user. Presence is a multidimensional construct that describes the extent to which users believe and feel that they exist in the environment simulated by VR (e.g., kitchen; Diemer et al., 2015) rather than in their true physical location (e.g., clinic/lab; Witmer and Singer, 1998). Presence may be influenced by the quality of the visual scene, method of interaction/interface with the virtual environment, and other factors. Immersive VR assessments of everyday function that elicit a high degree of presence in the user might demonstrate greater ecological and predictive validity of everyday function than non-immersive tasks (Shahrbanian et al., 2012; Parsons, 2015). Although immersive systems afford greater “presence,” they also introduce unique challenges. One challenge, which is particularly salient for older adults, is managing the interface between the user and the surrounding virtual environment, because the immersive context increases the complexity of the task. Using a head-mounted display (HMD), Nolin et al. (2013) and Banville et al. (2017, 2018) implemented an immersive VR task that required participants to use the computer keyboard and mouse to sort everyday objects – a task that would be quite easy for older adults in real-life. Results showed that that older participants took more time to navigate within the virtual environment and to complete the sorting task. Also, older participants were more variable in the time required to accomplish the sorting task as compared to younger participants. These findings underscore the importance of the comfort and ease of the interface, which should feel familiar to the user and optimize mobility. Many immersive VR hardware solutions have been introduced, such as data gloves or controllers, some with haptic feedback; however, they generally prove to be too expensive and require substantial set up time. New, low-cost and ready-to-use devices, such as advanced controllers, could keep costs and administration time low and promote presence in the user during the interaction (Caggianese et al., 2019).

Advanced controllers (hereafter controllers) include buttons and tactile surfaces that are manipulated by the participant. Controllers offer indirect tracking of the position and orientation of the participant’s body. In contrast, egocentric sensors (hereafter sensors) are head-mounted small sensing devices used to detect and track the users’ hands from images acquired from the users' point of view, directly transforming hands and finger movements into interactions with virtual objects. Both controllers and sensors allow the user to see the movement of her/his hands while being immersed in a virtual environment. A recent study comparing the most frequently used controllers (HTC Vive Controllers) and sensors (Leap Motion) with three simple manipulation tasks (i.e., select, position and rotate virtual objects) in eight participants aged 30–40 years showed an advantage for Vive Controllers, which were more stable, accurate, and easier to learn than the Leap Motion sensor (Caggianese et al., 2019).

The aim of the present study was to evaluate the feasibility of a fully immersive VR environment for the assessment of everyday function in older adults. We present a detailed case report of an elderly woman (Tina) who was selected because she represents a typical older adult with no particular computer or technological expertise and an average level of education. Tina was observed while performing an everyday activity in an immersive VR context with two different types of interfaces (controller vs. sensor). VR performance was compared against performance of the same task with real objects outside of the VR system. Comparisons were made on several dimensions, including (1) quality of task performance (e.g., order of task steps, errors, use and speed of hand movements); (2) subjective impression (e.g., attitudes, presence, cybersickness, and usability), and (3) physiological markers of stress.

MATERIALS AND METHODS

The Participant

Tina is a 91-year-old, single women living independently in Northern Italy in a community-residence for older adults. Tina was born in Italy and is a native Italian speaker. At the time of the study she reported that she was functioning independently and had no current or past neurological or psychiatric disorders or other major medical illness (e.g., dementia, brain injury, schizophrenia, depression, etc.). She demonstrated no sensory or motor deficits that precluded interaction with a Head Mounted Display and controllers/sensors. Tina was recruited as a volunteer through an announcement made at her residence.

Procedure

The study was approved by the Ethical Committee in the Department of Psychology of Developmental and Socialization Processes at "Sapienza" University of Rome. All procedures were completed in a single 2- to 3-h session that included the following (in order of administration): (1) informed consent obtained by the participant, (2) screening interview, (3) training on the Virtual Reality Action Test (VRAT) with controllers, (4) testing on the VRAT with controllers followed by presence and attitudes questionnaires, (5) testing on the Real Action Test followed by presence and attitudes questionnaires (6) VRAT sensor training; (7) VRAT sensor testing followed by presence and attitudes questionnaire, and (8) cognitive tests and questionnaires of mood, anxiety and everyday function.

Performance-Based Functional Tests

The breakfast task was administered across all platforms: Real Action Test and Virtual Reality Action Test (with two different controllers). The breakfast task was selected because it has been widely studied as part of the Naturalistic Action Test (NAT), a performance-based test developed to evaluate the cognitive difficulties associated with the completion of everyday activities in people with neurologic impairment (Schwartz et al., 2002). The breakfast task requires participants to prepare a slice of toast with butter and jelly and a cup of coffee with milk and sugar while seated at a table containing a toaster, two knives, one spoon, butter in butter dish, sugar in a bowl, bottle of milk, mug filled with warm water, bread, instant coffee, jelly jar, and a napkin at the central workspace. The shape of the table and the spatial arrangement of objects was informed by procedures in the NAT manual[1].

The breakfast task was administered in real and two different VR conditions (described below). In each condition, Tina was instructed to complete the task in silence, as quickly as possible, and without making errors. She was asked to make her movements as clear as possible and to tell the examiner when she was finished. Performance was recorded for scoring. Physiological and kinematic data were obtained while the participant completed the breakfast task according to the procedures described below.

Real Action Test (RAT)

The RAT required the participant to complete the breakfast task without feedback using real objects. The participant performed the RAT while wearing a smart band and wireless controllers (described below) attached to her arms to acquire kinematic and physiological data (see **Figure 1**).

Virtual Reality Action Test (VRAT)

The VRAT is a VR version of the breakfast task designed to maximize ecological validity by simulating a real kitchen and household objects. In this respect, the VRAT environment is characterized by a high degree of realism, including accurate 3D

[1] https://mrri.org/wp-content/uploads/2016/01/NATManual.pdf

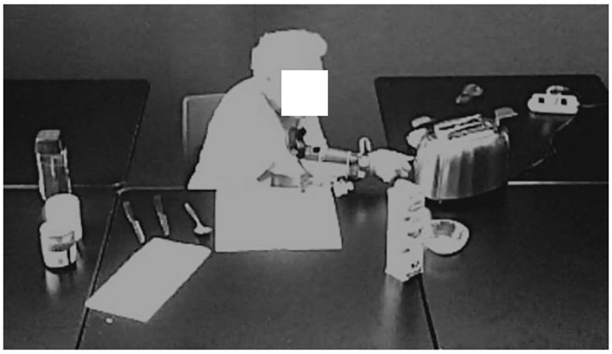


FIGURE 1 | The subject (Tina) performing the Real Action Test (RAT).

models and spatial audio. The VRAT includes automatic, real-time collection of movement data, as well as physiological and kinematic parameters (described below).

VRAT Apparatus and Controller Conditions

The VRAT system runs on a MSI Trident Gaming Desktop, with 8GB RAM and a GTX 1060 graphic card. The HTC Vive head mounted display[2] provides users with a fully immersive virtual environment. The HTC Vive visual system is based on two OLED displays for a total resolution of 2160 × 1200 pixels with a 110-degree FoV and a frequency of 90 Hz. The VR software was developed with Unity3D[3], a game development platform which provides native VR support.

Interaction in the VRAT was enabled through two different input devices: (1) *controllers* – the participant used HTC Vive controllers that provided tactile feedback through vibration; (2) *sensors* – a wearable egocentric sensor, the Leap Motion Controller[4], enabled interaction through movements of the participant's own hands. Performance with the two different devices were tested in different conditions.

Controllers: were worn during performance of the RAT and the VRAT-controller conditions. During the RAT, participants did not interact with the controller; it was used only to collet kinematic data. However, in the VRAT, the controller was used to interact with the VR environment while the participant was in a seated position using interaction metaphors similar to those used in real-life. To make the interaction as familiar and natural as possible, we implemented the Virtual Hand metaphor (Ruddle, 2005), in which the user's hand motions are directly mapped to the virtual hand movements. When the virtual hand reaches an object, the object is highlighted to inform the user through visual feedback that it is selected and interactable. To interact with a virtual object in the VRAT, the user is instructed to press the trigger button once the object is highlighted/selected. To end the interaction, the user is instructed to release the trigger. One advantage of the controller is that the participant is able to be tracked even when the user's hands are not visible within the user field of view, allowing a wider measurement area. Controllers also provide users with tactile feedback through vibrations of varying intensity. However, interactions with virtual objects occur through a tool that the user must always hold in the hands, even when they are not interacting with any object, reducing the naturalness of the interaction.

Virtual Reality Action Test sensors: were used during performance of the VRAT-sensor condition. In this condition, the participant interacted with virtual objects using Leap sensors by performing a pinch gesture (i.e., moving thumb and index fingers closer until they come into contact). To release the virtual object(s) the pinch gesture is relaxed. The Leap sensor allows the user to interact with virtual objects with their own hands, without having to wear gloves or hold controllers. Unlike the controllers, the sensor is able to track the main joints of the user's hand and replicate them in the virtual environment, increasing the hand representation and the sense of presence. However, the interaction area is limited to the tracking area of the sensor and the user's field of view. The sensor is mounted in front of the headset; therefore, the user must keep their hands in their field of view to interact with virtual objects. Furthermore, tracking may fail if the hand is occluded by the user's other hand or an obstacle/object in the real world.

Participants completed the RAT and both VRAT conditions while wearing a smart bracelet (Microsoft band 2) that was designed to obtain physiological measures of stress (described below).

Software Architecture

The system was designed as a multiplayer platform: one player is the participant, who performs the task within the virtual environment, and the other player is the examiner, who configures the test, and monitors, in real time, the scores and physiological parameters of the participant. The system includes a VR module that maps the data acquired by the HMD and input devices into the corresponding virtual actions within the virtual kitchen. The game logic of the breakfast task, including the physical features and behavior of each virtual element on the table, is coded in the VR module. An error checking module has been developed for automatically detecting an error by the participant. For each participant action during the task, the error checking module considers the virtual environment state, and through a specified set of rules, interprets the participant action as either an error or correct action. Each time the participant commits an error, it notifies the logger module. The logger module acquires data from various sources (error checking module, HMD, input devices) and synchronizes them under a single time value, making it possible to link all of the separate data streams (i.e., knowing the physiological state of the participant when she/he commits an error). All information is saved as. csv files at the end of the test. The examiner interface allows the examiner to manage the test from the control panel and view errors committed by the participant as well as physiological values in real time.

VR Training

Before each VR condition, the participant completed a brief training session with the system. Training included four mini-tasks that comprised elements of the breakfast task: (1) toast a slice of bread; (2) spread the jelly on toast; (3) add instant coffee to cup; (4) add milk to cup. The examiner controlled the presentation of each mini-task from a monitoring position.

Quality of Task Performance

Although the VRAT includes the error monitoring module, performance quality and accuracy on the RAT and two VRAT conditions were evaluated by trained coders who viewed recordings of the participant's performances. The following error scores were collected for each of the three conditions (RAT, VRAT-controller, VRAT-sensor):

Total overt errors – incorrect actions (commission), the failure to complete a step (omission), and off-task actions (additions)

[2] https://www.vive.com/eu/

[3] https://unity.com/

[4] https://www.leapmotion.com/

were recorded and assigned a code according to the error taxonomy shown in **Table 1** (Schwartz et al., 2002).

Total micro-errors – subtle, inefficient but not overtly incorrect actions; this category of errors was added to the overt error taxonomy to improve detection of subtle, inefficient behaviors in healthy people and those with mild cognitive difficulties.

Clumsy-motor imprecision errors during the execution of an accurate task step.

Code sheets with an exhaustive list of overt/micro-errors were used to promote inter-rater reliability and are included in **Supplementary Material**.

In addition to errors, human coders evaluated video recordings for accomplishment, time to completion and the order of task steps as follows:

Accomplishment score – an accomplishment point was assigned for each task step of the breakfast task completed without error (range = 0–16).

Overall performance score – this score combines accomplishment score with the sum of a subset of key, overt errors (Schwartz et al., 2002).

Completion Time – was recorded in seconds; timing began when the first step was initiated and ended when the participant indicated that she was finished with the task.

Order of Task Steps – In addition to coding errors and completion time, the order in which the participant completed each task step was recorded to examine similarities/differences across the RAT and VRAT conditions.

Kinematic measures were obtained by the input devices used in the RAT and VRAT conditions. During the RAT and VRAT-controller conditions, the participant wore wireless controllers, and during the VRAT-sensor condition, the participants movements were recorded by Leap Motion. Kinematic data was obtained to measure the precise movements of both the right and left hands, with an accuracy in millimeters (100 Hz). Instantaneous velocity measures greater than three meters per second were excluded to avoid noisy data due to hand tracking problems in the VRAT-sensor condition. For each condition, the following kinematic measures were obtained:

- Total hand movement, in meters.
- Average speed of the hands, in meters per second, computed as total hand movement divided by completion time.

TABLE 1 | Error Taxonomy Used Code Performance on the RAT and VRAT conditions.

Error type		Definition	Examples
Omission		Number of steps that are not performed	Does not add coffee grounds to coffee; does not add stamp to envelope
Commission	Substitution	Similar, alternate object is used in place of target object	Spreads butter on toast with spoon instead of knife
	Sequence	Anticipation of a step; steps or subtasks performed in reverse order	Butter on bread without toasting; applies jelly on bread then applies butter
	Perseveration	A step is performed more than once or for an excessive amount of time	Adds butter/jelly repeatedly to toast
Action-Additions		Performance of an action not readily interpreted as a task step	Puts toast in creamer
Micro-errors		Initiating and terminating an incorrect action before the error is completed by reaching for, touching or picking up an object	Reaches toward, touches or moves salt but never uses the salt in during the task
Clumsy		Correct step is performed but with difficulty due to motor imprecision	Coffee jar slips out of hand

Subjective Impressions

Immediately following each condition (RAT, VRAT-controller, VRAT-sensor), the participant used a five-point scale to describe her reaction to the test condition on the following five items/dimensions: useless/useful, not pleasant/pleasant, boring/funny, tiring/resting, stressing/relaxing. Item scores were aggregated into a single score, ranging from 5 to 25, for which higher values indicated more positive attitudes about the test condition. This scale was created by the authors of the study according to procedures described by Ajzen (1991); see **Supplementary Material**.

Physiological Measures of Stress

To compare indicators of stress during each testing condition, physiological data were recorded via a smart bracelet (Microsoft band 2)[5] worn by the participant while completing the RAT and both VRAT conditions. Kubios software (Tarvainen et al., 2014) was used to obtain the following variables:

Heart rate (bpm, 1 Hz),
Galvanic Skin Response (kohms, 0,2/5 Hz),
R–R interval (i.e., time between heart beats; seconds, variable frequency),
skin temperature (degrees centigrade, 0,033 Hz).

To correct for artifacts, particularly in the measure of heart rate variability (RR interval), a threshold-based algorithm was applied that compares every RR interval value against a local average interval, obtained by median filtering the RR interval time series. RR interval values that differ from the local average of a specified threshold value (i.e., 0.45 s) are marked as artifact and replaced using cubic spline interpolation.

Physiologic variables (i.e., Heart rate, Galvanic Skin Response, Skin temperature) were used to calculate an index of cardiovascular system stress, called Baevsky's stress index (Baevsky and Berseneva, 2008). The Baevsky's stress index is strongly linked to sympathetic nervous activity and

[5] https://support.microsoft.com/it-it/help/4000323/band-hardware-sensors

increases during stressful situations. Physiologic data were stored on .csv files and although they may be combined with the test start time to synchronize physiological and kinematic information, for the current study, physiologic data were aggregated and averaged for each test condition to obtain an overall stress index per condition (RAT, VRAT-controller, VRAT-sensor).

VRAT Presence, Cybersickness, and Usability

The following questionnaires were administered immediately following performance on the VRAT-controllers and VRAT-sensors conditions.

Presence Questionnaire (PQ)

The Italian version of PQ was administered to the participant in this study (Scheuchenpflug et al., 2003). The PQ required the participant to use a seven-point scale to rate her experience with each condition on 28 items focused on the following features: Realism (7 items); Possibility to act (4 Items); Quality of interface (3 Items); Possibility to examine (3 items); Self-evaluation of performance (2 Items) (Witmer and Singer, 1998; Slater, 2002; Witmer et al., 2005). Strong internal reliability has been reported (0.88) for the total score.

Cybersickness Symptoms

The Virtual Reality Symptom Questionnaire (VRSQ), developed by Ames in 2005 (Ames et al., 2005), was administered immediately after the VRAT-controllers condition and the VRAT-sensors condition to evaluate symptoms of cybersickness, a type of motion sickness caused by exposure to VR. The questionnaire assesses eight general physical side effects (general discomfort, fatigue, boredom, drowsiness, headache, dizziness, concentration difficulties, and nausea) and five visual effects (tired eyes, aching eyes, eyestrain, blurred vision, and difficulties focusing) on a seven-point scale (0–6), with 0-scores indicating no symptoms and higher scores indicating more severe symptoms. In the validation study, only symptoms that met a minimum correlation coefficient value of 0.2 with the total score were included in the final measure. The Italian version of the VRSQ (Solimini et al., 2011) was used with the participant in this study.

System Usability Scale (SUS)

The SUS is a 10-item measure that required the participant to use a five-point scale ranging from strongly disagree to strongly agree to indicate the extent to which they agree/disagree with positive and negative statements about the VRAT-controller and VRAT-sensor conditions (Brooke, 1996). SUS responses were transformed to a single score ranging from 0 to 100, with higher scores reflecting more favorable usability. The SUS is considered a robust measure of system usability (Bangor et al., 2008), even with a small sample size (Tullis and Stetson, 2004). The Italian version of the SUS was used in this study (Borsci et al., 2009).

Mood, Anxiety, and Cognition

Questionnaires of mood and anxiety symptoms, disposition toward immersive tendencies, and cognitive and functional abilities as well as neuropsychological tests of global and specific cognitive abilities were administered by a trained psychologist (AC). When available, Italian validated versions of questionnaires/tests were used; other measures were translated using a back-translation procedure (see **Table 2**).

Analysis Plan

Descriptive analyses of questionnaires and cognitive tests were performed to characterize the participant. Cognitive test scores also were evaluated by calculating the standardized (Z) score for the participant relative to normative data, using samples that were comparable to the age and education level of the participant. The following formula was used to calculate the Z-score (participant's raw test score – mean of the normative sample/E.S. of the normative sample).

Descriptive data from the RAT, VRAT-controllers, and VRAT-sensors were obtained to compare performance across the testing conditions on measures of (1) the quality of task performance (e.g., errors, accomplishment, time to completion, order of task steps, errors, use and speed of hand movements, etc.); (2) subjective impressions (e.g., attitudes, presence, cybersickness, and usability), and (3) physiological markers of stress.

RESULTS

Characterization of the Participant

Mood Status

Tina's report of depression (Geriatric Depression Scale = 4) and anxiety (Geriatric Anxiety Scale = 12) symptoms was well within the non-clinical range (cut-off scores: GDI > 11; GAI > 17) (Yesavage et al., 1982; Segal et al., 2010; Gould et al., 2014; Galeoto et al., 2018; Gatti et al., 2018).

Cognitive Testing

Raw cognitive test scores along with age- and education-adjusted normative-based Z-scores are reported in **Table 3**. Tina's overall cognitive status, as measured by the MMSE was well within the range of healthy, non-demented people. Scores on most tests of specific abilities fell within the average range, including tests of verbal episodic memory, processing speed, executive functions, and verbal fluency. She performed in the high average range on a verbal test of executive function and in the low average range on a test of visual episodic memory (immediate and delayed free recall).

On questionnaires of cognitive and functional abilities, Tina reported no significant change in her cognitive abilities as compared to 10 years ago [The ECOG SF12 = 1.75, cut-off score = 2.30 (Farias et al., 2008)] and minimal functional difficulties within the normal range [FAQ (score = 6) and the ADL-PI (score = 22)].

On a questionnaire pertaining to one's personal disposition toward immersion (ITQ), Tina reported an average level of

TABLE 2 | Mood and neuropsychological tests performed to characterize the participant.

Variable	Test	Original scale citation	Italian scale used for the study	Validity/Reliability of the instrument
Depression	Geriatric Depression Scale (GDS)	Yesavage et al., 1982	Galeoto et al., 2018	Cronbach's Alpha scored 0.84 in the Italian validated study (Galeoto et al., 2018)
Anxiety	Geriatric Anxiety Scale	Segal et al., 2010	Gatti et al., 2018	Cronbach's Alpha of the Italian scale was = 0.88 (Gatti et al., 2018)
Cognitive abilities	The Everyday Cognition scale short form 12 (ECOG SF12)	Farias et al., 2011	Back Translation procedure has been made for the study purposes	E-Cog has been reported to have high internal consistency (α = 0.96). Additionally, the scale demonstrates good test–retest reliability (*r* = 0.82) (Farias et al., 2011)
Functional activities	Functional Activity Questionnaire (FAQ)	Pfeffer et al., 1982	Stancati and Salussi, 2001	The scale has been reported to have high internal consistency (α > 0.90) (Pfeffer et al., 1982)
Daily living activities	The Activities of Daily Living-Prevention Instrument (ADL-PI)	Galasko et al., 2006	Back Translation procedure has been made for the study purposes	Test–Retest reliability: was *r* = 0.74 (Galasko et al., 2006)
Education	Brief Intelligence Test (Test di Intelligenza Breve; TIB)	Colombo et al., 2002	Original scale is in Italian	Cronbach's Alpha scored 0.91 (Colombo et al., 2002)
Visual memory	Brief Visual Memory Test Revised (BVMT – R)	Benedict et al., 1996	Argento et al., 2016	Test–retest reliability coefficients ranged from 0.60 for Trial 1 to 0.84 for Trial 3 (Argento et al., 2016)
Verbal fluency	Category Fluency	Sivan and Benton, 1984	Novelli et al., 1986	Test–retest reliability coefficients for the scale was > 0.75 (Kingery et al., 2011)
Processing speed	Trail Making Test-Part B	Armitage, 1946	Gaudino et al., 1995	Validity of the test has been extensively discussed and confirmed (for an extensive review see Sánchez-Cubillo et al., 2009)
Working memory	Digit Span backward	Wechsler et al., 2008	Monaco et al., 2013	The test reliability scored 0.89 (Orsini and Pezzuti, 2015)
Processing speed and visual perception	Symbol search	Wechsler et al., 2008	Orsini and Pezzuti, 2015	The test reliability scored 0.88 (Orsini and Pezzuti, 2015)
Personal disposition toward immersion	Immersive Tendencies Questionnaire	Witmer and Singer, 1998	Scheuchenpflug et al., 2003	The scale reliability scored 0.81 (Witmer and Singer, 1998)

immersion in terms of ability to focus and to become deeply involved in activities (Witmer and Singer, 1998).

Comparisons Across the RAT, VRAT-Controllers, and VRAT-Sensors

Performance Quality

As shown in **Table 4**, Tina made few errors on the breakfast task across all conditions, with most errors on the VRAT-controllers condition. She made no clumsy errors on the RAT, but an equal number of clumsy errors on both VRAT conditions. The Overall Performance Score, which considers accomplishment and the performance of key overt errors was identical across the conditions. Time to completion, also shown in **Table 4**, revealed a longer completion time for the VRAT – controllers than the other two conditions.

A qualitative analysis of the order in which steps were performed showed remarkable consistency. Task steps were performed in the following order across all three conditions: take bread, place bread in toaster, turn on toaster, wait for bread to

TABLE 3 | Characterization of the participant adjusted for age and education.

Test	Subtest	Raw score	Standard score (*z*-score)	Qualitative descriptor
MMSE		29/30		Within normal limits
BVMT[1]				
	Trial 1	3	−4.66	Impaired
	Trial 2	4	−3.79	Impaired
	Trial 3	8	−1.21	Low Average
	Learning trial	5	−0.18	Average
	Delayed recall trial	5	−2.95	Impaired
RAVLT[2]				
	Total score	45	1.52	Average
	Delayed recall	8	0.20	Average
	Recognition hits	15	2.25	Average
Symbol search[3]		19	1.67	Average
TMT-B[4]		298.21	0.48	Average
DIGIT SPAN backward[5]		5	3.37	High average
Fluency global[6] (Categories: car brand, animal, fruit)		52	2.52	Average
Questionnaires pertaining to cognition, everyday function, and immersive tendencies FAQ		6		
ECOG-Short Form		1.75		
ADL PI		22		
ITQ				
	Focus subscale	31		
	Involvement subscale	21		

BVMT, Brief Visual Memory Test Revised; RAVLT, Rey Auditory Learning Test; TMT-B, Trail Making Test-Part B; FAQ, Functional Activity Questionnaire; ECOG. SF-12, The Everyday Cognition scale short form 12; ADL-PI, The Activities of Daily Living-Prevention Instrument; ITQ, Immersive Tendencies Questionnaire. [1] Normative data for the Italian population (Argento et al., 2016); [2] Normative data (Carlesimo et al., 2002); [3] Normative data for the Italian population (Orsini and Pezzuti, 2015); [4] Normative data for the Italian population (Giovagnoli et al., 1996); [5] Normative data for the Italian population (Monaco et al., 2013); [6] Normative data for the Italian population (Novelli et al., 1986).

toast, remove bread from toaster, add butter to toast, add jelly to toast, add coffee to mug, add milk to mug, add sugar to mug. The final step of stirring the coffee was completed only in the RAT. Tina did not stir the virtual coffee mug in either the VRAT-controller or VRAT-sensor condition; this was coded as an overt (omission) error in both of the VRAT conditions.

Kinematic Results

Hand movements and average hand speed are also shown in **Table 4**. The same pattern of hand movement distance and speed was observed across all conditions – the right hand made more and faster movements than the left hand. There were few differences across conditions, except for a greater reliance on the right hand in the VRAT-controller condition.

A heatmap showing the paths of the right and left hand during each condition is shown in **Figure 2**. Note that the heatmap for the RAT was superimposed on a virtual display for presentation purposes only. The participant actually completed the RAT using real objects as shown in **Figure 1**. The heat maps illustrate subtle differences across conditions. In the RAT, the participant used both hands to perform the steps (i.e., using her left hand to grab the milk bottle, the butter dish and sugar bowl), with each hand performing tasks in the corresponding hemispace. In the VRAT conditions, particularly in the VRAT-controller condition, the participant used the dominant, right hand more frequently, even when completing subtasks in the opposite (left) hemispace.

Physiological Markers

As expected, the lowest stress index was obtained during the RAT (stress index = 4.1); followed by the VRAT-controller (stress index = 4.9) and VRAT-sensor (stress index = 6.2). This result suggests that the participant felt more comfortable with controllers rather than in the sensor condition without the controllers (**Table 4**).

Subjective Impressions

As shown in **Table 4**, Tina reported the most positive attitude toward the VRAT-controllers (Total = 25/25) and the RAT (Total = 24/25). She indicated the lowest score for the VRAT-sensor condition (16/25), as she reported that the VRAT-sensor condition was less "pleasant," "funny," "resting," and "relaxing" (each scored 3 out of 5).

Measures of presence, cybersickness, and usability were obtained after each of the VRAT conditions. As shown in **Table 4**, Tina reported a stronger feeling of presence in the VRAT-controllers than in the VRAT-sensors condition (PQ). Scores for each of the PQ subscales, except the "quality of interface" scale were all higher in the VRAT-controller condition (see **Table 4**). Tina reported no symptoms of cybersickness on VRSQ for either condition (Ames et al., 2005). Finally, Tina reported higher usability ratings for the VRAT-controllers than the VRAT-sensors condition.

DISCUSSION

This paper reports the detailed analysis of a 91-year old woman's (Tina) performance of a real (RAT) and immersive VR breakfast task (VRAT) to evaluate the feasibility of immersive VR for the assessment of everyday function in older adults. Two different VR interfaces were examined: controllers and sensors. Results showed similarities in performance quality, stress, and subjective reports between the RAT and both VRAT conditions, as well as positive ratings and no cybersickness for either VR condition. Taken together the results demonstrate the feasibility of immersive VR for function assessment in older adults and suggest the potential of the validity of this method.

Our results clearly demonstrate the feasibility of immersive VR for function assessment, even in an older adult with very limited computer experience, no prior VR exposure, average educational experiences, and mild cognitive difficulties.

TABLE 4 | Quality task performance, kinematic, physiological data and system usability.

	RAT	VRAT-controllers	VRAT-sensors
Performance analysis			
Accomplishment (%)	100	100	100
Total overt errors	0	1	1
Total micro-errors	3	6	0
Total clumsy errors	0	7	7
Overall Performance (Max = 6)	6	6	6
Completion Time	108.72	203.74	165.81
Kinematic analysis			
Total hand movement (m)			
Right hand	22.91	43.97	28.77
Left hand	11.3	6.71	11.6
Total hand speed (m/s)			
Right hand	0.21	0.21	0.17
Left hand	0.1	0.03	0.07
Physiological data – Mean (SD)			
Baevsky's stress index	4.1	4.9	6.2
Heart rate (bpm)	72.34 (1.54)	68.79 (3.63)	78.15 (2.29)
Galvanic skin response (kohms)	2467 (200)	4714 (820)	893 (44)
Skin temperature (celsius degree)	35.16 (0.07)	34.99 (0.13)	35.22 (0.02)
System usability			
System usability (SUS)		62.5/100	50/100
Sense of presence			
Sense of presence global score (PQ)		113/126	100/126
Realism – subscale		39/49	37/49
Possibility to act – PQ subscale		27/28	18/28
Quality of the interface – PQ subscale		16/21	18/21
Possibility to examine – PQ subscale		19/21	17/21
Self-Evaluation of the performance – PQ subscale		12/14	10/14

SUS, System Usability Scale; PQ, Presence Questionnaire.

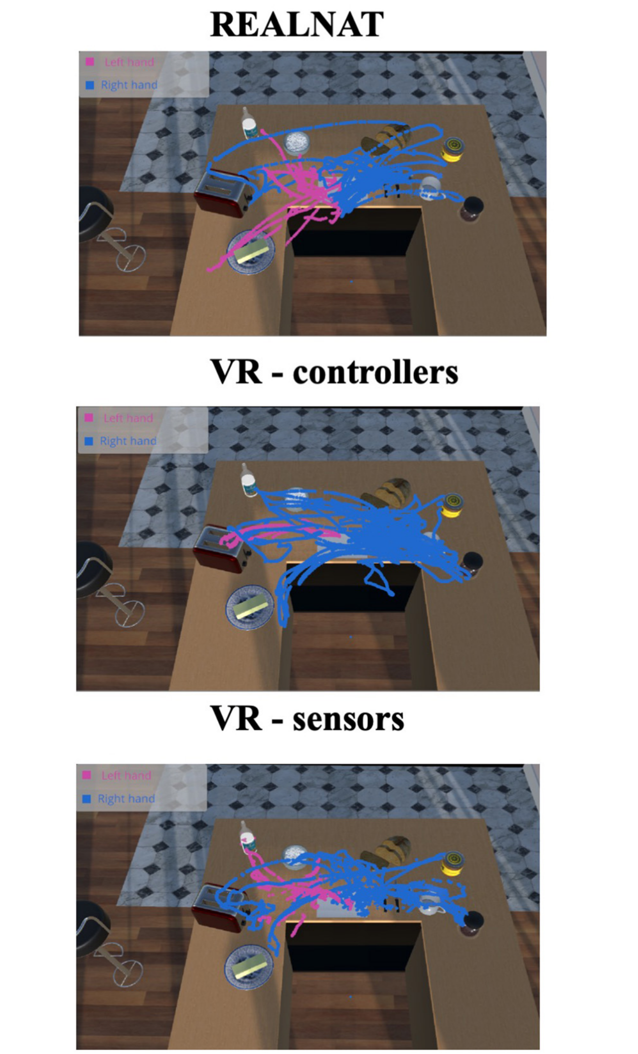


FIGURE 2 | Hands heat map for the three different experimental conditions.

The participant was capable of using controllers and sensors to manipulate objects in a purposeful and goal-directed manner in the VR paradigm. She reported no cybersickness and even indicated that interactions in the VR environment were pleasant and relaxing.

Our results also suggest the potential validity of the VR paradigm, as overall performance and accomplishment scores were similar, and task steps were completed in exactly the same order across conditions, even though there were numerous opportunities for variation in the order of steps (e.g., coffee could have been made before toast and the order of cream and sugar and butter and jelly was not fixed). Kinematic analyses also were generally comparable between the real (RAT) and the VRAT-sensor condition, and the participant reported positive attitudes toward real (RAT) and both VRAT tasks. These similarities are striking considering that immersive VR was completely unfamiliar to the participant.

Some important differences between the real and VR paradigms were observed and should inform future research. For example, the participant required less time and demonstrated a lower stress index while completing the real breakfast task (RAT). She also demonstrated fewer clumsy errors in the real task as compared to both VRAT conditions. These differences suggest that the real condition was considerably easier for the participant. Training with the VR controllers and sensors was minimal in the present study, and the participant had no prior experience with VR. Future studies that use VR with older adults should consider including more training to determine whether increased familiarity with the VR environment and practice with VR controllers/sensors may further reduce differences between real and virtual everyday task performance.

In contrast to past research with healthy participants showing advantages with controllers (Caggianese et al., 2019), our results do not clearly indicate which VR interface is best for the assessment of function in older adults, as each interface showed different strengths and weaknesses. When using the controllers, the participant made more micro-errors, and kinematic analyses showed a pattern of hand use that was dissimilar from performance on the real task, such that she appeared to favor her dominant (right) hand for completing the tasks in the VRAT-controller condition. However, she subjectively reported that she preferred the controllers, with higher ratings for usability and positive attitude toward the VRAT-controllers condition. Physiological indicators also reflected lower stress when she used the controllers (VRAT-controllers) than when she used the sensors (VRAT-sensors). By contrast, with the sensors, the participant showed a more natural pattern of use of the right and left hands (see kinematic data). Taken together, the results suggest that if problems in precisely controlling movements in the sensor interface could be addressed in future research, the sensor interface has potential to offer more accurate and naturalistic assessments of everyday function for older adults than controllers.

There are several limitations to acknowledge. First, the extent to which the results are influenced by order effects cannot be determined from this single case report. Future studies should control for and examine task order and practice effects on virtual and real everyday tasks. Future studies with more participants are needed to determine whether our results are generalizable and to establish the full psychometric properties of the VRAT.

In conclusion, our results support the feasibility of immersive VR as a tool to evaluate everyday function in older adults considering also the evaluated safety of the technology as suggested by a recent meta-analysis (Kourtesis et al., 2019). The results also provide guidance on considerations for VR interfaces (sensors vs. controllers). Because of its strong potential to offer objective, sensitive and standardized assessment of everyday function in older adults and a wide range of clinical populations future research on VR assessments is needed to identify optimal interfaces and procedures, compare the utility against non-immersive VR methods (Allain et al., 2014; Giovannetti et al., 2018), and ultimately establish the psychometric properties of immersive VR measures of everyday function. Moreover, the potential for immersive VR systems to offer interventions that might improve everyday functioning and promote independence should be explored (Banville et al., 2018; Foloppe et al., 2018).

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethical Committee – Department of Psychology of Developmental and Socialization Processes, Sapienza University of Rome, Rome, Italy. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in the manuscript.

AUTHOR CONTRIBUTIONS

All authors conceived and planned the experiments, contributed to the interpretation of the results, provided critical feedback, and helped to shape the research, analysis, and manuscript. AC, FeG, FrG, and PN carried out the experiments. PN and LG programmed the software. AC managed and analyzed data. TG and SS coded visual data. AG and TG took the lead in writing the manuscript.

FUNDING

This project has been supported by the Sbarro Health Research Organization (www.shro.org). TG's time on this project was funded by a grant from the National Institute on Aging Grant R21AG060422.

ACKNOWLEDGMENTS

We would like to thanks Ms. Angela Romano in coordinating the data collection procedure and recruiting.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00123/full#supplementary-material

REFERENCES

Ajzen, I. (1991). The theory of planned behavior. *Organ. Behav. Hum. Decis. Process.* 50, 179–211. doi: 10.1016/0749-5978(91)90020-T

Allain, P., Foloppe, D. A., Besnard, J., Yamaguchi, T., Etcharry-Bouyx, F., Le Gall, D., et al. (2014). Detecting everyday action deficits in Alzheimer's disease using a nonimmersive virtual reality kitchen. *J. Int. Neuropsychol. Soc.* 20, 468–477. doi: 10.1017/S1355617714000344

Ames, S. L., Wolffsohn, J. S., and Mcbrien, N. A. (2005). The development of a symptom questionnaire for assessing virtual reality viewing using a head-mounted display. *Optom. Vis. Sci.* 82, 168–176. doi: 10.1097/01.OPX.0000156307.95086.6

Argento, O., Smerbeck, A., Pisani, V., Magistrale, G., Incerti, C. C., Caltagirone, C., et al. (2016). Regression-based norms for the brief visuospatial memory test-revised in Italian population and application in MS patients. *Clin. Neuropsychol.* 30, 1469–1478. doi: 10.1080/13854046.2016.1183713

Armitage, S. G. (1946). An analysis of certain psychological tests used for the evaluation of brain injury. *Psychol. Monogr.* 60, 41–48. doi: 10.1037/h0093567

Baevsky, R. Ì, and Berseneva, A. P. (2008). Methodical Recommendations use kardivar System for Determination of the Stress Level and Estimation of the Body Standards of Measurements and Physiological Interpretation. Available at: https://pdfs.semanticscholar.org/74a2/92bfafca4fdf1149d557348800fcc1b0f33b.pdf (accessed August 16, 2019).

Bangor, A., Kortum, P. T., and Miller, J. T. (2008). An empirical evaluation of the system usability scale. *Int. J. Hum. Comput. Interact* 24, 574–594. doi: 10.1080/10447310802205776

Banville, F., Couture, J. F., Verhulst, E., Besnard, J., Richard, P., and Allain, P. (2017). "Using virtual reality to assess the elderly: The impact of human-computer interfaces on cognition," in *Lecture Notes in Computer Science*, (Berlin).

Banville, F., Lussier, C., Massicotte, E., Verhulst, E., Couture, J. F., Allain, P., et al. (2018). Lecture Notes in Computer Science, ln. *Validation of a Sorting Task Implemented in the Virtual Multitasking Task-2 and Effect of Aging.* (Berlin).

Benedict, R. H. B., Groninger, L., Schretlen, D., Dobraski, M., and Shpritz, B. (1996). Revision of the brief visuospatial memory test: studies of normal performance, reliability, and, validity. *Psychol. Assess* 8, 145–153. doi: 10.1037/1040-3590.8.2.145

Borsci, S., Federici, S., and Lauriola, M. (2009). On the dimensionality of the System usability scale: a test of alternative measurement models. *Cogn. Process* 10, 193–197. doi: 10.1007/s10339-009-0268-9

Brooke, J. (1996). "SUS - A quick and dirty usability scale Industrial usability evaluation," in *Usability Evaluation In Industry*, eds P. W. Jordan, B. T. I. L. McClelland, and B. Weerdmeester, (Boca Raton, LA: CRC Press).

Caggianese, G., Gallo, L., and Neroni, P. (2019). "The vive controllers vs. leap motion for interactions in virtual environments: a comparative evaluation," in *Smart Innovation, Systems and Technologies*, eds H. Robert, and J. C. Lakhmi, (Berlin: Springer).

Carlesimo, G. A., Buccione, I., Fadda, L., Graceffa, A., Mauri, M., Lorusso, S., et al. (2002). Normative data of two memory tasks: short-Story recall and Rey's Figure. *Nuova Riv. di Neurol.* 2, 1–13.

Chirico, A., Lucidi, F., De Laurentiis, M., Milanese, C., Napoli, A., and Giordano, A. (2016). Virtual reality in health system: beyond entertainment. a mini-review on the efficacy of VR during cancer treatment. *J. Cell. Physiol.* 231, 275–287. doi: 10.1002/jcp.25117

Cipresso, P., Albani, G., Serino, S., Pedroli, E., Pallavicini, F., Mauro, A., et al. (2014). Virtual multiple errands test (VMET): a virtual reality-based tool to detect early executive functions deficit in parkinson's disease. *Front. Behav. Neurosci.* 8:405. doi: 10.3389/fnbeh.2014.00405

Cipresso, P., and Serino, S. (2014). *Virtual Reality: Technologies, Medical Applications and Challenges*. Hauppauge, NY: Nova Science Publishers, Incorporated.

Cipresso, P., Serino, S., and Riva, G. (2016). Psychometric assessment and behavioral experiments using a free virtual reality platform and computational science. *BMC Med. Inform. Decis. Mak.* 16:37. doi: 10.1186/s12911-016-0276-5

Colombo, L., Sartori, G., and Brivio, C. (2002). Stima del quoziente intellettivo tramite l'applicazione del TIB (test breve di intelligenza). *G. Ital. Psicol.* 3, 613–637.

Diemer, J., Alpers, G. W., Peperkorn, H. M., Shiban, Y., and Mühlberger, A. (2015). The impact of perception and presence on emotional reactions: a review of research in virtual reality. *Front. Psychol.* 6:26. doi: 10.3389/fpsyg.2015.00026

Farias, S. T., Mungas, D., Harvey, D. J., Simmons, A., Reed, B. R., and Decarli, C. (2011). The measurement of everyday cognition: development and validation of a short form of the Everyday Cognition scales. *Alzheimer's Dement.* 7, 593–601. doi: 10.1016/j.jalz.2011.02.007

Farias, S. T., Mungas, D., Reed, B. R., Cahn-Weiner, D., Jagust, W., Baynes, K., et al. (2008). The measurement of everyday cognition (ECog): scale development and psychometric properties. *Neuropsychology* 22, 531–544. doi: 10.1037/0894-4105.22.4.531

Foloppe, D. A., Richard, P., Yamaguchi, T., Etcharry-Bouyx, F., and Allain, P. (2018). The potential of virtual reality-based training to enhance the functional autonomy of Alzheimer's disease patients in cooking activities: A single case study. *Neuropsychol. Rehabil.* 28, 709–733. doi: 10.1080/09602011.2015.1094394

Galasko, D., Bennett, D. A., Sano, M., Marson, D., Kaye, J., and Edland, S. D. (2006). ADCS Prevention Instrument Project: Assessment of instrumental activities of daily living for community-dwelling elderly individuals in dementia prevention clinical trials. *Alzheimer Dis. Assoc. Disord.* 20(4 Suppl. 3), S152–S169. doi: 10.1097/01.wad.0000213873.25053.2b

Galeoto, G., Sansoni, J., Scuccimarri, M., Bruni, V., De Santis, R., Colucci, M., et al. (2018). A psychometric properties evaluation of the Italian version of the geriatric depression scale. *Depress. Res. Treat.* 2018:1797536. doi: 10.1155/2018/1797536

Gatti, A., Gottschling, J., Brugnera, A., Adorni, R., Zarbo, C., Compare, A., et al. (2018). An investigation of the psychometric properties of the Geriatric Anxiety Scale (GAS) in an Italian sample of community-dwelling older adults. *Aging Ment. Heal.* 22, 1170–1178. doi: 10.1080/13607863.2017.1347141

Gaudino, E. A., Geisler, M. W., and Squires, N. K. (1995). Construct validity in the trail making test: what makes Part B harder? *J. Clin. Exp. Neuropsychol.* 17, 529–535. doi: 10.1080/01688639508405143

Giovagnoli, A. R., Del Pesce, M., Mascheroni, S., Simoncelli, M., Laiacona, M., and Capitani, E. (1996). Trail making test: normative values from 287 normal adult controls. *Ital. J. Neurol. Sci.* 17, 305–309. doi: 10.1007/BF01997792

Giovannetti, T., Bettcher, B. M., Brennan, L., Libon, D. J., Burke, M., Duey, K., et al. (2008a). Characterization of everyday functioning in mild cognitive impairment: a direct assessment approach. *Dement. Geriatr. Cogn. Disord.* 18, 787–798. doi: 10.1159/000121005

Giovannetti, T., Bettcher, B. M., Brennan, L., Libon, D. J., Kessler, R. K., and Duey, K. (2008b). Coffee with jelly or unbuttered toast: commissions and omissions are dissociable aspects of everyday action impairment in Alzheimer's Disease. *Neuropsychology* 22, 235–245. doi: 10.1037/0894-4105.22.2.235

Giovannetti, T., Libon, D. J., Buxbaum, L. J., and Schwartz, M. F. (2002). Naturalistic action impairments in dementia. *Neuropsychologia.* 40, 1220–1232. doi: 10.1016/S0028-3932(01)002299

Giovannetti, T., Richmond, L. L., Seligman, S. C., Seidel, G. A., Iampietro, M., Bettcher, B. M., et al. (2013). "A process approach to everyday action assessment," in *The Boston Process Approach to Neuropsychological Assessment: A Practitioner's Guide*, eds L. Ashendorf, R. Swenson, and D. Libon (Oxford University Press), 355–379.

Giovannetti, T., Yamaguchi, T., Roll, E., Harada, T., Rycroft, S. S., Divers, R., et al. (2018). The Virtual Kitchen Challenge: preliminary data from a novel virtual reality test of mild difficulties in everyday functioning. *Aging, Neuropsychol. Cogn.* 26, 823–841. doi: 10.1080/13825585.2018.1536774

Giovannetti, T., Yamaguchi, T., Roll, E., Harada, T., Rycroft, S. S., Divers, R., et al. (2019). The virtual kitchen challenge: preliminary data from a novel virtual reality test of mild difficulties in everyday functioning. *Aging, Neuropsychol. Cogn.* 26, 823–841. doi: 10.1080/13825585.2018.1536774

Gold, D. A., Park, N. W., Murphy, K. J., and Troyer, A. K. (2015). Naturalistic action performance distinguishes amnestic mild cognitive impairment from healthy aging. *J. Int. Neuropsychol. Soc.* 21, 419–428. doi: 10.1017/S135561771500048X

Gould, C. E., Segal, D. L., Yochim, B. P., Pachana, N. A., Byrne, G. J., and Beaudreau, S. A. (2014). Measuring anxiety in late life: a psychometric examination of the geriatric anxiety inventory and geriatric anxiety scale. *J. Anxiety Disord.* 28, 804–811. doi: 10.1016/j.janxdis.2014.08.001

Indovina, P., Barone, D., Gallo, L., Chirico, A., De Pietro, G., and Giordano, A. (2018). Virtual reality as a distraction intervention to relieve pain and distress during medical procedures. *Clin. J. Pain* 34, 858–877. doi: 10.1097/AJP.0000000000000599

Kessler, R. K., Giovannetti, T., and MacMullen, L. R. (2007). Everyday action in schizophrenia: performance patterns and underlying cognitive mechanisms. *Neuropsychology.* 21, 439–447. doi: 10.1037/0894-4105.21.4.439

Kingery, L., Bartolic, E., Meyer, S., Perez, M., Bertzos, K., Feldman, H., et al. (2011). Test-Retest Reliability of the MMSE, ADAS-Cog, and executive function tests in a phase II mild to moderate Alzheimer's disease study. *Alzheimer's Dement* 7, S249. doi: 10.1016/j.jalz.2011.05.711

Kourtesis, P., Collina, S., Doumas, L. A. A., and MacPherson, S. E. (2019). Technological competence is a pre-condition for effective implementation of virtual reality head mounted displays in human neuroscience: a technological review and meta-analysis. *Front. Hum. Neurosci* 13:342. doi: 10.3389/fnhum.2019.00342

Monaco, M., Costa, A., Caltagirone, C., and Carlesimo, G. A. (2013). Forward and backward span for verbal and visuo-spatial data: standardization and normative data from an Italian adult population. *Neurol. Sci.* 13:342. doi: 10.1007/s10072-012-1130-x

Nolin, P., Banville, F., Cloutier, J., and Allain, P. (2013). Virtual reality as a new approach to assess cognitive decline in the elderly. *Acad. J. Interdiscip. Stud* 2, 612–616.

Novelli, G., Papagno, C., Capitani, E., Laiacona, M., Vallar, G., and Cappa, S. F. (1986). Tre test clinici di ricerca e produzione lessicale. Taratura su soggetti normali. . *Arch. Psicol. Neurol. Psichiatr* 47, 477–506.

Orsini, A., and Pezzuti, L. (2015). *WAIS IV Contributo Alla Taratura Italiana (70 - 90 anni)*, ed. O. S. Giunti, (Italy:OrganizzazioniSecialiFirenze).

Parsons, T. D. (2015). Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. *Front. Hum. Neurosci.* 9:660. doi: 10.3389/fnhum.2015.00660

Pfeffer, R. I., Kurosaki, T. T., Harrah, C. H., Chance, J. M., and Filos, S. (1982).). Measurement of functional activities in older adults in the community. *J. Gerontol.* 37, 323–329. doi: 10.1093/geronj/37.3.323

Ruddle, R. (2005). 3D User interfaces: theory and practice - REVIEW. *Teleoperators Virtual Environ.* 14, 117–118.

Rycroft, S. S., Giovannetti, T., Divers, R., and Hulswit, J. (2018). Sensitive performance-based assessment of everyday action in older and younger adults. *Aging, Neuropsychol. Cogn.* 25, 259–276. doi: 10.1080/13825585.2017.1287855

Sánchez-Cubillo, I., Periáñez, J. A., Adrover-Roig, D., Rodríguez-Sánchez, J. M., Ríos-Lago, M., Tirapu, J., et al. (2009). Construct validity of the Trail making test: role of task-switching, working memory, inhibition/interference control, and visuomotor abilities. *J. Int. Neuropsychol. Soc.* 15, 438–450. doi: 10.1017/S1355617709090626

Scheuchenpflug, R., Ruspa, C., and Quattrocolo, S. (2003). "Presence in virtual driving simulators," in *Human Factors in the Age of Virtual Reality* , eds D. de Waard, K. A. Brookshuis, S. M. Breker, and W. B. Verwey, (Düren: Shaker), 143–148.

Schwartz, M. F. (2006). The cognitive neuropsychology of everyday action and planning. *Cogn. Neuropsychol.* 23, 202–221. doi: 10.1080/02643290500202623

Schwartz, M. F., Segal, M., Veramonti, T., Ferraro, M., and Buxbaum, L. J. (2002). The Naturalistic action test: a standardised assessment for everyday action impairment. *Neuropsychol. Rehabil.* 12, 311–339. doi: 10.1080/09602010244000084

Segal, D. L., June, A., Payne, M., Coolidge, F. L., and Yochim, B. (2010). Development and initial validation of a self-report assessment tool for anxiety among older adults: the Geriatric Anxiety Scale. *J. Anxiety Disord.* 24, 709–714. doi: 10.1016/j.janxdis.2010.05.002

Shahrbanian, S., Ma, X., Aghaei, N., Korner-bitensky, N., and Simmonds, M. J. (2012). Use of virtual reality (immersive vs. non immersive) for pain management in children and adults: a systematic review of evidence from randomized controlled trials. *Eur. J. Exp. Biol.* 2, 1408–1422.

Sivan, A. B., and Benton, A. L. (1984). Problems and conceptual issues in neuropsychological research in aging and dementia. *J. Clin. Neuropsychol.* 6, 57–63. doi: 10.1080/01688638408401196

Slater, M. (2002). Measuring presence: a response to the witmer and singer presence questionnaire. *Presence Teleoperators Virtual Environ.* 8, 560–565. doi: 10.1162/105474699566477

Solimini, A. G., Mannocci, A., and di Thiene, D. (2011). A pilot application of a questionnaire to evaluate visually induced motion sickness in spectators of tri-dimensional (3D) movies. *Ital. J. Public Health*. 12:779.

Stancati, A., and Salussi, F. (2001). *Scale di Valutazione e Malattie Reumatiche*. IBS.it: Assago, 1885.

Tarvainen, M. P., Niskanen, J. P., Lipponen, J. A., Ranta-aho, P. O., and Karjalainen, P. A. (2014). Kubios HRV - Heart rate variability analysis software. *Comput. Methods Programs Biomed.* 113, 210–220. doi: 10.1016/j.cmpb.2013.07.024

Tullis, T. S., and Stetson, J. N. (2004). "A comparison of questionnaires for assessing website usability," in *UPA 2004 Presentation*, (Boston, MA).

Wechsler, D., Coalson, D., and Raiford, S. (2008). *Wechsler Adult Intelligence Testin Fourth Edition Technical and Interpretive Manual.* San Antonio, TX: San AntonioPearson.

Witmer, B. G., Jerome, C. J., and Singer, M. J. (2005). The factor structure of the Presence Questionnaire. *Presence Teleoperators Virtual Environ.* 14, 298–312. doi: 10.1162/105474605323384654

Witmer, B. G., and Singer, M. J. (1998). Measuring presence in virtual environments: a presence questionnaire. *Presence Teleoperators Virtual Environ.* 7, 225–240. doi: 10.1162/105474698565686

Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., Adey, M., et al. (1982). Development and validation of a geriatric depression screening scale: a preliminary report. *J. Psychiatr. Res.* 17, 37–49. doi: 10.1016/0022-3956(82)90033-4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 14 February 2020
doi: 10.3389/fpsyg.2020.00197

The Impact of Test and Sample Characteristics on Model Selection and Classification Accuracy in the Multilevel Mixture IRT Model

Sedat Sen[1]* *and Allan S. Cohen*[2]

[1] *College of Education, Harran University, Şanliurfa, Turkey,* [2] *College of Education, University of Georgia, Athens, GA, United States*

Edited by:
Jason C. Immekus,
University of Louisville, United States

Reviewed by:
Luciana Pagliosa Carvalho Guedes,
Universidade Estadual do Oeste do Paraná, Brazil
Peida Zhan,
Zhejiang Normal University, China
Yong Luo,
Educational Testing Service, United States
Chung-Ying Lin,
The Hong Kong Polytechnic University, Hong Kong

***Correspondence:**
Sedat Sen
sedatsen06@gmail.com

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *22 August 2019*
Accepted: *28 January 2020*
Published: *14 February 2020*

Citation:
Sen S and Cohen AS (2020) The Impact of Test and Sample Characteristics on Model Selection and Classification Accuracy in the Multilevel Mixture IRT Model. Front. Psychol. 11:197. doi: 10.3389/fpsyg.2020.00197

The standard item response theory (IRT) model assumption of a single homogenous population may be violated in real data. Mixture extensions of IRT models have been proposed to account for latent heterogeneous populations, but these models are not designed to handle multilevel data structures. Ignoring the multilevel structure is problematic as it results in lower-level units aggregated with higher-level units and yields less accurate results, because of dependencies in the data. Multilevel data structures cause such dependencies between levels but can be modeled in a straightforward way in multilevel mixture IRT models. An important step in the use of multilevel mixture IRT models is the fit of the model to the data. This fit is often determined based on relative fit indices. Previous research on mixture IRT models has shown that performances of these indices and classification accuracy of these models can be affected by several factors including percentage of class-variant items, number of items, magnitude and size of clusters, and mixing proportions of latent classes. As yet, no studies appear to have been reported examining these issues for multilevel extensions of mixture IRT models. The current study aims to investigate the effects of several features of the data on the accuracy of model selection and parameter recovery. Results are reported on a simulation study designed to examine the following features of the data: percentages of class-variant items (30, 60, and 90%), numbers of latent classes in the data (with from 1 to 3 latent classes at level 1 and 1 and 2 latent classes at level 2), numbers of items (10, 30, and 50), numbers of clusters (50 and 100), cluster size (10 and 50), and mixing proportions [equal (0.5 and 0.5) vs. non-equal (0.25 and 0.75)]. Simulation results indicated that multilevel mixture IRT models resulted in less accurate estimates when the number of clusters and the cluster size were small. In addition, mean Root mean square error (RMSE) values increased as the percentage of class-variant items increased and parameters were recovered more accurately under the 30% class-variant item conditions. Mixing proportion type (i.e., equal vs. unequal latent class sizes) and numbers of items (10, 30, and 50), however, did not show any clear pattern. Sample size dependent fit indices BIC, CAIC, and SABIC performed poorly for the smaller level-1 sample size. For the remaining conditions, the SABIC index performed better than other fit indices.

Keywords: item response theory, mixture item response model, multilevel data, model selection, classification accuracy

INTRODUCTION

Item response theory (IRT; Lord and Novick, 1968) models have been used extensively for a variety of testing situations. However, traditional IRT models assume a single homogenous population which may be violated in some real data situations with multiple albeit latent subpopulations. Mixture extensions of IRT models have been proposed to account for heterogeneity due to these latent populations (Mislevy and Verhelst, 1990; Rost, 1990). Mixture IRT models combine a latent class model and an IRT model in a single model. Combining both models provides both qualitative and quantitative results simultaneously about the test and examinees by accounting for both categorical latent variables (i.e., latent classes) and continuous latent variables (i.e., factors) (e.g., Rost, 1990). Mixture IRT models have been used frequently due to their utility for measuring individual differences, when distinct subpopulations are present in the overall population (see Sen and Cohen, 2019, for a review of applications of mixture IRT models).

The single-level mixture IRT models are like multigroup item response models (Bock and Zimowski, 1997) in that groups are treated as manifest. Groups are taken as latent classes, however, in mixture IRT models. These models are useful for heterogeneous samples, although they do not account for the dependencies present in a multilevel (hierarchical) structure, such as are common in educational and psychological data. Ignoring the hierarchical structure with lower-level units aggregated in higher-level units has been shown to yield less accurate results because of violation of the local independence assumption (Lee et al., 2018). The hierarchical structure should be considered, in other words, in analyses of data from multilevel clusters. In this regard, multilevel mixture IRT models have been developed to account for possible dependencies, such as can arise due to cluster or multistage sampling (Vermunt, 2007). The dependency in multilevel data structures can be modeled in a straightforward way in a multilevel framework. These models can then be used to obtain information at both the individual (i.e., within) level and group (i.e., between) level. Students or examinees can be used to represent within-level and classrooms or schools can be used to represent between-level classes. Within-level latent classes capture the association between the responses at the student-level unit while between-level latent classes capture the association between the students within school-level units (Vermunt, 2003; Cho and Cohen, 2010).

As described in Lee et al. (2018), the two-parameter multilevel mixture item response model can be written as:

$$\text{logit}\left[P\left(Y_{jki}=1|\theta_{jkg},\theta_k,C_{jk}\right)\right]=\alpha_{ig.W}\theta_{jkg}+\alpha_{i.B}\theta_k-\beta_{ig}, \quad (1)$$

where Y_{jki} represents the responses of person j nested within the kth cluster (k = 1...,K) to item i, C_{jk} is a within-level latent classification variable where Cj = 1,., g,.,G for person j nested within cluster k, $\alpha_{ig.W}$ represents a within-level item discrimination parameter, $\alpha_{i.B}$ represents between-level item discrimination parameter, β_{ig} is a class-specific item location parameter, θ_{jkg} is a class-specific within-level continuous latent variable σ_g^2 and θ_k represents a between-level continuous latent variable. Both θ_{jkg} and θ_k are assumed to follow normal distributions with a mean of zero and variance σ_g^2 and τ^2, respectively.

The multilevel mixture IRT models have interested researchers due to their utility for correctly accounting for dependencies among the data in multilevel data structures (Vermunt, 2008; Cho and Cohen, 2010; Tay et al., 2011; Bacci and Gnaldi, 2012, 2015; Varriale and Vermunt, 2012; Cho et al., 2013; Finch and Finch, 2013; Bennink et al., 2014; Jilke et al., 2015; Liu et al., 2018). Cho and Cohen (2010), Finch and Finch (2013), and Bennink et al. (2014) describe applications of different types of multilevel mixture IRT models for detection of differential item functioning (DIF). Bacci and Gnaldi (2012, 2015), and Vermunt (2008) analyzed educational data sets using multilevel mixture IRT models. Examples of other studies using multilevel mixture IRT models are analysis of self-reported emotions (Tay et al., 2011) and measurement non-equivalence (Jilke et al., 2015).

The exploratory use of multilevel mixture IRT modeling is based on the comparison of alternative models using relative fit indices such as the Akaike Information Criterion (AIC;Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978) indices. The successful applications of these models partly depend on selecting the correct model and its classification accuracy. Several studies have been conducted on model selection and classification accuracy issues with different mixture IRT models (Li et al., 2009; Preinerstorfer and Formann, 2012; Choi et al., 2017; Lee et al., 2018; Sen et al., 2019). Most of these studies focused on single-level mixture IRT models. Simulation studies conducted by Li et al. (2009) and Preinerstorfer and Formann (2012) suggested that BIC performed best among the model selection indices selected in dichotomous mixture IRT models. Similar results were reported by Sen et al. (2019) for multilevel mixture Rasch models. Lee et al. (2018) found BIC to better perform than AIC in selecting the correct multilevel model compared to a single level model. Previous studies on single level mixture IRT models reported that performances of model selection indices and the classification accuracy of these models can be affected by several factors including percentage of class-variant items, magnitude of item difficulty differences, pattern of item difficulty differences, mixing proportion of latent classes (Choi et al., 2017). Choi et al. (2017) found that AIC, corrected AIC (AICC; Sugiura, 1978), BIC, and sample-size adjusted BIC (SABIC; Sclove, 1987) performed differently depending on the percentage of class-variant items and the magnitude and pattern of item difficulty differences under a two-class structure. There appear to be no studies yet reported, however, examining these issues in multilevel extensions of mixture IRT models. Thus, the current study aims to investigate the effects of various class distinction features on the model selection, classification accuracy and quality of parameter recovery in multilevel mixture IRT models. The current study focused on the effects of class distinctive features on fitting a multilevel mixture 2-parameter logistic IRT model (Multilevel Mix2PL). Although the graded response model (GRM; Samejima, 1969) is common in psychological studies, the 2PLM essentially represents a simpler case of the

GRM that; it was used as a starting point for investigating the research questions posed in the current study. To this end, this study investigated the following three research questions:

(1) How do the different test characteristics affect the quality of parameter estimates in multilevel mixture IRT models?
(2) How do these different characteristics affect classification accuracy in multilevel mixture IRT models?
(3) How do the model selection indices perform in the presence of these different characteristics?

MATERIALS AND METHODS

A Monte Carlo simulation study was conducted to answer the three research questions. Details of the simulation study are given below.

Design of the Simulation Study

Data were simulated based on the dichotomous multilevel mixture IRT model (Lee et al., 2018) having two between-level and two within-level latent classes (labeled here as CB2C2). The generating parameters for the study were obtained from estimates of an empirical data set. Item threshold values obtained from this data set were used in data generation (see **Supplementary Data Sheet S2**). All data sets were generated with the Mplus 7.4 software package (Muthén and Muthén, 1998–2015) using the Mplus syntax provided by Lee et al. (2018) (see **Supplementary Data Sheet S1**). Different data sets were generated for a varying number of conditions using the MONTE CARLO simulation implemented in Mplus. The following conditions were simulated: number of items (10, 30, and 50), mixing proportions (equal and not equal), percentage of class variant items (30, 60, and 90%), number of clusters (50 and 100), and cluster size (10 and 50).

Ten-item test was used to represent a short test condition, a 30-item test was used to represent a medium test length and a 50-item test was used to represent a long test. Two different mixing proportions were included to investigate the effect of different mixing proportions, π: equal mixing proportions ($\pi_1 = \pi_2 = 0.5$) and unequal mixing proportions ($\pi_1 = 0.75$, $\pi_2 = 0.25$). Items with the same item threshold parameters across latent classes are considered class-invariant items, and items having unequal threshold parameters are considered class-variant items. Given that the number of class-variant items has been shown to affect number of detected latent class (Choi et al., 2017), different percentages of class-variant items were manipulated in this simulation study. The percentage of class-variant items manipulated in the simulation study was 30, 60, and 90% of items on the simulated tests. Number of clusters and cluster size have also been found to affect multilevel mixture IRT results (Lee et al., 2018). Thus, the numbers of clusters manipulated in the simulation study were 50 and 100 and the cluster sizes manipulated in the simulation study were 10 and 50. Overall, 72 conditions were simulated in this study (3 numbers of items × 2 mixing proportions × 3 class variant item percentages × 2 number of clusters × 2 cluster size). One hundred replications were generated for each condition.

Estimation

Four different models were estimated: CB1C2, CB2C2, CB2C3 and CB3C3, CB is the notation for level-2 and C is the notation for level-1. Thus, CB1C2 represents a model with one level-two class and two level-one classes, CB2C2 represents a model with two level-one classes and two level-two classes, CB2C3 represents a model with level-two classes and three level-one classes, etc. The true (i.e., generating) model in this simulation study was the CB2C2 model, i.e., a multilevel mixture item response model with two within-level and two between-level latent classes. Thus, misspecified models were the CB1C2, CB2C3 and CB3C3 models. The total number of runs was 28,800 (=100 replications × 4 models × 72 conditions). Marginal maximum-likelihood estimation with the MLR estimator option was used as implemented in Mplus for estimation of the multilevel mixture IRT models. The following Mplus options were used: TYPE = TWOLEVEL MIXTURE; ALGORITHM = INTEGRATION; PROCESSORS = 2;. The Mplus syntax for model estimation is provided in the **Supplementary Data Sheet S1**.

Evaluation Measures (RMSE-Model Selection)

Item Parameter Recovery Analysis

Root mean square error (RMSE) statistics were calculated, after item parameter estimates were placed onto the scale of the generating parameters, to examine the recovery of the generating parameters. RMSE was calculated between item threshold parameters of the true model and the estimated model using $\sqrt{\sum_{r=1}^{R}\left(\hat{\lambda}_i - \lambda\right)^2/R}$, where r represents the rth replication ($r = 1,\ldots,R$).

Label switching can be a concern with mixture IRT estimation. Estimated latent classes can be switch across different replications. As an example, between-level latent class 2 on one data set can potentially correspond to between-level class 1 on another data set. Therefore, results for each data set were monitored to detect and, if necessary, to correct label switching. Threshold values obtained from the class were then used to appropriately calculate RMSE values.

Classification Accuracy Rate

In the mixture IRT framework, each respondent has an estimated posterior probability for membership in each latent class. Each respondents is assigned to a single class based on their highest estimated posterior probability value. As described in Lee et al. (2018, p. 143), for each person j nested within cluster k, the posterior probability for membership in each latent class, P_{jkg}, can be calculated as follows:

$$P_{jkg} = \frac{\hat{\pi}_g \cdot \prod_{i=1}^{I} \left(P\left(y_{jki} = 1 | \tilde{\theta}_{jkg}, \tilde{\theta}_k, C_{jk} \right) \right)^{y_{jki}} \left[1 - P\left(y_{jki} = 1 | \tilde{\theta}_{jkg}, \tilde{\theta}_k, C_{jk} \right) \right]^{1-y_{jki}}}{\sum_{g=1}^{G} \hat{\pi}_g . \hat{\pi}_g \cdot \prod_{i=1}^{I} \left(P\left(y_{jki} = 1 | \tilde{\theta}_{jkg}, \tilde{\theta}_k, C_{jk} \right) \right)^{y_{jki}} \left[1 - P\left(y_{jki} = 1 | \tilde{\theta}_{jkg}, \tilde{\theta}_k, C_{jk} \right) \right]^{1-y_{jki}}},$$

where Y_{jki} represents the responses of person j nested within kth cluster to item i, and k represents cluster k (k = 1,.,K), C_{jk} is a categorical latent variable at the within level, $\hat{\pi}_g$ is an estimated mixing proportion, $\tilde{\theta}_{jkg}$ is a class-specific within-level predicted score, and θ_k represents a between-level predicted score. The P_{jkg} values sum to 1 for each person (i.e., $\sum_{g=1}^{G} P_{jkg} = 1$).

Simulated examinees were assigned to specified latent classes during data generation. It is necessary to determine whether these examinees were classified into the same latent classes after model estimation. Posterior probabilities for membership of each examinee were calculated using the CPROBABILITIES option of the SAVEDATA command in Mplus. Classification accuracy rate was calculated for each condition. The correct detection rate was defined as the correct classification of the latent class membership for each examinee. Generated and simulated class memberships were compared and a percentage was computed across the 100 replications for each condition. Thus, agreement was recorded when an examinee assigned to the first class (Class 1) during data generation was also classified into Class 1 after estimation.

Model Selection

Unlike multigroup IRT models, the latent classes in mixture IRT models are not known *a priori* in an exploratory analysis as they are unobserved. In an exploratory analysis, different numbers of latent classes are specified as candidate models and estimated for a given data set. The most commonly used criteria for model selection in IRT models are based on either a likelihood ratio test or information criterion indices. Nylund et al. (2007) note that the likelihood ratio test is not appropriate for model selection for mixture IRT models. Thus, information criterion indices were used for model selection in this study.

Information criterion indices are based on some form of penalization of the loglikelihood. The penalization is used to adjust for the selection of over-parameterized models. Let L be the likelihood function obtained from maximum likelihood estimation and P be the penalty term. The following is a general form for information criterion indices:

$$-2\log L + P$$

The performances of AIC, BIC, consistent AIC (CAIC; Bozdogan, 1987), and SABIC were investigated in this study as these are generally the more commonly used indices in mixture IRT applications (Sen and Cohen, 2019). Each of these indices applies a different penalty function to the $-2\log L$ term. Thus, the definitions of the relative fit indices in this study are as follows:

$$\text{AIC} = -2\log L + 2d,$$

$$\text{BIC} = -2\log L + d.\ln(N),$$

$$\text{CAIC} = -2\log L + d.\left[\ln(N) + 1\right],$$

$$\text{SABIC} = -2\log L + d.\ln[(N+2)/24],$$

Where, N represents the number of examinees and d represents the number of parameters. Smaller numbers for these fit indices indicate better fit. Performances of these indices were examined by calculating the proportion of correct selections for each model. To evaluate correct model selections, the data sets generated based on CB2C2 model were analyzed with four different models (i.e., CB1C2, CB2C2, CB2C3, and CB3C3). The correct detection rate was defined as the correct detection of the simulated CB2C2 model with the correct number of within- and between-level latent classes.

RESULTS

Parameter Recovery

Table 1 presents mean RMSE values for each condition. The labels indicate the condition under which the data were generated. For example, the label E5010 indicates that the CB2C2 data were generated for equal mixing proportions for 50 clusters and with a cluster size of 10. That is, number of level-2 units is 50 and number of level-1 is 10. The NE label indicates unequal mixing proportion conditions. Results of each condition are presented for 10-, 30-, and 50-item test lengths and 30, 60, and 90% of class variant items. Mean RMSE values for item threshold estimates ranged from 0.092 to 2.927.

As shown in **Table 1**, the mean RMSE values decreased as the cluster size and number of examinees for level-1 increased. Similarly, mean RMSE values increased as the percentage of class-variant items increased. As expected, greater accuracy was observed with the higher number of clusters and cluster size conditions. Type of mixing proportion (equal vs. unequal) and number of items (10, 30, and 50) did not show any clear pattern of recovery.

Classification Accuracy

As with latent class models, mixture IRT models assign each examinee to one of the latent classes based on class probability values. The class memberships created during the data generation were compared with the estimated class memberships. A classification accuracy rate was calculated for each condition

TABLE 1 | Mean RMSE values of item threshold estimates for the CB2C2 Model.

Simulation condition	Percent of class variant items								
	10 Items			30 Items			50 Items		
	30	60	90	30	60	90	30	60	90
E5010	1.335	1.812	1.949	0.454	0.993	1.333	0.562	1.231	2.007
E5050	0.256	0.325	0.829	0.118	0.732	0.977	0.107	0.985	1.268
E10010	0.752	0.830	1.099	0.213	0.766	1.007	0.199	1.006	1.458
E10050	0.164	0.191	0.767	0.083	0.724	0.965	0.075	0.977	1.260
NE5010	1.087	1.213	1.401	1.873	2.653	2.710	1.435	1.860	2.927
NE5050	0.400	0.596	1.010	0.328	0.751	1.087	0.134	0.988	1.321
NE10010	0.803	1.377	1.565	1.289	1.621	1.928	0.548	1.120	1.766
NE10050	0.335	0.376	0.859	0.328	0.734	1.070	0.092	0.979	1.262

E, Equal proportion; NE, Non-equal proportions; E5010 reprents a condition with equal mixing proportions under 50 clusters and with a cluster size of 10.

TABLE 2 | Classification accuracy rates for CB2C2 Model.

Simulation condition	10 Items			30 Items			50 Items		
	30	60	90	30	60	90	30	60	90
E5010	37.35	38.20	31.43	43.19	24.14	44.66	69.11	80.04	69.38
E5050	45.13	58.58	38.69	57.86	45.05	38.85	82.29	89.02	86.92
E10010	30.82	42.54	27.87	44.18	27.58	58.00	70.04	83.34	78.93
E10050	35.39	61.53	30.18	61.15	47.43	37.79	82.09	89.02	87.12
NE5010	37.00	37.42	30.93	28.50	27.27	26.69	65.86	74.70	45.70
NE5050	52.94	57.05	45.03	38.71	26.58	29.01	85.04	90.50	88.61
NE10010	34.79	47.14	36.97	26.61	32.31	32.50	72.86	85.12	66.97
NE10050	60.27	57.42	32.45	31.13	12.31	15.87	85.85	90.64	86.52

E, Equal proportion; NE, Non-equal proportions; E5010 reprents a condition with equal mixing proportions under 50 clusters and with a cluster size of 10.

between generated values and estimated values based on the same model. Classification accuracy rates are shown in **Table 2**. These rates ranged from 12.31 to 90.64%. **Table 2** shows that the classification accuracy rates increase as the number of items increases. The highest rates occurred for the 50-item conditions while the lowest rates were observed with 10-item conditions. Only the 30-item conditions with 60% of class-variant items did not follow this pattern. This condition actually yielded lower rates than the 10-item counterparts (i.e., 10-item conditions with 60% of class variant items). Equal mixing proportion conditions yielded smaller accuracy rates than unequal mixing proportion conditions for almost each percentage of class-variant items and test length conditions. As shown in **Table 2**, conditions with 60% of class-variant items yielded higher accuracy rates than conditions with 30 and 90% of class-variant items under 10- and 50-item condition. However, this was not the case with the 30-item conditions. The cluster size seemed to influence the classification accuracy rates. The conditions with the smaller level-1 sample size (i.e., 10) yielded lower accuracy rates than the conditions with the higher level-1 sample size (i.e., 50). Similarly, the number of clusters appeared to influence classification accuracy rates. The conditions with 50 clusters yielded lower accuracy rates than the conditions with 100 clusters. As expected, increases in the number of items, number of clusters and cluster size had a positive effect on classification accuracy.

Model Selection

AIC, BIC, CAIC, and SABIC values were calculated for each condition. The number of correct selections was calculated as the number of detections of the CB2C2 (i.e., the generating) model over 100 iterations. The frequencies of correct model selections are shown in **Tables 3–5** for each of the information indices.

The numbers of correct detections for 10-item conditions are presented in **Table 3**. Correct detection frequencies ranged between 0 and 100 out of 100 replications in the 10-item conditions. As shown in **Table 3**, BIC, CAIC, and SABIC performed better than AIC index for the conditions with level-1 sample size of 50 (i.e., E5050, E10050, NE5050, and NE10050). The number of correct detections of the BIC and CAIC indices for the smaller number of level-1 sample size conditions were all either very low or zero except for unequal mixing proportion condition with 100 clusters and level-1 sample size of 10 (i.e., NE10010). The SABIC index performed better than BIC index for almost all conditions. BIC and CAIC performed less well than the SABIC for the small level-1 sample. However, the level-1 sample size did not appear to have any effect on the performance of AIC. The percentage of class-variant items appeared to influence

TABLE 3 | Number of correct detections over 100 replications for 10-Item conditions.

	AIC			BIC			SABIC			CAIC		
	30	**60**	**90**	**30**	**60**	**90**	**30**	**60**	**90**	**30**	**60**	**90**
E5010	82	52	65	3	0	2	59	31	48	2	0	0
E5050	82	76	97	100	100	100	100	100	100	97	99	98
E10010	86	67	67	21	0	3	84	58	65	7	0	1
E10050	57	70	89	80	100	100	77	100	100	77	97	97
NE5010	70	57	69	1	0	2	51	26	41	0	0	0
NE5050	91	79	90	100	80	100	100	87	100	97	77	95
NE10010	86	74	73	11	1	2	78	42	70	5	0	2
NE10050	75	38	92	100	70	100	100	59	100	100	73	97

E, Equal proportion; NE, Non-equal proportions; E5010 reprents a condition with equal mixing proportions under 50 clusters and with a cluster size of 10; AIC, Akaike Information Criterion; BIC, Bayesian Information Criterion; CAIC, Consistent AIC; SABIC, Sample size adjusted BIC.

TABLE 4 | Number of correct detections over 100 replications for 30-Item conditions.

	AIC			BIC			SABIC			CAIC		
	30	**60**	**90**	**30**	**60**	**90**	**30**	**60**	**90**	**30**	**60**	**90**
E5010	53	55	47	28	0	0	99	97	66	11	0	0
E5050	56	72	37	100	100	100	100	100	100	100	100	100
E10010	48	34	48	99	53	0	100	99	66	97	20	0
E10050	59	77	41	99	100	100	99	100	100	98	100	100
NE5010	28	38	25	0	2	0	11	6	100	0	0	0
NE5050	18	65	53	81	66	8	97	99	83	66	33	1
NE10010	16	47	31	0	0	0	13	6	1	0	0	0
NE10050	5	63	39	100	99	92	85	99	99	100	98	85

E, Equal proportion; NE, Non-equal proportions; E5010 reprents a condition with equal mixing proportions under 50 clusters and with a cluster size of 10; AIC, Akaike Information Criterion; BIC, Bayesian Information Criterion; CAIC, Consistent AIC; SABIC, Sample size adjusted BIC.

TABLE 5 | Number of correct detections over 100 replications for 50-Item conditions.

	AIC			BIC			SABIC			CAIC		
	30	**60**	**90**	**30**	**60**	**90**	**30**	**60**	**90**	**30**	**60**	**90**
E5010	58	79	78	0	0	1	54	30	2	0	0	0
E5050	67	66	77	100	100	100	100	100	100	100	100	90
E10010	67	76	92	1	0	0	93	89	21	0	0	0
E10050	69	65	65	100	100	97	100	100	94	100	100	98
NE5010	57	49	31	0	0	0	23	3	0	0	0	0
NE5050	77	74	76	100	89	36	100	99	100	99	78	12
NE10010	60	73	97	0	0	0	53	26	0	0	0	0
NE10050	92	91	68	100	100	100	100	100	100	100	100	98

E, Equal proportion; NE, Non-equal proportions; E5010 reprents a condition with equal mixing proportions under 50 clusters and with a cluster size of 10; AIC, Akaike Information Criterion; BIC, Bayesian Information Criterion; CAIC, Consistent AIC; SABIC, Sample size adjusted BIC.

the correct detection rates based on four fit indices. The 60% conditions yielded lower correct detection rates for almost every condition. The effects of mixing proportion type (equal vs. unequal), however, did not show any clear pattern.

The number of correct detections for the 30-item conditions ranged between 0 and 100 (see **Table 4**). As shown in **Table 4**, BIC, CAIC, and SABIC performed better than AIC for the sample size of 50 (i.e., E5050, E10050, NE5050, and NE10050). As was the case for the 10-item conditions, the numbers of correct detections of the BIC and CAIC indices for smaller number of level-1 sample size conditions were all either very low or zero for the E5010 and E10010 conditions. SABIC performed better than BIC and CAIC for most conditions except for NE10050 condition under 30% of class-variant items. The small level-1 sample size (i.e., 10) appeared to influence the performance of BIC and CAIC compared to SABIC. However, the level-1 sample size did not show any clear pattern for the performance of AIC. The percentage of class-variant items appears to influence the correct detection rates based on four fit indices. The 60% conditions yielded lower correct detection rates for most of the conditions. The effects of mixing proportion type (equal vs. unequal), however, did not show any clear pattern.

Correct detection frequencies (see **Table 5**) ranged between 0 and 100 in the 50-item conditions. As shown in **Table 5**, BIC, CAIC, and SABIC performed better than AIC for the conditions with the level-1 sample size of 50 (i.e., E5050, E10050, NE5050, and NE10050). AIC performed better than BIC, CAIC, and SABIC, however, for the conditions with the level-1 sample size of 10 (i.e., E5010, E10010, NE5010, and NE10010). As was the case with the 10- and 30-item conditions, the numbers of correct detections of the BIC and CAIC indices for smaller level-1 sample size conditions were all either very low or zero for the 50-item conditions. SABIC performed better than BIC and CAIC for most conditions except for E10050 for the 90% class-variant items condition. The small level-1 sample size (i.e., 10) appears to influence the performance of BIC and CAIC compared to SABIC. The level-1 sample size, however, did not show any clear pattern for AIC. Similarly, the percentage of class-variant items and the effects of type of mixing proportion (i.e., equal vs. unequal) did not show any clear pattern.

SUMMARY AND DISCUSSION

This simulation study examined the accuracy of parameter estimates and classifications under different multilevel and mixture conditions. The simulation factors in this research were chosen to represent different class-distinction features in multilevel mixture IRT modeling, in which the percentage of class-variant items, the number and magnitude of clusters, and the number of items varied for the structure with two level-1 and two level-2 classes (i.e., CB2C2 model). In addition, this study also investigated the differential performance of the four information criteria (AIC, BIC, CAIC, and SABIC) for model selection with different multilevel mixture IRT model applications.

Findings from the simulation study indicated that greater accuracy was observed with the higher number of clusters (i.e., 100 clusters) and cluster size (i.e., 50 simulated examinees) conditions, as well as the lower (30%) percentage of class-variant item conditions. When the number of clusters and the cluster sizes were small, the applications of multilevel mixture IRT models can be problematic with respect to the accuracy of item parameter estimates. These findings were consistent with previous research by Lee et al. (2018) which found that the multilevel mixture IRT model does not perform well for small sample sizes.

Findings regarding classification accuracy rates showed that the classification accuracy rates increased as the number of items increased. Equal mixing proportion conditions yielded smaller accuracy rates than unequal mixing proportion conditions for most percentages of class-variant items and test length conditions. The numbers of clusters and cluster size appeared to influence classification accuracy rates. The smaller cluster size (i.e., 10 examinees) and smaller number of clusters (i.e., 50 clusters) yielded lower accuracy rates. Similarly, the number of clusters appeared to influence classification accuracy rates. As expected, increases in the number of items, number of clusters and cluster size had a positive effect on classification accuracy.

Differential performances of the AIC, BIC, CAIC, and SABIC were observed under the different study conditions. Overall, SABIC performed better than BIC or CAIC for the small level-1 sample (i.e., 10) conditions, and for the conditions with the higher sample size at level-1 (i.e., 50). BIC and CAIC failed to select the true model for conditions with the smaller level-1 sample size. Overall, BIC and CAIC indices showed similar performances under the different data conditions. The SABIC appears to be the better than BIC and CAIC for the smaller level-1 sample size. These findings were consistent with Choi et al. (2017) that showed the superiority of SABIC over other relative fit indices. AIC also appeared to perform better than SABIC, however, under some conditions (i.e., NE5010, NE10010 conditions with 10-, 30- and 50-items and E5010, E10010 conditions with 10- and 50-items). Thus, results suggest that no uniformly superior single information criterion index of the four indices studied here was consistently the best model selection index under each of the simulated conditions here.

Multilevel mixture IRT models and relative fit indices used for model selection perform better with higher number of clusters and cluster sizes. The percentage of class-variant items also appeared to have an effect on accuracy of model estimates and on performance of model selection indices. Given these findings, it is important to note that model selection also needs to pay attention to substantive theory as well as to multiple fit indices rather than relying on a single fit index for model selection. The present study shares similar limitations to those of other simulation studies using similar conditions in the study design (e.g., Choi et al., 2017; Lee et al., 2018).

DATA AVAILABILITY STATEMENT

Datasets generated for E5010 conditions of this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

Both authors contributed equally to the data analyses and reporting parts.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00197/full#supplementary-material

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Control* 19, 716–723. doi: 10.1109/tac.1974.1100705

Bacci, S., and Gnaldi, M. (2012). Multilevel mixture IRT models: an application to the university teaching evaluation. *Anal. Mod. Complex Data Behav. Soc. Sci.* 38, 2775–2791.

Bacci, S., and Gnaldi, M. (2015). A classification of university courses based on students' satisfaction: an application of a two-level mixture item response model. *Qual. Quant.* 49, 927–940. doi: 10.1007/s11135-014-0101-0

Bennink, M., Croon, M. A., Keuning, J., and Vermunt, J. K. (2014). Measuring student ability, classifying schools, and detecting item bias at school level based on student-level dichotomous items. *J. Educ. Behav. Stat.* 39, 180–201.

Bock, R. D., and Zimowski, M. F. (1997). "Multiple group IRT," in *Handbook of modern item response theory*, eds W. J. van der Linden, and R. K. Hambleton, (New York, NY: Springer-Verlag), 433–448. doi: 10.1007/978-1-4757-2691-6_25

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* 52, 345–370. doi: 10.1007/bf02294361

Cho, S.-J., and Cohen, A. S. (2010). A multilevel mixture model with applications to DIF. *J. Educ. Behavi Stat.* 35, 336–370. doi: 10.3102/1076998609353111

Cho, S.-J., Cohen, A. S., and Bottge, B. A. (2013). Detecting intervention effects using a multilevel latent transition analysis with a mixture IRT model. *Psychometrika* 78, 576–600. doi: 10.1007/s11336-012-9314-0

Choi, I. H., Paek, I., and Cho, S. J. (2017). The impact of various class-distinction features on model selection in the mixture Rasch model. *J. Exp. Educ.* 85, 411–424. doi: 10.1080/00220973.2016.1250208

Finch, W. H., and Finch, M. E. H. (2013). Investigation of specific learning disability and testing accommodations based differential item functioning using a multilevel multidimensional mixture item response theory model. *Educ. Psychol. Meas.* 73, 973–993. doi: 10.1177/0013164413494776

Jilke, S., Meuleman, B., and Van de Walle, S. (2015). We need to compare, but how? Measurement equivalence in comparative public administration. *Public Adm. Rev*. 75, 36–48. doi: 10.1111/puar.12318

Lee, W. Y., Cho, S. J., and Sterba, S. K. (2018). Ignoring a multilevel structure in mixture item response models: impact on parameter recovery and model selection. *Appl. Psychol. Meas.* 42, 136–154. doi: 10.1177/0146621617711999

Li, F., Cohen, A. S., Kim, S.-H., and Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Appl. Psychol. Meas.* 33, 353–373. doi: 10.1177/0146621608326422

Liu, H., Liu, Y., and Li, M. (2018). Analysis of process data of PISA 2012 computer-based problem solving: application of the modified Multilevel Mixture IRT model. *Front. Psychol.* 9:1372. doi: 10.3389/fpsyg.2018.01372

Lord, F. M., and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Mislevy, R. J., and Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika* 55, 195–215. doi: 10.1007/bf02295283

Muthén and Muthén, L. K., and Muthén and Muthén, B. O. (1998–2015) *Mplus Users Guide*, 7th Edn. Los Angeles, CA: Author. doi: 10.1007/bf02295283

Nylund, K. L., Asparouhov, T., and Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: a monte carlo simulation study. *Struct. Equ. Model.* 14, 535–569.

Preinerstorfer, D., and Formann, A. K. (2012). Parameter recovery and model selection in mixed Rasch models. *Br. J. Math. Stat. Psychol.* 65, 251–262. doi: 10.1111/j.2044-8317.2011.02020.x

Rost, J. (1990). Rasch models in latent classes: an integration of two approaches to item analysis. *Appl. Psychol. Meas.* 14, 271–282. doi: 10.1177/014662169001400305

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychom. Monogr*. 34, 1–97. doi: 10.1007/BF03372160

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika* 52, 333–343. doi: 10.1007/bf02294360

Sen, S., and Cohen, A. S. (2019). Applications of mixture IRT models: a literature review. *Meas.: Interdiscip. Res. Perspect.* 17, 177–191. doi: 10.1080/15366367.2019.1583506

Sen, S., Cohen, A. S., and Kim, S. H. (2019). Model selection for multilevel mixture Rasch models. *Appl. Psychol. Meas.* 43, 272–289. doi: 10.1177/0146621618779990

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun. Stat.Theory Methods, A* 7, 13–26. doi: 10.1080/03610927808827599

Tay, L., Diener, E., Drasgow, F., and Vermunt, J. K. (2011). Multilevel mixed-measurement IRT analysis: an explication and application to self-reported emotions across the world. *Organ. Res. Methods* 14, 177–207. doi: 10.1177/1094428110372674

Varriale, R., and Vermunt, J. K. (2012). Multilevel mixture factor models. *Multivar. Behav. Res.* 47, 247–275. doi: 10.1080/00273171.2012.658337

Vermunt, J. K. (2003). Multilevel latent class models. *Sociol. Methodol.* 33, 213–239. doi: 10.1111/j.0081-1750.2003.t01-1-00131.x

Vermunt, J. K. (2007). "*Multilevel mixture item response theory models: an application in education testing*," in *Proceedings of the* 56th session of the International Statistical Institute, (Lisbon), 2228.

Vermunt, J. K. (2008). Multilevel latent variable modeling: an application in education testing. *Aus. J. Stat.* 37, 285–299.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 19 February 2020
doi: 10.3389/fpsyg.2020.00016

Validation of Two Short Personality Inventories Using Self-Descriptions in Natural Language and Quantitative Semantics Test Theory

Danilo Garcia[1,2,3]*, *Patricia Rosenberg*[1], *Ali Al Nima*[1,2], *Alexandre Granjard*[1,2], *Kevin M. Cloninger*[1,4] *and Sverker Sikström*[5]*

[1] Blekinge Center of Competence, Karlskrona, Sweden, [2] Department of Psychology, University of Gothenburg, Gothenburg, Sweden, [3] Department of Behavioral Science and Learning, Linköping University, Linköping, Sweden, [4] Anthropedia Foundation, St. Louis, MO, United States, [5] Department of Psychology, Lund University, Lund, Sweden

OPEN ACCESS

Edited by:
Pietro Cipresso,
Italian Auxological Institute (IRCCS),
Italy

Reviewed by:
Radosław Rogoza,
Cardinal Stefan Wyszyński University
in Warsaw, Poland
Atsushi Oshio,
Waseda University, Japan

***Correspondence:**
Danilo Garcia
danilo.garcia@icloud.com
Sverker Sikström
sverker.sikstrom@psy.lu.se

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: 29 September 2019
Accepted: 07 January 2020
Published: 19 February 2020

Citation:
Garcia D, Rosenberg P, Nima AA,
Granjard A, Cloninger KM and
Sikström S (2020) Validation of Two
Short Personality Inventories Using
Self-Descriptions in Natural Language
and Quantitative Semantics Test
Theory. Front. Psychol. 11:16.
doi: 10.3389/fpsyg.2020.00016

Background: If individual differences are relevant and prominent features of personality, then they are expected to be encoded in natural language, thus manifesting themselves in single words. Recently, the quantification of text data using advanced natural language processing techniques offers innovative opportunities to map people's own words and narratives to their responses to self-reports. Here, we demonstrate the usefulness of self-descriptions in natural language and what we tentatively call Quantitative Semantic Test Theory (QuSTT) to validate two short inventories that measure character traits.

Method: In Study 1, participants ($N_1 = 997$) responded to the Short Character Inventory, which measures self-directedness, cooperativeness, and self-transcendence. In Study 2, participants ($N_2 = 2373$) responded to Short Dark Triad, which measures Machiavellianism, narcissism, and psychopathy. In both studies, respondents were asked to generate 10 self-descriptive words. We used the Latent Semantic Algorithm to quantify the meaning of each trait using the participants' self-descriptive words. We then used these semantic representations to predict the self-reported scores. In a second set of analyses, we used word-frequency analyses to map the self-descriptive words to each of the participants' trait scores (i.e., one-dimensional analysis) and character profiles (i.e., three-dimensional analysis).

Results: The semantic representation of each character trait was related to each corresponding self-reported score. However, participants' self-transcendence and Machiavellianism scores demonstrated similar relationships to all three semantic representations of the character traits in their respective personality model. The one-dimensional analyses showed that, for example, "loving" was indicative of both high cooperativeness and self-transcendence, while "compassionate," "kind," and "caring" was unique for individuals high in cooperativeness. The words "kind" and "caring" indicated low levels of Machiavellianism and psychopathy, whereas "shy" or "introvert" indicated low narcissism. We also found specific keywords that unify or that make the individuals in some profiles unique.

Conclusion: Despite being short, both inventories capture individuals' identity as expected. Nevertheless, our method also points out some shortcomings and overlaps between traits measured with these inventories. We suggest that self-descriptive words can be quantified to validate measures of psychological constructs (e.g., prevalence in self-descriptions or QuSTT) and that this method may complement traditional methods for testing the validity of psychological measures.

Keywords: character, identity, quantitative semantic test theory, narrative self, personality

INTRODUCTION

Human personality can be defined as the dynamic organization, within the person, of biopsychosocial systems that regulate adaptation to a changing environment (Cloninger et al., 1993; see also Cloninger et al., 2019). This includes systems of self-government that modulate cognitions, emotions, impulse control, and social relationships. In this context, specific personality traits are responsible for how the individual perceives and thinks about oneself, other people, and the world as a whole (Cloninger, 2004, 2009), which are aspects that are strongly associated to physical, mental, social, and spiritual health (Vaillant and Vaillant, 1990; World Health Organization [WHO], 2001; Cloninger, 2003, 2004; VanderWeele, 2017). The measuring of personality is often done using self-reports, something that is not without controversy regarding conceptualization and measure accuracy (cf. Cloninger et al., 2019). For instance, although trait models of personality stem from natural self-descriptive language (Leising et al., 2014), the validation of inventories that measure personality and most psychological constructs is often done using Classical Test Theory (CTT) and more recently using Item Response Theory (IRT) rather than natural language. This is important because individual differences are expected to be encoded in natural language if they are relevant and prominent features of personality, thus, manifesting themselves in single words (cf. the psycholexical hypothesis; John et al., 1988). These single words might be used in self-descriptions, which in turn reflect people's temperament and own concept of the self or character, including the perception of her/his identity (Adams et al., 2012). In one study, for example, researchers found 624 adjectives that laypeople used when freely generating words to describe people they know (Leising et al., 2014). What is more, the adjectives that these participants rated as more important were found more frequently in an independent large text corpus of 500 million words of online communication. Hence, suggesting that the words people frequently use to describe personality might indeed be valid to describe human temperament and character (cf. Garcia et al., 2015).

Despite the fact that CTT and IRT are good methods for the validation of measures, there are some limitations. For instance, CTT methods are dependent on the number of items and on the sample's size and other features, so any changes to these features can strongly affect both item and the total psychometric properties of the scale. Moreover, IRT methodology does not address, for example, the issue of social desirability or response style (Oishi, 2007). We argue that using, for example, the words people use to describe themselves might serve as a new tool to validate measures of personality and other psychological phenomena. One obstacle, however, has been that advanced methodological techniques are necessary to actually use freely generated self-descriptive words in such analyses. Researchers have only recently started using these techniques in the social sciences (see Leising et al., 2014; Sikström and Garcia, 2019). Indeed, despite the fact that lexical models of personality have their basis in natural language, self-descriptive words have not been mapped to specific personality constructs to distinguish meaningful patterns that explain people's behavior and tendencies (for a review, see Uher, 2013). Importantly, at times, researchers look for short measures for the assessment of personality, which might compromise validity. Moreover, regarding personality, different measures can be used that are, for example, stated as representing a dark side of personality rather than just personality. Thus, making psychometric scrutiny regarding these short measures even more important, if we do not want to risk ending up with "quick and dirty measures" that lack a comprehensive theory (cf. Wong and Roy, 2018) and suffer of "jingle-jangle" fallacy[1] (cf. Kelley, 1927; Block, 1995).

More recently, the quantification of text data using advanced natural language processing techniques offers innovative opportunities to map people's own words and narratives to their responses to self-reports' scales. Here, we demonstrate the usefulness of what we tentatively call Quantitative Semantics Test Theory (QuSTT) to validate two short inventories that measure character traits. We use the Latent Semantic Analysis algorithm, which is not only a method but also a theory for how humans acquire, induct, and represent meaning and knowledge (Landauer and Dumais, 1997; Landauer, 2008). By applying this statistical computation on a large text corpus, researchers can extract and represent the meaning of words based on the context in which it co-occurs with other words. We expected that the quantified meaning of words that an individual uses to intentionally describe herself/himself may predict her/his level in different personality traits. We aim to exemplify this by mapping the words that participants use to their responses in each scale and also to personality profiles. Before

[1] Jingle refers to two constructs with equivalent labels that really reflect different phenomena, whereas jangle refers to when one construct is given multiple names (Kelley, 1927; Block, 1995).

stating any further expectations, we present the personality models in each study.

Light Character Traits: Self-Directedness, Cooperativeness, and Self-Transcendence

Cloninger proposed in his model of personality (Cloninger et al., 1993) four dimensions of temperament and three dimensions of character. Here, we focus on character, which can be defined as what the individual makes of her/himself intentionally or individual differences in values, goals, and self-conscious emotions, such as, hope, love, and faith (Cloninger, 2004). We do this partially for practical reasons; the shortest measure derived to measure these dimensions assesses only the three character traits, but also because the light character traits stand in contrast to the Dark Triad traits (Garcia and Rosenberg, 2016). The three character traits are the following: (1) *self-directedness*, which refers to the person's level of self-determination and tendency to self-control, self-sufficiency, self-acceptance, responsibility, and reliableness; (2) *cooperativeness*, accounts for individual differences in social acceptance, tolerance toward others, and tendency to be a helpful and empathic person; and (3) *self-transcendence*, which refers to the person's tendency to experience self-forgetfulness, spiritual acceptance, and to be patient and imaginative (Cloninger et al., 1993; Köse, 2003). In this context, Cloninger developed the Temperament and Character Inventory for the assessment of personality according to his biopsychosocial model[2] (Cloninger et al., 1993; see also Garcia et al., 2017). The original long version comprises 240 items that operationalize the four temperament dimensions and the three character dimensions, while the inventory that we investigate here is a short version that measures the character traits using 15 items (i.e., the Short Character Inventory).

As the long version, this short version was designed to be applicable to large normal populations without being stigmatizing or pathologizing. Furthermore, instead of assuming that personality can be decomposed into independent dimensions, Cloninger based his personality model and inventories on complex interactions, such as gene–gene and gene–environment (Cloninger, 2004; Zwir et al., 2018a,b, 2019). Thus, personality is a dynamic complex adaptive system. In other words, on a daily basis a person is adapting not only to the environment but also to the emotions and cognitions within her/himself. This notion of personality as whole system unit has been suggested to be best studied by analyzing "common types" or profiles, see **Figure 1** (Bergman and Magnusson, 1997; Cloninger et al., 1997; Bergman and Wångby, 2014; Zwir et al., 2018a,b, 2019). For instance, perceptual aberrations such as superstitious or magical thinking and vulnerability to overvalued ideas or psychosis is a product of excessive imagination (i.e., high self-transcendence) in combination with lack of solid reality testing (i.e., low self-directedness) (Smith et al., 2008). Moreover, individuals who report high levels in all three character traits (i.e., "Creative" profile) or high levels in self-directedness and cooperativeness, but low in self-transcendence ("Organized" profile) report the highest levels of health, well-being, longevity, and functionality (Cloninger, 2004). Creative people are expected to see life as being filled with opportunities to learn from mistakes (i.e., high self-directedness), to work in the service of others (i.e., high cooperativeness), and to grow in awareness (i.e., high self-transcendence) around life as a whole and what is beyond human existence (Cloninger, 2004). In contrast, people with an "Apathetic" profile are low in all three traits of character, so they often think "life is hard, people are mean, and then you just die!" Not surprisingly, they are unhappy, alienated, and physically unhealthy and fearful of death with high rates of mental and physical disorders (Cloninger, 2004) (see **Figure 1**).

The Dark Triad: Machiavellianism, Narcissism, and Psychopathy

Peoples' propensities to amoral behavior, manipulativeness, opportunism, selfishness, callousness, and self-centeredness are suggested to be reflected in individual differences in three dark character traits: Machiavellianism, narcissism, and psychopathy (Paulhus and Williams, 2002). At a general level, this outlook of separateness (cf. Cloninger, 2004, 2007, 2013) expressed by any of these dark traits also express uncooperativeness as one common aspect of a vicious character (e.g., Garcia and Rosenberg, 2016; Moshagen et al., 2018) and different levels of other personality tendencies (Vernon et al., 2008). At the conceptual level, individuals high on Machiavellianism are cold, manipulative, and have a sarcastic worldview (Christie and Geis, 1970; Jones and Paulhus, 2014). Individuals high on narcissism lack empathy, have fantasies of enormous power, beauty and success, have low self-esteem, and are exhibitionistic and exploitative (Raskin and Hall, 1979). In other words, they regard themselves as better, smarter, more dominant and superior than others but at the same time tend to be sensitive to criticism and with a need for constant reassurance. Individuals high on psychopathy show low empathy, low anxiety, are impulsive, and thrill seeking (Hare, 1985). Although individuals high in Machiavellianism and psychopathy can be described using the same terms (e.g., manipulative and callous), those high on psychopathy are impulsive, reckless, aggressive, and lack the same convincing social skills that individuals high on Machiavellianism display (Hawley, 2003). Individuals high on narcissism are also expected to display callousness and manipulation, but they are expected to show self-enhancement as well. Accordingly, these malevolent traits, often labeled the Dark Triad (Paulhus and Williams, 2002), are addressed as overlapping constructs that can be measured separately, since they are considered to be distinctive enough (see Persson, 2019 for another point of view). Behavioral studies, for example, show that while Machiavellianism and psychopathy predict cheating when it required an intentional lie, psychopathy predicted cheating when punishment was a serious risk and individuals high in Machiavellianism cheated under high risk, but only if they were ego depleted (Jones and Paulhus, 2017; see also Crysel et al., 2013; Jones, 2014). Hence, as for the light character traits, the dark character traits might also be seen, at least in theory, as one dynamic complex adaptive

[2] http://anthropedia.org

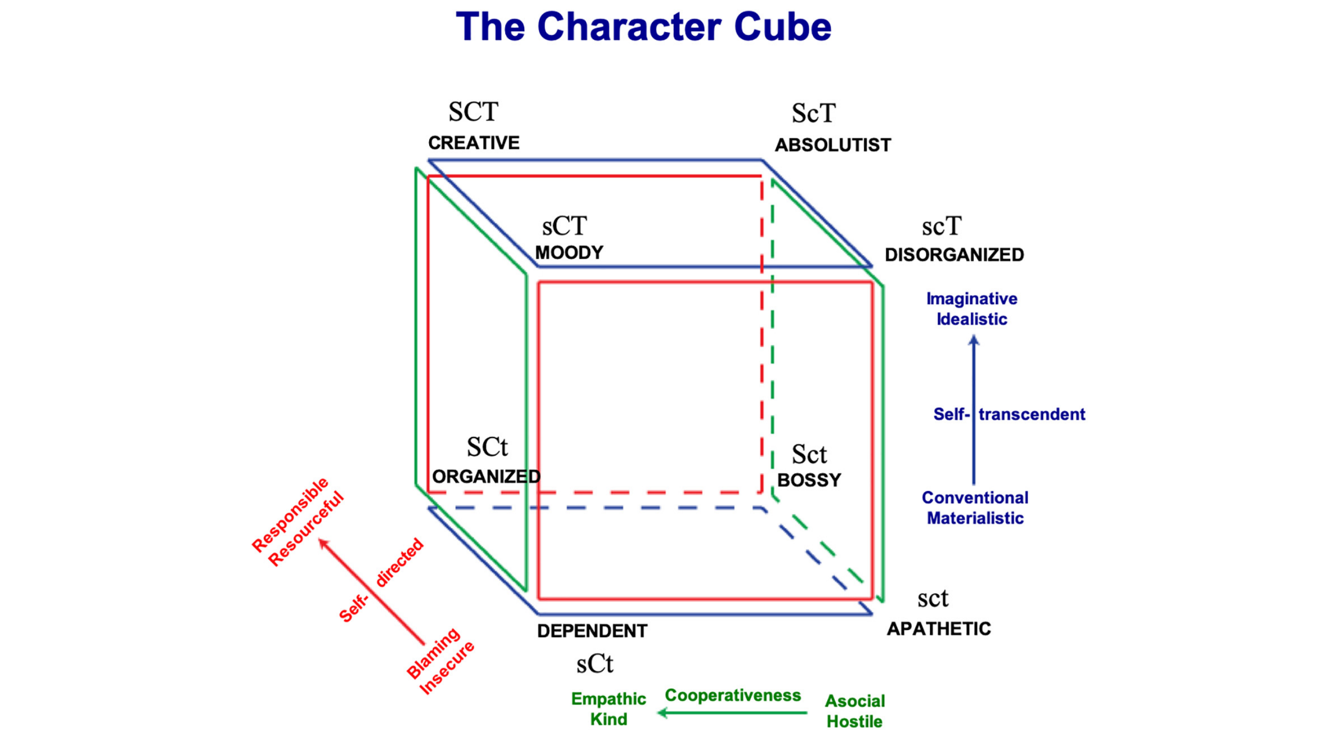


FIGURE 1 | The character cube representing the eight possible combinations of high and low scores in Cloninger's light character traits. Reprinted with permission from Anthropedia Foundation. S/s, high/low self-directedness; C/c, high/low cooperativeness, T/t, high/low self-transcendence.

system rather than three single traits. In this line of thinking, Garcia (Garcia and Rosenberg, 2016; Garcia, 2018) suggested, analogous to Cloninger's "light" character cube (Cloninger, 2004), the Dark Cube, which comprises the eight possible combinations of high/low scores in the three malevolent traits (see **Figure 2**; Garcia and Rosenberg, 2016; Garcia and Gonzàlez, 2017; Garcia, 2018; Garcia et al., 2018).

At the operationalization level, factor-analytic studies using short measures of the Dark Triad (27 items or less) have shown that narcissism and psychopathy load on the same factor (Furnham and Crump, 2005; Garcia and Rosenberg, 2016; Kajonius et al., 2016; Persson et al., 2017, 2019). On this basis, some researchers have suggested a dyad rather than a triad (e.g., Garcia and Rosenberg, 2016), and others even suggest that, at least based on the analyses of short measures, the three traits can be described well by individuals' response to a single item measuring their tendency to exploit others (e.g., Kajonius et al., 2016). We argue that the mapping of words and their meaning to short scales' scores might shed some light to validate if the scales target different malevolent character traits.

Quantitative Semantics Test Theory (QuSTT)

We have argued that since psychological phenomena is expressed in natural language (e.g., psycholexical hypothesis), if reliably quantified, the mere words people use to express, for example, their personality, can be used to validate self-report scales of the construct at hand. We quantified the words that people use when asked to describe who they are with 10 words, using the Latent Semantic Analysis algorithm. The analyses were conducted in semanticexcel[3], which is a web-based program for the analyses of quantitative semantics developed by Sverker Sikström at Lund University, Sweden (for details, see Garcia and Sikström, 2013a,b, 2014; Garcia et al., 2015; Sikström and Garcia, 2019). Here, we just present a brief overview of how semantic representations are generated, how the self-descriptive words generated by the participants are linked to this representation and then regressed on participants' own character traits scores, and how we map the self-descriptive words to the character traits scores. This whole procedure stands as the basis of QuSTT.

Creating a Semantic Representation of the English Language

Semanticexcel comprises semantic representations of several languages, including English, Spanish, Swedish, etc. The representation of English used here was generated using Google N-grams[4], which might be the largest possible available English

[3] www.semanticexcel.com

[4] "In the fields of computational linguistics and probability, an n-gram is a contiguous sequence of n items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words, or base pairs according to the application. The n-grams typically are collected from a text or speech corpus. When the items are words, n-grams may also be called shingles" (Retrieved from http://en.wikipedia.org/wiki/N-gram).

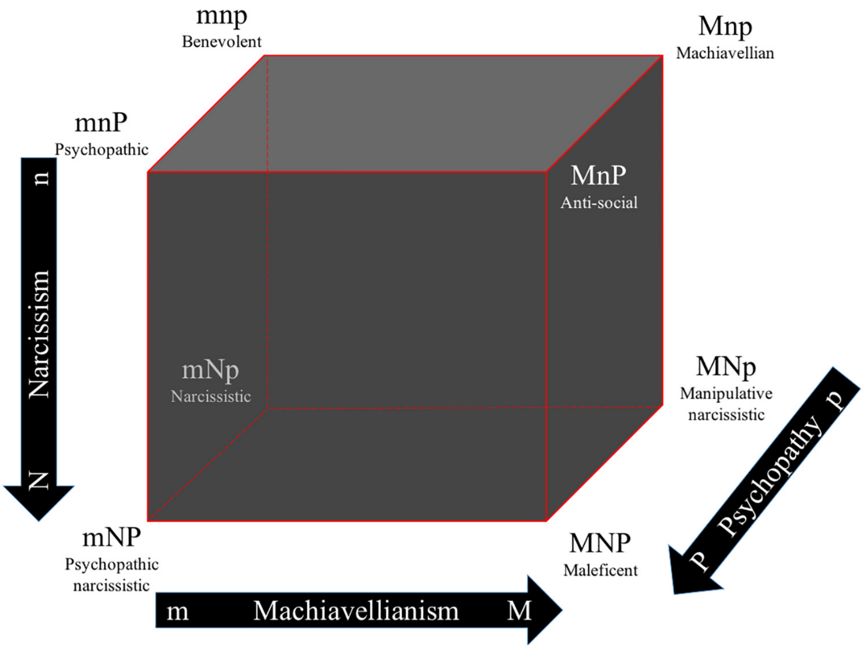


FIGURE 2 | The dark cube as an analogy to Cloninger's character cube, showing all eight possible combinations of high/low scores in Machiavellianism, narcissism, and psychopathy. Adapted with permission from C. R. Cloninger. Originally published in: Garcia and Rosenberg (2016) The dark cube: dark and light character profiles. M/m, high/low Machiavellianism; N/n, high/low narcissism; P/p, high/low psychopathy.

text corpus[5] (see also Lin et al., 2012). First, using semanticexcel, the researcher generates a matrix where rows correspond to unique single words and each column corresponds to the 5-gram context to the words in the corpus. The rows for the English corpus used here consisted of the 120,000 most frequent words, whereas the columns consisted of the contexts of the 10,000 most common words. The contexts of the words were generated from the 5-gram of Google N-grams database, that is, for each 5-gram that each word had, the context consisted of four other words. Thus, cells in this matrix represent the frequency of occurrence of a word (rows) within a context of a word (columns). For example, the word "grateful" may have a frequency f_1 in the context "aiding" and a frequency f_2 in the context "accidents." In this way, every word is represented by an array of frequencies of occurrence in each related context to a word. A basic assumption is that words with similar meaning tend to occur in the same contexts (cf. Landauer and Dumais, 1997; Landauer et al., 2007; Landauer, 2008). This implies that the vectors representing similar words are expected to point in similar direction. However, to get a good semantic representation, this word-by-context sample matrix needs to be compressed to a smaller word-by-semantic dimension matrix, where this smaller matrix tends to create a more generalized semantic representation. We conducted this data compression using singular value decomposition, a widespread dimensionality-reduction technique similar to principal component analysis. The resulting matrix is called a semantic space, which describes the semantic relatedness between words. In our analysis, the resulting semantic representation consisted of 120,000 words, where each word is represented in a vector consisting of 512 dimensions. In the present study, using semanticexcel, we simply added the vectors representing each of the 10 self-descriptive words generated by the participants. Hence, each participant's set of 10 words obtains a quantified semantic representation based on the sum of the vectors corresponding to each of the participant's words. For a more elaborated description, see Sikström and Garcia (2019).

Predicting Participants' Character Traits Scores Based on the Semantic Representation of Their Own Self-Descriptive Words

Semanticexcel uses multiple linear regressions ($Y = c \times X$), with the semantic representations as input (X, i.e., a participants × semantic dimensions matrix), to train the regression coefficients (c, i.e., a vector corresponding to the weights of each semantic dimension) to predict participants' self-reported scores in each of the personality traits (Y). One multiple linear regression was conducted for each trait score. An N-leave (where N is 10% of the total dataset) out-cross validation procedure is used to evaluate the results from the multiple linear regression so that the-to-be predicted data point is removed from the training set (where the coefficients of the multiple linear regression are generated) and where these coefficients are applied to make a prediction on the left-out test data point. Thus, 10 (N) new training and testing sets are made for cross-validation. To avoid overfitting, a subset of the dimensions in a semantic representation is used, where fitting with too many parameters in relation to the number test data points may yield poor

[5] https://books.google.com/ngrams

generalization to test dataset. This subset is selected by selecting the first (N) dimension in semantic representation and then optimizing the number of dimensions (N) used by an additional 10% leave-out procedure. Furthermore, the maximum number of dimensions used is set to one half of the total number of predicted data points. In short, semanticexcel generates the predicted values by applying the regression coefficients (c) from the training dataset on the test dataset. To evaluate whether participants' personality trait scores are significantly predicted by the semantic representation of the 10 generated words, the personality trait scores are simply correlated with the predicted values. A significant positive correlation (one-tailed) indicates that the semantic representation predicts the outcome variable (i.e., the participants' score in each of the personality traits).

Mapping the Frequency of Self-Descriptive Words and Self-Reported Personality Traits

Each word's frequency was correlated to participants' scores in each of the personality traits. To present these results, for each personality measure, we conducted one-dimensional correlations (i.e., one trait at a time) and three-dimensional correlations (i.e., interactions between high and low scores in the three character traits for each personality model). Preliminary analyses of the one-dimensional correlations presented in **Figures 3**, **5** were earlier published elsewhere (Garcia and Sikström, 2019).

The Present Study

In the present study, we used quantitative semantics to validate two short personality inventories, the Short Character Inventory and the Short Dark Triad. This method allowed us to extract and represent the meaning of words based on the context in which they co-occur with other words. We expected that the quantified meaning of words that an individual use to intentionally describe herself/himself may predict her/his level in different personality traits, thus, allowing the validation of each trait measurement. We also mapped the self-presentation words to responses in each scale and also to any interaction between the traits within each personality model (i.e., light character profiles and dark character profiles).

Ethics Statement

Ethics approval was not required at the time the research was conducted as per national regulations. The consent of the participants was obtained by virtue of survey completion after they were provided with all relevant information about the research (e.g., anonymity, possibility to withdraw at any time, etc.).

STUDY 1: LIGHT CHARACTER

Method

Participants and Procedure

The participants were recruited from Mechanical Turk (MTurk)[6]. In the initial stage, we informed the participants that the survey was anonymous, voluntary, and that they could stop the survey at any time. The participants received a small compensation/reward of USD 0.50. for participating and were requested, through the Amazon system, to be residents of the United States and to have American English as their mother tongue. We added two control questions to control for automatic responses (i.e., This is a control question, please answer "neither agree or disagree"). Three out of 1,000 participants failed to respond correctly to this question; thus, the final sample comprised 997 participants (age M= 34.13, SD = 11.92; 363 male, 634 female).

Instruments

The 10 Words Personality Inventory

This instrument was designed to request participants to freely generate words they use for self-description (Garcia and Sikström, 2015, 2019). It contains one question, asking the participants to generate 10 words that describe her/his personality ("Please describe your personality using ten words").

The Short Character Inventory

C. R. Cloninger designed the Short Character Inventory for Time Magazine as a brief version of the Temperament and Character Inventory that is easy to administer for testing relationships among personality variables in large groups (Cloninger et al., 1993). We obtained permission from C. R. Cloninger to include the inventory in the present study. The inventory contains 15 items, all present in the original long version, which are rated on a five-point Likert scale (1 = definitely false, 5 = definitely true). Examples of the items are the following: "Each day I try to take another step toward my goals" (self-directedness; Cronbach's α = 0.56), "I enjoy getting revenge on people who hurt me" (cooperativeness, reversed item, Cronbach's α = 0.54), and "Sometimes I have felt like I was part of something with no limits or boundaries in time and space" (self-transcendence, Cronbach's α = 0.57).

Results and Discussion

Semantic Representations and Self-Reported Scores of Light Character Traits

The semantic representations of the characters created using the self-descriptive words correlated significantly with the corresponding values of the self-reported traits: self-directedness: r = 0.33, p < 0.0001; cooperativeness: r = 0.28, p < 0.0001; and self-transcendence: r = 0.16, p < 0.0001 (black cells in **Table 1**). The intracorrelations between the self-reported scores (dark gray cells in **Table 1**) and the intracorrelations between the light character traits semantic representations (light gray cells in **Table 1**) showed a different pattern. There were significantly higher correlations (ranging between 0.46 and 0.50) between the semantic representations of the traits compared to the correlations between the self-reported scores (ranging between

[6] MTurk is an online system by Amazon.com (www.mturk.com/mturk/welcome) that provides access to a wide range of participants for research and other tasks. Each participant receives a payment for his/her work, and the amount varies depending on the size of the assignment. According to Goodman et al. (2013), 16 of America's top 30 universities use MTurk to collect data. Rand (2011) verified that MTurk's demographic answers are correct, and Buhrmester et al. (2011) have validated the psychometric properties of the answers in relation to data collected among undergraduate students and clinical samples.

TABLE 1 | Correlations between the semantic representation and the self-reported scores of the light character traits.

Malevolent Character		Self-reported scores			Semantic representation		
		Self-directedness	Cooperativeness	Self-transcendence	Self-directedness	Cooperativeness	Self-transcendence
Self-reported Scores	Self-directedness	–			0.33***	0.19***	0.19***
	Cooperativeness	0.29***	–		0.18***	0.28***	0.23***
	Self-transcendence	0.10**	0.16***	–	0.14***	0.18***	0.16***
Semantic Representation	Self-directedness				–		
	Cooperativeness				0.46***	–	
	Self-transcendence				0.46***	0.50***	–

****p < 0.0001; **p < 0.001. Black cells, correlations between semantic representations and self-reported scores of light character; dark gray cells, correlations between self-reported scores of light character; light gray cells, correlations between semantic representations of light character.*

0.10 and 0.29): for the correlation between self-directedness–cooperativeness was $z = -4.43$, $p < 0.001$; for self-directedness–self-transcendence was $z = -8.85$, $p < 0.001$; for cooperativeness–self-transcendence was $z = -8.65$, $p < 0.001$. Thus, these suggest that the semantic representations may not be able to discriminate between the character traits or that the items in the scales prime participants to generate words with similar meaning. This was more accentuated for the trait of self-transcendence, where the self-reported score correlated to an almost equal degree to all three semantic representations of the three light character traits: 0.14 with the semantic representation of self-directedness; 0.18 with the semantic representation of cooperativeness; and 0.16 with the semantic representation of self-transcendence. That being said, the fact that the semantic representations were so strongly related to each other, while the self-reported scores were not, suggests that the quantification of the self-descriptive words might fail to capture the nuances targeted by the scales. Other algorithms might be necessary to allow a better validation (see among others Larsen et al., 2008; Arnulf et al., 2019).

Self-Descriptive Words and Self-Reported Scores of Light Character Traits

We conducted a correlation analysis between participants' scores in each of the traits and the participant's frequency of occurrence of each of the self-descriptive words (**Figure 3**). The 997 participants generated 1,436 words that appeared one time or more in the dataset, that is, they were "unique words." Because the number of participants were quite large, we could find significant effect although some correlations were somewhat low (e.g., $r = 0.11$); thus, the p values were corrected for multiple comparisons using Holm's correction.

The number of times that participants have generated significant words in Study 1 are found in **Supplementary Table S1**. In the first analysis, one-dimensional Pearson correlations, we found one word associated with both self-directedness and self-transcendence character trait scores, namely, "happy" ($n = 180$). Accordingly, Cloninger (2004, 2007, 2013) has, in a series of studies, showed that both of these character traits are associated to happiness and positive affect and emotions. Moreover, one communal word was positively associated with participants' scores in cooperativeness and self-transcendence: "loving" ($n = 257$). The words "caring" ($n = 320$, which is the most commonly generated word, corresponding to 22% of the participants responses) and "kind" ($n = 251$), and "compassionate" ($n = 89$) were indicative only of cooperativeness. Both these traits are expressions of a person's relation to others and the world around. Self-transcendence specifically is associated with humanistic and oceanic feelings; thus, the world "loving" might express more of a universal feeling, while "kind," "caring," and "compassionate" might refer to one's relationship to others. For high levels of self-directedness, two words were indictive: "outgoing" ($n = 150$) and "strong" ($n = 116$). Both words are in line with high self-directedness (Cloninger, 2004). In addition, low self-directedness was indicated by words such as "anxious" ($n = 63$), "shy" ($n = 123$), "lazy," "quiet" ($n = 157$), "reserved" ($n = 77$), and "introverted" ($n = 72$), hence suggesting that the self-directedness scale measures both

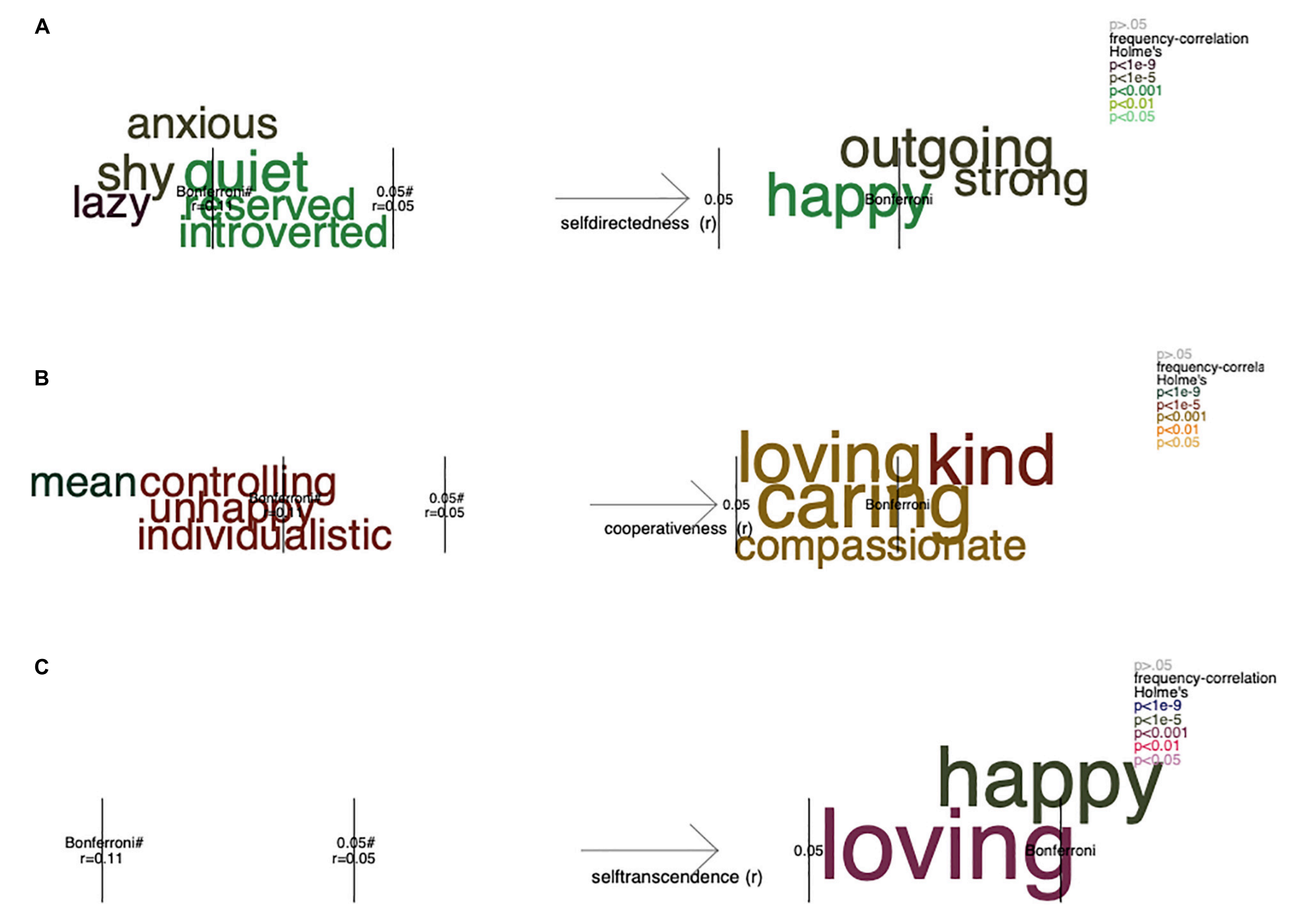


FIGURE 3 | One-dimensional analysis: the frequency of the self-descriptive words that significantly correlated with participants' scores in self-directedness **(A)**, cooperativeness **(B)**, and self-transcendence **(C)**. The figure shows, on the *x*-axis, color-coded words that significantly discriminate between the high and the low value of the scale. The area outside of the inner gray lines represents significant differences without correction for multiple comparisons ($p = 0.05$), and the areas outside of the outer gray lines represents significant values following Holm's correction for multiple comparisons, where the number of significant words are $n = 8$ for self-directedness **(A)**, $n = 6$ for cooperativeness **(B)**, and $n = 2$ for self-transcendence **(C)**. The font size represents the frequency of occurrence of the words. The total number of unique words was 1,436, so that the percentage of unique significant words ranged from 0.14 to 0.56%. Significance testing are made by Pearson correlation to scores in each light character trait. Preliminary analyses for the results presented here were earlier published in Garcia and Sikström (2019).

degree of responsibility ("lazy") and extroversion/introversion ("reserved," "quiet," "introverted"). Finally, low self-directedness has been found to be associated to mental illness (Cloninger, 2004), which here was indicated by the relationship to self-describing oneself as "anxious." Indeed, other studies (e.g., De Fruyt et al., 2000) using self-reported scores have found self-directedness to correlate to neuroticism ($r = -0.63$), extraversion ($r = 0.29$), and conscientiousness ($r = 0.45$).

We used the theorized eight profiles within the "Light" Character Cube (Cloninger, 2004) as the framework of the three-dimensional analyses (see **Figure 4**): SCT "creative," SCt "organized," ScT "absolutist," Sct "bossy," sCT "moody," sCt "dependent," scT "disorganized," and sct "apathetic." As expected individuals with an "apathetic" profile described themselves with words typical of a person with an immature character and high ill-being, for example, "sarcastic," "mean," "lazy," and "anxious." In contrasts, individuals with the opposite profile (i.e., "creative") described themselves with words such as "kind," "caring," "loving," "happy," "warm," and "compassionate." Indeed, the combination of being highly self-directed, cooperative, and self-transcendent (i.e., "creative" character profile) facilitates a person getting in a state of calm alertness, thus allowing her/him to discover creative solutions that are adaptive for her/him, other people, and humanity at large (Cloninger et al., 2016). In contrast, people who are low in all three character traits (i.e., "apathetic" profile) feel that "life is hard, people are mean, and then you die." (Cloninger, 2004). In other words, they feel victimized and helpless (low self-directedness and low cooperativeness) and are injudicious (low self-transcendence) and distrustful (low cooperativeness and low self-transcendence). Consequently, they experience frequent negative emotions and rare positive emotions (Cloninger, 2004). Individuals with a "bossy" profile were denoted by the word "strong." Accordingly, Cloninger (2004) has described people with this profile as

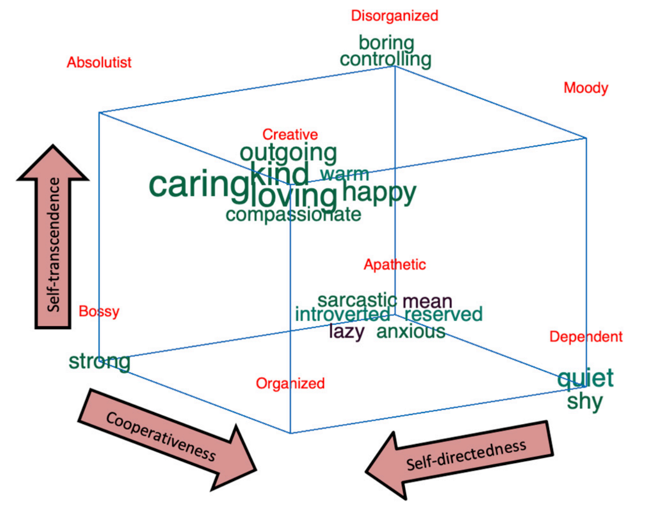


FIGURE 4 | Three-dimensional analysis: the self-descriptive words mapped to the interactions between all three character traits, that is, character profiles. The analyses plot the self-descriptive words as a cube, where the corners of each cube represent words indicative of high or low values of the three character traits following Holm's correction of multiple comparisons. Each of the eight corners of the cube represent the eight possible combinations that a word is significant for a high or low value in the three portrayed traits. For example, if a word is significant for a high value in all three traits, then it is placed in the SCT "creative" corner, whereas if it is significant for a low value of all three traits, it is placed in the sct "aphetic" corner. For details on the three axes, see the footnote in **Figure 3**.

domineering (high self-directedness and low cooperativeness), logical (high self-directedness and low self-transcendence), and distrustful (low cooperativeness and low self-transcendence). They often give orders without listening to other people to gain a shared perspective because they are distrustful. Hence, using the word "strong" to describe the self makes sense in this context. Furthermore, Cloninger (2004) describes individuals with a "disorganized" profile as often being preoccupied with unrealistic fantasies and experiencing frequent distortions of reality, such as illusions and superstitions. It is unclear if the self-descriptive words associated with this profile (i.e., "boring" and "controlling") validate this specific character combination. In contrast, the self-descriptive words associated with a "dependent" profile ("quiet" and "shy") are a relatively good description of a person that is submissive (low self-directedness and high cooperativeness), injudicious (low self-directedness and low self-transcendence), and conventional (high cooperativeness and low self-transcendence). This creates an insecure dependent relationship in which they are not self-reliant (Cloninger, 2004).

However, three of the profiles were not associated with any specific self-descriptive words. Thus, these specific character combinations (i.e., SCt "organized," ScT "absolutist," and sCT "moody") might be less valid using the Short Character Inventory. Indeed, in recent genetic studies (Zwir et al., 2018a,b, 2019), Cloninger and colleagues have shown that the natural building blocks of personality are multifaceted profiles of the whole person, not individual traits, something that can hardly be accurately calculated using a short self-reported measure.

STUDY 2: DARK CHARACTER

Method

Participants and Procedure

As for Study 1, participants in Study 2 were recruited through MTurk, and we followed exactly the same protocol for the data collection. The 10 Words Personality Inventory was also used in Study 2 to ask participants to describe their personality using words. As for Study 1, we added two control questions to control for automatic responses (e.g., This is a control question, please answer "neither agree or disagree"), which eliminated 100 participants (4.04% internal dropout) from the final cohort: 2,373 participants, 845 of which were men (M = 33.37, SD = 11.52) and 1,527 were women (M = 35.44, SD = 12.78).

Instruments

The 10 Words Personality Inventory

This instrument was designed to request participants to freely generate self-descriptive words (Garcia and Sikström, 2015, 2019). It contains one question, asking the participants to generate 10 words that describe her/his personality ("Please describe your personality using ten words").

The Short Dark Triad

We used the Short Dark Triad (Jones and Paulhus, 2014) to measure the three dark traits: Machiavellianism, narcissism, and psychopathy. The Short Dark Triad comprises 27 items, nine per trait, that are rated on a five-point Likert scale (1 = strongly disagree to 5 = strongly agree). Examples of the items are the following: "Most people can be manipulated" (Machiavellianism; Cronbach's α = 0.76), "People see me as a natural leader" (narcissism; Cronbach's α = 0.76), and "Payback needs to be quick and nasty" (psychopathy; Cronbach's α = 0.73).

Results and Discussion

Semantic Representations and Self-Reported Scores of Malevolent Character Traits

The semantic representations of the malevolent characters created using the self-descriptive words correlated with the corresponding values of the self-reported dark traits: Machiavellianism: r = 0.19, p < 0.0001; narcissism: r = 0.35, p < 0.0001; and Psychopathy: r = 0.35, p < 0.0001 (see black cells in **Table 2**). The intracorrelations between the self-reported scores (dark gray cells in **Table 2**) and the intracorrelations between the dark traits semantic representations (black cells in **Table 2**) showed almost the same pattern: a higher correlation between Machiavellianism and psychopathy (r = 0.52 between self-reported scores and r = 0.58 for semantic representations; z = −2.97, p < 0.001), a more moderate correlation between narcissism and psychopathy (r = 0.39 between self-reported scores and r = 0.44 for semantic representations; z = −2.08, p < 0.05), and a lower correlation between Machiavellianism and narcissism (r = 0.34 between self-reported scores and r = 0.16

TABLE 2 | Correlations between the semantic representation and the self-reported scores of the dark traits.

Malevolent Character		Self-reported scores			Semantic representation		
		Machiavellianism	**Narcissism**	**Psychopathy**	**Machiavellianism**	**Narcissism**	**Psychopathy**
Self-reported Scores	Machiavellianism	–			0.19***	0.04	0.23***
	Narcissism	0.34***	–		0.04	0.35***	0.16***
	Psychopathy	0.52***	0.39***	–	0.23***	0.16***	0.35***
Semantic Representation	Machiavellianism				–		
	Narcissism				0.16***	–	
	Psychopathy				0.58***	0.44***	–

*Black cells, correlations between semantic representations and self-reported scores of malevolent character; dark gray cells, correlations between self-reported scores of malevolent character; light gray cells, correlations between semantic representations of malevolent character. ***p < 0.0001.*

for semantic representations; $z = 6.63$, $p < 0.001$). Nevertheless, there were some inconsistencies. For instance, the relationship between the semantic representation of Machiavellianism and the psychopathy score ($r = 0.23$) was similar ($z = 1.44$, $p = 0.08$) to the correlation between the semantic representation of Machiavellianism and the Machiavellianism score ($r = 0.19$), that is, suggesting that Machiavellianism was less accurately assessed by either the semantic representation or the self-reported score. What is more, accordingly to recent research (e.g., Persson, 2019), Machiavellianism should be unified with psychopathy, which here is expressed by the similar correlations between the Machiavellianism self-reported score and the semantic representation of psychopathy compared to the correlation between the Machiavellianism self-reported score and the semantic representation of Machiavellianism.

Self-Descriptive Words and Self-Reported Scores of Dark Character Traits

We conducted a correlation analysis between participants' scores in each of the traits and the participant's frequency of occurrence of each of the self-descriptive words. The 2,373 participants generated 25,698 words, 2,367 of these appeared one time or more in the dataset; that is, they were "unique words." In the first analysis (**Figure 5**), one-dimensional correlations, we found three communal words negatively associated with participants' scores in Machiavellianism and psychopathy: "kind," "caring," and "loving." In addition, only the word "aggressive" was positively related to all three dark traits. This is in line with the unification argument and past research suggesting a common, uncooperative, or disagreeable core among individuals expressing any or all of these malevolent tendencies (e.g., Paulhus and Williams, 2002; Lee and Ashton, 2005; Jakobwitz and Egan, 2006; Garcia and Rosenberg, 2016).

Furthermore, there were three words that were negatively related only to psychopathy (i.e., "friendly," "warm," and "compassionate") and three words negatively related only to narcissism ("shy," "quiet," and "introverted"). Interestingly, all other words that were positively related to the dark traits were unique for each trait; for Machiavellianism, "sarcastic" and "lazy;" for narcissism, "charismatic," "leader," "intelligent," and "confident," "fun," "outgoing," "strong," "charming," and "brave;" and for psychopathy, "mean," "rugged," "vicious," "tiresome," "exceptional," "abrasive," "domineering," "awesome," "gritty," "lustful," "cool," "mean," "smooth," "angry," "Christ," "joking," "dirty," "distracted," "arrogant," "sexy," "greedy," "hurting," "troubled," "dangerous," and "aggravated" (see **Figure 5**). This finding is in line with our expectations regarding unique expressions of malevolent tendencies expressed as nuances of (un)cooperativeness—for example, the less frequent use of the word "compassionate" vs. "loving" and "kind," which was unique for individuals high in psychopathy; the frequent use of the word "sarcastic" that was common among those high in Machiavellianism vs. the frequent use of the word "mean" that was more commonly used by individuals high in psychopathy.

The number of times that participants have generated significant words are found in **Supplementary Table S2**. From this table, we can see how often the participants generated words that are indicative of a trait. For example, for the trait of being high in Machiavellianism, 139 participants generated the word sarcastic, 100 lazy, and 22 aggressive. Words with positive valence tend to be generated more frequently than words with negative valence. Thus, words that were indicative of low levels of the dark traits are more commonly expressed than those that were indicative of high levels of the dark traits. For example, the words "fun" (n = 377), "outgoing" (n = 346), "sarcastic" (n = 135), "leader" (n = 47), "charismatic" (n = 35), and "mean" (n = 25) were less frequently used than "caring" (n = 774), "kind" (n = 618), "quiet" (n = 379), and "warm" (n = 156), "shy" (n = 315), and "introvert" (n = 168). Indeed, people tend to self-enhance (i.e., the desire of maximizing the positivity of self-views) and self-protect (i.e., the desire and preference for minimizing the negativity of self-views) in their self-presentations (Rosse et al., 1998; Rowatt et al., 1998) even when there is apparently no reason to appear more desirable (Tice et al., 1995; see also Amato et al., in press). However, individuals high in any of the Dark Triad traits seem to do less so, more specifically with regard to communal self-presentations. Although, we already can see in this first analysis that some words and nuances of cooperative self-presentation words discriminate between participants' scores in each of the three dark traits, we continued with the three-dimensional analysis to control for covariance between the traits.

We used the theorized eight profiles within the Dark Cube (Garcia and Rosenberg, 2016) as the framework of the three-dimensional analysis. The results are displayed in **Figure 6** and consist of words that significantly correlated with at least one of the three dimensions, following Holm's correction for multiple

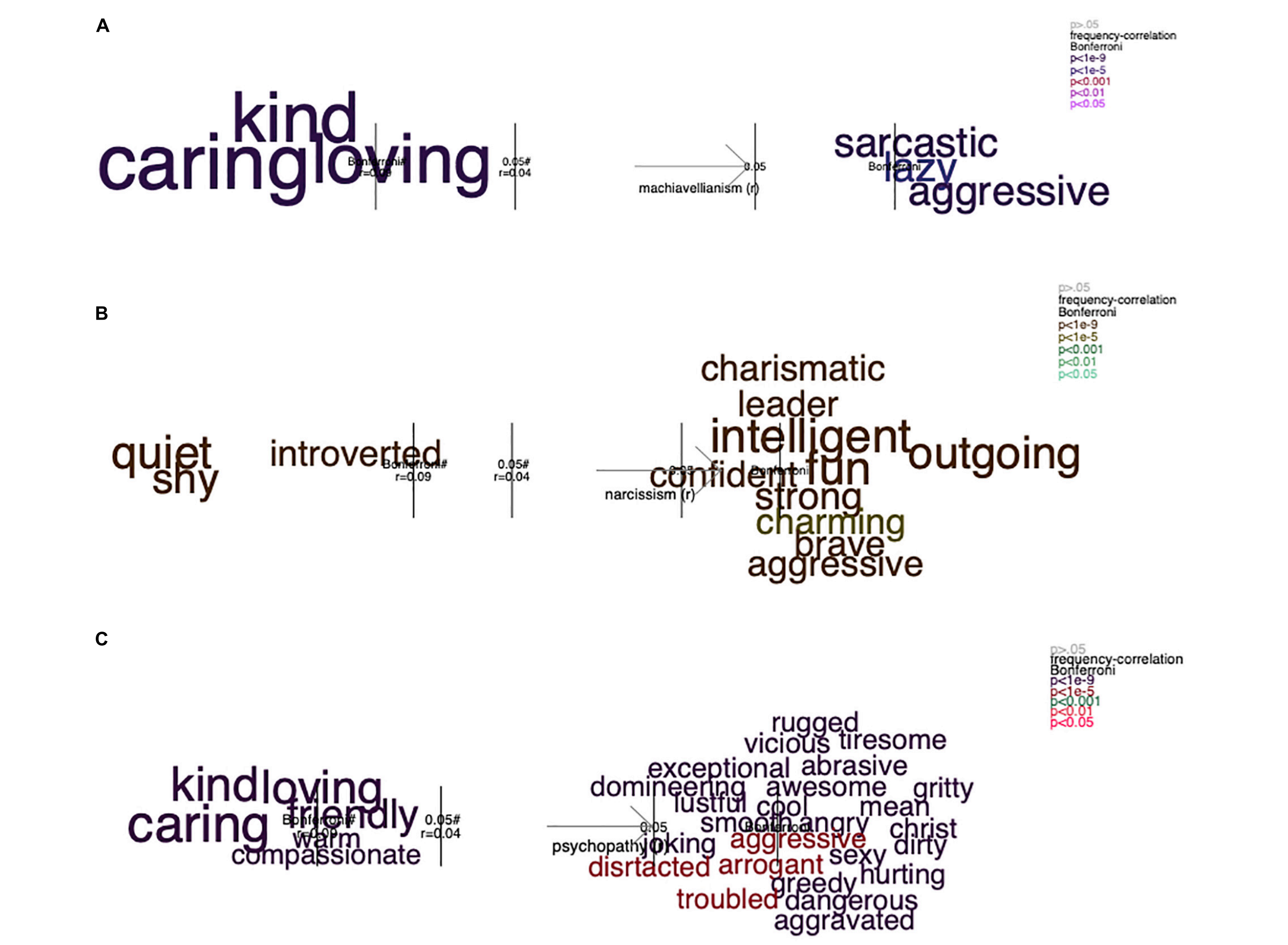


FIGURE 5 | One-dimensional analysis: the frequency of the self-descriptive words that significantly correlated with participants' scores in Machiavellianism **(A)**, narcissism **(B)**, and psychopathy **(C)**. The figure shows, on the *x*-axis, color-coded words that significantly discriminate between the high and the low values in the dark character traits. The area outside of the inner gray lines represents significant differences (*p* = 0.05), and the areas outside of the outer gray lines represents significant values following Holm's corrections for multiple comparisons. The font size represents the frequency of occurrence of the words. The *x*-axis represents the full range of the scores in Machiavellianism **(A)**, narcissism **(B)**, and psychopathy **(C)**. For additional details, see the figure note of **Figure 6**. Preliminary analyses for the results presented here were earlier published in Garcia and Sikström (2019).

comparisons. These words were located in one of the eight corners of the cube, depending on whether they were more or less common on each of the three dimensions. Individuals with a benevolent profile (i.e., low on all three traits) used the words "warm," "shy," "kind," "friendly," "compassionate," and "caring" more frequently in their self-presentations. This is, again, reinforcing the unification argument suggesting a common, uncooperative, or disagreeable core among individuals expressing any or all of these malevolent tendencies (e.g., Paulhus and Williams, 2002; Lee and Ashton, 2005; Jakobwitz and Egan, 2006; Garcia and Rosenberg, 2016).

Individuals high in Machiavellianism and low in both narcissism and psychopathy (i.e., Machiavellian profile) used words such as "quiet" and "introvert" less frequently. Together with the one-dimensional analysis, this suggests that individuals low in narcissism do present themselves as "quiet" and "introverted" but only if they at the same time are low in psychopathy and high in Machiavellianism. Conversely, individuals low in Machiavellianism and psychopathy but high in narcissism (i.e., narcissistic profile) used "loving" less frequently and "strong" more frequently. Indeed, highly narcissistic individuals manipulate others to gain self-validation, regardless if they hurt someone in doing so (Watson et al., 1984), which here is expressed as them presenting themselves as "strong." In addition, low levels of narcissism seem to be associated to being "loving" only when the individual is low in the other two malevolent traits, but to being "quite" and "introvert" when the individual is high in Machiavellianism and low in psychopathy.

Individuals with psychopathic (high in psychopathy and low in the other two) or manipulative–narcissistic profiles (high in both Machiavellianism and narcissism and low in

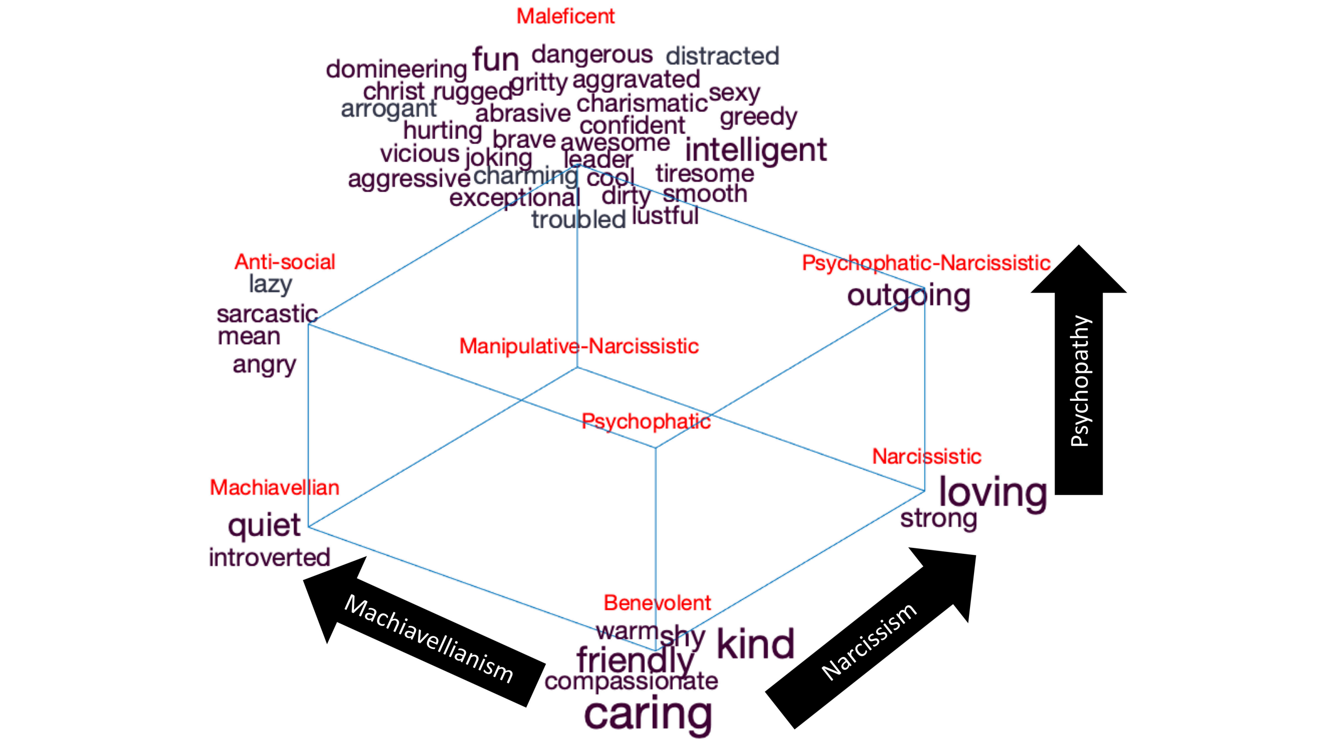


FIGURE 6 | Three-dimension analysis: the self-descriptive words mapped to the interactions between all three dark character traits, that is, dark character profiles. The figure shows words where the frequency of occurrences significantly correlates with the scores on Machiavellianism (*x*-axis; 6, or 0.26% of the unique words, are significant after Holm's correction for multiple comparisons 214 data points that are significant without correction for multiple comparisons of a total of 2,277 data points, including the comparison dataset), narcissism (*y*-axis; 13 words, or 0.57% of the unique words, are significant after Holm's correction for multiple comparisons 225 data points that are significant without correction for multiple comparisons of a total of 2,277 data points, including the comparison dataset) or psychopathy (*z*-axis; 31 words, or 1.4%, are significant after Holm's correction for multiple comparisons 278 data points that are significant without correction for multiple comparisons of a total of 2,277 data points). Significance testing were made by Pearson correlation to the dark traits scores. The value on the *x*-axis and the *y*-axis correlates $r = 0.22$, $p = 0.0000$. The value on the *x*-axis and the *z*-axis correlates $r = 0.45$, $p = 0.0000$. The value on the *y*-axis and the *z*-axis correlates $r = 0.29$, $p = 0.0000$. The words are plotted as word clouds on the corners of the three-dimensional Dark Cube representing these dark traits. The font size represents the frequency of occurrence of the words. The Dark Cube was adapted with permission from C. R. Cloninger, and it was originally published in Garcia and Rosenberg (2016).

psychopathy) seem to be harder to spot by only the use of self-presentations since none of the words correlated significantly with any of these profiles, while those with a psychopathic-narcissistic profiles (high in narcissism and psychopathy and low in Machiavellianism) expressed being "outgoing," and those individuals with an antisocial profile (high in Machiavellianism and psychopathy and low in narcissism) expressed being "lazy," "sarcastic," "mean," and "angry." Together with the one-dimensional analysis, this suggest that high Machiavellianism can be expressed by being, for example, "lazy" and "sarcastic" but only when psychopathy is high and narcissism is low. Likewise, psychopathy is expressed as being "mean" but only when Machiavellianism is high and narcissism is low. Indeed, past research suggest that individuals high in Machiavellianism and psychopathy are also low in self-discipline and that they also lack sense of duty (i.e., "lazy") (Paulhus and Williams, 2002). Last but not the least, the Maleficent profile (i.e., high in all three dark traits) was expressed with most of the words, thus depicting a dark and malevolent character (see **Figure 6**).

CONCLUSION

In the present set of studies, we used quantitative semantics to validate two short personality inventories, the Short Character Inventory and the Short Dark Triad. This method allowed us to extract and represent the meaning of words based on the context in which they co-occur with other words. We predicted that the quantified meaning of words that individuals use to describe themselves intentionally may predict their scores in different personality traits, thus allowing the validation of each trait measurement. We also mapped the self-presentation words to responses in each scale and also to any interaction between the traits within each personality model (i.e., light and dark character profiles).

Limitations and Final Remarks

Despite the limitations of our data collection method through MTurk (e.g., Buhrmester et al., 2011; Goodman et al., 2013), our study showed that the traits measured by both inventories are associated to the meaning of words people use for self-description. At the general level, each self-reported score was related to the semantic representation of each respective character trait. However, participants' self-transcendence (Study 1) and Machiavellianism scores (Study 2) demonstrated similar relationships to all three semantic representations of the character traits in their respective personality model. That being said, many of the correlations were relatively low, which might be explained by the fact that individuals were not explicitly asked to describe specific traits with their own words but their personality *per se*. Instead, the one-dimensional analyses of specific words were more informative in the validation of specific traits. Indeed, some words were indicative of both high and low levels of the character traits in each model. At the three-dimensional level, we found specific keywords that unify or that make the individuals in some profiles unique. Nevertheless, some of the profiles were not associated to any specific words. For instance, in recent genetic studies (Zwir et al., 2018a,b, 2019; Cloninger et al., 2019), Cloninger and colleagues have shown that the natural building blocks of personality are multifaceted profiles of the whole person, not individual traits. Something that can hardly be accurately calculated using short self-reported measures. Last but not the least, the measure for the light character traits is an extremely shortened version of Cloninger's Temperament and Character Inventory, and the Dark Triad measure is far from being the best measure of malevolent character. This is certainly a problem for the measures used here (e.g., the measure for light character had Cronbach's alphas that did not exceed 0.60). This is of course, partially, due to the low number of items.

In sum, despite being short, it seems like both inventories capture individuals' identity as it could be expected. Nevertheless, our method also points out some shortcomings and overlaps between traits measured with these two short personality inventories. Hence, we suggest that self-descriptive words can be quantified to validate measures of psychological constructs (e.g., using self-descriptive words in natural language and QuSTT) and that this method may complement traditional methods for testing the validity of psychological measures. Finally, since it is beyond the scope of the present study, future studies need to address the fundamental question of how the mapped words might be the base of a trait description of individuals who are high and low in different character traits. For example, as our results show, is a person high in Machiavellianism best described as sarcastic, lazy, and aggressive?

> *"I tried to gain an idea of the number of the more conspicuous aspects of the character by counting in an appropriate dictionary the words used to express them. I examined many pages of its index here and there as samples of the whole, and estimated that it contained fully one thousand words expressive of character, each of which has a separate shade of meaning, while each shares a large part of its meaning with some of the rest (Galton, 1884, p. 181)."*

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

Ethics approval was not required at the time the research was conducted as per national regulations. The consent of the participants was obtained by virtue of survey completion after they were provided with all relevant information about the research.

AUTHOR CONTRIBUTIONS

DG, PR, KC, and SS contributed to the conception and design of the study. DG, PR, AN, and AG collected the data. SS and DG performed the statistical analysis. DG and PR wrote the first draft of the manuscript. All authors wrote the sections of the manuscript, contributed to the manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by grants from the Swedish Research Council (Dnr. 2015-01229), Vinnova (Dnr. 2018-02007), and Kamprad Foundation (Dnr. 20180281). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00016/full#supplementary-material

REFERENCES

Adams, B. G., Van de Vijver, F. J. R., and De Bruin, G. P. (2012). Identity in South Africa: examining self-descriptions across ethnic groups. *Int. J. Intercult. Relat.* 36, 377–388. doi: 10.1016/j.ijintrel.2011.11.008

Amato, C., Sikström, S., and Garcia, D. (in press). "Tell Me Who You Are"(-)latent semantic analysis for analyzing spontaneous self-presentations in different situations. *Test. Psychom. Methodol. Appl. Psychol.*

Arnulf, J. K., Dysvik, A., and Larsen, K. R. (2019). Measuring semantic components in training and motivation: a methodological introduction to the semantic theory of survey response. *Human Resour. Dev. Q.* 30, 17–38. doi: 10.1002/hrdq.21324

Bergman, L. R., and Magnusson, D. (1997). A person-oriented approach in research on developmental psychopathology. *Dev. Psychopathol.* 9, 291–291.

Bergman, L. R., and Wångby, M. (2014). The person-oriented approach: a short theoretical and practical guide. *Eesti Haridusteaduste Ajak.* 2, 29–49. doi: 10.12697/eha.2014.21.02b

Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychol. Bull.* 117, 187–215. doi: 10.1037/0033-2909.117.2.187

Buhrmester, M. D., Kwang, T., and Gosling, S. D. (2011). Amazon's mechanical turk: a new source of inexpensive, yet high-quality, data? *Perspect. Psychol. Sci.* 6, 3–5. doi: 10.1177/1745691610393980

Christie, R., and Geis, F. L. (1970). *Studies in Machiavellianism*. New York, NY: Academic Press.

Cloninger, C. R. (2003). A psychobiological model of temperament and character: TCI. *Yeni Symp.* 41, 86–97.

Cloninger, C. R. (2004). *Feeling Good: The Science of Well-Being*. New York, NY: Oxford University Press.

Cloninger, C. R. (2007). Spirituality and the science of feeling good. *South. Med. J.* 100, 740–743.

Cloninger, C. R. (2009). The evolution of human brain functions: the functional structure of human consciousness. *Aust. N. Z. J. Psychiatry* 43, 994–1006. doi: 10.1080/00048670903270506

Cloninger, C. R. (2013). What makes people healthy, happy, and fulfilled in the face of current world challenges. *Mens Sana Monogr.* 11, 16–24. doi: 10.4103/0973-1229.109288

Cloninger, C. R., Cloninger, K. M., and Mezzich, J. E., (2016). "Holistic framework for ill health and positive health," in *Person Centered Psychiatry*, eds J. E. Mezzich, M. Botbol, G. N. Christodoulou, C. R. Cloninger, and I. M. Salloum, (New York, NY: Springer).

Cloninger, C. R., Cloninger, K. M., Zwir, I., and Keltigangas-Järvinen, L. (2019). The complex genetics and biology of human temperament: a review of traditional concepts in relation to new molecular findings. *Transl. Psychiatry* 9:290. doi: 10.1038/s41398-019-0621-4

Cloninger, C. R., Svrakic, D. M., and Przybeck, T. R. (1993). A psychobiological model of temperament and character. *Arch. Gen. Psychiatry* 50, 975–990.

Cloninger, C. R., Svrakic, N. M., and Svrakic, D. M. (1997). Role of personality self-organization in development of mental order and disorder. *Dev. Psychopathol.* 9, 881–906. doi: 10.1017/S095457949700148X

Crysel, L. C., Crosier, B. S., and Webster, G. D. (2013). The dark triad and risk behavior. *Pers. Individ. Diff.* 54, 35–40.

De Fruyt, F., De Wiele, L. V., and Van Heeringen, C. (2000). Cloninger's psychobiological model of temperament and character and the five-factor model of personality. *Pers. Individ. Diff.* 29, 441–452. doi: 10.1016/s0191-8869(99)00204-4

Furnham, A., and Crump, J. (2005). Personality traits, types, and disorders: an examination of the relationship between three self-report measures. *Eur. J. Pers.* 19, 167 –184.

Galton, F. (1884). Measurement of character. *Fortnightly Review* 36, 179–185.

Garcia, D. (2018). "Dark Cube," in *Encyclopedia of Personality and Individual Differences*, eds V. Zeigler-Hill, and T. Shackelford, (Cham: Springer), 1–6.

Garcia, D., Anckarsäter, H., Kjell, O. N. E., Archer, T., Rosenberg, P., Cloninger, C. R., et al. (2015). Agentic, communal, and spiritual traits are related to the semantic representation of written narratives of positive and negative life events. *Psychol. Well Being* 5, 1–20. doi: 10.1186/s13612-015-0035-x

Garcia, D., and Gonzàlez, F. R. (2017). The dark cube: dark profiles character profiles and OCEAN. *PeerJ* 5:e3845. doi: 10.7717/peerj.3845

Garcia, D., Lester, N., Cloninger, K. M., and Cloninger, C. R. (2017). "The temperament and character inventory (TCI)," in *Encyclopedia of Eersonality and Individual Differences*, eds V. Zeigler-Hill, and T. Shackelford, (Cham: Springer), 1–3. doi: 10.1007/978-3-319-28099-8_91-1

Garcia, D., and Rosenberg, P. (2016). The dark cube: dark and light character profiles. *PeerJ* 4:e1675. doi: 10.7717/peerj.1675

Garcia, D., Rosenberg, P., Gonzaìlez, F. R., and Rapp-Ricciardi, M. (2018). Dark time matter: dark character profiles and time perspective. *Psychology* 9, 63–79. doi: 10.4236/psych.2018.91005

Garcia, D., and Sikström, S. (2013a). A collective theory of happiness: words related to the word happiness in swedish online newspapers. *Cyberpsychol. Behav. Soc. Netw.* 16, 469–472. doi: 10.1089/cyber.2012.0535

Garcia, D., and Sikström, S. (2013b). Quantifying the semantic representations in adolescents' memories of positive and negative life events. *J. Happiness Stud.* 14, 1309–1323. doi: 10.1007/s10902-012-9385-8

Garcia, D., and Sikström, S. (2014). The dark side of facebook – dark triad of personality predicts semantic representation of status updates. *Pers. Individ. Diff.* 67, 92–94. doi: 10.1016/j.paid.2013.10.001

Garcia, D., and Sikström, S. (2015). Friend or worker? Descriptions of one's personality in linkedin. international society for the study of individual differences meeting. London, Ontario, Canada. *Pers. Individ. Diff.* 101, 480. doi: 10.1016/j.paid.2016.05.144

Garcia, D., and Sikström, S. (2019). "The ten words personality inventory (10WPI)," in *Encyclopedia of Personality and Individual Differences*, eds V. Zeigler-Hill, and T. Shackelford, (Cham: Springer), 1–6. doi: 10.1007/978-3-319-28099-8_2314-1

Goodman, J. K., Cryder, C. E., and Cheema, A. (2013). Data collection in a flat world: the strengths and weaknesses of mechanical turk samples. *J. Behav. Decis. Mak.* 26, 213–224. doi: 10.1002/bdm.1753

Hare, R. D. (1985). Comparison of procedures for the assessment of psychopathy. *J. Consult. Clin. Psychol.* 53, 7–16 doi: 10.1037/0022-006x.53.1.7

Hawley, P. (2003). Prosocial and coercive configurations of resource control in early adolescence: a case for the well-adapted Machiavellian. *J. Dev. Psychol.* 49, 279–309. doi: 10.1353/mpq.2003.0013

Jakobwitz, S., and Egan, V. (2006). The dark triad and normal personality traits. *Pers. Individ. Diff.* 2, 331–339. doi: 10.1016/j.paid.2005.07.006

John, O. P., Angleitner, A., and Ostendorf, F. (1988). The lexical approach to personality: a historical review of trait taxonomic research. *Eur. J. Pers.* 2, 171–203. doi: 10.1002/per.2410020302

Jones, D. J., and Paulhus, D. L. (2014). Introducing the short dark triad (SD3): a brief measure of dark personality traits. *Assessment* 21, 28–41. doi: 10.1177/1073191113-514105

Jones, D. N. (2014). Predatory personalities as behavioral mimics and parasites: mimicry-deception theory. *Perspect. Psychol. Sci.* 9, 445–451. doi: 10.1177/1745691614535936

Jones, D. N., and Paulhus, D. L. (2017). Duplicity among the dark triad: three faces of deceit. *J. Pers. Soc. Psychol.* 113, 329–342. doi: 10.1037/pspp0000139

Kajonius, P. J., Persson, B. N., Rosenberg, P., and Garcia, D. (2016). The (mis)measurement of the dark triad dirty dozen: exploitation at the core of the scale. *PeerJ* 4:e1748. doi: 10.7717/peerj.1748

Kelley, T. L. (1927). *Interpretation of Educational Measurements*. New York, NY: World Book Co.

Köse S. (2003). A psychobiological model of temperament and character. *Yeni Symp*. 41, 86–97. doi: 10.1001/archpsyc.1993.01820240059008

Landauer, T. K. (2008). "LSA as a theory of meaning." in *Handbook of Latent Semantic Analysis*, eds T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, (Mahwah, NJ: Lawrence Erlbaum Associates).

Landauer, T. K., and Dumais, S. (1997). A solution to plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104, 211–240. doi: 10.1037/0033-295x.104.2.211

Landauer, T. K., McNamara, D. S., Dennis, S., and Kintsch, W. (2007)*Handbook of Latent Semantic Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates. doi: 10.1037/0033-295x.104.2.211

Larsen, K. R., Nevo, D., and Rich, E. (2008). "Exploring the semantic validity of questionnaire scales," in *Proceedings from the 41st Hawaii International Conference on System Sciences* (Waikoloa, HI: IEEE).

Lee, K., and Ashton, M. C. (2005). Psychopathy, machiavellianism, and narcissism in the five-factor model and the HEXACO model of personality structure. *Pers. Individ. Diff.* 38, 1571–1582. doi: 10.1016/j.paid.2004.09.016

Leising, D., Scharloth, J., Lohse, O., and Wood, D. (2014). What types of terms do people use when describing an individual's personality? *Psychol. Sci.* 25, 1787–1794. doi: 10.1177/0956797614541285

Lin, Y., Michel, J.-B., Lieberman Aider, E., Orwant, J., Brockman, W., and Petrov, S. (2012). "Syntactic annotations for the google books ngram corpus," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 169–174.

Moshagen, M., Hilbig, B. E., and Zettler, I. (2018). The dark core of personality. *Psychol. Rev.* 125, 656–688. doi: 10.1037/rev0000111

Oishi, S. (2007). "The application of structural equation modeling and item response theory to cross-cultural positive psychology research," in *Series in Positive Psychology. Oxford Handbook of Methods in Positive Psychology*, eds A. D. Ong, and M. H. M. van Dulmen, (New York, NY: Oxford University Press), 126–138.

Paulhus, D. L., and Williams, K. M. (2002). The dark triad of personality: narcissism, Machiavellianism, and psychopathy. *J. Res. Pers.* 36, 556–563. doi: 10.1016/s0092-6566(02)00505-6

Persson, B., Kajonius, P., and Garcia, D. (2017). Testing construct independence in the short dark triad using item response theory. *Pers. Individ. Diff.* 117, 74–80. doi: 10.1016/j.paid.2017.05.025

Persson, B. N. (2019). *The Latent Structure of the Dark Triad: Unifying Machiavellianism and Psychopathy*. Doctoral thesis, University of Turku, Finland, UA.

Persson, B. N., Kajonius, P. J., and Garcia, D. (2019). Revisiting the structure of the short dark triad. *Assessment* 26, 3–16. doi: 10.1177/107391117701192

Rand, D. G. (2011). The promise of mechanical turk: how online labor markets can help theorists run behavioral experiments. *J. Theor. Biol.* 299, 172–179. doi: 10.1016/j.jtbi.2011.03.004

Raskin, R., and Hall, C. S. (1979). A narcissistic personality inventory. *Psychol. Rep.* 45:590

Rosse, J. G., Stecher, M. D., Miller, J. L., and Levin, R. A. (1998). The impact of response distortion on pre-employment personality testing and hiring decisions. *J. Appl. Psychol.* 83, 634–644. doi: 10.1037/0021-9010.83.4.634

Rowatt, W. C., Cunninghan, M. R., and Druen, P. B. (1998). Deception to get a date. *Pers. Soc. Psychol. Bull.* 24, 1228–1242. doi: 10.1177/01461672982411009

Sikström, S., and Garcia, D. (2019). *Statistical Semantics – Methods and Applications*. Cham: Springer.

Smith, M. J., Cloninger, C. R., Harms, M. P., and Csernansky, J. G. (2008). Temperament and character as schizophrenia-related endophenotypes in non-psychotic siblings. *Schizophr. Res.* 104, 198–205. doi: 10.1016/j.schres.2008.06.025

Tice, D. M., Butler, J. L., Muraven, M. B., and Stillwell, A. M. (1995). When modesty prevails: differential favorability of self-presentation to friends and strangers. *J. Pers. Soc. Psychol.* 69, 1120–1138. doi: 10.1037/0022-3514.69.6.1120

Uher, J. (2013). Personality psychology: lexical approaches, assessment methods, and trait concepts reveal only half of the story. *Int. Psychol. Behav. Sci.* 47, 1–55. doi: 10.1007/s12124-013-9230-6

Vaillant, G. E., and Vaillant, C. O. (1990). Natural history of male psychological health, XII: a 45-year study of predictors of successful aging at age 65. *Am. J. Psychiatry* 147, 31–37. doi: 10.1176/ajp.147.1.31

VanderWeele, T. J. (2017). On the promotion of human flourishing. *Proc. Nat. Acad. Sci. U.S.A.* 114, 8148–8156. doi: 10.1073/pnas.1702996114

Vernon, P. A., Villani, V. C., Vickers, L. C., and Harris, J. A. (2008). A behavioral genetic investiga-tion of the dark triad and the big 5. *Pers. Individ. Diff.* 44, 445–452. doi: 10.1016/j.paid.2007.09.007

Watson, P. J., Grisham, S. O., Trotter, M. V., and Biderman, M. D. (1984). Narcissism and empathy: validity evidence for the narcissistic personality inventory. *J. Pers. Assess.* 48, 301–305. doi: 10.1207/s15327752jpa4803_12

Wong, P. T. P., and Roy, S. (2018). “Critique of positive psychology and positive interventions,” in *The Routledge International Handbook of Critical Positive Psychology*, eds N. J. L. Brown, T. Lomas, and J. Eiroa-Orosa, (New York, NY: Routledge), 142–160. doi: 10.4324/9781315659794-12

World Health Organization [WHO] (2001). *Mental Health: New Understanding, New Hope*. Geneva: World Health Organization.

Zwir, I., Arnedo, J., Del-Val, C., Pilkki-Råback, L., Konte, B., Yang, S. S., et al. (2018a). Uncovering the complex genetics of human character. *Mol. Psychiatry* doi: 10.1038/s41380-018-0263-6

Zwir, I., Arnedo, J., Del-Val, C., Pilkki-Råback, L., Konte, B., Yang, S. S., et al. (2018b). Uncovering the complex genetics of human temperament. *Mol. Psychiatry* doi: 10.1038/s41380-018-0264-5

Zwir, I., Del-Val, C., Arnedo, J., Pilkki-Raåback, L., Konte, B., Yang, S. S., et al. (2019). Three genetic-environmental networks for human personality. *Mol. Psychiatry*. doi: 10.1038/s41380-019-0579-x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

METHODS
published: 06 March 2020
doi: 10.3389/fpsyg.2020.00308

A Comparison of IRT Observed Score Kernel Equating and Several Equating Methods

Shaojie Wang[1], *Minqiang Zhang*[1,2]* *and Sen You*[2]

[1] School of Psychology, South China Normal University, Guangzhou, China, [2] The Chinese Society of Education, Beijing, China

Item response theory (IRT) observed score kernel equating was evaluated and compared with equipercentile equating, IRT observed score equating, and kernel equating methods by varying the sample size and test length. Considering that IRT data simulation might unequally favor IRT equating methods, pseudo tests and pseudo groups were also constructed to make equating results comparable with those from the IRT data simulation. Identity equating and the large sample single group rule were both set as criterion equating (or true equating) on which local and global indices were based. Results show that in random equivalent groups design, IRT observed score kernel equating is more accurate and stable than others. In non-equivalent groups with anchor test design, IRT observed score equating shows lowest systematic and random errors among equating methods. Those errors decrease as a shorter test and a larger sample are used in equating; nevertheless, effect of the latter one is ignorable. No clear preference for data simulation method is found, though still affecting equating results. Preferences for true equating are spotted in random Equivalent Groups design. Finally, recommendations and further improvements are discussed.

Keywords: item response theory observed score kernel equating, classical test theory, item response theory, data simulation, criterion equating

OPEN ACCESS

Edited by:
África Borges,
University of La Laguna, Spain

Reviewed by:
Fco. Pablo Holgado-Tello,
National University of Distance Education (UNED), Spain
Luis Manuel Lozano,
University of Granada, Spain

Correspondence:
Minqiang Zhang
zhangmq1117@qq.com

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *14 November 2019*
Accepted: *10 February 2020*
Published: *06 March 2020*

Citation:
Wang S, Zhang M and You S (2020) A Comparison of IRT Observed Score Kernel Equating and Several Equating Methods. Front. Psychol. 11:308. doi: 10.3389/fpsyg.2020.00308

INTRODUCTION

Test Equating and Kernel Equating Method

Test equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably (Kolen and Brennan, 2014). In general, two types of equating methods exist. Those based on the classical test theory (CTT) including mean equating (ME), linear equating (LE), and equipercentile equating (EE). ME assumes that scores in two paralleled test forms with the same distance to respective mean scores are equivalent. In reality, test forms not only differ on mean scores but also can have distinct standard deviations. In order to improve it, LE further hypothesizes that scores with the same distance to the mean in the corresponding standard deviation unit in two test forms are equivalent. However, two paralleled test forms may differ from each other not only on the mean and standard deviation but also on the other higher central moments. When score distribution statistics (for example, *M*, *SD*, *Sk.*, *Ku.*, etc.) of two test forms are similar, scores in paralleled test forms with the same percentile rank are equivalent according to the philosophy of EE. It can be easily deduced that ME and LE are special cases of EE.

Looking back, the classical test theory, on which the CTT equating methods are based, has been generally acknowledged that both ability parameter (i.e., observed score) and item parameters (i.e., difficulty and discrimination) are dependent on each other, limiting its utility in practical test development (Hambleton and Jones, 1993). As Lord (1977) and Cook and Eignor (1991) stated, traditional observed score equating is not possible except when test forms are of exactly equal difficulty.

Then the item response theory (IRT) solves the CTT interdependency problem by combining ability and item parameters in one model. One of the widely used IRT models is the three-parameter logistic model (3PLM), which includes location (b), discrimination (a), and pseudo-guessing (c) parameters for items, and ability (θ) parameter for participants. In IRT equating, estimated parameters in two forms are first transformed onto the same scale (Marco, 1977; Haebara, 1980; Loyd and Hoover, 1980; Stocking and Lord, 1983). The sense behind scale transformation is that if an IRT model fits data satisfactorily; then, it still does when any linear transformation of the ability or location scale has been done (Kolen and Brennan, 2014). After that, the IRT true score equating (IRTTSE) and observed score equating (IRTOSE) methods are used to transform scaled parameters in two test forms to interpretable and understandable score relationships. In IRTTSE, true scores with the same θ_i in two test forms are equated. In IRTOSE, estimated distributions of sum scores in two forms are deduced by the IRT model, which then is equated by the EE philosophy. The IRT equating methods are proven to be more accurate and stable than the CTT methods (Hambleton and Jones, 1993; Han et al., 1997; De Ayala, 2013; Kolen and Brennan, 2014) and lays foundation for modern large-scale computer-based tests, such as adaptive test, cognitive diagnosis test, and so on (Educational Testing Service, 2010; Kastberg et al., 2013; OECD, 2017). However, there are still situations where IRT equating does not suit satisfactorily. One of these circumstances is that sometimes, only a small sample (for example, less than 500 cases) is available, which is very common in practice because of participant sampling. Here, the IRT parameter estimation often confronts convergence problems (Whitely, 1977; Wright, 1977; Hambleton and Jones, 1993; de la Torre and Hong, 2010). For example, in the 3PLM, suppose one test contains j items, then, $3j$ item parameters must be estimated. As parameters increase, the minimum number of cases needed to achieve acceptable convergence results and satisfying fitness indices dramatically climb, keeping other affecting parameters (person distribution, data characteristics, etc.) fixed (De Ayala, 2013). Over the past decades, some Bayesian methods, such as the MCMC estimation (Liu et al., 2008; Sheng, 2008; Yao, 2011; Mun et al., 2019), have been developed to reduce uncertainty in the IRT models by incorporating posterior information of the parameters. However, parameter estimation under a small sample condition is still not satisfactory enough due to its unavoidable uncertainty and instability (Swaminathan and Gifford, 1985, 1986). Thus, with biased parameter estimates at the calibration stage, more errors accumulate in the IRT equating when a sample size is small. Besides, many lumps and gaps occur in a small sample score distribution, also introducing equating errors (von Davier et al., 2004; Skaggs, 2005; Kim et al., 2006; Puhan et al., 2008).

Kernel equating (KE) was proposed and aimed at solving problems mentioned above from a different perspective. It is a unified approach to test equating based on a flexible family of equipercentile-like equating functions that contains LE as a special case (von Davier et al., 2004). It first pre-smooths univariate or bivariate score probabilities from a sample by fitting appropriate statistical models, which are usually log-linear ones, to raw data obtained in an equating design. The second is to estimate score probabilities on target population by design function (DF), which is an identity, linear, or other complex forms according to the equating design. To understand this critical component, the reader should know that in KE, raw data and pre-smoothed ones by log-linear model are stored in a matrix (or contingency table) with each column and row representing a possible score in two test forms, respectively, for Single Group design (SG), Counter-Balanced groups design (CB), and Non-Equivalent groups with Anchor Test design (NEAT). However, the input in the later procedure is a probability vector. So, DF is a matrix to transform a joint score distribution of two test forms into a marginal one. Especially, if data are collected in the random Equivalent Groups design (EG) with a univariate log-linear model, no further transformation is needed, and DF is an identity matrix. However, if data are collected in other designs, more sophisticated bivariate models are used. Therefore, in order to get a probability vector, complex matrices (DF) with elements including only 1 and 0 are necessary. The third is a continuization, where discrete cumulative distribution functions for test scores are transformed into continuous ones by kernel smoothing techniques. This process is achieved through a continuized random variable, which is a combination of three parts, including the original discrete score variable, a continuous random variable characterizing a smoothing kernel, and a parameter controlling the degree of smoothness. The fourth is to equate test forms by the general EE function defined under the KE framework. Finally, the standard error of equating (SEE) and standard error of equating difference (SEED) between equating functions are calculated as criteria for evaluating KE performance (von Davier et al., 2004). The same as in evaluating other equating methods, the SEE is an indicator of a random error caused by inferring population parameters by a sample data. The SEED is a distinctive criterion in KE, and it depicts the standard deviation of differences between two KE functions. According to von Davier et al. (2004), KE differences between -2SEED and 2SEED could be regarded as mainly coming from sample uncertainty than functions themselves. Attributing to its advantages of pre-smoothing and continuization of score distributions, KE has been testified and shown equivalent to or better than other equating methods, especially traditional ones, in the aspect of equating accuracy and stability (Chen, 2012; von Davier and Chen, 2013; Kim, 2014; Leôncio and Wiberg, 2017; Wedman, 2017; Arıkan and Gelbal, 2018; De Ayala et al., 2018).

By integrating IRTOSE and KE, Andersson et al. (2013) proposed the IRT observed score kernel equating (IRTKE) in a package “kequate” in an R environment. In the IRTKE, the IRT model is first fitted to a test data, where score probabilities are derived. One of the essential components for the IRTKE, asymptotic covariance matrix of score probabilities, is also calculated (Andersson, 2016). Then, score probabilities are used to estimate continuous approximations to discrete test score distributions by kernel continuization in order to perform IRTOSE. Later, several researchers investigated the IRTKE's performances and related topics. For example, Andersson (2016) derived an asymptotic standard error for IRTKE with polytomous items with the delta method, which was used in equating evaluation, especially in error estimation. Sample size, distribution misspecification, and anchor test length were manipulated in their study to explore the effects on the derived asymptotic standard error. Then, Andersson and Wiberg (2017) introduced the IRTKE in NEAT at length, and extended asymptotic covariance matrices to chained and poststratification equating conditions. They found that IRTKE offered small standard errors and biases under most circumstances. Further, Wiberg (2016) investigated how ability changes between two test administrations affected the IRTKE and other equating methods in NEAT. Lacking of true equating criterion in empirical data, they did not draw much conclusions about which method was better performed. Meanwhile, researchers put forward some new methods by combing KE with other methods, such as the local IRTKE, local KE (Wiberg et al., 2014), and linear IRTKE (Wiberg, 2016). To sum up, the newly proposed IRTKE has been theoretically validated for its superiority to other methods, but few simulated studies are carried out to verify its equating performances when compared with the CTT methods (such as EE) and IRT methods (such as IRTOSE), which is one of major objectives in this study.

Simulation Methods

In test equating, the Monto Carlo simulation procedure is frequently used to generate response data under IRT framework (Andersson, 2016; Andersson and Wiberg, 2017; De Ayala et al., 2018). First, item parameters (difficulty, discrimination, pseudo-guessing, etc.) are randomly drawn from a certain prior distribution, which is usually lognormal, normal, or uniform distribution. Then, the response probability of answering an item right is computed according to the IRT model. Finally, if the probability is larger than a random number drawn from the uniform distribution, this person is scored 1, else 0. As illustrated roughly above, a simulation based on the IRT (simplified as the IRT method later) gives researchers much freedom to manipulate the item and person relationships by setting and changing their different prior distributions. Thus, various equating conditions could be controlled in experiments, and true values are known in advance, both of which are important to psychometric simulation. So, the IRT simulation, indeed, helps. However, there is always another concern about the possible unfairness to certain equating methods caused by the IRT, itself (Harris and Crouse, 1993; Godfrey, 2007; Choi, 2009; Norman Dvorak, 2009; Wiberg and González, 2016; Andersson and Wiberg, 2017; Kim et al., 2017; De Ayala et al., 2018). In detail, a simulation study backgrounded on the IRT may be partial to some relevant equating methods, such as IRTOSE and IRTTSE, and disadvantage others. As one manipulation procedure used mainly in equating studies, selecting real responses to items from empirical test data to construct pseudo-tests and pseudo-group (PTPG) simulation might alleviate this concern, which was first used by Petersen et al. (1982). In their study, 54 subsamples each with 1,577 participants were created by selecting cases from real test data to form random, similar, and dissimilar samples in ability. PTPG simulation directly constructs pseudo test forms and pseudo groups satisfying certain requirements without relying on IRT; thus, it is more neutral to the comparison of equating methods to some extent. Other studies involving PTPG exist (Powers and Kolen, 2011, 2012; Sinharay, 2011; Kim and Lu, 2018). One of their limitations is that repetition was not used; thus, random error could not be separated from total error. Further, Hagge and Kolen (2011, 2012) used PTPG to investigate how differences in proficiency between old and new equating groups, relative difficulty of multiple-choice and constructed-response items, format representativeness of common-item set, and equating methods affected the results. A new idea proposed was that simulation procedures were repeated 500 times, and criterion equatings were averaged as a benchmark to evaluate differences between equating methods. Therefore, the traditional frequently used IRT simulation method in test equating and the more neutral PTPG simulation method were manipulated and compared simultaneously, in order to shed light on interpretations of equating results impartially.

Criterion Equating

As its name indicates, criterion equating (also called true equating) is the baseline for equating evaluation. Kolen and Brennan (2014) summarized four equating criteria, which included criterion based on error in estimating equating relationships, equating in a circle, group invariance, and criterion based on equity property. This study focuses on equating errors. To calculate them, criterion equating needs to be defined in advance. One of the true equating relationships considered in this study is based on the large-sample single group (LSSG). Suppose one operational test has enough items and representative samples, where pseudo tests and pseudo groups could be extracted, which has been introduced before. Then, a true equating relationship can be founded based on the entire examinee samples. The logic behind is to treat all examinees as population after pseudo tests are constructed. However, another problem still exists about which equating function is used to calculate equated values. EE, IRT, KE, or IRTKE? One function might favor equating results under a similar theoretical framework (Qu, 2007; Ricker and von Davier, 2007; Choi, 2009; Chen, 2012; Wiberg and González, 2016). That is, the criteria calculated by the EE reference might lead the EE, KE, even IRTKE to smaller errors compared with the IRT, as these methods are exactly EE, itself, or its extension. The criteria calculated by other references may cause similar problems. Therefore, a reference, which is fairer and more equal to all equating methods, is needed. Identity equating (IE) treats identity function as true equating, where form Y equivalent to a form X score is set equal to the form X score, and no further transformation is needed at all. When test specification, design, data collection, and quality control procedures are adequate, IE would lead to less errors than other equating methods. In sum, to avoid it, five true equatings (IE, EE, IRT, KE, and IRTKE) were used in this study to detect criterion equating preference by comparing the results from LSSG reference with those from IE reference.

Therefore, in this study, four equating methods, including EE, IRT, KE, and IRTKE, are compared under circumstances where sample size and test length are manipulated. Meanwhile, the preference caused by the simulation method and criterion equating are also tested using two simulation methods and specifying two sorts of criterion equatings. The structure of this article is as follows. Independent variables, simulation procedures, and evaluation indices are introduced in the first part. Then come the results in EG and NEAT. Finally, discussion, conclusion, and further directions are provided.

MATERIALS AND METHODS

Data

The raw data used in simulation were from a large-scale verbal test ADM12 as part of an entrance examination to college (González and Wiberg, 2017). Form I and form II for verbal test each contains 80 multiple-choice items and 10,000 records, which are binary scored. The basic statistics are listed in **Table 1**.

Independent Variables

Five factors were crossed: equating method, sample size, test length, simulation method, and criterion equating.

Equating Method

EE (chained equating in NEAT), IRTOSE, KE, and IRTKE were applied to simulated data, which represented equating methods under the framework of CTT, IRT, KE, and a combination of the latter two methods, respectively.

Sample Size per Group

Usually, 500 or more cases are required in the IRT data analysis in consideration of model fit and convergence (Hambleton and Jones, 1993). Therefore, 500, 1,000, and 2,500 test takers were considered in this study, which represented small-, moderate-, and large-sample conditions, respectively, in educational assessment.

TABLE 1 | Summary statistics for ADM12 verbal test.

Statistics	Form I	Form II
Sample size	8000	8000
Number of items	80	80
Min (possible min)	9 (0)	11 (0)
Max (possible max)	79 (80)	78 (80)
Mean	43.33	44.24
SD	12.66	12.59
Skewness	0.12	0.04
Kurtosis	−0.65	−0.65
Reliability	0.90	0.90
Correlation between form I and form II	0.71	

Test Length

Tests including 30 and 45 items were constructed separately. Meanwhile, in NEAT, the number of internal anchor items was fixed at 30% of the total items, indicating that 9 and 14 items were labeled as common between two test forms, respectively.

Simulation Method

The IRT method and the PTPG (pseudo-tests and pseudo-groups) method were compared.

Criterion Equating

The IE (identity equating) criterion and LSSG (large-sample single group) criterion were considered. So, in fact, five true equatings (IE, EE, IRT, KE, and IRTKE) were calculated for each equating method across 500 repetitions.

Therefore, 240 conditions (4 equating methods × 3 sample sizes × 2 test lengths × 2 simulation methods × 5 criterion equatings) were manipulated in this study.

Evaluation Indices

Local and global indices were considered. Equating performances at a single score point could be inferred from local indices. Besides, overall performances were formed by adding up local indices weighted by score frequencies across a whole score scale.

Local Indices

Local indices include absolute bias (AB), standard error of equating (SE), and root mean squared error (RMSE). AB is a representative of systematic error. $AB\left[e_Y\left(x_i\right)\right] = \left|\frac{1}{500}\sum_r e_{Yr}\left(x_i\right) - e_{YC}\left(x_i\right)\right|$, $e_{Yr}(x_i)$ stands for equating result for x_i in the rth repetition, and $e_{YC}(x_i)$ is the final true equating by averaging 500 repetitions of respective criterion equating function. SE reflects random error, usually caused by sampling, $SE\left[e_Y\left(x_i\right)\right] = \sqrt{\frac{1}{500}\sum_r\left[e_{Yr}\left(x_i\right) - \frac{1}{500}\sum_r e_{Yr}\left(x_i\right)\right]^2}$. Finally, the random error is added up with the systematic error to get the total error, $RMSE[e_Y(x_i)] = \sqrt{\left[\frac{1}{500}\sum_r e_{Yr}(x_i) - e_{YC}(x_i)\right]^2 + \frac{1}{500}\sum_r\left[e_{Yr}(x_i) - \frac{1}{500}\sum_r e_{Yr}(x_i)\right]^2}$.

Global Indices

Global indices include the weighted absolute bias (WAB), weighted standard error of equating (WSE), and weighted root mean squared error (WRMSE). As aforementioned, global indices are a summation of local indices according to the corresponding weight at each score point. Therefore, $WAB\left(e_Y\right) = \sum_i w_i AB\left[e_Y\left(x_i\right)\right]$, $WSE\left(e_Y\right) = \sum_i w_i SE\left[e_Y\left(x_i\right)\right]$, and$WRMSE\left(e_Y\right) = \sum_i w_i RMSE\left[e_Y\left(x_i\right)\right]$, where $w_i = N_i/N_T$, N_i, and N_T are the case numbers of x_i and the population, respectively.

Simulation Procedures

For the PTPG simulation, there were four steps in general. Step 1, in EG, items were randomly drawn from verbal test form I to construct the pseudo-tests X and Y without replacement. In NEAT, items for anchor test A were drawn first followed by the unique parts in tests X and Y. Note that the items in the whole test consist of anchor (common) items and unique items. Step 2, two groups of students were randomly selected to construct equating samples without replacement. To be mentioned, in NEAT, students were categorized into high- and low-ability groups according to the mean score of the test form II, and then two pseudo groups with ability differences were selected randomly. Step 3, pseudo tests X and Y were equated. Finally, steps 1 to 3 were repeated 500 times, and evaluation indices were calculated.

For the IRT simulation, a two-parameter logistic model was first fit to raw data to get the slope, location, and theta parameters. In step 2, response matrices were calculated for the pseudo items and pseudo students drawn by the PTPG procedures according to the formula of the two-parameter logistic model with parameters calculated in step 1. In step 3, pseudo tests X and Y were equated. In the end, steps 1 to 3 were repeated 500 times, and evaluation indices were calculated.

The R software version 3.5.0 (R Core Team, 2017) was used in the simulation and sample choosing. The EE, IRTOSE, KE, and IRTKE were performed with the package *equate*, *mirt* and *equateIRT*, and *kequate*, respectively (Chalmers, 2012; Andersson et al., 2013; Battauz, 2015; Albano, 2016). The related R code in this study could be found in the **Appendix**.

RESULTS

Overview of Simulated Data

To get a clear view on the simulated pseudo-tests and pseudo-groups, summary statistics for pseudo test X across replications are listed in **Tables 2**, **3**. Each row represents one condition where all 500 repeated samples are aggregated together to get a brief view of the simulated sample distribution. In EG, sample means from the PTPG are approximately three points higher than those from the IRT simulation, and SDs are approximately 0.5 point lower than those from the IRT simulation, which makes more scores from the PTPG centralize around the mean score compared with those from the IRT simulation. In NEAT, sample means from PTPG are approximately two and three points higher than those from the IRT simulation in the 30- and 45-item conditions, respectively, but the SDs are approximately 0.5 point lower than those from the IRT simulation, thus, also making more cases from PTPG dwell around the corresponding mean score. It is shown that the mean, SD, and other higher-order score statistics are similar with the IE and SG references, which makes results comparable under the same conditions. What is more, in EG, the mean score for the pseudo-test X in the 30-item condition is approximately eight points lower than that in the 45-item condition for the PTPG simulation, and approximately 6.5 points lower for the IRT simulation. In NEAT, the mean score for the pseudo-test X in the 30-item condition is approximately 10 points lower than that in the 45-item condition for the PTPG simulation, and approximately nine points lower for the IRT simulation. The results in EG and NEAT are to be described separately next.

TABLE 2 | Summary statistics for simulated samples in EG across replications.

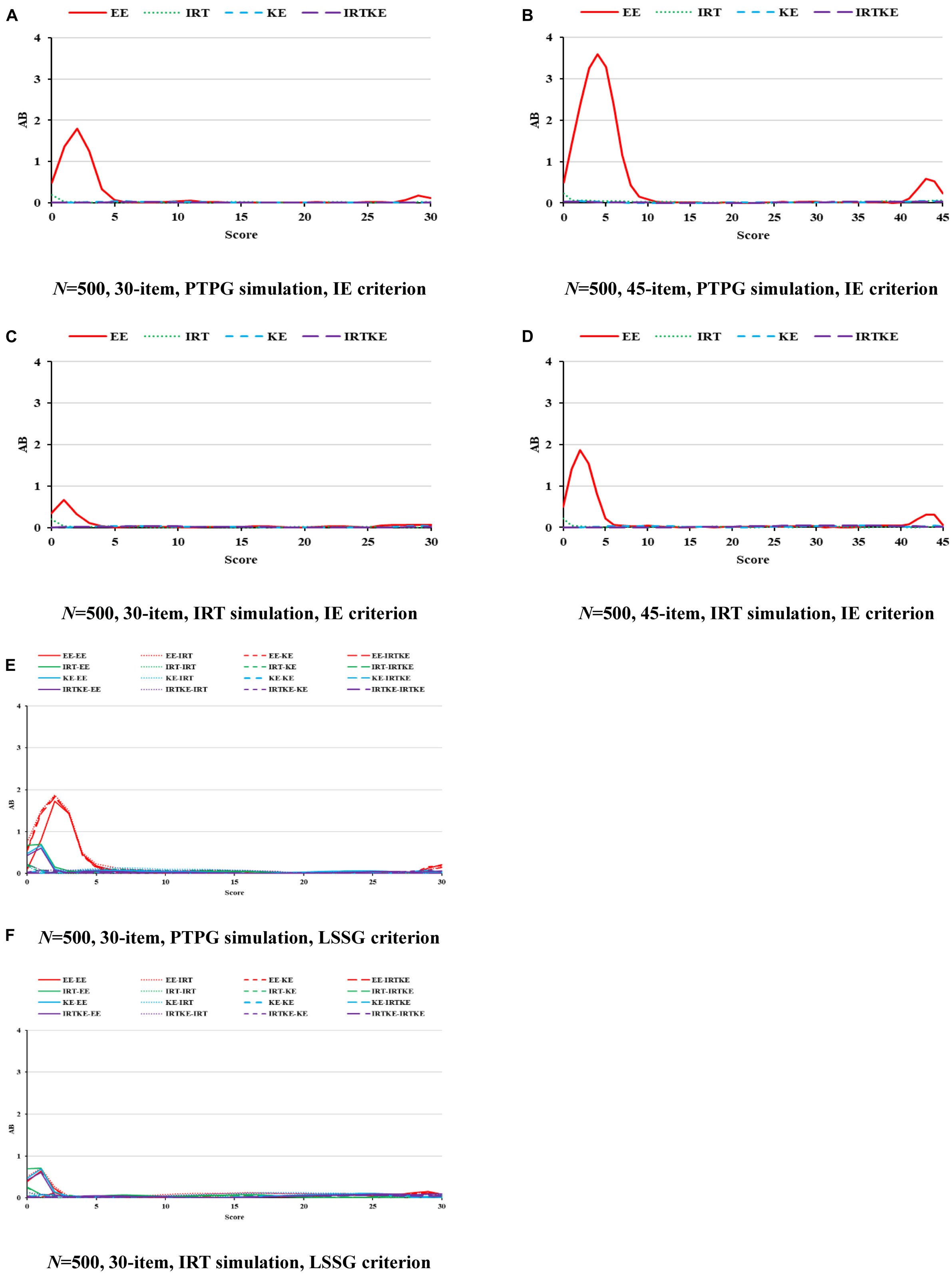
Simulation method	Criterion equating	Sample size-test length	*M*	*SD*	Min	Max	Sk	Ku
PTPG	IE	500–30	16.29	5.18	0	30	0.06	−0.59
		1000–30	16.28	5.19	0	30	0.06	−0.60
		2500–30	16.28	5.19	0	30	0.06	−0.60
		500–45	24.43	7.45	1	45	0.08	−0.63
		1000–45	24.42	7.45	1	45	0.08	−0.63
		2500–45	24.42	7.44	1	45	0.08	−0.63
	SG	500–30	16.28	5.19	1	30	0.06	−0.60
		1000–30	16.22	5.18	0	30	0.06	−0.60
		2500–30	16.28	5.19	0	30	0.05	−0.60
IRT	IE	500–30	13.35	5.58	0	30	0.45	−0.35
		1000–30	13.36	5.58	0	30	0.45	−0.35
		2500–30	13.35	5.57	0	30	0.45	−0.35
		500–45	20.04	8.06	0	45	0.49	−0.33
		1000–45	20.05	8.06	0	45	0.49	−0.33
		2500–45	20.05	8.06	0	45	0.49	−0.33
	SG	500–30	13.35	5.57	0	30	0.45	−0.35
		1000–30	13.35	5.58	0	30	0.45	−0.36
		2500–30	13.36	5.58	0	30	0.45	−0.36

TABLE 3 | Summary statistics for simulated samples in NEAT across replications.

Simulation method	Criterion equating	Sample size-test length	*M*	*SD*	Min	Max	Sk	Ku
PTPG	IE	500–30	19.78	4.04	0	30	−0.21	−0.11
		1000–30	19.78	4.03	0	30	−0.21	−0.10
		2500–30	19.78	4.03	0	30	−0.22	−0.11
		500–45	29.67	5.63	3	45	−0.19	−0.06
		1000–45	29.67	5.63	2	45	−0.19	−0.07
		2500–45	29.67	5.64	2	45	−0.19	−0.07
	SG	500–30	19.77	4.04	1	30	−0.22	−0.11
		1000–30	19.78	4.03	1	30	−0.22	−0.12
		2500–30	19.78	4.03	0	30	−0.22	−0.12
		500–45	29.68	5.65	2	45	−0.20	−0.04
		1000–45	29.68	5.64	2	45	−0.19	−0.05
		2500–45	29.67	5.63	2	45	−0.19	−0.07
IRT	IE	500–30	17.77	4.40	2	30	0.26	−0.40
		1000–30	17.78	4.40	2	30	0.27	−0.39
		2500–30	17.78	4.40	2	30	0.27	−0.39
		500–45	26.66	6.18	7	45	0.38	−0.36
		1000–45	26.66	6.19	8	45	0.38	−0.36
		2500–45	26.66	6.18	6	45	0.38	−0.36
	SG	500–30	17.77	4.40	4	30	0.27	−0.38
		1000–30	17.77	4.40	3	30	0.27	−0.39
		2500–30	17.78	4.39	2	30	0.27	−0.39
		500–45	26.65	6.18	8	45	0.38	−0.36
		1000–45	26.66	6.18	6	45	0.38	−0.37
		2500–45	26.67	6.18	7	45	0.38	−0.36

EG

In **Figure 1**, ABs are very small for all equating methods, except the EE results in low- and high-score ranges, especially in the former one, indicating that when the premise of test specification equivalence is satisfied in EG, equating methods with complicated assumptions and models, such as IRTOSE and



N=500, 30-item, PTPG simulation, IE criterion

N=500, 45-item, PTPG simulation, IE criterion

N=500, 30-item, IRT simulation, IE criterion

N=500, 45-item, IRT simulation, IE criterion

N=500, 30-item, PTPG simulation, LSSG criterion

N=500, 30-item, IRT simulation, LSSG criterion

FIGURE 1 | AB in EG. PTPG, Pseudo-Tests and Pseudo-Groups method; IRT, Item Response Theory method; IE, Identity Equating; LSSG, Large Sample Single Group; EE, Equipercentile Equating; IRT, IRT observed score equating; KE, Kernel Equating; IRTKE, IRT observed score Kernel Equating. In **(A–D)**, Red, green, blue, and purple lines represent results of EE, IRT, KE, and IRTKE respectively, calculated by IE criterion. In **(E,F)**, Red, green, blue, and purple lines represent results of EE, IRT, KE, and IRTKE respectively; continuous, dotted, short-dashed, and long-dashed lines represent results calculated by EE, IRT, KE, and IRTKE criterion respectively. Test with 45 items under LSSG reference condition was not considered.

TABLE 4 | Weighted absolute bias (WAB) in EG.

						LSSG															
		IE				**EE**				**IRT**				**KE**				**IRTKE**			
		EE	**IRT**	**KE**	**IRTKE**	**EE**	**IRT**	**KE**	**IRTKE**	**EE**	**IRT**	**KE**	**IRTKE**	**EE**	**IRT**	**KE**	**IRTKE**	**EE**	**IRT**	**KE**	**IRTKE**
PTPG	500–30	0.02	**0.01**	**0.01**	**0.01**	**0.02**	0.04	0.04	**0.02**	0.06	**0.01**	0.06	0.06	0.02	0.04	**0.01**	0.01	0.02	0.04	0.03	**0.01**
	1000–30	0.02	**0.01**	0.02	0.02	**0.01**	0.05	0.03	**0.01**	0.05	**0.00**	0.06	0.06	0.02	0.05	**0.01**	0.02	0.01	0.05	0.02	**0.00**
	2500–30	0.01	0.02	**0.01**	**0.01**	**0.00**	0.05	0.03	0.01	0.05	**0.00**	0.05	0.05	0.03	0.05	**0.00**	0.02	0.01	0.05	0.02	**0.00**
	500–45	0.03	0.02	**0.01**	**0.01**																
	1000–45	0.02	0.02	**0.01**	**0.01**																
	2500–45	**0.01**	**0.01**	**0.01**	**0.01**																
IRT	500–30	0.02	**0.01**	0.02	0.02	**0.02**	0.05	0.03	0.03	0.09	**0.03**	0.08	0.07	**0.02**	0.04	0.03	**0.02**	0.03	0.04	0.03	**0.02**
	1000–30	**0.01**	0.02	**0.01**	**0.01**	**0.03**	0.06	**0.03**	**0.03**	0.10	**0.02**	0.09	0.07	0.04	0.05	0.03	**0.02**	0.04	0.05	**0.03**	**0.03**
	2500–30	0.01	0.02	0.01	**0.00**	0.04	0.07	0.03	**0.02**	0.11	**0.01**	0.10	0.07	0.05	0.07	0.04	**0.02**	0.05	0.07	0.04	**0.02**
	500–45	0.02	**0.01**	0.02	0.02																
	1000–45	**0.01**	0.02	**0.01**	**0.01**																
	2500–45	0.01	0.01	0.01	**0.00**																

A number in bold font is the smallest value under each circumstance.

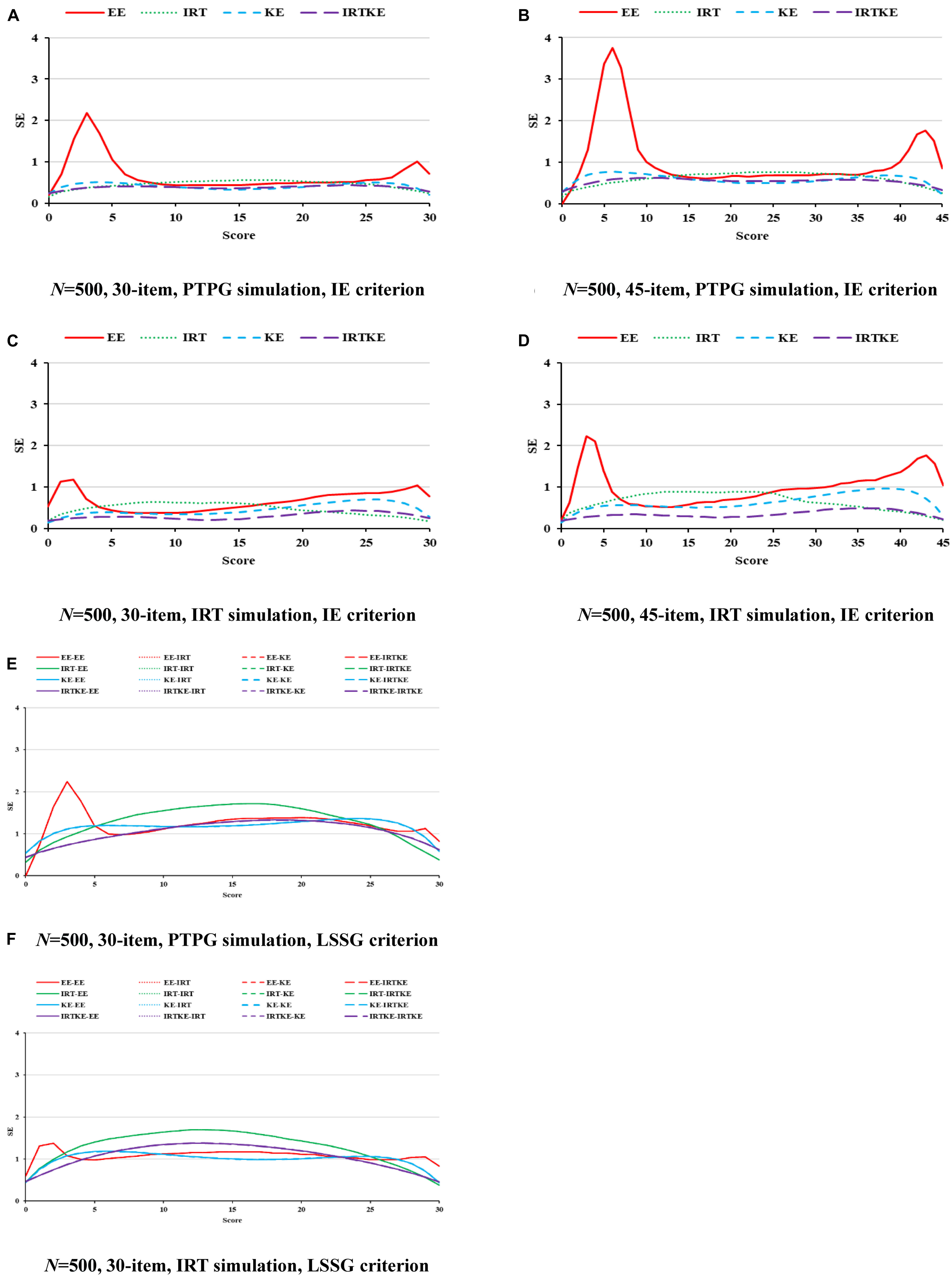
IRTTSE, are not necessary, since traditional simpler EE can give acceptable results. Nonetheless, EE should be used cautiously when equating is performed at extreme scores, where much less records lay. Because sample size plays a similar role under all conditions, and its effect on equating is summarized in **Table 4**, only figures for 500 test takers are shown, with others to be requested from the author for correspondence. Note that the test with 45 items under the LSSG reference condition was not considered here because 90 (45 + 45) items were needed to fulfill the LSSG's philosophy. The ABs change little when sample size and test length increase, usually by approximately 0.01 raw score, hardly affecting practical equating and decision making, according to the rule of Difference That Matter (DTM) (Dorans, 2004). WABs in **Table 4** also describe these trends. Besides, WABs calculated from same true equating are smaller than those from different ones. However, the difference between them is ignorable and insignificant. Results for the PTPG and IRT simulation methods coincide with each other to a high extent in regard to WABs. To sum up, equating methods perform alike in EG according to ABs and WABs.

As for the SEs in **Figure 2**, according to its formula, the same equating method from different true equating functions share identical SE values in the LSSG. Therefore, four lines could be detected, but 16 lines actually exist in **Figures 2E,F**. The IRTKE and KE are most stable, followed by IRTOSE, and finally EE, across whole scores under PTPG simulation circumstance. When the IRT simulation method is used, IRTKE performs better than the others based on the IE criterion, whereas KE prevails based on the LSSG criterion. Again, EE fluctuates more than the others, and two similar peaks in **Figure 1** appear again. In contrast to ABs, SEs are much larger, meaning that random error accounts more equating variabilities than systematic error does in EG. In addition, random error decreases when sample size becomes larger. A shorter test ensures lower SEs. However, those two trends caused by the change in sample size and test length are not significant. All trends mentioned above are quantified in **Table 5**.

Finally presented are the RMSEs and their weighted versions. Since trends are similar in the illustration of ABs and SEs, and RMSEs are formed by aggregating those two together, it is easy to comprehend this. Under the PTPG condition, the KE and IRTKE are spotted as the lowest total errors, whereas under IRT simulation condition, things get different. The IRTKE performs best with the IE reference, but the KE prevails when the LSSG is set as a reference. The EE behaves poorly when scores are very low or high in **Figure 3**. RMSEs get smaller as the sample size increases, and the test length decreases, whose changes are less than the DTM guideline. Furthermore, index values calculated from the IE reference are much lower than those from the LSSG reference. However, the criterion equating deviation is not spotted because the SEs overweigh the ABs overwhelmingly, and the former cannot show any more information. More details are shown in **Figure 3** and **Table 6**.

NEAT

When it comes to NEAT, things get different. In **Figures 4–6**, ABs, SEs, and RMSEs are much larger than those in EG, indicating that equating results in EG are more accurate and stable in this simulation study. In detail, for ABs in **Figure 4**, IRTOSE is the most accurate method, and the difference between it and the others is extremely large, meaning that when sample specifications, such as ability and score distribution, are not equivalent, IRTOSE does an excellent job, benefiting from its robustness to sample misspecification. Besides one peak, every plot has a valley near the high-score range. As shown in **Table 7**, WABs increase a lot when the test becomes longer, but show little improvement when the sample size changes. ABs from IRT simulation are larger than those from the PTPG simulation results; however, this trend is reversed when it comes to IRTOSE. Explicitly, WABs for IRTOSE from the IRT simulation are



N=500, 30-item, PTPG simulation, IE criterion

N=500, 45-item, PTPG simulation, IE criterion

N=500, 30-item, IRT simulation, IE criterion

N=500, 45-item, IRT simulation, IE criterion

N=500, 30-item, PTPG simulation, LSSG criterion

N=500, 30-item, IRT simulation, LSSG criterion

FIGURE 2 | SE in EG. PTPG, Pseudo-Tests and Pseudo-Groups method; IRT, Item Response Theory method; IE, Identity Equating; LSSG, Large Sample Single Group; EE, Equipercentile Equating; IRT, IRT observed score equating; KE, Kernel Equating; IRTKE, IRT observed score Kernel Equating. In **(A–D)**, Red, green, blue, and purple lines represent results of EE, IRT, KE, and IRTKE respectively, calculated by IE criterion. In **(E,F)**, Red, green, blue, and purple lines represent results of EE, IRT, KE, and IRTKE respectively; continuous, dotted, short-dashed, and long-dashed lines represent results calculated by EE, IRT, KE, and IRTKE criterion respectively. Test with 45 items under LSSG reference condition was not considered.

TABLE 5 | Weighted standard error of equating (WSE) in EG.

						LSSG															
		IE				EE				IRT				KE				IRTKE			
		EE	IRT	KE	IRTKE	EE	IRT	KE	IRTKE	EE	IRT	KE	IRTKE	EE	IRT	KE	IRTKE	EE	IRT	KE	IRTKE
PTPG	500–30	0.49	0.52	**0.39**	**0.39**	1.27	1.56	**1.23**	**1.23**	1.27	1.56	**1.23**	**1.23**	1.27	1.56	**1.23**	**1.23**	1.27	1.56	**1.23**	**1.23**
	1000–30	0.34	0.37	**0.27**	0.28	1.22	1.52	**1.19**	**1.19**	1.22	1.52	**1.19**	**1.19**	1.22	1.52	**1.19**	**1.19**	1.22	1.52	**1.19**	**1.19**
	2500–30	0.22	0.24	**0.18**	**0.18**	1.19	1.49	**1.18**	**1.18**	1.19	1.49	**1.18**	**1.18**	1.19	1.49	**1.18**	**1.18**	1.19	1.49	**1.18**	**1.18**
	500–45	0.71	0.71	**0.55**	0.56																
	1000–45	0.50	0.51	**0.39**	**0.39**																
	2500–45	0.32	0.32	**0.25**	**0.25**																
IRT	500–30	0.52	0.57	0.42	**0.27**	1.11	1.54	**1.07**	1.25	1.11	1.54	**1.07**	1.25	1.11	1.54	**1.07**	1.25	1.11	1.54	**1.07**	1.25
	1000–30	0.37	0.39	0.29	**0.19**	1.07	1.52	**1.05**	1.26	1.07	1.52	**1.05**	1.26	1.07	1.52	**1.05**	1.26	1.07	1.52	**1.05**	1.26
	2500–30	0.23	0.25	0.19	**0.12**	**1.03**	1.48	**1.03**	1.24	**1.03**	1.48	**1.03**	1.24	**1.03**	1.48	**1.03**	1.24	**1.03**	1.48	**1.03**	1.24
	500–45	0.76	0.81	0.61	**0.34**																
	1000–45	0.53	0.57	0.42	**0.23**																
	2500–45	0.34	0.37	0.28	**0.15**																

A number in bold font is the smallest value under each circumstance.

smaller than those from the PTPG simulation. In terms of criterion equating, IE tells us that IRTOSE is the best-performed method. However, the LSSG shows some vague opinions because the results are related to which equating function is used as true equating. For example, when the EE is chosen as the true equating, EE performs better than it does under other true equating conditions. This phenomenon is more evident in the PTPG simulation.

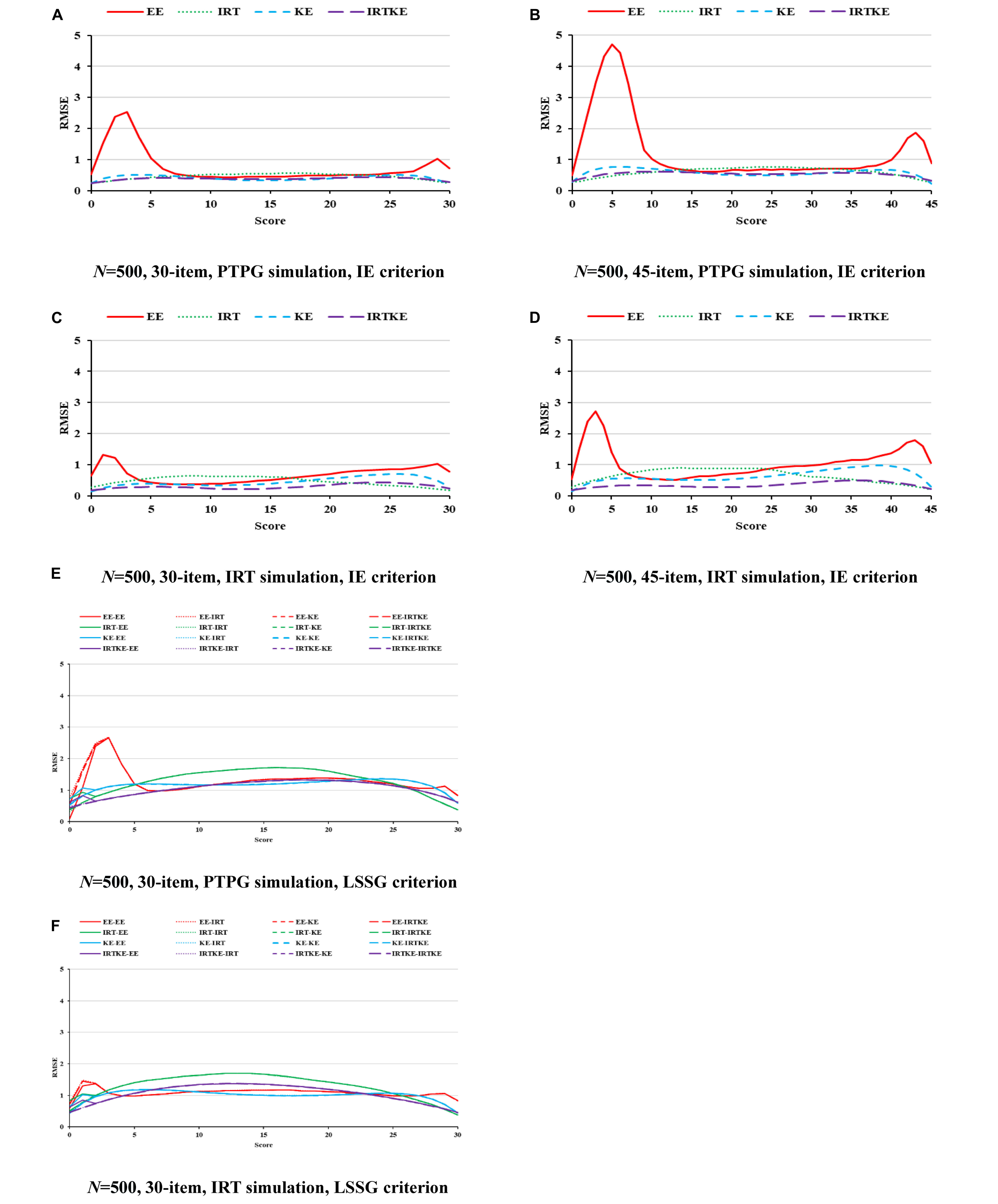
For the SEs in **Figure 5**, the IRTOSE, IRTKE, and KE are more stable than the EE, with the latter one showing two peaks. However, in the mid-score range where score frequencies are larger, all the equating methods resemble more. Another phenomenon worth mentioning is that the SEs for EE get close to 0 in the low and some high-score ranges (**Figure 5**, plots except A and E), attributing to the logic of EE transformation that scores with the same percentile rank are equivalent, even though the two samples are different in score distribution distinctly. So, it is not so much stable as inaccurate. The SEs become smaller when the sample size increases, and the test length decreases in **Table 8**. Again, only the test length contributes significantly to the SE change. The IRT data simulation favors the IRTOSE obviously as is the same case with the ABs. In short, the IRTKE and KE, especially the former one, are more stable than the others under IE reference condition, whereas the IRTOSE is more stable under the LSSG reference condition.

By illustrating the RMSEs and WRMSEs in **Figure 6** and **Table 9**, respectively, the IRTOSE is the best choice for equating in NEAT according to its least amount of total error, followed by KE and EE, the latter of which shows high peaks. The IRTKE leads to larger WRMSEs under most circumstances. In addition, the RMSEs become smaller when the sample size increases, and the test length decreases, but the changes are not significant according to the DTM rule. Again, except for the IRTOSE results, the others from the PTPG simulation are approximately 0.5 point higher than those from the IRT simulation. No clear difference is found between the IE and LSSG.

SUMMARY AND DISCUSSION

IRTKE and Other Equating Methods

IRTKE is a new method integrating the IRTOSE into the KE, taking advantage of the flexible and accurate IRT models fitted to the testing data (Andersson and Wiberg, 2017). Results show that the IRTKE and KE produce less random error and total error than other methods in most situations investigated in the EG, whereas in NEAT, the IRTOSE is superior to others in terms of equating errors, with the exception of random errors calculated with the IE reference. Since the IRTKE is a combination of the IRTOSE and KE, it is still surprising that the IRTOSE wins over the IRTKE by every index when abilities differ a lot in NEAT. We speculate that the IRTKE is rather a modification of the KE compared to that of the IRTOSE, which is proven by the result that the IRTKE and KE show more similarities. In addition, the IRTKE embraces more basic elements from the KE, such as continuization and equating, although it calculates score probabilities based on the IRT models. It is also found that the IRTOSE is proven to be a good choice when the sample size is large (more than 500 cases), which is considered to be a rough threshold where the IRT model fitting and parameter estimation can successfully converge (Hambleton and Jones, 1993; Kolen and Brennan, 2014). In general, increasing the sample size leads to lower total errors (represented by the RMSEs and WRMSEs in this study), but the accuracy improvements are not large enough to make a difference in equating practices, which contradicts former studies (Moses and Holland, 2007; Liang and von Davier, 2014). For example, the levels of the sample size manipulated were 200 and 2000, and 100, 200, and 1,000 in the Liang and von Davier study and the Moses and Holland study, respectively. Therefore, we have confidence in speculating that a larger sample size used in this study led to the stability of equating errors as it changes. Small sample conditions, such as the 200 and 500 cases, should be investigated in the future to explore the equating methods'



N=500, 30-item, PTPG simulation, IE criterion

N=500, 45-item, PTPG simulation, IE criterion

N=500, 30-item, IRT simulation, IE criterion

N=500, 45-item, IRT simulation, IE criterion

N=500, 30-item, PTPG simulation, LSSG criterion

N=500, 30-item, IRT simulation, LSSG criterion

FIGURE 3 | RMSE in EG. PTPG, Pseudo-Tests and Pseudo-Groups method; IRT, Item Response Theory method; IE, Identity Equating; LSSG, Large Sample Single Group; EE, Equipercentile Equating; IRT, IRT observed score equating; KE, Kernel Equating; IRTKE, IRT observed score Kernel Equating. In **(A–D)**, Red, green, blue, and purple lines represent results of EE, IRT, KE, and IRTKE respectively, calculated by IE criterion. In **(E,F)**, Red, green, blue, and purple lines represent results of EE, IRT, KE, and IRTKE respectively; continuous, dotted, short-dashed, and long-dashed lines represent results calculated by EE, IRT, KE, and IRTKE criterion respectively. Test with 45 items under LSSG reference condition was not considered.

TABLE 6 | Weighted root mean squared error (WRMSE) in EG.

										LSSG											
			IE				EE				IRT				KE				IRTKE		
		EE	**IRT**	**KE**	**IRTKE**	**EE**	**IRT**	**KE**	**IRTKE**	**EE**	**IRT**	**KE**	**IRTKE**	**EE**	**IRT**	**KE**	**IRTKE**	**EE**	**IRT**	**KE**	**IRTKE**
PTPG	500–30	0.49	0.52	**0.39**	**0.39**	1.27	1.56	**1.23**	**1.23**	1.27	1.56	**1.23**	**1.23**	1.27	1.56	**1.23**	**1.23**	1.27	1.56	**1.23**	**1.23**
	1000–30	0.34	0.37	**0.27**	0.28	1.22	1.52	**1.19**	**1.19**	1.22	1.52	**1.19**	**1.19**	1.22	1.52	**1.19**	**1.19**	1.22	1.52	**1.19**	**1.19**
	2500–30	0.22	0.24	**0.18**	**0.18**	1.19	1.50	**1.18**	**1.18**	1.19	1.49	**1.18**	**1.18**	1.19	1.50	**1.18**	**1.18**	1.19	1.50	**1.18**	**1.18**
	500–45	0.71	0.72	**0.55**	0.56																
	1000–45	0.50	0.51	**0.39**	**0.39**																
	2500–45	0.32	0.32	**0.25**	0.26																
IRT	500–30	0.52	0.57	0.42	**0.27**	1.11	1.54	**1.07**	1.25	1.11	1.54	**1.07**	1.26	1.11	1.54	**1.07**	1.25	1.11	1.54	**1.07**	1.25
	1000–30	0.37	0.39	0.29	**0.19**	1.07	1.52	**1.05**	1.26	1.07	1.52	**1.05**	1.26	1.07	1.52	**1.05**	1.26	1.07	1.52	**1.05**	1.26
	2500–30	0.23	0.25	0.19	**0.12**	**1.03**	1.48	**1.03**	1.24	1.04	1.48	**1.03**	1.25	**1.03**	1.48	**1.03**	1.24	**1.03**	1.48	**1.03**	1.24
	500–45	0.76	0.81	0.61	**0.34**																
	1000–45	0.53	0.57	0.42	**0.23**																
	2500–45	0.35	0.37	0.28	**0.15**																

A number in bold font is the smallest value under each circumstance.

performances under extreme conditions, though it may cause convergence problems. Another inconsistent phenomenon is that equating errors get larger when test forms are lengthened (Fitzpatrick and Yen, 2001; Godfrey, 2007; Norman Dvorak, 2009). Kim et al. (2017) investigated the performance of four approaches to handling structural zeros in NEAT equating where test length, proportion of common items, examinee ability effect size, and sample size were manipulated. Consistent with this study, they also found that evaluation statistics were smaller for shorter tests than for longer ones. They speculated that since the IRTOSE employed smoothed distributions using explicitly specified distributions of ability in the population of examinees, it gave an advantage to shorter tests. That is, with other conditions fixed, observed relative frequency distributions for simulated data sets became smoother for shorter test lengths and, thus, closer to the population relative frequency distributions. Besides, we infer that when other factors are fixed, the number of items allocated to a single score point decreases, thus, making the equating error increase (Akour, 2006). What is more, the percentage of the anchor items might affect the equating results, which was fixed at 30% in this study. In addition, the other extreme ratios of the anchor items to the total items are worth exploring. Nowadays, large-scale assessments containing far more than 50 items are usual, such as PISA, TIMSS, and so on. Nevertheless, limited to the 80-item ADM verbal test used, a long-test situation was not manipulated in this study, which could be considered to verify equating performances in the future.

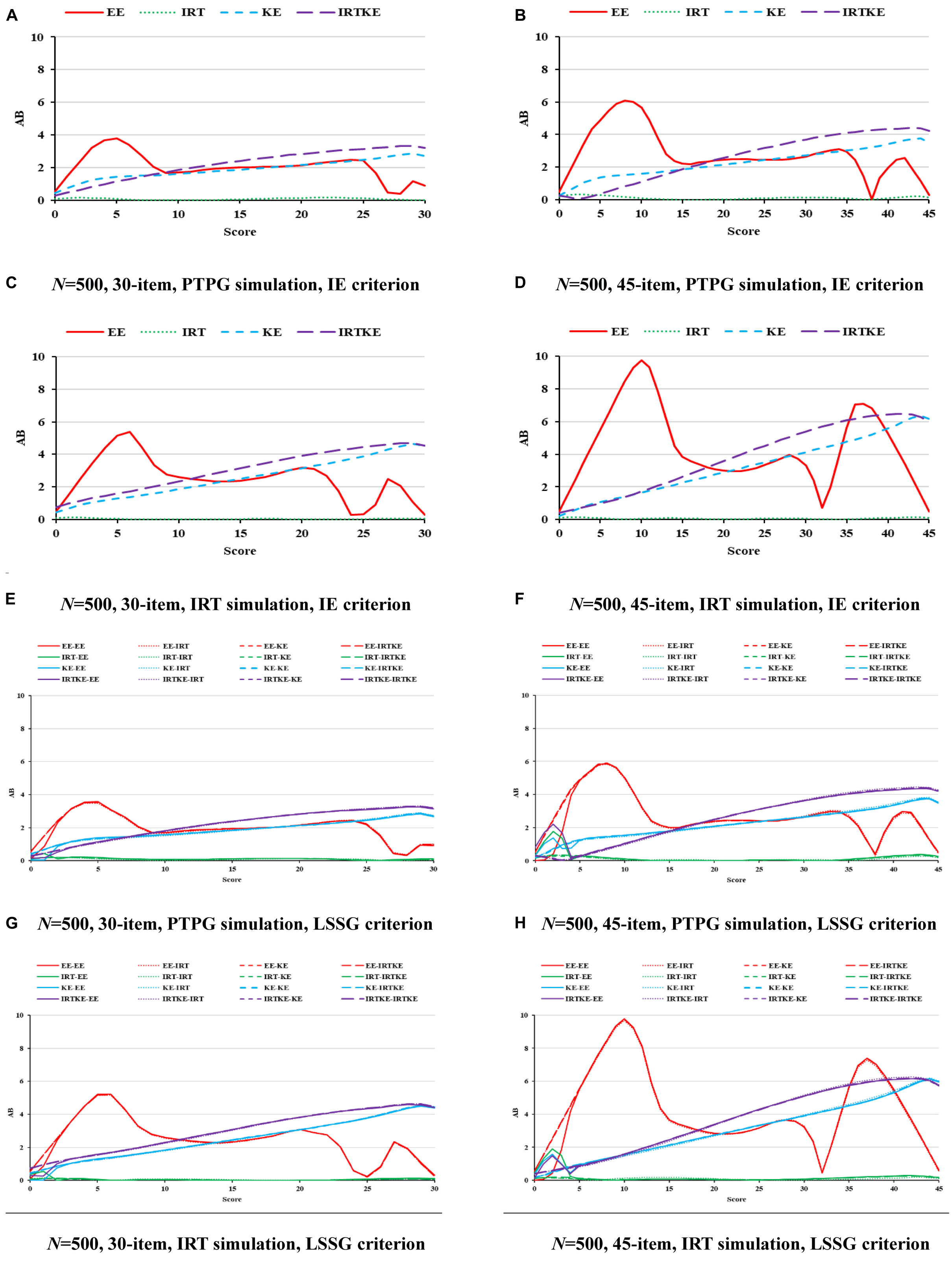
Data Simulation Preference

The phenomenon that data obtained from the IRT simulation favors the IRTOSE in NEAT is a signal of simulation method preference. Nevertheless, it is a relief that the spotted IRT preference does not affect the final comparative results among the equating methods because no matter which true equating is selected, the IRTOSE is the best performed, followed by the EE, KE, and IRTKE, which are also indicators of robustness of the IRT equating methods (Skaggs and Lissitz, 1986; Béguin, 2000; Kim and Kolen, 2006). The mechanism behind might be that the simulation methods make pseudo test score distributions different with each other, and thus, equating performances are not coincident. However, the IRT preference was not spotted in EG. We speculate that the idealized sample equivalence controlled by randomly selecting cases in EG made it happen. More researches could be conducted on the testing simulation method preference in EG when equivalence assumption is violated. It also alerts that more caution and proofs validating equating performance are required before making conclusions based on one single-simulation study, which is usually ignored. Further studies could be carried out on finding other fairer simulation procedures for equating method comparison. That content specifications were not controlled in test forms is another limitation in this study, which could be improved by taking the test content into consideration when pseudo tests are constructed.

Criterion Equating Preference

In order to investigate whether criterion equating plays a different role in equating evaluation or not, four equating methods (EE, IRT, KE, and IRTKE) and IE were chosen as true equatings. Following this logic, it was found that WABs favor equating results using the same true equating functions in EG. WSEs and WRMSEs do not show this preference. Because WSEs are identical under the same true equating, and random errors (SEs and WSEs) contribute more than systematic errors (ABs and WABs) to total errors (RMSEs and WRMSEs) in this study, it is not surprising that no clear criterion equating preference is found for WRMSE.

Based on simulation and the discussion above, several recommendations are summarized. First, when equating is conducted in EG, and the requirement of the ability equivalence between the two groups could be satisfied well, the IRTKE is strongly recommended owing to its much less random error caused by sampling. However, when equating groups show clear ability difference in NEAT, the IRTOSE might be a



N=500, 30-item, PTPG simulation, IE criterion

N=500, 45-item, PTPG simulation, IE criterion

N=500, 30-item, IRT simulation, IE criterion

N=500, 45-item, IRT simulation, IE criterion

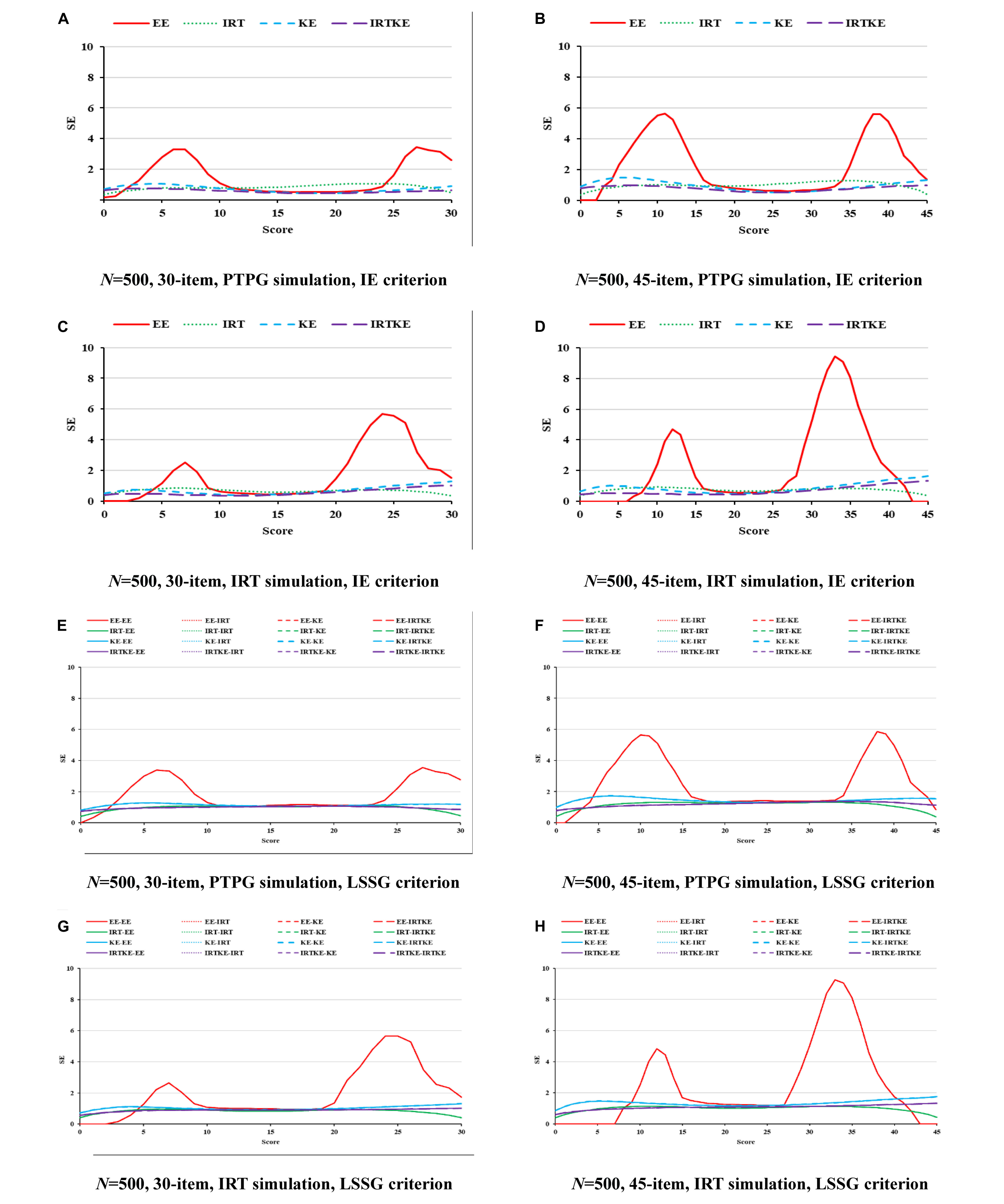
N=500, 30-item, PTPG simulation, LSSG criterion

N=500, 45-item, PTPG simulation, LSSG criterion

N=500, 30-item, IRT simulation, LSSG criterion

N=500, 45-item, IRT simulation, LSSG criterion

FIGURE 4 | AB in NEAT. PTPG, Pseudo-Tests and Pseudo-Groups method; IRT, Item Response Theory method; IE, Identity Equating; LSSG, Large Sample Single Group; EE, Equipercentile Equating; IRT, IRT observed score equating; KE, Kernel Equating; IRTKE, IRT observed score Kernel Equating. In **(A–D)**, Red, green, blue, and purple lines represent results of EE, IRT, KE, and IRTKE respectively, calculated by IE criterion. In **(E–H)**, Red, green, blue, and purple lines represent results of EE, IRT, KE, and IRTKE respectively; continuous, dotted, short-dashed, and long-dashed lines represent results calculated by EE, IRT, KE, and IRTKE criterion respectively.



N=500, 30-item, PTPG simulation, IE criterion

N=500, 45-item, PTPG simulation, IE criterion

N=500, 30-item, IRT simulation, IE criterion

N=500, 45-item, IRT simulation, IE criterion

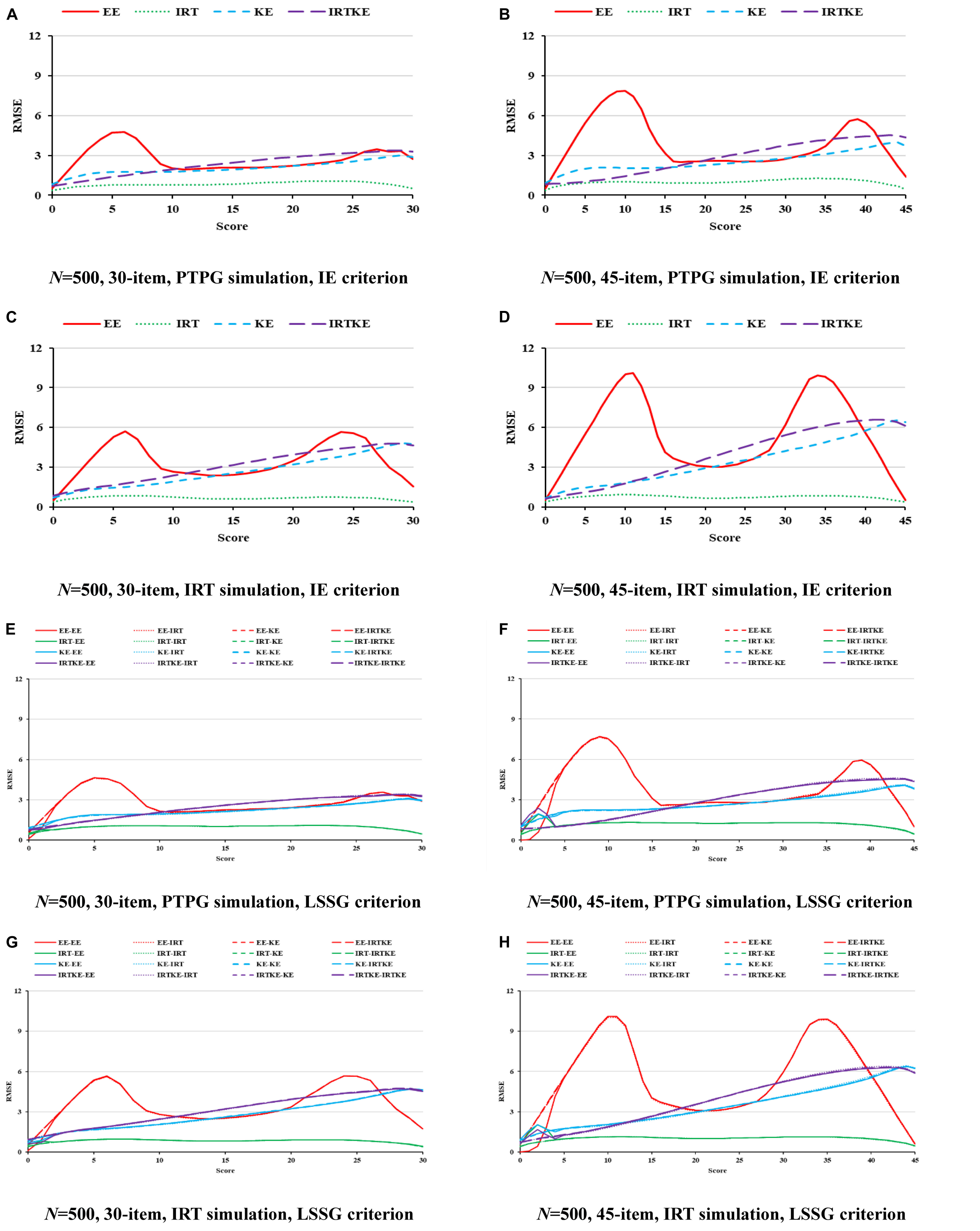
N=500, 30-item, PTPG simulation, LSSG criterion

N=500, 45-item, PTPG simulation, LSSG criterion

N=500, 30-item, IRT simulation, LSSG criterion

N=500, 45-item, IRT simulation, LSSG criterion

FIGURE 5 | SE in NEAT. PTPG, Pseudo-Tests and Pseudo-Groups method; IRT, Item Response Theory method; IE, Identity Equating; LSSG, Large Sample Single Group; EE, Equipercentile Equating; IRT, IRT observed score equating; KE, Kernel Equating; IRTKE, IRT observed score Kernel Equating. In **(A–D)**, Red, green, blue, and purple lines represent results of EE, IRT, KE, and IRTKE respectively, calculated by IE criterion. In **(E–H)**, Red, green, blue, and purple lines represent results of EE, IRT, KE, and IRTKE respectively; continuous, dotted, short-dashed, and long-dashed lines represent results calculated by EE, IRT, KE, and IRTKE criterion respectively.



N=500, 30-item, PTPG simulation, IE criterion

N=500, 45-item, PTPG simulation, IE criterion

N=500, 30-item, IRT simulation, IE criterion

N=500, 45-item, IRT simulation, IE criterion

N=500, 30-item, PTPG simulation, LSSG criterion

N=500, 45-item, PTPG simulation, LSSG criterion

N=500, 30-item, IRT simulation, LSSG criterion

N=500, 45-item, IRT simulation, LSSG criterion

FIGURE 6 | RMSE in NEAT. PTPG, Pseudo-Tests and Pseudo-Groups method; IRT, Item Response Theory method; IE, Identity Equating; LSSG, Large Sample Single Group; EE, Equipercentile Equating; IRT, IRT observed score equating; KE, Kernel Equating; IRTKE, IRT observed score Kernel Equating. In **(A–D)**, Red, green, blue, and purple lines represent results of EE, IRT, KE, and IRTKE respectively, calculated by IE criterion. In **(E–H)**, Red, green, blue, and purple lines represent results of EE, IRT, KE, and IRTKE respectively; continuous, dotted, short-dashed, and long-dashed lines represent results calculated by EE, IRT, KE, and IRTKE criterion respectively.

TABLE 7 | Weighted absolute bias (WAB) in NEAT.

						LSSG															
		IE				**EE**				**IRT**				**KE**				**IRTKE**			
		EE	**IRT**	**KE**	**IRTKE**	**EE**	**IRT**	**KE**	**IRTKE**	**EE**	**IRT**	**KE**	**IRTKE**	**EE**	**IRT**	**KE**	**IRTKE**	**EE**	**IRT**	**KE**	**IRTKE**
PTPG	500–30	2.08	**0.11**	2.15	2.79	2.05	**0.10**	2.13	2.76	2.06	**0.10**	2.13	2.76	2.06	**0.11**	2.13	2.76	2.06	**0.10**	2.13	2.76
	1000–30	2.16	**0.12**	2.15	2.78	2.12	**0.09**	2.13	2.75	2.13	**0.09**	2.13	2.76	2.13	**0.09**	2.13	2.76	2.13	**0.09**	2.13	2.76
	2500–30	2.19	**0.10**	2.15	2.77	2.17	**0.10**	2.12	2.74	2.17	**0.11**	2.13	2.75	2.17	**0.11**	2.13	2.75	2.17	**0.11**	2.13	2.75
	500–45	2.50	**0.10**	2.72	3.59	2.46	**0.06**	2.64	3.53	2.48	**0.06**	2.69	3.57	2.46	**0.06**	2.65	3.53	2.45	**0.05**	2.64	3.53
	1000–45	2.58	**0.12**	2.71	3.56	2.53	**0.06**	2.65	3.48	2.57	**0.04**	2.69	3.53	2.54	**0.06**	2.65	3.48	2.53	**0.05**	2.65	3.48
	2500–45	2.65	**0.12**	2.71	3.54	2.61	**0.05**	2.64	3.46	2.65	**0.04**	2.69	3.51	2.62	**0.04**	2.65	3.47	2.61	**0.03**	2.65	3.47
IRT	500–30	2.43	**0.03**	2.88	3.53	2.39	**0.02**	2.82	3.45	2.36	**0.04**	2.79	3.43	2.39	**0.02**	2.82	3.46	2.38	**0.02**	2.82	3.46
	1000–30	2.57	**0.03**	2.88	3.53	2.53	**0.03**	2.79	3.45	2.50	**0.05**	2.77	3.43	2.53	**0.02**	2.80	3.45	2.53	**0.03**	2.79	3.45
	2500–30	2.76	**0.02**	2.88	3.54	2.69	**0.03**	2.81	3.47	2.67	**0.05**	2.79	3.45	2.69	**0.02**	2.81	3.48	2.69	**0.03**	2.81	3.47
	500–45	3.50	**0.04**	3.72	4.72	3.39	**0.08**	3.53	4.47	3.37	**0.06**	3.55	4.49	3.39	**0.07**	3.53	4.47	3.39	**0.08**	3.53	4.47
	1000–45	3.42	**0.04**	3.72	4.69	3.32	**0.08**	3.56	4.49	3.30	**0.07**	3.57	4.50	3.32	**0.08**	3.56	4.49	3.31	**0.08**	3.55	4.49
	2500–45	3.52	**0.04**	3.75	4.68	3.41	**0.10**	3.57	4.50	3.40	**0.09**	3.58	4.51	3.41	**0.10**	3.57	4.50	3.40	**0.11**	3.57	4.50

A number in bold font is the smallest value under each circumstance.

TABLE 8 | Weighted standard error of equating (WSE) in NEAT.

						LSSG															
		IE				**EE**				**IRT**				**KE**				**IRTKE**			
		EE	**IRT**	**KE**	**IRTKE**	**EE**	**IRT**	**KE**	**IRTKE**	**EE**	**IRT**	**KE**	**IRTKE**	**EE**	**IRT**	**KE**	**IRTKE**	**EE**	**IRT**	**KE**	**IRTKE**
PTPG	500–30	0.84	0.97	0.52	**0.48**	1.38	1.05	1.11	**1.04**	1.38	1.05	1.11	**1.04**	1.38	1.05	1.11	**1.04**	1.38	1.05	1.11	**1.04**
	1000–30	0.70	0.90	0.49	**0.44**	1.24	**0.95**	1.09	1.04	1.24	**0.95**	1.09	1.04	1.24	**0.95**	1.09	1.04	1.24	**0.95**	1.09	1.04
	2500–30	0.53	0.86	0.46	**0.42**	1.13	**0.92**	1.08	1.02	1.13	**0.92**	1.08	1.02	1.13	**0.92**	1.08	1.02	1.13	**0.92**	1.08	1.02
	500–45	1.43	1.13	0.66	**0.63**	2.04	**1.26**	1.37	1.32	2.04	**1.26**	1.37	1.32	2.04	**1.26**	1.37	1.32	2.04	**1.26**	1.37	1.32
	1000–45	1.09	1.07	0.61	**0.57**	1.70	**1.20**	1.33	1.28	1.70	**1.20**	1.33	1.28	1.70	**1.20**	1.33	1.28	1.70	**1.20**	1.33	1.28
	2500–45	0.80	1.03	0.57	**0.53**	1.49	**1.15**	1.32	1.27	1.49	**1.15**	1.32	1.27	1.49	**1.15**	1.32	1.27	1.49	**1.15**	1.32	1.27
IRT	500–30	1.55	0.66	0.61	**0.53**	1.87	**0.86**	0.98	0.91	1.87	**0.86**	0.98	0.91	1.87	**0.86**	0.98	0.91	1.87	**0.86**	0.98	0.91
	1000–30	1.23	0.53	0.57	**0.51**	1.55	**0.75**	0.97	0.91	1.55	**0.75**	0.97	0.91	1.55	**0.75**	0.97	0.91	1.55	**0.75**	0.97	0.91
	2500–30	0.98	**0.43**	0.54	0.48	1.35	**0.71**	0.96	0.93	1.35	**0.71**	0.96	0.93	1.35	**0.71**	0.96	0.93	1.35	**0.71**	0.96	0.93
	500–45	2.67	0.74	0.74	**0.64**	2.99	**1.05**	1.26	1.12	2.99	**1.05**	1.26	1.12	2.99	**1.05**	1.26	1.12	2.99	**1.05**	1.26	1.12
	1000–45	2.24	**0.62**	0.71	**0.62**	2.55	**0.95**	1.22	1.08	2.55	**0.95**	1.22	1.08	2.55	**0.95**	1.22	1.08	2.55	**0.95**	1.22	1.08
	2500–45	1.73	**0.49**	0.67	0.55	2.14	**0.88**	1.20	1.08	2.14	**0.88**	1.20	1.08	2.14	**0.88**	1.20	1.08	2.14	**0.88**	1.20	1.08

A number in bold font is the smallest value under each circumstance.

TABLE 9 | Weighted root mean squared error (WRMSE) in NEAT.

						LSSG															
		IE				**EE**				**IRT**				**KE**				**IRTKE**			
		EE	**IRT**	**KE**	**IRTKE**	**EE**	**IRT**	**KE**	**IRTKE**	**EE**	**IRT**	**KE**	**IRTKE**	**EE**	**IRT**	**KE**	**IRTKE**	**EE**	**IRT**	**KE**	**IRTKE**
PTPG	500–30	2.36	**0.97**	2.22	2.83	2.55	**1.05**	2.40	2.95	2.55	**1.05**	2.41	2.95	2.55	**1.05**	2.41	2.95	2.55	**1.05**	2.41	2.95
	1000–30	2.31	**0.91**	2.21	2.82	2.49	**0.95**	2.39	2.95	2.49	**0.95**	2.40	2.95	2.49	**0.95**	2.40	2.95	2.49	**0.95**	2.40	2.95
	2500–30	2.27	**0.87**	2.20	2.80	2.45	**0.92**	2.38	2.93	2.45	**0.93**	2.39	2.93	2.45	**0.93**	2.39	2.93	2.46	**0.93**	2.39	2.93
	500–45	3.16	**1.14**	2.80	3.65	3.37	**1.27**	2.98	3.77	3.39	**1.26**	3.02	3.81	3.37	**1.27**	2.98	3.77	3.36	**1.27**	2.98	3.77
	1000–45	3.00	**1.07**	2.78	3.60	3.17	**1.21**	2.96	3.71	3.20	**1.21**	3.00	3.75	3.17	**1.21**	2.97	3.71	3.17	**1.21**	2.97	3.71
	2500–45	2.88	**1.04**	2.77	3.58	3.08	**1.16**	2.96	3.69	3.11	**1.15**	3.00	3.73	3.08	**1.15**	2.96	3.69	3.08	**1.15**	2.96	3.69
IRT	500–30	3.25	**0.66**	2.95	3.57	3.30	**0.86**	2.99	3.58	3.27	**0.86**	2.97	3.56	3.30	**0.86**	2.99	3.58	3.30	**0.86**	2.99	3.58
	1000–30	3.14	**0.53**	2.94	3.57	3.19	**0.75**	2.96	3.57	3.16	**0.76**	2.94	3.55	3.19	**0.75**	2.96	3.57	3.18	**0.75**	2.96	3.57
	2500–30	3.11	**0.43**	2.93	3.57	3.15	**0.71**	2.97	3.60	3.13	**0.71**	2.95	3.58	3.15	**0.71**	2.98	3.60	3.15	**0.71**	2.97	3.60
	500–45	4.91	**0.75**	3.80	4.76	4.90	**1.06**	3.76	4.62	4.87	**1.06**	3.77	4.63	4.90	**1.06**	3.76	4.62	4.89	**1.06**	3.75	4.61
	1000–45	4.59	**0.62**	3.79	4.73	4.57	**0.96**	3.76	4.62	4.55	**0.95**	3.78	4.64	4.57	**0.96**	3.77	4.62	4.56	**0.96**	3.76	4.62
	2500–45	4.37	**0.49**	3.81	4.71	4.34	**0.89**	3.77	4.63	4.33	**0.89**	3.78	4.65	4.34	**0.89**	3.77	4.63	4.33	**0.89**	3.77	4.63

A number in bold font is the smallest value under each circumstance.

wise choice because it relates to far less systematic error than the other methods. Second, in the view of data simulation preference, the PTPG is suitable for comparative studies of test equating, especially for those including methods under distinct theoretical backgrounds. In contrast, researchers should be alert and cautious about the conclusions when comparing the IRT and the other equating methods based on the IRT simulation. Similar recommendations are made on the selection of criterion equating. The final conclusion about equating study and its further application must be based on solid proofs and comprehensive and unbiased criteria, which cannot be overemphasized.

Further researches could focus on several topics. First, for simplicity, only dichotomous items were considered in this study. However, polytomous and mixed-format ones could detect and evaluate more sophisticated and higher-level abilities in educational tests. Therefore, equating results under these conditions should be tested. Second, considering that two or more items with identical contents and psychometric specifications would be unrealistic in practical tests, items were drawn without replacement in this study, as were students (or respondent cases). Since drawing with replacement is also one usual option in data simulation, future research could try it. Third, note that raw score distributions used in this study are close to normal distribution, and equating performances under other distributions, such as binomial distribution and χ^2 distribution should also be considered. On the other hand, besides raw data, when simulated pseudo tests are not conformed to normal distribution, how well would equating methods perform? In addition, the effect of the IRT data-model misfit on equating is also worthy of investigation. Finally, besides multiple-choice question, various types of items exist, such as constructed response, fill-in-the-blank, and matching questions. So, equating comparison with mixed-format tests is also a realistic topic to discuss.

DATA AVAILABILITY STATEMENT

The raw datasets analyzed for this study can be found in the homepage of Associate Professor Jorge González Burgos, http://www.mat.uc.cl/~jorge.gonzalez/index_archivos/EquatingRbook.htm.

AUTHOR CONTRIBUTIONS

SW and MZ designed the study. SW processed the data and wrote the manuscript. MZ and SY guided the data processing and manuscript writing. All authors revised the manuscript, read and approved the final manuscript.

FUNDING

This study was funded by the National Social Science Foundation of China (Grant No. BHA180141).

ACKNOWLEDGMENTS

The authors wish to thank Dr. Björn Andersson for his help with some R codes, and Prof. Jorge González and Marie Wiberg for permission of using raw datasets.

REFERENCES

Akour, M. M. M. (2006). *A Comparison of Various Equipercentile and Kernel Equating Methods Under the Random Groups Design*. Ph.D. dissertation, University of Iowa, Iowa City.

Albano, A. D. (2016). Equate: an R package for observed-score linking and equating. *J. Stat. Softw.* 74, 1–36. doi: 10.18637/jss.v074.i08

Andersson, B. (2016). Asymptotic standard errors of observed-score equating with polytomous IRT Models. *J. Educ Meas.* 53, 459–477. doi: 10.1111/jedm.12126

Andersson, B., Bränberg, K., and Wiberg, M. (2013). Performing the kernel method of test equating with the package kequate. *J. Stat. Softw.* 55, 1–25. doi: 10.18637/jss.v055.i06

Andersson, B., and Wiberg, M. (2017). Item response theory observed-score kernel equating. *Psychometrika* 82, 48–66. doi: 10.1007/s11336-016-9528-7

Arıkan, ÇA., and Gelbal, S. (2018). A comparison of traditional and kernel equating methods. *Int. J. Assess. Tools Educ.* 5, 417–427. doi: 10.21449/ijate.409826

Battauz, M. (2015). equateIRT: an R package for IRT test equating. *J. Stat. Softw.* 68, 1–22. doi: 10.18637/jss.v068.i07

Béguin, A. A. (2000). *Robustness of Equating High-Stakes Tests*. Ph.D. dissertation, University of Twente, Enschede.

Chalmers, R. P. (2012). mirt: a multidimensional item response theory package for the R environment. *J. Stat. Softw.* 48, 1–29. doi: 10.18637/jss.v048.i06

Chen, H. (2012). A comparison between linear IRT observed-score equating and levine observed-score equating under the generalized kernel equating framework. *J. Educ. Meas.* 49, 269–284. doi: 10.1111/j.1745-3984.2012.00175.x

Choi, S. I. (2009). *A Comparison of Kernel Equating and Traditional Equipercentile Equating Methods and the Parametric Bootstrap Methods for Estimating Standard Errors in Equipercentile Equating*. Ph.D. dissertation, University of Illinois at Urbana-Champaign, Champaign, IL.

Cook, L. L., and Eignor, D. R. (1991). IRT equating methods. *Educ. Meas. Issues Pract.* 10, 37–45. doi: 10.1111/j.1745-3992.1991.tb00207.x

De Ayala, R. J. (2013). *The Theory and Practice of Item Response Theory*. New York, NY: Guilford Publications.

De Ayala, R. J., Smith, B., and Norman Dvorak, R. (2018). A comparative evaluation of kernel equating and test characteristic curve equating. *Appl. Psychol. Meas.* 42, 155–168. doi: 10.1177/0146621617712245

de la Torre, J., and Hong, Y. (2010). Parameter estimation with small sample size a higher-order IRT model approach. *Appl. Psychol. Meas.* 34, 267–285. doi: 10.1177/0146621608329501

Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *J Educ Meas.* 41, 43–68. doi: 10.1111/j.1745-3984.2004.tb01158.x

Educational Testing Service, (2010). *Linking TOEFL iBT Scores to IELTS Scores: A Research Report*. Princeton, NJ: Educational Testing Service.

Fitzpatrick, A. R., and Yen, W. M. (2001). The effects of test length and sample size on the reliability and equating of tests composed of constructed-response items. *Appl. Meas Educ.* 14, 31–57. doi: 10.1207/S15324818AME1401_04

Godfrey, K. E. (2007). *A Comparison of Kernel Equating and IRT True Score Equating Methods*. Ph.D. dissertation, The University of North Carolina at Greensboro, Greensboro, NC.

González, J., and Wiberg, M. (2017). *Applying Test Equating Methods: Using R*. Cham: Springer.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Jpn. Psychol. Res.* 22, 144–149. doi: 10.4992/psycholres1954.22.144

Hagge, S. L., and Kolen, M. J. (2011). "Equating mixed-format tests with format representative and non-representative common items," in *Mixed-Format Tests: Psychometric Properties with a Primary Focus on Equating (volume 1). (CASMA Monograph Number 2.1)*, eds M. J. Kolen, and W. Lee (Iowa City, IA: CASMA, The University of Iowa), 95–135.

Hagge, S. L., and Kolen, M. J. (2012). "Effects of group differences on equating using operational and pseudo-tests," in *Mixed-Format Tests: Psychometric Properties with a Primary Focus on Equating (volume 2). (CASMA Monograph Number 2.2)*, eds M. J. Kolen, and W. Lee (Iowa City, IA: CASMA, The University of Iowa), 45–86.

Hambleton, R. K., and Jones, R. W. (1993). An NCME instructional module on: comparison of classical test theory and item response theory and their applications to test development. *Educ. Meas. Issues Pract.* 12, 38–47. doi: 10.1111/j.1745-3992.1993.tb00543.x

Han, T., Kolen, M., and Pohlmann, J. (1997). A comparison among IRT true-and observed-score equatings and traditional equipercentile equating. *Appl. Meas. Educ.* 10, 105–121. doi: 10.1207/s15324818ame1002_1

Harris, D. J., and Crouse, J. D. (1993). A study of criteria used in equating. *Appl Meas. Educ.* 6, 195–240. doi: 10.1207/s15324818ame0603_3

Kastberg, D., Roey, S., Ferraro, D., Lemanski, N., and Erberber, E. (2013). *US TIMSS and PIRLS 2011 Technical Report and User's Guide. NCES 2013-046*. Washington, DC: National Center for Education Statistics.

Kim, H. J., Brennan, R. L., and Lee, W. C. (2017). Structural zeros and their implications with log-linear bivariate presmoothing under the internal-anchor design. *J. Educ. Meas.* 54, 145–164. doi: 10.1111/jedm.12138

Kim, H. Y. (2014). *A Comparison of Smoothing Methods for the Common Item Nonequivalent Groups Design*. Ph.D. dissertation, University of Iowa, Iowa city.

Kim, S., and Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Appl. Meas. Educ.* 19, 357–381. doi: 10.1207/s15324818ame1904_7

Kim, S., and Lu, R. (2018). The pseudo-equivalent groups approach as an alternative to common-item equating. *ETS Res. Rep Ser.* 2018, 1–13. doi: 10.1002/ets2.12222

Kim, S., von Davier, A. A., and Haberman, S. (2006). An alternative to equating with small samples in the non-equivalent groups anchor test design. *ETS Res. Rep. Ser.* 2006, 1–40. doi: 10.1002/j.2333-8504.2006.tb02033.x

Kolen, M. J., and Brennan, R. L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices*. Berlin: Springer Science & Business Media.

Leôncio, W., and Wiberg, M. (2017). "Evaluating equating transformations from different frameworks,". in *Proceedings of the Annual Meeting of the Psychometric Society* (Cham: Springer), 101–110. doi: 10.1007/978-3-319-77249-3_9

Liang, T., and von Davier, A. A. (2014). Cross-validation: an alternative bandwidth-selection method in kernel equating. *Appl. Psychol. Meas.* 38, 281–295. doi: 10.1177/0146621613518094

Liu, Y., Schulz, E. M., and Yu, L. (2008). Standard error estimation of 3PL IRT true score equating with an MCMC method. *J. Educ. Behav. Stat.* 33, 257–278. doi: 10.3102/1076998607306076

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *J. Educ. Meas.* 14, 117–138. doi: 10.1111/j.1745-3984.1977.tb00032.x

Loyd, B. H., and Hoover, H. D. (1980). Vertical equating using the Rasch model. *J. Educ. Meas.* 17, 179–193. doi: 10.1111/j.1745-3984.1980.tb00825.x

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *J. Educ. Meas.* 14, 139–160. doi: 10.1111/j.1745-3984.1977.tb00033.x

Moses, T., and Holland, P. (2007). Kernel and traditional equipercentile equating with degrees of presmoothing. *ETS Res. Rep. Ser.* 2007, 1–39. doi: 10.1002/j.2333-8504.2007.tb02057.x

Mun, E. Y., Huo, Y., White, H. R., Suzuki, S., and de la Torre, J. (2019). Multivariate higher-order IRT model and MCMC algorithm for linking individual participant data from multiple studies. *Front. Psychol.* 10:1328. doi: 10.3389/fpsyg.2019.01328

Norman Dvorak, R. L. (2009). *A Comparison of Kernel Equating to the Test Characteristic Curve Method*. Ph.D. dissertation, University of Nebraska, Lincoln, NE.

OECD (2017). *PISA 2015 Technical Report*. Paris: OECD Publishing.

Petersen, N. S., Marco, G. L., and Stewart, E. E. (1982). "A test of the adequacy of linear score equating models," in *Test Equating*, eds P. W. Holland, and D. B. Rubin (New York: Academic Press Inc), 71–135.

Powers, S. J., and Kolen, M. J. (2011). "Evaluating equating accuracy and assumptions for groups that differ in performance," in *Mixed-Format Tests: Psychometric Properties with a Primary Focus on Equating (volume 1). (CASMA Monograph Number 2.1)*, eds M. J. Kolen, and W. Lee (Iowa City, IA: CASMA, The University of Iowa), 137–175.

Powers, S. J., and Kolen, M. J. (2012). "Using matched samples equating methods to improve equating accuracy," in *Mixed-Format Tests: Psychometric Properties with a Primary Focus on Equating (volume 2). (CASMA Monograph Number 2.2)*, eds M. J. Kolen, and W. Lee (Iowa City, IA: CASMA, The University of Iowa), 87–114.

Puhan, G., Moses, T., Grant, M., and McHale, F. (2008). An alternative data collection design for equating with very small samples. *ETS Res. Rep. Ser.* 2008, 1–35. doi: 10.1002/j.2333-8504.2008.tb02097.x

Qu, Y. (2007). *The Effect of Weighting in Kernel Equating Using Counter-balanced Designs*. Ph.D. dissertation, Michigan State University, East Lansing, MI.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ricker, K. L., and von Davier, A. A. (2007). The impact of anchor test length on equating results in a nonequivalent groups design. *ETS Res. Rep. Ser.* 2007, 1–19. doi: 10.1002/j.2333-8504.2007.tb02086.x

Sheng, Y. (2008). Markov Chain Monte Carlo estimation of normal ogive IRT models in MATLAB. *J. Stat. Softw.* 25, 1–15. doi: 10.18637/jss.v025.i08

Sinharay, S. (2011). "Chain equipercentile equating and frequency estimation equipercentile equating: comparisons based on real and simulated data," in *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland. Lecture Notes in Statistics 202*, eds N. J. Dorans, and S. Sinharay (New York, NY: Springer), 203–219. doi: 10.1007/978-1-4419-9389-2_11

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *J Educ. Meas.* 42, 309–330. doi: 10.1111/j.1745-3984.2005.00018.x

Skaggs, G., and Lissitz, R. W. (1986). IRT test equating: relevant issues and a review of recent research. *Rev. Educ. Res.* 56, 495–529. doi: 10.3102/00346543056004495

Stocking, M. L., and Lord, F. M. (1983). Developing a common metric in item response theory. *Appl. Psychol. Meas.* 7, 201–210. doi: 10.1177/014662168300700208

Swaminathan, H., and Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika* 50, 349–364. doi: 10.1007/BF02294110

Swaminathan, H., and Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika* 51, 589–601. doi: 10.1007/BF02295598

von Davier, A. A., and Chen, H. (2013). The kernel levine equipercentile observed-score equating function. *ETS Res. Rep. Ser.* 2013, 1–27. doi: 10.1002/j.2333-8504.2013.tb02345.x

von Davier, A. A., Holland, P. W., and Thayer, D. T. (2004). *The Kernel Method of Test Equating*. Berlin: Springer Science & Business Media.

Wedman, J. (2017). *Theory and Validity Evidence for a Large-Scale Test for Selection to Higher Education*. Ph.D. dissertation, Umeå University, Umeå.

Whitely, S. E. (1977). Models, meanings and misunderstandings: some issues in applying Rasch's theory. *J. Educ. Meas.* 14, 227–235. doi: 10.1111/j.1745-3984.1977.tb00040.x

Wiberg, M. (2016). Alternative linear item response theory observed-score equating methods. *Appl. Psychol. Meas.* 40, 180–199. doi: 10.1177/0146621615605089

Wiberg, M., and González, J. (2016). Statistical assessment of estimated transformations in observed-score equating. *J. Educ. Meas.* 53, 106–125. doi: 10.1111/jedm.12103

Wiberg, M., van der Linden, W. J., and von Davier, A. A. (2014). Local observed-score kernel equating. *J. Educ. Meas.* 51, 57–74. doi: 10.1111/jedm.12034

Wright, B. D. (1977). Misunderstanding the Rasch model. *J. Educ. Meas.* 14, 219–225. doi: 10.1111/j.1745-3984.1977.tb00039.x

Yao, L. (2011). Multidimensional linking for domain scores and overall scores for nonequivalent groups. *Appl. Psychol. Meas.* 35, 48–66. doi: 10.1177/0146621610373095

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

R codes for simulation (N = 500, 30 items, PTPG simulation, IE criterion, in NEAT).

```
# import raw data
load ('ADM12.Rda')
vt.form1 < - ADM12 [, 81 : 160]
vt.form2 < - ADM12 [, 241 : 320]
vt.form1 < - as.matrix (vt.form1)
vt.form2 < - as.matrix (vt.form2)
# create matrices for result storing
ee.result < - irt.result < - ke.result < - matrix (NA, 31, 500)
wro.irtke < - NULL
irtke.result < - matrix (NA, 31, 1)
sum_500 < - matrix (NA, 500, 500)
# repeat sampling and equating 500 times
for (i in 1 : 500){
# divide students into two groups (with high & low abilities) and sample randomly
form2.mean < - mean (rowSums (vt.form2))
form1.high < - vt.form1 [which (rowSums (vt.form2) > form2.mean),]
form1.low < - vt.form1[which (rowSums (vt.form2) < form2.mean),]
items < - sample (1 : 80, 30, FALSE)
high.sam < - sample (1 : nrow (form1.high), 500, FALSE)
low.sam < - sample (1 : nrow (form1.low), 500, FALSE)
# sample items to construct pseudo tests and responses
x.sam < - form1.high [high.sam, items]
y.sam < - form1.low [low.sam, items]
x.sam < - as.matrix (x.sam)
y.sam < - as.matrix (y.sam)
# EE
library (equate)
xa.score < - apply (x.sam [, 22 : 30], 1, sum)
ya.score < - apply (y.sam [, 22 : 30], 1, sum)
x.score < - apply (x.sam [, 1 : 30], 1, sum)
sum_500 [, i] < - x.score
y.score < - apply (y.sam [, 1 : 30], 1, sum)
neat.x < - cbind (x.score, xa.score)
neat.y < - cbind (y.score, ya.score)
neat.x1 < - freqtab (x = neat.x, scales = list (0 : 30, 0 : 9))
neat.y1 < - freqtab (x = neat.y, scales = list (0 : 30, 0 : 9))
ee < - equate (neat.x1, neat.y1, type = "equip", method = "chained")
ee.eq < - ee $ concordance
ee.result [, i] < - ee.eq [, 2]
# IRTOSE
library (equateIRT)
library (mirt)
colnames (x.sam) < - c (paste ("x", 1 : 21, sep = ""), paste ("c", 1 : 9, sep = ""))
colnames (y.sam) < - c (paste ("y", 1 : 21, sep = ""), paste ("c", 1 : 9, sep = ""))
x.2pl < - mirt (x.sam, 1, itemtype = "2PL")
y.2pl < - mirt (y.sam, 1, itemtype = "2PL")
par.x < - import.mirt (x.2pl, display = FALSE, digits = 3)
par.y < - import.mirt (y.2pl, display = FALSE, digits = 3)
par.x < - as.matrix (par.x $ coef)
par.y < - as.matrix (par.y $ coef)
row.names (par.x) < - c (paste ("x", 1 : 21, sep = ""), paste ("c", 1 : 9, sep = ""))
row.names (par.y) < - c (paste ("y", 1 : 21, sep = ""), paste ("c", 1 : 9, sep = ""))
par.xy < - list (par.x, par.y)
mod.2pl < - modIRT (coef = par.xy, ltparam = FALSE, lparam = FALSE)
coef.ab < - direc (mod1 = mod.2pl [1], mod2 = mod.2pl [2], method = "Stocking-Lord")
irtose.eq < - score (coef.ab, method = "OSE", se = FALSE, scores = 0 : 30)
irt.result [, i] < - irtose.eq [, 2]
# KE
library (kequate)
ker.x < - kefreq (in1 = x.score, xscores = 0 : 30, in2 = xa.score, ascores = 0 : 9)
ker.y < - kefreq (in1 = y.score, xscores = 0 : 30, in2 = ya.score, ascores = 0 : 9)
pre.x < - glm (frequency ∼ I (X) + I (X^2) + I (A) + I (A^2) + I (X) : I (A), family = "poisson", data = ker.x, x = TRUE)
pre.y < - glm (frequency ∼ I (X) + I (X^2) + I (A) + I (A^2) + I (X) : I (A), family = "poisson", data = ker.y, x = TRUE)
ke.x < - kequate ("NEAT_CE", 0 : 30, 0 : 30, 0 : 9, pre.x, pre.y)
ke.eq < - getEq (ke.x)
ke.result [, i] < - ke.eq
# IRTKE
x.sam_ke < - x.sam [, c (1 : 30, 22 : 30)]
y.sam_ke < - y.sam [, c (1 : 30, 22 : 30)]
temp.irtke < - tryCatch (
{irtose ("CE", x.sam_ke, y.sam_ke, 0 : 30, 0 : 30, 0 : 9)},
error = function (e) { return (NULL) }
)
if (is.null (temp.irtke)) {wro.irtke < - c (wro.irtke,i)}
else {
irtkeequ < - temp.irtke @ equating
irtke.result < - cbind (irtke.result, irtkeequ [, 1])
}
}
# calculate evaluation criteria
identity.ref < - matrix (NA, 31, 500)
for (i in 1 : 500){
identity.ref [, i] < - 0 : 30
}
#calculate bias^2
bias2 < - matrix (NA, 31, 4)
irtke.result < - irtke.result [, -1]
colnames (bias2) < - c ("ee", "irt", "ke", "irtke")
bias2[, 1] < - (rowMeans (ee.result - identity.ref))^2
bias2[, 2] < - (rowMeans (irt.result - identity.ref))^2
bias2[, 3] < - (rowMeans (ke.result - identity.ref))^2
bias2[, 4] < - (rowMeans (irtke.result - identity.ref))^2
#calculate ab
ab < - sqrt (bias2)
#calculate VAR
```

```
vars < - matrix (NA, 31, 4)
colnames (vars) < - c ("ee","irt","ke","irtke")
vars [, 1] < - rowMeans ((ee.result - rowMeans (ee.result))ˆ2)
vars [, 2] < - rowMeans ((irt.result - rowMeans (irt.result))ˆ2)
vars [, 3] < - rowMeans ((ke.result - rowMeans (ke.result))ˆ2)
vars [, 4] < - rowMeans ((irtke.result - rowMeans (irtke.result))ˆ2)
#calculate se
se < - sqrt (vars)
#calculate mse and rmse
mse < - matrix (NA, 31, 4)
mse < - bias2 + vars
rmse < - sqrt (mse)
#calculate wab, wse, wrmse
num_sum < - as.data.frame (table (sum_500))
fre_sum < - num_sum [, 2]/250000
wrmse < - wab < - wse < - numeric(4)
for (i in 1:4){
wrmse [i] < - rmse [, i] %*% fre_sum
wab [i] < - ab [, i] %*% fre_sum
wse [i] < - se [, i] %*% fre_sum
}
```

ORIGINAL RESEARCH
published: 24 March 2020
doi: 10.3389/fpsyg.2020.00446

Toward a Machine Learning Predictive-Oriented Approach to Complement Explanatory Modeling. An Application for Evaluating Psychopathological Traits Based on Affective Neurosciences and Phenomenology

Pasquale Dolce[1], *Davide Marocco*[2]*, *Mauro Nelson Maldonato*[3] *and Raffaele Sperandeo*[4]

[1] Department of Public Health, University of Naples Federico II, Naples, Italy, [2] Department of Humanistic Studies, University of Naples Federico II, Naples, Italy, [3] Department of Neuroscience and Reproductive and Odontostomatological Sciences, University of Naples Federico II, Naples, Italy, [4] SiPGI Postgraduate School in Gestalt Integrated Psychotherapy, Torre Annunziata, Italy

OPEN ACCESS

Edited by:
Pietro Cipresso,
Italian Auxological Institute (IRCCS),
Italy

Reviewed by:
Lucia Billeci,
Italian National Research Council, Italy
Cosimo Tuena,
Italian Auxological Institute (IRCCS),
Italy

***Correspondence:**
Davide Marocco
davide.marocco@unina.it

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *07 November 2019*
Accepted: *26 February 2020*
Published: *24 March 2020*

Citation:
Dolce P, Marocco D, Maldonato MN and Sperandeo R (2020) Toward a Machine Learning Predictive-Oriented Approach to Complement Explanatory Modeling. An Application for Evaluating Psychopathological Traits Based on Affective Neurosciences and Phenomenology. Front. Psychol. 11:446. doi: 10.3389/fpsyg.2020.00446

This paper presents a procedure that aims to combine explanatory and predictive modeling for the construction of new psychometric questionnaires based on psychological and neuroscientific theoretical grounding. It presents the methodology and the results of a procedure for items selection that considers both the explanatory power of the theory and the predictive power of modern computational techniques, namely exploratory data analysis for investigating the dimensional structure and artificial neural networks (ANNs) for predicting the psychopathological diagnosis of clinical subjects. Such blending allows deriving theoretical insights on the characteristics of the items selected and their conformity with the theoretical framework of reference. At the same time, it permits the selection of those items that have the most relevance in terms of prediction by therefore considering the relationship of the items with the actual psychopathological diagnosis. Such approach helps to construct a diagnostic tool that both conforms with the theory and with the individual characteristics of the population at hand, by providing insights on the power of the scale in precisely identifying out-of-sample pathological subjects. The proposed procedure is based on a sequence of steps that allows the construction of an ANN capable of predicting the diagnosis of a group of subjects based on their item responses to a questionnaire and subsequently automatically selects the most predictive items by preserving the factorial structure of the scale. Results show that the machine learning procedure selected a set of items that drastically improved the prediction accuracy of the model (167 items reached a prediction accuracy of 88.5%, that is 25.6% of incorrectly classified), compared to the predictions obtained using all the original items (260 items with a prediction accuracy of 74.4%). At the same time, it reduced the redundancy of the items and eliminated those with less consistency.

Keywords: machine learning, predictive modeling, explanatory modeling, item selection, neural networks, psychopathological assessment

INTRODUCTION

Statistical modeling is traditionally separated into two different cultures. One uses an explanation-oriented approach to science, the explanatory modeling that Breiman (2001) defines as "data modeling culture." The other uses a prediction-oriented approach, defined by Breiman as "algorithmic modeling culture." In the former approach, data is assumed to be drawn from a given stochastic model, researchers are interested in testing the hypothesized "true" relationship between two or more variables and the mechanisms governing their intercorrelation, and the main objective is to reproduce model parameters using statistical inference and to improve the explanatory power of models. In the second approach, the data-generating process is unknown, and researchers are interested in finding an algorithm capable of recognizing different patterns hidden in data, which then gives the best prediction for the output values through the input values of new observations (Shmueli, 2010). However, in many disciplines, particularly in psychology and social sciences, statistical modeling for explanation is the predominant, if not the exclusive approach. Conversely, in domains like bioinformatics and natural language processing, algorithmic modeling is predominant (Breiman, 2001).

Beyond a confirmatory approach with the corresponding inferential assumptions (often not met in the real world), predictive modeling can help establish theoretically grounded models that have high predictive power (Sarstedt et al., 2014) and increase the efficiency and reproducibility of a researcher's analysis (Yarkoni and Westfall, 2017). Psychology research may improve comprehensively by exploiting the potentiality of Machine Learning and Artificial Intelligence algorithms while maintaining the data modeling culture.

Psychology research needs to be grounded in a common theoretical framework of reference, which is the initial stage of the research design. The credibility of a research study is generally derived from the quality of this initial stage of the design. Consequently, psychology research should not steer toward a prediction-based orientation to the detriment of an approach that aims at testing model relationships in an explanatory sense. Even in a predictive-oriented approach, hypothesis formulation is a crucial step and it is always the investigator who chooses the statistical methods better suited for the related theoretical and empirical models. Results depend crucially on the user's knowledge of the domain they are investigating (Pessa, 2004). In the presence of complex theories, moreover, testing a pre-determined system of hypotheses may become problematic in terms of model assumptions and interpretation. In such a case, a discovery-oriented process should be envisioned (Wold, 1985), where the investigator should be able to exploit the appropriate statistical and computational however methodologies to convert data and models into actionable insights to support such theories and for prediction purposes (Breiman, 2001; Lauro, 2019). Indeed, machine learning approaches to clinical psychology and psychiatry may focus on large multidimensional data sets to improve the decisions associated with diagnosing and treating people who have been diagnosed with mental illness using ordinary clinical methods (Dwyer et al., 2018).

In an evolved vision of the use of artificial intelligence methods in the context of psychopathology, scholars have the unprecedented opportunity to integrate complex brain, behavior and genes patterns to develop precision psychiatry. Indeed, growing evidence suggests that the classification of psychiatric patients derived from these approaches may better predict treatment outcomes than ordinary DSM/ICD-based diagnoses (Bzdok and Meyer-Lindenberg, 2018).

Another interesting use of machine learning is for demonstrating the reliability of a scale and testing for convergence validity with other variables. Instead of using traditional techniques, predictive models can achieve the same results but in a much more efficient way, computing the out-of-sample prediction accuracy of the scale with respect to one or several other measures (Du et al., 2014; Yarkoni and Westfall, 2017).

Indeed, predictive modeling can be used instrumentally to complement explanatory modeling in order to further scientific knowledge (Breiman, 2001; Shmueli, 2010; Yarkoni and Westfall, 2017; Azzolina et al., 2019). The use of the two approaches should be complementary rather than competitive. A proper combination of the two approaches may lead to the use of a wide variety of statistical and computational tools, by exploiting the strengths of both approaches through a single method in order to have stronger grounds for theory testing, knowledge discovery, prediction and decision-making, for example, for the assessment and diagnosis of psychopathology.

In line with these considerations, we think that a methodology that highlights the features of predictive modeling in terms of model building and assessment may be welcomed in psychology research and other social science disciplines, which can only benefit from these methodological developments.

The present work focuses on the psychopathological and behavioral dimensions that play the role of main nosographic organizers of psychiatric diagnosis, to improve the precision with which the classification of patients in specific diagnostic categories is carried out.

The study intends to present a new methodology for approaching prediction in a psychopathological diagnosis context applied to the construction of a novel diagnostic scale, by preserving the psychometric properties of the models as they are traditionally approached from an explicative point of view.

The current psychopathological diagnosis relies on syndromic models that we have inherited from authors such as Kraepelin and Bleuler, who operated in a pre-neuroscientific era. It follows that many psychiatric disorders are classified through obsolete concepts that do not consider the knowledge we currently have of the brain and the basic emotional systems that comprise its deepest part (Lane and Sher, 2015; Montag et al., 2017). Especially in humans, it has become increasingly evident that the phylogenetically more recent cortical structures, to which the awareness of experience links, have improved the adaptation of fundamental emotional processes to social contexts, but have not replaced the weight of emotions in the organization of social life (Panksepp et al., 2017). This evidence can have a significant impact on the psychopathological investigation that can now focuses on emotionality and affective regulation systems (Stanghellini, 2019). Indeed, the present work introduces concepts derived from affective neuroscience into psychopathological diagnostics, which up to now have been largely underestimated for the study of psychic disorders and can improve the naturalistic value and stability of psychiatric nosography.

In particular, in this paper, we propose a procedure for the selection and analysis of the items to be included in a novel scale for the evaluation of psychopathological traits based on affective neurosciences and phenomenology, which combines explanatory psychometric measurements, such as factorial coherence and construct validity, with measurements of the predictivity of the instrument carried out through machine-learning methods.

The proposed procedure identifies a well-fitting, in terms of validity and reliability of the factor structure, and a predictive yet parsimonious model among competitive ones. Indeed, parsimonious and well-fitting models exhibit higher predictive abilities and are more likely to be scientifically replicable and explainable (Sharma et al., 2019).

The main objective is to maximize the predictive ability of the model while maintaining the psychometric properties and factorial structure of the scales. A machine learning procedure is applied to identify the best predictor items for the presence of pathological variants of the personality to find the set of items that maximize the predictive ability of the model. The factorial structure is then evaluated through principal component analysis (PCA).

The model evaluation will consider the performance of the model in terms of both explanatory power and predictive accuracy. Measurements of explanatory power are typically in-sample metrics and refer to how well the proposed model (in this case, the model of the factor structure) accounts for the covariances between items. For predictive power, out-of-sample metrics are used, which are computed through a cross-validation procedure.

Theory Reference

The Relationship Between Emotions and Mental Disorders

The self-report diagnostic test described in this paper is rooted in both phenomenological and neuroscientific views of emotions. In this integrated perspective, emotions present three inseparable functions: the production of socially adequate behavior; the regulation of internal homeostasis; the production of a conscious mental state characterized by adaptive values (e.g., good or bad, unpleasant or pleasant) that are salient for the subject (Maldonato et al., 2018; Sperandeo et al., 2018b).

In these functions, emotions are the basis of rational processes. As shown by numerous authors, subjects with lesions of basic emotional systems show profound impairment in their decision-making activity and are substantially incapable of responding rationally to life events (Stanghellini et al., 2016; LeDoux and Hofmann, 2018). Below we will describe the two perspectives of reading that clarify the emergence of psychopathology from affective processes in a complementary and integrable way.

For current affective neuroscience, human minds express several phylogenetically ancient emotional processes. Basic emotional tendencies have great significance for psychopathology and we consider it extremely important for the study of psychic disorders. These systems are present in all mammals but, of course, the vast cognitive capacities of humans add unique dimensions to emotional consciousness. The interweaving of cognitive and affective capacities, and in particular the aspects of memory, can make human beings particularly sensitive to psychiatric disorders. Through cognitive processes of emotional amplification, humans can sustain emotional arousal for a long time after the precipitating causes have passed. In this way, our cognitive functions can become critical agents in the creation of emotional problems. Intense emotional excitement sustained and unregulated by ruminative tendencies can interfere with our thinking patterns, even intensify, and energize our cognitive concerns by producing a deleterious vicious circle. Thanks to our remarkable cognitive abilities, we create complex mental lives, with intrapsychic tensions typical of our species. Our vast ability to look far into our memory and imagine terrible future problems pushes us to sustain the emotional excitement generated internally and to encounter psychic disorders much more than other mammals. Prolonged emotional excitement can also lead to prolonged turbulence in our bodies, producing various psychosomatic disorders and disorders in our daily quality of life (Clynes and Panksepp, 2013).

From the phenomenological perspective, emotions precisely determine the motivation for movement. They are functional states of our organism that motivate actions; they provide orientation in life by making sure that attention moves in a particular direction and attributes specific meanings and values to the world. Recognizing this aspect of emotions allows us to elevate them from mere biological reactions or mental phenomena to fundamental expressions of the "lived body," representing the moment in which the psychobiological dimensions of experience are articulated (Messas et al., 2018; Sperandeo et al., 2019).

Emotions allow us to see reality from a specific perspective. The analysis of the mental states of an angry person and a frightened person allows us to understand the differences in their respective life perspectives. Therefore, the subject's way of experiencing the world reflects his or her state of mind, so it follows that emotions are the primary way to understand a person and his or her psychopathology. Finally, emotions play a fundamental role in the development of sociality, inter-subjectivity and empathy. When a child perceives his mother's happy face, he or she automatically reproduces her facial expression; through this reflection, he feels his mother's happiness. It is an inter-corporeity produced by a perceptual-motor process, which is the very essence of the emotional phenomenon. In the absence of emotions, the world appears unreal and distant, devoid of interest and meaning. The objects that belong to the world appear to be a collection of meaningless things of which one can have a non-practical theoretical knowledge. Emotions are the motivation for performing actions, and without them, there is no motivation to move and thus no action. The absence of emotions implies the loss of vital contact with reality, everything in the world appears equivalent and devoid of salience so that neither movement, nor choice, nor meaning is possible (Stanghellini, 2019).

In our opinion, emotions – understood in their entirety as effective experiences, adaptive behaviors, and autonomous and self-regulating processes – are the basis for the emergence of psychopathological phenomena (Solms and Panksepp, 2012). The main clinical manifestations currently classified by the adult psychiatric nosography are personality disorders, pathologies resulting from mental trauma and stressful events, mood disorders, somatic symptom disorders and anxiety disorders. Negative emotions such as fear, suffering, anger are present in all of these disorders, but currently, an adequate nomenclature to describe these relationships has not agreed. Studying psychopathology from the perspective of the emotional events of a subject is therefore difficult because it cannot follow paths traced and shared in the scientific community. It is precisely for this reason that the development of an innovative vision appears to be indispensable.

The Panksepp Model of Emotions

In this paper, we present the development of a self-report diagnostic tool for the exploration of the psychopathological manifestations that emerge from the emotional affective processes organized in the medial part of the brain. For this purpose, we have used the model of basic emotional systems as described by Panksepp and Biven (2012). According to this approach, in mammals' brains, there are at least seven emotional neuronal circuits (fear, rage, sexual impulses, care, anxiety of separation and social bond, playfulness, and a general system of lust and seeking) from which behaviors, autonomic processes and conscious affective states emerge which are essential for one's interpersonal relationships.

When these systems are activated, individuals experience intense feelings, recall memories, implement behaviors of adaptation to the environment, and activate hormonal processes and vegetative regulation. The basic emotional systems at the beginning of childhood psychological development are weakly linked to the objects of the world. The basic affective tools that evolution has provided emerge in the development of the brain without an initial intrinsic connection to the events of the world. It is through life experiences, both individual and cultural that these connections are forged. Even if these emotionally evaluated systems are clustered into constellations of positive and negative affections, it seems unlikely that only two primary types of affective feelings are the raw materials from which all other affections within the brains of mammals are created. Indeed, affection is not interpreted as an independent sensory function of the brain but is based on tendencies toward action.

Considerable evidence arising from animal brain research suggests that at least seven basic emotional systems are concentrated in the subcortical regions of the brain and are located essentially in the same regions of the brain in all mammals.

A brief description of each basic emotional systems is presented below.

The SEEKING system must be conceptualized as a primary action system that helps to realize emotional drives, to seek nourishment and to realize expectations. This system operates in both positive and negative emotional situations (e.g., security seeking) and helps to maintain the fluency of the behavior as well as supporting learning and other cognitive activities (Ikemoto and Panksepp, 1999).

The FEAR system associates anxiety and the tendency to escape from the many dangers present in our world. The RAGE system supports the defense and the achievement of objectives. The LUST system supports libidinal appetites. The CARE system supports the protection and care of offspring. The GRIEF (Panic) system aims at preventing the loss of protective figures. The PLAY system aims at developing sociality (Panksepp, 2014).

MATERIALS AND METHODS

Study Population

As part of the ordinary psycho-diagnostic evaluation procedure, 604 adult patients have been enrolled in the clinical centers of SiPGI, a specialization school in psychotherapy. The questionnaire described below was administered to subjects who agreed to participate in the study.

Personality disorders were found in 196 (32.5%) patients out of the 604. Subjects in the depressive, manic or acute psychotic phase and subjects with cognitive deficits and head injuries with detectable parenchymal lesions were excluded. The diagnosis was made using the Italian version of the personality diagnostic interviews associated with DSM-5: The Structured Clinical Interview for DSM-5 Personality Disorders (SCID-5-PD). It is one of the most used tools for the diagnosis of personality disorders in clinical and research areas and has demonstrated excellent reproducibility and clinical validity (Somma et al., 2017). The subjects that did not meet the diagnostic criteria for any nosographic category were classified as healthy, and all others were classified as unhealthy.

Characteristic of patients, as shown in **Table 1**, are the following: 273 males (45.2%) and 331 females (54.8%); average age of 33.96 ± 11.34, 342 (56.5%) were unmarried, 223 (36.9%) married, 32 (5.3%) divorced and 8 (1.3%) widow; 161 (26.7%) patients were graduated, 336 (55.6%) had high/secondary school, 100 (16.6%) middle school and 7 (1.2%) elementary school; 393 (65.1%) were employed, 197 (32.6%) unemployed, and 14 (2.3%) retired. No statistically significant differences between the two groups (healthy vs. unhealthy) were found for all the variables, except for marital status ($p = 0.012$).

Measures

For the structuring of the questionnaire, a group of six experts in psycho-diagnostics, under the supervision of two of the authors of this work, produced a list of 260 items that – according to them – describe the dimensions of the seven basic emotional systems within the main psychic pathologies and personalities currently framed in the classification systems.

The questions are formulated to obtain dichotomous answers (yes/no), avoiding the frequency and intensity of the phenomenon under investigation within the same descriptions, limited exclusively to the detection of its presence or absence.

TABLE 1 | Characteristics of patients.

Characteristic	Total *n* = 604	Healthy *n* = 408	Unhealthy *n* = 196	*p*-Value
Sex				0.551
Male	273 (45.2%)	181 (44.4%)	92 (46.9%)	
Female	331 (54.8%)	227 (55.6%)	104 (53.1%)	
Age	33.96 ± 11.3	34.52 ± 11.4	32.78 ± 10.9	0.076
Marital status				
Unmarried	341 (56.5%)	129 (65.8%)	212 (52%)	0.012
Married	223 (36.9%)	58 (29.6%)	165 (40.4%)	
Divorced	32 (5.3%)	8 (4.1%)	24 (5.9%)	
Widow	8 (1.3%)	1 (0.5%)	7 (1.7%)	
Education				
Graduated	161 (26.7%)	50 (25.5%)	111 (27.2%)	0.962
High/secondary school	336 (55.6%)	112 (57.1%)	224 (54.9%)	
Middle school	100 (16.6%)	32 (16.3%)	68 (16.7%)	
Elementary school	7 (1.2%)	2 (1%)	5 (1.2%)	
Occupational position				
Employed	393 (65.1%)	124 (63.3%)	269 (65.9%)	0.789
Unemployed	197 (32.6%)	129 (31.6%)	129 (31.6%)	
Retired	14 (2.3%)	10 (2.5%)	10 (2.5%)	

Data are reported as number of patients (%) or mean (± standard deviation), as appropriate. p-Values are based on Student's t-test, χ^2 test or Fisher's exact test, as appropriate. Note that for some characteristics frequencies over categories do not sum to the total number of patients, because there were some missing values.

The items are organized into three distinct areas:

- 157 items are related to the "emotional characteristics" present in the personality disorder area. Many of these questions are presented in order to detect the non-pathological psychic phenomenon. In line with Panksepp model of basic motivational systems, most questions investigate emotional experiences and behaviors. Other questions investigate physical sensations while a small group of questions looks for the subject's opinions to detect the impact of cortical functions on emotional systems.
- 24 questions explore the presence of "dissociative phenomena" commonly present in the area of post-traumatic pathologies. In this group of questions, only the presence of dissociative phenomena in the three dimensions of depersonalization-derealization, dissociated mental states and dissociative amnesia is sought.
- 79 questions explore the main "psychopathological traits." These questions also explicitly refer to the presence or absence of a pathological phenomenon.

The division into three areas (emotional characteristics, dissociative phenomena, psychopathological traits) of the items arises from the theoretical assumption that the processes of sensitization or desensitization of the seven basic emotional systems produce a type of symptomatology (described in the group of items belonging to the emotional characteristics) that is different from that determined by the cognitive reworking of the emotional states (described in the group of items belonging to the psychopathological traits). Both symptoms are distinguishable from the dissociative one in which the traits of emotions produced by the system of anger and fear spread and invade the structures of awareness (Trull et al., 2015; Sperandeo et al., 2018a).

Statistical Analysis and Multi-Step Machine Learning Procedure

Preliminary analyses concerned the handling of missing data was performed. Missing data were assumed to be missing completely at random (MCAR). The multiple imputation method for incomplete multivariate data was performed for the imputation process, using the predictive mean matching method built in the R package "mice" (Van Buuren and Groothuis-Oudshoorn, 2011).

As for the explanatory side of the work, to evaluate the factorial structure of the scales and assess its psychometric properties a PCA and orthogonal Varimax rotation was performed.

For the predictive side, which relies on machine learning techniques, artificial neural networks (ANN) are applied as a classifier to maximize the predictive power of the model. To this end, multi-layer ANNs were trained with *resilient backpropagation algorithm* (Riedmiller, 1994) to classify subjects as healthy or unhealthy, considering all items of the scale (see **Figure 1**).

Resilient backpropagation (RPROP) is a fast and effective learning algorithm that uses the direction of the error gradient (i.e., the sign of the change) for calculating the weight change,

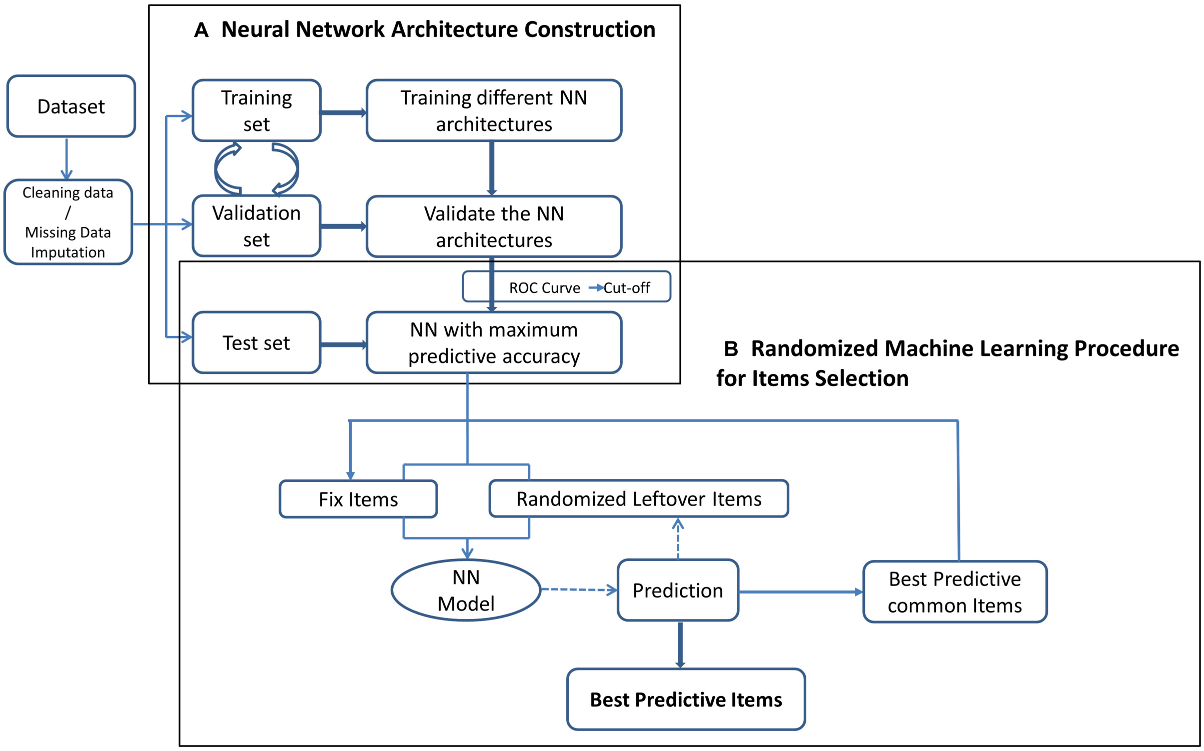

A Neural Network Architecture Construction
Dataset
Cleaning data / Missing Data Imputation
Training set
Training different NN architectures
Validation set
Validate the NN architectures
ROC Curve
Cut-off
Test set
NN with maximum predictive accuracy
B Randomized Machine Learning Procedure for Items Selection
Fix Items
Randomized Leftover Items
NN Model
Prediction
Best Predictive common Items
Best Predictive Items

FIGURE 1 | Flow chart of proposed predictive-oriented machine learning procedure.

rather than the actual magnitude of the partial derivative, as in the traditional backpropagation.

Resilient backpropagation calculates an individual delta Δ_{ij}, for each connection, which determines the size of the weight update. The calculation of delta at any given time of the learning process follows the rule:

$$\Delta_{ij}^{t}\begin{cases}\eta^{+}\times\Delta_{ij}^{t-1}, & if\ \dfrac{\partial E^{t-1}}{\partial w_{ij}}\times\dfrac{\partial E^{t}}{\partial w_{ij}}>0\\ \eta^{-}\times\Delta_{ij}^{t-1}, & if\ \dfrac{\partial E^{t-1}}{\partial w_{ij}}\times\dfrac{\partial E^{t}}{\partial w_{ij}}<0\\ \Delta_{ij}^{t}, & \text{otherwise}\end{cases}$$

where $0 < \eta^{-} < 1$ and $\eta^{+} > 1$.

Synaptic weights (w_{ij}^{t}) are updated according the usual formula:

$$w_{ij}^{t}=w_{ij}^{t-1}+\Delta w_{ij}^{t}$$

The output neuron activation o_j of the ANN is calculated based on the neuron net-input x_j, according to the following functions:

$$x_j=i_i w_{ij}-b_j$$

$$o_j=\frac{1}{1+e^{-x_j}}$$

where i_i is the i-*th* input, b_j is the bias of the j-*th* post-synaptic neuron and w_{ij} is the weights matrix connecting presynaptic to post-synaptic neurons.

For the actual ANN training computation we used the "neuralnet" R package (Günther and Fritsch, 2010).

The construction and the subsequent exploitation of the ANN predictive power for item selection purposes was carried out in two stages.

In a first stage, a series of fully connected ANNs with 260 input nodes (i.e., one for each item of the scale), one single output node (encoding healthy or unhealthy predictions) and a variable number of hidden units, ranging from 0 to 50, were tested. The parameters were fixed for all architectures: learning rate factors η^{-} and η^{+} were set at 0.5 and 1.2, respectively; synaptic weights were randomly initialized from a normal distribution in the rage $[-4, 4]$; the stopping criteria for the error function was 0.0005; and the maximum number of iterations was fixed in 5000 epochs. At this stage, a cross-validation procedure was used to select the best neural network architecture, i.e., the more effective number of hidden nodes, in terms of prediction accuracy. A Monte Carlo Cross-validation procedure has been chosen to avoid over-fitting in the following way: at first, from the entire set of the available data, a *test set* was extracted. In the test set we maintained the same number of patients in the two groups (healthy and unhealthy). Thus, we randomly selected the 20% of patients among unhealthy ones. Then, we selected the same number of patients among the healthy ones. Consequently, the test set was composed of about 13% of all the patients.

Subsequently, at each step of the training procedure, the remaining data were halved into two different sets: the *training set* (80% of remaining patients), which is used to find a set of good weights and bias values by comparing the desired output with the one produced by the ANN – thus for calculating the actual error – and the *validation set* (20% of remaining patients), which is used to evaluate at runtime the progress of the learning process. The *test set* is eventually used to assess the quality of the resulting ANN in terms of out-of-sample prediction accuracy at the end of the training.

Receiver operating characteristic (ROC) analysis was applied to find the optimal output node threshold, i.e., the one that gives the best diagnostic accuracy for the model (Woods and Bowyer, 1997). The (0, 1) criterion was used to select the optimal threshold, giving maximum sensitivity and specificity. This procedure assures a better prediction accuracy among groups of subjects, even if the groups are not balanced. Model performance was measured on the test data using the area under the curve (AUC) and classification error rate.

At a second stage, a knowledge-based randomized machine learning procedure was applied to identify the best predictor items for mental disorders, i.e., the set of items that maximize the predictive ability of the model. This procedure started by defining a set of items that are theoretically relevant and are never dropped from the neural network's inputs (this is the knowledge-based part of the procedure). Then, predictions were obtained adding new items randomly sampled from the set of the remaining items. The items in common across all the "best" solutions in terms of prediction accuracy, were then considered as fixed for the following step, together with the theoretically relevant items. Then, items were again randomly sampled from the set of the remaining items until the algorithm figured out which set of items achieves the best prediction accuracy. Finally, the factorial structure of the select items was evaluated through principal PCA. The entire procedure is depicted in **Figure 1**.

The final model evaluation considers the performance of the model in terms of both explanatory power and predictive accuracy.

All computations and statistical analyses were performed using the R software environment for statistical computing.

RESULTS

Principal Component Analysis on All Items

For all items of the scale, only the 0.1% of the data were missing and were assumed to be MCAR.

Principal component analysis was performed separately for each of the three areas, selecting seven components for each area, according to theory, because the purpose of this analysis was not to extract components, but rather to examine the coherence of the scale and the extent to which the results of the two analysis (respectively, the one with all the items and the one with only the selected items) differ.

As will be evident below, the explained variability of components appears relatively low for each area. However, it should be noted that PCA was applied to binary variables. Even though PCA on binary data provides a plausible low-dimensional representation (Gower, 1966; Jolliffe, 2002), the obtained principal components, like the components computed using multiple correspondence analysis (MCA) of categorical data, are just fractional coordinates in a smooth Euclidean space mapping, and scale indeterminacy arises. Scale change leads to the so-called low percentage of inertia problem since eigenvalues tend to zero and the variance explained by the components would be severely underestimated. Therefore, the percentage of the explained variance gives a pessimistic view of the proportion to which the extracted components account for the variation of the data and simple scale adjustment of the solution can give a more precise estimate (Benzécri, 1979; Lebart et al., 1995; Greenacre and Blasius, 2006). For these reasons, explained variance components may still be very informative, as in the case of this study, which allows us to interpret the PCA results correctly.

As shown in **Table 2**, for the area of "emotional characteristics presents" (136 items) the seven components cumulatively explain 25% of the variance. The first component better explains 44 items in which the "yes" answers describe a condition of hypersensitization of the system of grief. The second component represents 18 items, in which the "yes" answers describe the good functioning of the care system. The third component explains the12 items in which the "yes" answers describe a hypersensitization of the system of fear. The fourth component consists of 18 items in which the "yes" answers describe the correct functioning of the search system. The fifth component explains the 18 items in which the "yes" answers describe the good functioning of the game system. The sixth component is composed of 11 items in which the "yes" answers describe a hypersensitization of the system of anger. The seventh component is composed of 15 items in which the "yes" answers describe a hypersensitization of the system of lust.

Table 3 shows the seven components selected from the items related to the area of psychopathological traits (75 items). The first component better explains 17 items in which the "yes" answers describe pathological traits determined by the hypersensitivity of the grief system. The second component is composed of 16 items in which the "yes" answers describe pathological traits determined by the hypoactivity of the Seeking system. The third component is composed of 11 items in which the "yes" answers describe pathological traits determined by the hypoactivity of the care system. The fourth component is composed of 10 items in which the "yes" answers describe pathological traits determined by the hypersensitivity of the fear

TABLE 2 | PCA – area of emotional characteristics.

Component n. items	Eigenvalue	% explained variance	Cumulative % explained variance
(1) PANIC 44 items	10.67	6.76	6.755
(2) CARE 18 items	5.31	3.36	10.114
(3) FEAR 12 items	5.31	3.36	13.472
(4) SEEK 18 items	5.27	3.34	16.809
(5) PLAY 18 items	5.27	3.34	20.144
(6) RAGE 11 items	4.75	3.00	23.15
(7) LUST 15 items	3.75	2.37	25.52

system. The fifth component is composed of 12 items in which the "yes" answers describe pathological traits determined by the hyperactivity of the system of anger. The sixth and seventh components are composed of 7 and 2 items, respectively, in which the "yes" answers describe pathological traits determined by the hypoactivity of the game system and pleasure.

Table 4 shows the two components that emerged from the area of dissociative phenomena consisting of a total of 22 items. The component of depersonalization-derealization and the composition of dissociative amnesia are composed respectively of 12 and 10 items in which the "yes" answers describe the two typical ways of altering the cognitive functions produced by the uncoordinated hyperactivity of the basic emotional systems.

The group called "emotional characteristics" composed of 136 items has 15 with negative loadings, 55 with very low loadings (less than 0.4) and 38 with low loadings (less than 0.5). The component called "Seek" has 14 items 5 of them are negative. Moreover, in the component called "Panic," there are 10 items that show significantly high values even in other components. This component, composed of 44 items, has only 8 items with high loadings (greater than 0.5).

Two out of the 75 items in the group called "psychopathological traits" are negative, 40 have very low loadings, and 26 low loadings. Twelve out of the 23 items in the group called "dissociative phenomena" have very low loadings and 4 low loadings.

Neural Network Architecture Construction

As described above, a multiple-layer ANN was trained with backpropagation to classify subjects with and without the presence of pathological variants of the personality, considering all the items as inputs. The first stage of the procedure, as described in Section "Materials and Methods," selected as the best predictive model, an ANN with 25 nodes in the hidden layer. The best result was reached in 546 epochs of training. The limit of 5000 epochs was never reached for all the architectures trained.

TABLE 3 | PCA – area of psychopathological traits.

Component n. items	Eigenvalue	% explained variance	Cumulative % explained variance
(1) PANIC 17 items	6.408	8.215	8.215
(2) SEEK 16 items	5.454	6.992	15.208
(3) CARE 11 items	5.08	6.512	21.72
(4) FEAR 10 items	4.396	5.635	27.355
(5) RAGE 12 items	4.199	5.384	32.739
(6) PLAY 7 items	3.89	4.988	37.727
(7) LUST 2 items	1.931	2.476	40.203

TABLE 4 | PCA – area of dissociative phenomena.

Component n. items	Eigenvalue	% explained variance	Cumulative % explained variance
(1) Depersonalization 12 items	7.83	32.6	32.6
(2) Amnesia 10 items	4.47	18.61	51.21

Then, ROC analysis was applied to find the optimal threshold. The resulted threshold was 0.088. With this parameter fixed, on the out-of-sample *test set* the ANN achieved a classification error of 0.2564, meaning a prediction accuracy equal to 74.4% (i.e., 25.6% of incorrectly classified) and an AUC equal to 0.778. The corresponding ROC curve is shown as a dotted line in **Figure 2**. In particular, the classification error rate was equal to 28.2% for patients with the presence of pathological variants of the personality and 23.1% for patients without the presence of pathological variants of the personality.

Knowledge-Based Randomized Machine Learning Procedure for Items Selection

The selected ANN architecture, together with the weights and the found optimal threshold, was then used for the item selection procedure. Predictions and classification error rates were computed using only the *test set*.

Twenty-one items were chosen as theoretically relevant and therefore, always considered as fixed inputs of the ANN. These items are descriptive of (a) mood disorders both in the depressive and manic sense, (b) alterations of the content of thought and (c) dis-perceptive phenomena. They were chosen for their fundamental link with psychopathology.

A multi-step procedure was needed to find the optimal solution, i.e., the set of items that achieves the best prediction accuracy. At the first step, 5 million combinations of items were randomly sampled from the set of the remaining items. Then, the items that appear in the solution with the lowest classification error rate were selected and added up to theoretically relevant items (thus, both sets were considered as fixed inputs for the subsequent steps). At the second step, another 5 million combinations of items were randomly sampled from the set of remaining items and the common items across all the "best" solutions were again selected and considered as fixed items for

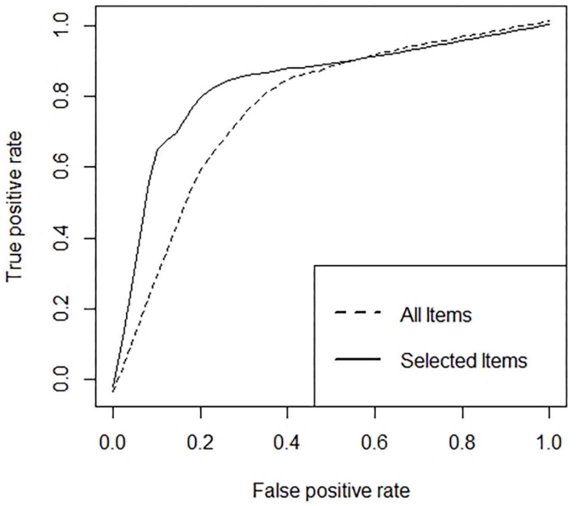


FIGURE 2 | Roc curves obtained using all the items and the items selected by the proposed procedure.

the subsequent steps. This procedure was repeated until the classification error rate of the "best" solution did not improve.

The entire selection procedure took about 10 h to complete on a parallel implementation of R running over 2 processors on a Windows 10 Pro 64-bit platform equipped with a i5-7200u intel processor and 8 GB of RAM.

Figure 3 represents the number of items for the best-parsimonious solution and (a) the number of common items obtained at each step and (b) the corresponding classification error. It clearly shows that the best solution is reached in four steps. Solutions after the 4th step are all worse (data not shown).

At step 4, the best prediction accuracy was achieved by a combination of 167 items, the 21 theoretically relevant ones and 146 selected by the randomized machine learning procedure. Among the selected items, 98 items relate to emotional characteristics, 15 relate to dissociative phenomena and 33 relate to psychopathological traits.

The prediction accuracy on the test set was equal to 88.5% (i.e., 11.5% incorrectly classified) and an AUC equal to 0.849. The corresponding ROC curve is shown as a solid line in **Figure 2**. In particular, the classification error rate was equal to 15.4% for patients with the presence of pathological variants of the personality and 7.7% for patients without the presence of pathological variants of the personality.

Principal Component Analysis on the Selected Items

Table 5 shows the results of the PCA performed on the group of items (78 items) in the area of emotional characteristics selected by the neural network. The seven components from the PCA (globally explaining 28.7% of the variance) appear to be perfectly consistent with the reference theory discussed in the previous sections. The items of the specific components describe behaviors and affective mental contents provided in the Panksepp model. The first component explains 23 Items in which the "yes" answers describe a condition of hypersensitization of the system of grief. People described by these items tend to be blocked by a continuous state of anguish that annihilates them and leads them into a state of depression. The second component consists of 12 items in which the "yes" answers describe the good functioning of the care system. The people described in these items know how to take care of others and the system to which they belong. The third component is composed of 9 items in which the "yes" answers describe a hypersensitization of the seeking system. The people described by this component are optimistic, open to seeking and focused on achieving their goals. The fourth component consists of 14 items in which the "yes" answers describe how well the play system works. The people described by this component can socialize and enjoy the experiences of life. The fifth component is composed of 8 items in which the "yes" answers describe a hypersensitization of the system of lust. The people described by these items live in the continuous fantasy of satisfying their libidinal urges. The sixth component is composed of 5 items in which the "yes" answers describe a hypersensitization of the system of anger. The people described by these items are intolerant and aggressive. The seventh component is composed of 7 items in which the "yes" answers describe a hypersensitization of the system of fear. The people described by this component exert control over their world because they have associated numerous dangers with the activation of this emotional system.

Table 6 shows the seven components that emerged from the area of psychopathological traits selected from the neural network, which is composed of 68 items. The first component is composed of 17 items in which the "yes" answers describe pathological traits determined by the hypersensitization of the grief system. The people described in this component can be self-destructive and hetero-destructive. The second component consists of 12 items in which the "yes" answers describe pathological traits determined by the hypofunction of the care system. The people described here are unable of taking care of their environment and the people around them, who they feel to be dangerous and intrusive.

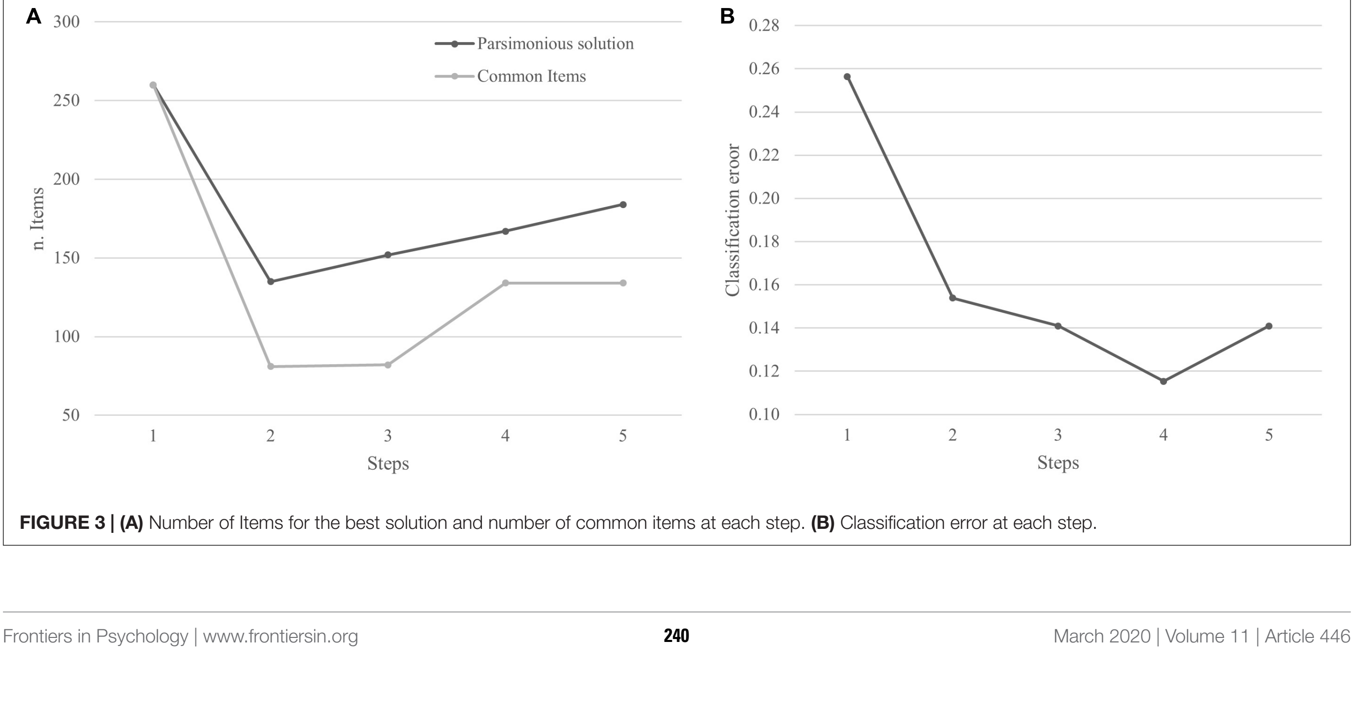


FIGURE 3 | (A) Number of Items for the best solution and number of common items at each step. **(B)** Classification error at each step.

TABLE 5 | PCA loadings – Area of emotional characteristics in the pool of selected items (28.67% of explained variance – Kaiser-Meyer-Olkin = 0.839).

PANIC (Items = 23, Eigenvalue = 11.05,% explained variance = 8.57)	
è insicuro d'avanti ai problemi?	0.62
rinuncia facilmente alle cose perché si preoccupa dei rischi?	0.588
quando è stanco ha bisogno (o chiede aiuto) agli altri?	0.551
si sente inferiore agli altri?	0.528
deve impegnarsi molto per avere fiducia in se stesso?	0.519
attende che gli altri le risolvano i problemi	0.507
ha paura che le sue cose vadano male?	0.5
fa fatica a guarire da un malanno?	0.5
pensa di avere problemi al cervello?	0.492
la sua vita è priva di senso?	0.49
è pessimista?	0.479
ha bisogno di riposare durante la giornata?	0.444
rinuncia facilmente difronte a compiti impegnativi?	0.441
è preoccupato difronte a situazioni nuove?	0.411
fa fatica a comprendere le persone?	0.41
ha cattive abitudini che vorrebbe cambiare?	0.409
la preoccupano gli imprevisti?	0.409
le sue scelte sono determinate dagli altri?	0.365
ignora quale sia lo scopo della sua vita?	0.358
si entusiasma lentamente per le novità?	0.355
quando fa degli errori se la cava da solo?	0.339
si sente carico di energia per tutta la giornata?	−0.427
è molto sicuro di se?	−0.496
CARE (Items = 12, Eigenvalue = 4.83,% explained variance = 3.74, cumulative% explained variance = 12.31)	
è connesso spiritualmente agli altri?	0.597
ha mai avuto esperienze paranormali?	0.518
ha mai fatto intense esperienze spirituali?	0.495
sente un legame profondo con la natura?	0.493
quando è concentrato molto perde la cognizione del tempo e dello spazio?	0.471
è talmente preso dalle sue attività da perdere il contatto con la realtà?	0.429
ha idee creative quando si lascia andare all'ozio?	0.396
gli altri la definiscono distratto?	0.332
la vita dipende da una forza spirituale al di sopra di noi?	0.33
è accomodante con gli altri?	0.329
sa di avere un 'sesto senso'?	0.321
è costante nelle cose che fa?	0.307
SEEK (Items = 9, Eigenvalue = 4.75,% explained variance = 3.68, cumulative% explained variance = 15.99)	
si definirebbe ottimista?	0.495
inventa storie o dice bugie solo per divertimento?	0.467
è tranquillo sul suo futuro?	0.436
evita situazioni o attività che la irritano?	0.408
le sono indifferenti i complimenti?	0.387
sa mentire bene?	0.361
ritiene importante i legami di amicizia?	0.35
è a suo agio anche con persone sconosciute?	0.331
affronta le difficoltà prendendole come sfide?	0.309
PLAY (Items = 14, Eigenvalue = 4.66,% explained variance = 3.61, cumulative% explained variance = 19.61)	
soffre se vede altri soffrire?	0.556
tende ad aiutare gli altri?	0.51
ama collaborare con gli altri	0.491
tende a collaborare con gli altri?	0.449
è empatico e disponibile?	0.436
è altruista anche con chi l'ha trattata male?	0.392
reagisce agli eventi coerentemente con i suoi valori?	0.389
riflette molto prima di prendere una decisione?	0.388
trova qualcosa di poetico anche nelle piccole cose?	0.387
agisce secondo le sue abitudini?	0.359
ha molte buone abitudini quotidiane?	0.354
si commuove davanti a prodotti artistici?	0.336
investe molta energia per fare le cose?	0.329
sta male se perde delle amicizie?	0.324
LUST (Items = 8, Eigenvalue = 4.28,% explained variance = 3.32, cumulative% explained variance = 22.92)	
desidererebbe essere più bello di chiunque altro?	0.57
le piacerebbe essere il più intelligente di tutti?	0.568
vorrebbe essere più potente di chiunque altro?	0.56
le piacerebbe essere il più forte di tutti?	0.512
le piace fare shopping?	0.422
le piacerebbe non-invecchiare mai	0.381
le piacerebbe fermare il tempo?	0.372
abbandona facilmente se non è sicuro di ottenere ciò che vuole?	0.306
RAGE (Items = 5, Eigenvalue = 3.91,% explained variance = 3.03, cumulative% explained variance = 25.95)	
non-tollera chi la pensa diversamente da lei?	0.516
è intollerante nei confronti di chi è diverso da lei?	0.51
si spazientisce quando gli altri non-sono d'accordo con lei?	0.485
impone agli altri il suo modo di fare le cose?	0.427
è molto fortunato/a	0.339
FEAR (Items = 7, Eigenvalue = 3.51,% explained variance = 2.72, cumulative% explained variance = 28.67)	
Tende a risparmiare molto?	0.436
tende a nascondere le sue emozioni?	0.427
ha difficoltà ad aprirsi con gli amici?	0.424
riflette a lungo su ciò che è giusto e ciò che è sbagliato?	0.383
riflette intensamente prima di decidere?	0.368
tende generalmente a risparmiare denaro?	0.342
Mantiene il controllo delle sue emozioni?	0.307

The third component is composed of 13 items in which the "yes" answers describe pathological traits determined by the hypoactivity of the seeking system. The people described in these items are basically incapable of activating themselves to satisfy their desires and feel life as a strenuous physical effort.

The fourth component is composed of 8 items in which the "yes" answers describe pathological traits determined by the hypersensitivity of the fear system. The people described by these items are continuously in a state of anxiety and defense from dangers. The fifth component is composed of 10 items in which the "yes" answers describe pathological traits determined by the hyperactivity of the rage system. The sixth component is composed of 5 items in which the "yes" answers describe pathological traits determined by the hyposensitivity of the play system. The people described by these items are incapable of adequate socialization. The seventh component

TABLE 6 | PCA loadings – Area of psychopathological traits in the pool of selected items (40.78% of explained variance – Kaiser-Meyer-Olkin = 0.941).

PANIC (Items = 17, Eigenvalue = 6.47,% explained variance = 8.29)	
le persone non le sono amiche?	0.657
si accorge che gli altri la guardano e/o parlano male di lei?	0.621
gli altri non-apprezzano il suo lavoro?	0.609
ha un senso di fastidio se gente la guarda?	0.547
ha idee che nessuno condivide?	0.541
si sente incompreso?	0.54
ha l'impressione che gli altri si approfittino di lei?	0.539
è sensibile alle critiche e alle offese	0.511
ha scarsa fiducia negli altri?	0.496
critica facilmente la gente?	0.484
si sente inferiore agli altri o inadeguato?	0.457
è imbarazzato in presenza di altre persone?	0.47
Pensa che alcune persone sono responsabili dei malesseri che prova?	0.417
si sente a disagio quando è in compagnia?	0.406
litiga spesso con le persone?	0.402
è timido verso le persone di sesso opposto?	0.39
pensa di stare scontando una pena?	0.315
CARE (Items = 12, Eigenvalue = 5.49,% explained variance = 7.04, cumulative% explained variance = 15.33)	
è incapace di portare a termine un compito?	0.559
si sente la mente vuota?	0.551
ogni cosa le richiede uno sforzo?	0.546
trascura cose importanti della sua vita?	0.516
si sente inutile?	0.486
ha problemi di memoria?	0.485
ha difficoltà a prendere decisioni?	0.475
ha scarsi interessi?	0.44
si sente senza speranza?	0.438
ritiene di dover sempre finire ciò che ha iniziato?	0.431
si colpevolizza facilmente?	0.426
si sente lontano dalle altre persone?	0.383
SEEK (Items = 13, Eigenvalue = 4.92,% explained variance = 6.31, cumulative% explained variance = 21.6)	
ha dolori muscolari?	0.667
si sente fisicamente debole?	0.661
soffre di mal di schiena?	0.612
ha gli arti appesantiti?	0.598
ha nausea o mal di stomaco?	0.561
si affatica facilmente?	0.478
ha palpitazioni o cuore in gola?	0.476
passa rapidamente da sensazioni di freddo a sensazioni di caldo?	0.468
ha un nodo alla gola?	0.464
affatica facilmente?	0.458
le capita di sentirsi venir meno?	0.432
pensa di avere una grave malattia fisica o mentale?	0.4
soffre di cefalea?	0.396
FEAR (Items = 8, Eigenvalue = 4.85,% explained variance = 6.22, cumulative% explained variance = 27.86)	
evita alcuni oggetti. situazioni o luoghi perché la spaventano?	0.709
ha paura di viaggiare su un mezzo di trasporto	0.671
ha paura di uscire da solo?	0.643
ha dei momenti di terrore o panico	0.582
ha paura di tutto senza un valido motivo?	0.536
ha paura?	0.525
si sente a disagio tra la folla?	0.488
è a disagio quando è solo?	0.414
RAGE (Items = 10, Eigenvalue = 4.02,% explained variance = 5.15, cumulative% explained variance = 33)	
sente l'impulso di distruggere le cose?	0.628
si arrabbia tanto?	0.556
sente l'impulso a colpire o a far male a qualcuno?	0.523
rompe oggetti e grida facilmente?	0.501
ha dei pensieri che non-sono suoi?	0.495
alcune persone controllano i suoi pensieri?	0.489
pensa al suicidio?	0.475
sente voci o rumori che altri non-sono in grado di sentire?	0.465
alcune persone percepiscono il suo pensiero	0.44
ha la sensazione di essere preso in trappola?	0.415
PLAY (Items = 5, Eigenvalue = 3.17,% explained variance = 4.07, cumulative% explained variance = 37.8)	
è insofferente e irritato?	0.594
è una persona nervosa?	0.47
si preoccupa facilmente per qualsiasi cosa?	0.434
si sente triste?	0.418
è teso o sulle spine?	0.401
LUST (Items = 3, Eigenvalue = 2.49,% explained variance = 3.19, cumulative% explained variance = 40.27)	
ha scarso appetito?	0.578
piange facilmente?	0.438
si sente solo anche se è in compagnia di altre persone?	0.402

(Continued)

TABLE 6 | Continued

consists of 3 items and describes people with a hyperactivity of the pleasure system.

Table 7 shows the area of dissociative phenomena composed of a total of 15 items selected from the neural network divided into three components. Although three components have emerged, composed of 6, 5, and 4 items respectively, they describe depersonalization/disorganization and dissociative amnesia, two typical ways of altering the cognitive functions produced by the intrusion of emotionality into conscious experiences.

The tables mentioned above present the items in the Italian language, as the original and only language of the diagnostic scale is Italian. For the benefit of not Italian speakers, an English translation of the selected items is provided in the **Supplementary Appendix**. However, it should be noted that the English version provided has never been validated nor used with English speaking subjects and it is only intended as language aid. Moreover, the items presented and translated do not sum up to 167, as previously indicated, as 6 of them did not load on any component and were discarded.

In addition to the strict consistency of the components with theoretical reference model, the items that make up the components of each of the three areas have a marked internal consistency as documented by a Cronbach α value of 0.900 for the area of emotional characteristics, of 0.889 for the area of dissociative phenomena and alfa value of 0.953 for the area of psychopathological traits.

Only 2 out of the 78 items in the group called "emotional characteristics" have negative loadings, 32 have very low loadings

TABLE 7 | PCA loadings – Area of dissociative phenomena in the pool of selected items (40.55% of explained variance – Kaiser-Meyer-Olkin = 0.931).

DEPERSONALIZATION (Items = 6, Eigenvalue = 3.7,% explained variance = 15.41)	
le capita di vedere il modo come se fosse attraverso un vetro?	0.727
le capita di sentirsi una persona diversa da come normalmente è?	0.629
le è capitato di sentire come sconosciuti i luoghi che le sono familiari?	0.617
ha mai sentito i suoi sogni come se fossero reali?	0.607
le è capitato di non-riconoscere la sua immagine allo specchio?	0.585
sente nella testa voci che commentano i suoi pensieri eι/o le dicono cosa fare?	0.511
A**MNESIA (Items = 5, Eigenvalue = 3.63,% explained variance = 15.11, cumulative% explained variance = 30.52)**	
le capita di non-sapere se ha fatto una cosa o se ha solo pensato di farla?	0.761
si è accorto aver fatto cose che non-ricordava di aver fatto?	0.721
le capita di possedere oggetti che non-ricorda di aver acquistato?	0.497
le è capitato di rivivere eventi già vissuti?	0.436
incontra persone che la conoscono ma che lei non-riconosce?	0.409
IMAGINATIVE ABSORPTION (Items = 4, Eigenvalue = 2.41,% explained variance = 10.03, cumulative% explained variance = 40.55)	
le è capitato di non-riconoscere persone che le sono familiari?	0.785
le capita di accorgersi di essersi vestito senza ricordarsi di averlo fatto?	0.563
ha dimenticato eventi importanti nella sua vita?	0.561
le capita di trovarsi in luoghi che non-ricorda di aver raggiunto?	0.361

and 26 low loadings. No items have high values in more than one component.

One out of the 51 items in the group called "psychopathological traits" is negative, only 4 have very low loadings and 11 low loadings. Only 2 out of the 15 items in the group called "dissociative phenomena" have low loadings.

CONCLUSION

In this work, we have presented a procedure that aims at combining explanatory and predictive modeling for the construction of novel psychometric questionnaires based on psychological and neuroscientific theoretical grounding, especially with regards to the aspect of the item selection, in a way that considers both the explanatory power of the theory and the predictive power of modern computational techniques. Such combination allows deriving theoretical insights on the characteristics of the items selected and their conformity with the theoretical framework of reference. At the same time, it permits the selection of items that have the most relevance in terms of prediction by therefore considering the relationship of the items with the actual psychopathological diagnosis, helping to construct a diagnostic tool that both conforms with the theory and with the individual characteristics of the population at hand, by providing insights on the power of the scale in precisely identifying out-of-sample pathological subjects.

The proposed randomized machine learning procedure selected a set of items that drastically improved the prediction accuracy of the model, compared to the predictions obtained using all the original items. At the same time, it reduced the redundancy of the items and eliminated those with less consistency.

Moreover, comparing the results obtained applying PCA on all the items and the results obtained using only the set of items selected by the ANN, clear differences emerge in the distribution and consistency of the items among the different components. The hypothesized latent structure is indeed only partially confirmed by the analysis of all items of the test. However, on the group of selected items, it clearly emerges a greater coherence in the components obtained by the PCA, better confirming the hypothesized latent structure.

The methodology exploits the relationships and the inner consistency that link the theoretical assumptions and the experience of the psychopathology, by showing that focusing on the prediction of the diagnosis and the pathology phenomena can also help to support the explanatory modeling of those phenomena.

By looking at the relationship between the items selected by the procedure and the proposed theoretical framework, by following the psychopathological model identified, it is consistent that some systems produce adaptation problems if they are hyperactive (for example the panic systems of fear and anger produce malaise only if they are active) and other systems are maladaptive if hypoactive. Such dynamic is captured by the "yes" or "no" answers within the questionnaire.

The components that emerge from the group of "emotional characteristics" describe maladaptive processes that are expressed at a non-verbal level of consciousness and do not require the intervention of cortical functions of judgment or conscious evaluation of events (Solms and Panksepp, 2012). In our opinion they can represent the emotional substrate of personality disorders.

The selected items that belong to the group "psychopathological traits" describe maladaptive phenomena that require the intervention of cognitive evaluation and belong to that group of behaviors, psychic functions, emotional states and contents of thought unanimously considered as psychiatric symptoms. In our opinion, the components that emerged in this group describe the action of the conscious mind on basic emotional states. This group of components can represent the emotional dimension of the psychopathology of mental disorders (Panksepp, 2014).

The items belonging to the group "dissociative phenomena" present three components that describe the destructuring of the self-experience and episodic memory. This psychopathological manifestation is due to traumatic events that can occur in every moment of the person's life acutely and intensely or with less intensity for a very long time (Lanius et al., 2014).

In conclusion, we believe that the present methodology has the potential to offer an approach for the construction of new psychometric scales or the reorganization of existing ones, by focusing on the predictive power of the scale in accordance with observable phenomena, in conjunction with the traditional dimensional approach that characterizes many modern psychometric tools.

In the exemplar case presented in this work, we are aware that additional investigations are required for a compelling validation of the proposed psychometric questionnaire, to demonstrate its robustness further and support its use in real psychodiagnostic settings. At the same time, the methodology could be likewise applied to the restructuring of existing and already validated psychometric scales. This work, envisioned for the future, might further support the validity of such methodology. Moreover, we will try to combine predictive and validity metrics in a unified procedure to balance the validity and predictive performance of models, toward the definition of prediction-based validity principls and tools. Nevertheless, we believe that its application to scale constructions, as in the present case, might already demonstrate the potential of the proposed approach.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethical Committee of Psychological Research of the Department of Humanistic Studies, University of Naples Federico II. The patients/participants provided their written informed consent to participate in this study. Patients agreed to participate in the study, then signed the consent form currently required by Italian law and were informed that the data collected would be treated anonymously and would not result in changes in their diagnostic and therapeutic course of treatment.

AUTHOR CONTRIBUTIONS

PD: methodology design, methodology implementation, running experimentation, and writing technical sections of the manuscript and results. DM: methodology design, contribution on methodology implementation, data analysis, running experimentation, and contribution in writing technical sections of the manuscript and results. MM: design and implementation of psychodiagnostic tool, contribution in methodology design, contribution in data collection, and contribution in writing psychological sections of the manuscript. RS: design and implementation of psychodiagnostic tool, contribution in methodology design, data collection, data analysis, and writing psychological sections of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00446/full#supplementary-material

REFERENCES

Azzolina, D., Baldi, I., Barbati, G., Berchialla, P., Bottigliengo, D., and Bucci, A. (2019). Machine learning in clinical and epidemiological research: isn't it time for biostatisticians to work on it? *Epidemiol. Biostatist. Public Health* 16:e13245. doi: 10.2427/13245

Benzécri, J.-P. (1979). Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire: addendum et erratum à (Bin. Mult.). *Les Cahiers l'Analy. Données* 4, 377–378.

Breiman, L. (2001). Statistical modeling: the two cultures. *Statist. Sci.* 16, 199–215.

Bzdok, D., and Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: opportunities and challenges. *Biol. Psychiatr.* 3, 223–230. doi: 10.1016/j.bpsc.2017.11.007

Clynes, M., and Panksepp, J. (2013). *Emotions and Psychopathology*. Berlin: Springer.

Du, S., Tao, Y., and Martinez, A. M. (2014). Compound facial expressions of emotion. *Proc. Natl. Acad. Sci. U.S.A.* 111, 1454–1462.

Dwyer, D. B., Falkai, P., and Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Ann. Rev. Clin. Psychol.* 14, 91–118. doi: 10.1146/annurev-clinpsy-032816-045037

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53, 325–338. doi: 10.1093/biomet/53.3-4.325

Greenacre, M., and Blasius, J. (eds) (2006). *Multiple Correspondence Analysis and Related Methods*. New York, NY: Chapman and Hall/CRC, doi: 10.1201/9781420011319

Günther, F., and Fritsch, S. (2010). Neuralnet: training of neural networks. *R. J.* 2, 30–38.

Ikemoto, S., and Panksepp, J. (1999). The role of nucleus accumbens dopamine in motivated behavior: a unifying interpretation with special reference to reward-seeking. *Brain Res. Rev.* 31, 6–41. doi: 10.1016/s0165-0173(99)00023-5

Jolliffe, I. (2002). *Principal Component Analysis*, 2nd Edn, New York, NY: Springer.

Lane, S. P., and Sher, K. J. (2015). Limits of current approaches to diagnosis severity based on criterion counts: an example with DSM-5 alcohol use disorder. *Clin. Psychol. Sci.* 3, 819–835. doi: 10.1177/2167702614553026

Lanius, U. F., Paulsen, S. L., and Corrigan, F. M. (eds) (2014). *Neurobiology and Treatment Of Traumatic Dissociation: Towards an Embodied Self*. Berlin: Springer Publishing Company.

Lauro, C. (2019). *Prolegomena to Any Future Statistics, That Will Be Able to Present Itself as a (Data) Science*. Manchester: Manchester University Press. Available online at: https://www.linkedin.com/pulse/definition-data-science-carlo-lauro/

Lebart, L., Morineau, A., and Piron, N. (1995). *Statistique Exploratoire Multidimensionelle*. Paris: Dunod.

LeDoux, J. E., and Hofmann, S. G. (2018). The subjective experience of emotion: a fearful view. *Curr. Opin. Behav. Sci.* 19, 67–72. doi: 10.1016/j.cobeha.2017.09.011

Maldonato, N. M., Sperandeo, R., Moretto, E., and Dell'Orco, S. (2018). A non-linear predictive model of borderline personality disorder based on multilayer perceptron. *Front. Psychol.* 9:447. doi: 10.3389/fpsyg.2018.00447

Messas, G., Tamelini, M., Mancini, M., and Stanghellini, G. (2018). New perspectives in phenomenological psychopathology: its use in psychiatric treatment. *Front. Psychiatr.* 9:466. doi: 10.3389/fpsyt.2018.00466

Montag, C., Widenhorn-Müller, K., Panksepp, J., and Kiefer, M. (2017). Individual differences in Affective Neuroscience Personality Scale (ANPS) primary emotional traits and depressive tendencies. *Comprehens. Psychiatr.* 73, 136–142. doi: 10.1016/j.comppsych.2016.11.007

Panksepp, J. (2014). The fundamental substrates of human emotions. *About Body* 47, 383–393.

Panksepp, J., and Biven, L. (2012). *The Archaeology Of Mind: Neuroevolutionary Origins Of Human Emotion*. New York, NY: Norton.

Panksepp, J., Lane, R. D., Solms, M., and Smith, R. (2017). Reconciling cognitive and affective neuroscience perspectives on the brain basis of emotional experience. *Neurosci. Biobehav. Rev.* 76, 187–215. doi: 10.1016/j.neubiorev.2016.09.010

Pessa, E. (2004). *Statistica Con le Reti Neurali*. Roma: Di Renzo.

Riedmiller, M. (1994). Advanced supervised learning in multi-layer perceptrons - From backpropagation to adaptive learning algorithms. *Comput. Stand. Interf.* 16, 265–278. doi: 10.1016/0920-5489(94)90017-5

Sarstedt, M., Ringle, C., Henseler, J., and Hair, J. (2014). On the emancipation of PLS-SEM: a commentary on rigdon (2012). *Long Range Plan* 47, 154–160. doi: 10.1016/j.lrp.2014.02.007

Sharma, P. N., Sarstedt, M., Shmueli, G., Kim, K. H., and Thiele, K. O. (2019). PLS-based model selection: the role of alternative explanations in information systems research. *J. Assoc. Inform. Syst.* 20:4.

Shmueli, G. (2010). To explain or to predict? *Statist. Sci.* 25, 289–310. doi: 10.1214/10-sts330

Solms, M., and Panksepp, J. (2012). The "Id" knows more than the "Ego" admits: neuropsychoanalytic and primal consciousness perspectives on the interface between affective and cognitive neuroscience. *Brain Sci.* 2, 147–175. doi: 10.3390/brainsci2020147

Somma, A., Borroni, S., Maffei, C., Besson, E., Garbini, A., Granozio, S., et al. (2017). Inter-rater reliability of the Italian translation of the structured clinical interview for DSM-5 personality disorders (SCID-5-PD): a study on consecutively admitted clinical adult participants. *J. Psychopathol.* 23, 105–111.

Sperandeo, R., Maldonato, M., Moretto, E., and Dell'Orco, S. (2019). "Executive functions and personality from a systemic-ecological perspective," in *Cognitive Infocommunications, Theory and Applications*, Vol. 13, *Topics in Intelligent Engineering and Informatics*, eds R. Klempous, J. Nikodem, and P. Baranyi (Cham: Springer).

Sperandeo, R., Monda, V., Messina, G., Carotenuto, M., Maldonato, N. M., Moretto, E., et al. (2018a). Brain functional integration: an epidemiologic study on stress-producing dissociative phenomena. *Neuropsychiatr. Dis. Treat.* 14:11. doi: 10.2147/NDT.S146250

Sperandeo, R., Moretto, E., Iennaco, D., and Maldonato, N. M. (2018b). "Using complex networks to model, simulate and understand the dynamics of psychotherapeutic processes: an experimental study proposal," in *Proceedingsa of the 2018 9th IEEE International Conference on Cognitive Infocommunications*, (Piscataway, NJ: IEEE).

Stanghellini, G. (2019). *The Oxford Handbook of Phenomenological Psychopathology*. Oxford: Oxford University Press.

Stanghellini, G., Aragona, M., Doerr-Zegers, O., Musalek, M., and Madeira, L. (2016). Phenomenology of emotions. *Eur. Psychiatr.* 33:S41.

Trull, T. J., Lane, S. P., Koval, P., and Ebner-Priemer, U. W. (2015). Affective dynamics in psychopathology. *Emot. Rev.* 7, 355–361.

Van Buuren, S., and Groothuis-Oudshoorn, K. (2011). MICE: multivariate imputation by chained equations in R. *J. Statist. Softw.* 45, 1–67.

Wold, H. (1985). "Partial least squares," in *Encyclopedia of Statistical Sciences*, eds S. Kotz, and N. L. Johnson, (New York, NY: Wiley), 581–591.

Woods, K., and Bowyer, K. W. (1997). Generating ROC curves for artificial neural networks. *IEEE Trans. Med. Imaging* 16, 329–333.

Yarkoni, T., and Westfall, J. (2017). Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* 12, 1100–1122. doi: 10.1177/1745691617693393

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 28 April 2020
doi: 10.3389/fpsyg.2020.00663

Doping Use in High-School Students: Measuring Attitudes, Self-Efficacy, and Moral Disengagement Across Genders and Countries

Laura Girelli[1], Elisa Cavicchiolo[2]*, Fabio Alivernini[2], Sara Manganelli[2], Andrea Chirico[3], Federica Galli[3], Mauro Cozzolino[1] and Fabio Lucidi[3]

[1] Department of Human, Philosophical, and Educational Sciences, University of Salerno, Salerno, Italy, [2] National Institute for the Evaluation of the Education System (INVALSI), Rome, Italy, [3] Department of Social and Developmental Psychology, Sapienza University of Rome, Rome, Italy

OPEN ACCESS

Edited by:
Pietro Cipresso,
Italian Auxological Institute (IRCCS),
Italy

Reviewed by:
Carlo Chiorri,
University of Genoa, Italy
James Soland,
University of Virginia, United States

***Correspondence:**
Elisa Cavicchiolo
elisa.cavicchiolo@invalsi.it

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: 23 December 2019
Accepted: 19 March 2020
Published: 28 April 2020

Citation:
Girelli L, Cavicchiolo E, Alivernini F, Manganelli S, Chirico A, Galli F, Cozzolino M and Lucidi F (2020) Doping Use in High-School Students: Measuring Attitudes, Self-Efficacy, and Moral Disengagement Across Genders and Countries. Front. Psychol. 11:663. doi: 10.3389/fpsyg.2020.00663

The main aim of this research was to test the factorial validity and measurement invariance across genders and countries of a set of instruments designed to assess high-school students' attitudes, self-regulatory efficacy, and moral disengagement with regard to doping. A second aim was to examine the criterion and predictive validity of these scales. In total, 402 high-school students from Italy, Romania, and Turkey (40.0, 25.1, and 34.9%, respectively; *M* age 14.78 years old; SD = 1.04; 52.8% females) completed questionnaires measuring attitudes toward doping, self-regulatory efficacy in refraining from doping, doping-specific moral disengagement, and intention to use doping substances. A confirmatory factor analysis (CFA) supported our expectations with regard to the factor structure of the scales. Multigroup CFAs provided evidence for the full equivalence of the measures across males and females and partial equivalence of the measures across the three countries. The results of the latent mean comparison showed that male students had lower levels of self-regulatory efficacy than females and that Romanian and Turkish students had higher levels of moral disengagement and lower level of self-regulatory efficacy than Italian students. Finally, the results of a structural equation modeling supported the hypothesis that the proposed model predicted students' intentions to use doping, thus generally confirming the criterion and the predictive validity of the measures. These findings suggested the validity of a set of instruments measuring attitudes toward doping, self-regulatory efficacy to refrain from doping, and doping-specific moral disengagement in high-school students from a cross-gender and a cross-cultural perspective and provided meaningful estimates of the differences in the three factors between males and females as well as between Italian, Romanian, and Turkish high-school students.

Keywords: high-school students, self-regulatory efficacy, moral disengagement, measurement invariance, gender differences, cross-cultural differences

INTRODUCTION

The use of doping is recognized as a relevant issue in sport. A growing body of literature indicates that not only elite athletes use and abuse doping substances but also those who engage in amateur and recreational sports (Dunn and White, 2011) sometimes to an even greater extent than professional athletes (Wanjek et al., 2007). For this reason, doping has been identified as a rising public health problem. Furthermore, a rise in doping has been detected in the young, whether they are athletes or not (LaBotz and Griesemer, 2016), a tendency which is becoming apparent at increasingly young ages (LaBotz and Griesemer, 2016; Nicholls et al., 2017). The term "doping" generally indicates the use of illegal performance- and appearance-enhancing substances (PAES; Mallia et al., 2013), but several studies in reference to various different sports have demonstrated the extensive consumption of legal PAES, such as proteins, amino acids, creatine, etc. (Bell et al., 2004). Even though these substances are legal, they may act as a "gateway" to doping practices (Lucidi et al., 2017). Also the use of legal PAES appears to be increasing among young people (Bell et al., 2004; Hoffman et al., 2008; LaBotz and Griesemer, 2016). For example, a study conducted in the US revealed that, among high-school students, 38.8% of boys and 18.2% of girls reported protein supplement use; furthermore, although students who regularly practiced sports used these substances more frequently, also 18.2% of other students frequently consumed protein supplements (Eisenberg et al., 2012). Therefore, for both legal and illegal PAES, high-school students are one of the groups that are more at risk (Dodge and Hoagland, 2011; Dunn and White, 2011; Eisenberg et al., 2012).

A growing body of research has investigated the factors that affect the use of doping in athletes and non-athletes (Lucidi et al., 2008; Petróczi and Strauss, 2015; Mallia et al., 2016; Kavussanu et al., 2019). A recent meta-analysis identified positive attitudes toward doping, morality, and self-efficacy to resist from doping as being some of the strongest psychological predictors of doping intentions and behaviors (Ntoumanis et al., 2014). Although these factors have been extensively investigated in the context of doping in competitive sport, so far, no studies have specifically examined the measurement invariance of these scales across genders and across countries in a population of non-athletes. Valid instruments, which are equivalent across males and females and across countries, are essential in order to make the prediction of doping intention and behavior more accurate. In the subsequent paragraphs, we will define the constructs under examination. We will then outline the importance of possessing valid instruments for measuring them in order to facilitate doping prevention.

According to the Theory of Planned Behavior (TPB; Ajzen, 1991), one of the leading socio-cognitive theories, the term "attitude" refers to the degree to which individuals have a favorable or unfavorable evaluation of a behavior. One's attitude toward doping therefore consists of a positive or negative evaluation of its use either for performance enhancement or for esthetic reasons. Research conducted on the basis of the TPB has demonstrated that attitudes toward doping are effective in predicting doping intentions and behavior (Lucidi et al., 2004, 2008; Wiefferink et al., 2008; Goulet et al., 2010; Lazuras et al., 2010; Mallia et al., 2016). This applies to various different groups, such as elite athletes (Lazuras et al., 2010), non-professional athletes (Wiefferink et al., 2008), and students (Lucidi et al., 2008; Zelli et al., 2010; Mallia et al., 2013). These results are therefore generalizable across different populations and contexts.

According to the Social Cognitive Theory (Bandura, 1997), "perceived self-efficacy" refers to the beliefs that individuals hold about their capacity to achieve their personal goals and to overcome difficulties. According to Bandura, self-efficacy must be tailored to the particular domain of functioning or conduct that is being investigated (Bandura, 1997). Thus, as in the context of doping, social normative pressures, such as the influence of significant others, may have a significant role with respect to the use/abuse of illegal substances (Kindlundh et al., 1999), individuals' beliefs about their own ability to resist to them are fundamental. Therefore, "doping-specific self-regulatory efficacy" refers to one's ability to resist social pressure toward doping and to avoid or cope with situations in which doping occurs more often. Studies have shown that, in addition to attitudes, self-regulative efficacy toward doping is effective in predicting doping intentions and self-reported doping use (Lucidi et al., 2008; Lazuras et al., 2010; Zelli et al., 2010; Barkoukis et al., 2013; Mallia et al., 2013, 2015).

Within the Social Cognitive Theory (Bandura, 1991), a construct that has recently been receiving increasing attention is that of moral disengagement (Kavussanu et al., 2016), which is a process of convincing oneself that ethical standards do not apply in a particular context, by suspending or deactivating the mechanism of self-condemnation and self-sanction. One's internal moral standards can thus be activated or inhibited by mechanisms of self-justification. In these cases, people may not feel obliged to make decisions that conform to their normal moral standards. Moral disengagement in the context of doping refers to the "self-serving self-regulatory process that allows people to dope while still believing they are acting morally" (Lucidi et al., 2017, pp. 2). It constitutes a moral justification for doping, for example, by comparing it with more extremely inhumane actions or when substance use is not perceived as being under the individual's own control. Many studies have demonstrated that moral disengagement influences adolescents' intention to use doping and doping substances effective use (Lucidi et al., 2008; Kavussanu et al., 2016).

Despite the fact that measures of attitudes, self-regulatory efficacy, and moral disengagement toward doping have been used extensively in research into doping (Kavussanu et al., 2016, 2019; Mallia et al., 2016; Lucidi et al., 2017) and much support has been found for their internal consistency and reliability, no studies have been published that establish the validity of these scales for non-athletes. This is important since various studies have shown that the use of illegal PAES is a widespread issue (Mallia et al., 2013), which is not limited to athletes and is particularly relevant for adolescents (Barkoukis et al., 2016). Moreover, the use of doping substances poses a significant threat in many other area of adolescents' lives, as it has been associated with behaviors that pose a high health risk, such as the abuse of alcohol and illicit drugs (e.g., DuRant et al., 1995; Kindlundh et al., 1999), and it is often seen as having an effect on variables related to young people adjustment, such as academic achievement (DuRant et al., 1995; Kindlundh et al., 1999). It is therefore particularly important to prevent doping in adolescents because at this age, individuals are more susceptible to normative influences (Ntoumanis et al., 2014) and their attitudes are shaped. According to various studies (Smith and Stewart, 2010; Barkoukis et al., 2016), our values with regard to sports, as well as doping, are associated with social norms that shape our attitudes toward the use of doping substances. It is therefore highly recommended to support interventions targeting adolescents' perceptions of sport values, social norms, and attitudes toward PAES use in sports. Interventions which aim to improve adolescents' perceptions of sporting values, social norms, and attitudes toward the use of PAES in sports therefore need to be supported. For this purpose, we need to possess valid instruments for the screening and evaluation of interventions focused on combating doping among young people.

In addition, no study hitherto has analyzed the measurement invariance of these instruments across genders and countries. Doping is influenced by the interplay of several factors, such as personal characteristics and social contexts (e.g., social norms) (Ntoumanis et al., 2014) that can vary for several reasons (e.g., different countries, laws, or individual backgrounds). In order to evaluate and compare anti-doping interventions, it is therefore important to use measures that can be applied regardless of the context or individual-specific characteristics. Limited studies have investigated the validity of several different instruments in a sample of athletes (Kavussanu et al., 2016; Mallia et al., 2016). In particular, Kavussanu et al. (2016) investigated the validity and the measurement invariance across genders of a 12-item measure moral disengagement toward doping which was specifically devised for the context of sport, whereas Mallia et al. (2016) developed and validated measures of self-regulatory efficacy in refraining from doping and moral disengagement toward doping in team contexts, as well as examining the measurement invariance of the scales across different countries. As previously pointed out, the factorial validity of these scales in the population of non-athletes has so far not been analyzed. Furthermore, it is necessary to determine whether doping-related measures of attitudes, self-efficacy, and moral disengagement are valid in terms of their equivalence across males and females and across different cultures. The present study was designed to make up for this lack, and it represents the first assessment of a set of measures of psychosocial determinants of doping use, as well as the first examination of their measurement invariance across males and females and across countries in a sample of non-athletes.

The Present Study

The aim of this research was to test the factorial validity of a set of instruments measuring attitudes, self-efficacy, and moral disengagement toward doping in a sample of non-athletes and to test the measurement invariance of these scales across genders and countries, i.e., to determine the extent to which individuals of different groups interpret the items of a measure in an equivalent way (Cheung and Rensvold, 2002), a factor that is essential in order to reliably compare these groups. When there is a satisfactory level of measurement invariance, any differences that are detected between the groups reflect genuine differences in the variables and rather than variations in the responses that are merely due to a different interpretation or understanding of the items in the questionnaires. Previous research has indicated the presence of gender differences in attitudes, self-regulatory efficacy, and moral disengagement in several populations, including high-school students (Lucidi et al., 2008) and team sport athletes (Boardley and Kavussanu, 2008). Typically, females report less positive attitudes toward doping, as well as higher levels of self-regulatory efficacy and of moral disengagement than males (Lucidi et al., 2008). Since gender differences are common regarding these factors, it is fundamental to investigate whether the differences measured are simply due to different interpretations of the items, which means that it is necessary to test the measurement invariance of these scales across males and females. Previous studies have also identified cross-cultural differences in these variables in different populations. For example, Mallia et al. (2016) found that German team-athletes considered themselves more able to resist social pressure with regard to doping use than did the Greek and Italian team-athletes, while the Greeks reported that they lowered their personal moral standards to a greater extent than their Italian and German peers. Hence, in order to understand whether different responses between the groups reflect actual differences in the variables examined and not just differing interpretations of the items, it is also necessary to investigate measurement invariance across countries. For this reason, the present investigation included participants from three different countries: Italy, Romania, and Turkey. By comparing latent means, we closely examined the differences within the two groups—one based on gender, the other on nationality—with regard to the measures used. Finally, in order to examine the criterion and the predictive validity of the scales, the present study also examined the associations of the measures of attitudes, self-efficacy, and moral disengagement to participants' reported intentions to use doping substances in the near future.

MATERIALS AND METHODS

Participants and Procedures

The participants were high-school students (n = 402; M age 14.78 years, SD = 1.04; age range 14–18 years; 52.8% females) who participated in a project designed to promote life skills in young people in order to promote health. The project involved students from three European countries: Italy (n = 161; 40%; M age = 13.89 years; SD = 0.88), Romania (n = 101; 25.1%; M age: 15.28 years, SD = 0.72) and Turkey (n = 140; 34.8%; M age = 15.47 years, SD = 0.53). In accordance with ethical guidelines, a description of the study with a consent form to be signed was sent to the students' parents and voluntary participation was only requested from those students whose parents had given their consent. All of the participants were informed about confidentiality and anonymity and their right to withdraw from the study at any time. The participants were invited to complete an online questionnaire lasting about 20 min during classroom hours. All of the questionnaires were anonymous.

Measures

The questionnaire, based on the instruments developed by Lucidi et al. (2008), included measures of attitudes toward doping, doping-specific self-regulatory efficacy, moral disengagement and intention to use doping substances. As the original versions of the scales were in Italian, all the scales were translated from Italian into Romanian and Turkish using the standardized back translation procedures (Hambleton and Patsula, 1998). Thus, the questionnaire, which had already been translated into Romanian and Turkish by a professional translator, was translated back into the original language (Italian) by another expert translator in order to ensure that the original meaning of the questions had not been changed in any way. Descriptive statistics and reliability coefficients for all the measures used in the study are reported in **Table 1**. The measures in Italian, Romanian, Turkish, and English are included in the **Supplementary Materials**.

Attitudes Toward Doping

The respondents' attitudes toward doping were measured by five items, with responses on five-point semantic differential scales with the bipolar adjectives: "*useless/useful*," "*foolish/wise*," "*undesirable/desirable*," "*negative/positive*," "*harmful/beneficial*," asking participants to express the extent to which the "use of illegal substances to improve sporting performance or physical appearance would be for you. . ."

Doping-Specific Self-Regulatory Efficacy

Self-regulatory efficacy toward doping was measured by six items referring to the extent to which participants felt confident in avoiding or coping with situations or circumstances in which doping use is more likely on a 5-point Likert-type scale ranging from *not at all capable* (1) to *completely capable* (5).

TABLE 1 | Descriptive statistics, reliability, and zero-order correlations among all the key variables of the study.

	Mean (SD)	Cronbach's alpha	Zero-order correlations			
			(1)	(2)	(3)	(4)
(1) Attitudes toward doping	1.89 (0.99)	0.82				
(2) Self-regulatory efficacy	5.45 (1.62)	0.85	−0.40**			
(3) Moral disengagement toward doping	2.00 (0.87)	0.72	0.34**	−0.32**		
(4) Intention to use doping substances	1.48 (0.84)	0.82	0.58**	−0.38**	0.41**	

***p < 0.01.*

Doping Moral Disengagement

The participants' moral disengagement was measured by six items addressing the mechanisms of moral disengagement that are relevant to doping. For example, the item "compared to the damaging effects of alcohol and tobacco, the use of illicit substances is not so bad" refers to the mechanism of advantageous comparison, while the items "it is not right to condemn those who use illicit substances to improve their body, since many do so" refers to the mechanism of displacing or diffusing responsibility. No items measured the attribution of blame or dehumanization as these processes are not pertinent in the field of doping research (Lucidi et al., 2008). For each item, students rated their agreement on a five-point Likert scale ranging from *I do not agree at all* (1) to *I completely agree* (7).

Intention

Intention was assessed by three items measuring the likelihood of using doping substances in the following 3 months (i.e., "How strong is your intention to use illegal substances to improve your sporting performance or your physical appearance in the next 3 months?"). The responses were recorded on a 5-point Likert scale ranging from *not strong at all* (1) to *very strong* (5).

Analyses

Testing the Measurement Model

Confirmatory factor analyses (CFA) of attitudes, self-regulatory efficacy, and moral disengagement were initially conducted using MPLUS software (Version 8; Muthén and Muthén, 2017). An initial CFA was conducted in order to examine the hypothesis that each set of items measured only one latent factor (i.e., the model implied three factors: attitude, doping-specific self-regulatory efficacy, and moral disengagement) and that the three factors were correlated with each other. The measurement model is displayed in **Figure 1**. Model parameters were estimated using the maximum likelihood (ML) estimation method, and the quality of the measurement model was examined through multiple fit indices: comparative fit index (CFI), Tucker–Lewis index (TLI), root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR) (Hu and Bentler, 1999), and chi-square (χ^2)/df ratio (Tabachnick and Fidell, 2006). Cutoff values of 0.90 or above for the CFI were considered indicative of adequate model fit (Marsh et al., 2004).Values of 0.08 or less for the RMSEA and the SRMR were deemed satisfactory for well-fitting models (Marsh et al., 2004). A value of two or less for the χ^2/df ratio is considered a good indicator of model fit (Tabachnick and Fidell, 2006); however, Kline (1998) suggested that a χ^2/df ratio of three or less is also a reasonably good indicator of model fit.

Evaluating Measurement Invariance Across Genders and Countries and Estimating Latent Mean Differences

Subsequently, in order to test the hypothesis of measurement invariance across genders and countries (i.e., that measurement structure of the instruments applies equally well to males and females, as well as to each country), a second series of CFAs of the

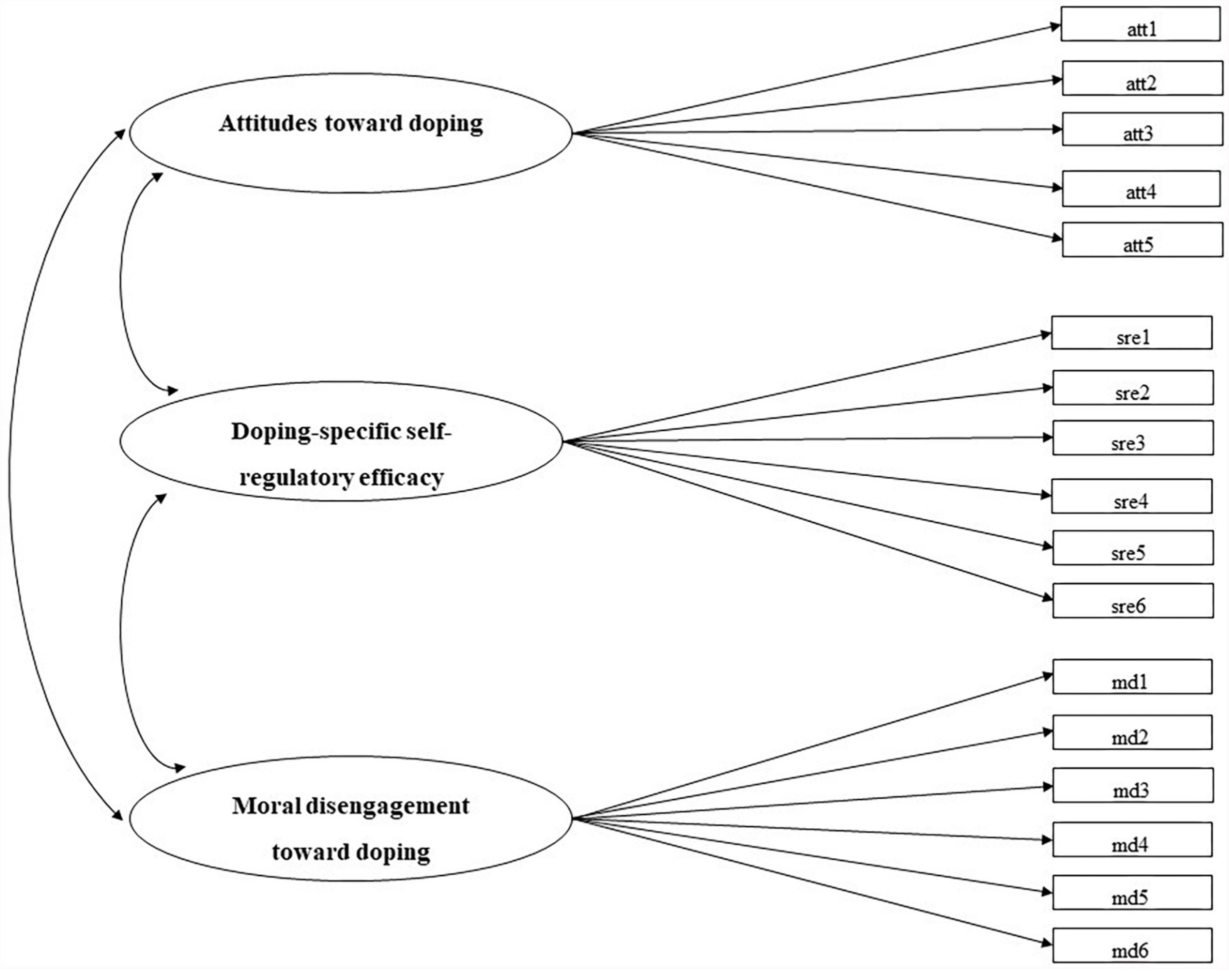


FIGURE 1 | The measurement model for the three instruments used in the study.

model were performed. In line with the literature (Byrne, 2008), these multigroup CFAs tested the configural equivalence (to ascertain that the number of factors and their loading pattern are invariant across groups), the measurement or metric equivalence (to ascertain that all the factor loadings are invariant across groups), and the scalar equivalence (to ascertain that all the item intercepts are invariant across groups). The fit of the models was compared using the change in CFI values (ΔCFI $\leq$ 0.01) according to Cheung and Rensvold (2002). In addition, in order to test differences with regard to the factors considered in the study, latent mean differences were estimated across genders and countries for each of the three factors—attitudes, self-regulatory efficacy, and moral disengagement—by fixing the latent factor means for one group (i.e., the reference group) to zero and freely estimating the latent factor means for the other groups. Latent mean differences were estimated separately for genders and countries.

Testing the Structural Equation Model

In order to evaluate the criterion and predictive validity of the measures, we tested the hypothesis that adolescents' attitudes, self-regulatory efficacy, and moral disengagement were predictive of doping intention in a Structural Equation Model (SEM). As for the CFA, model parameters were estimated using the ML estimation method using MPLUS software (Version 8; Muthén and Muthén, 2017), and the quality of the structural model was examined through multiple fit indices: CFI, TLI, RMSEA, SRMR (Marsh et al., 2004), and χ^2/df ratio (Tabachnick and Fidell, 2006). The same cutoff values used for the evaluation of the model fit of the CFA were used for the SEM.

RESULTS

Descriptive Statistics, Reliability, and Correlations

Table 1 shows descriptive statistics, reliability, and zero-order correlations of all the key variables of the study for the total sample. Cronbach's alpha coefficients indicated an acceptable to good reliability for the scores on all of the scales. Zero-order correlations showed that doping intention was positively correlated with attitudes and with moral disengagement toward doping and negatively correlated with doping-specific self-regulatory efficacy.

The Measurement Invariance of the Scales Across Genders and Countries

Configural, metric, and scalar measurement invariance of the model was tested across genders and countries. **Table 2** shows the goodness-of-fit indexes for all the models tested. With respect to measurement invariance across genders, the comparison of the configural invariance model with the metric invariance model showed that the difference in the CFI was smaller than the cutoff criterion (ΔCFI = 0.009), providing support for the metric invariance of the scales across genders. Furthermore, the comparison of the metric invariance model with the scalar invariance model provided support for the full scalar invariance of the scales across genders (ΔCFI = 0.005). With regard to measurement invariance across countries, the comparison of the configural invariance model with the metric invariance model confirmed the metric invariance of the scales across Italy, Romania, and Turkey as the difference in the CFI was smaller than the cutoff criterion (ΔCFI = 0.005). However, the comparison between the scalar invariance model and the metric invariance model indicated that there was not a complete scalar equivalence. In fact, when the model was revised to include the constraints of the item intercepts, the difference in the CFI was bigger than the cutoff criterion. When we examined the modification indices obtained, we found that for six of the items, the intercepts were not statistically equivalent across countries. Accordingly, the equality constraints for these items were released (i.e., the item intercepts were freely estimated) and a second multigroup CFA was then performed. The results of this analysis suggested a partial scalar equivalence across the three national samples, as indicated by an improved ΔCFI (0.008). **Table 3** shows the standardized factor loadings and the reliability for each group—for males and females and for Italian, Romanian, and Turkish—as well as the items that were non-invariant across the countries.

Differences in Attitudes, Self-Regulatory Efficacy, and Moral Disengagement Between Genders and Countries

Table 4 shows the results of the analysis of the latent mean differences across genders and countries with regard to the factors considered in the study. Male students showed lower levels of self-regulatory efficacy than females, but there were no statistically significant differences between males and females regarding attitudes and moral disengagement, although males tended to have more positive attitudes toward doping (p = 0.08). Both Romanian and Turkish students showed higher levels of moral disengagement and lower levels of self-regulatory efficacy than Italian students. There were no statistically significant differences between the three groups with regard to attitudes, although Romanian students tended to have more positive attitudes toward doping than Italian students, but this finding did not reach a level of statistical significance (p = 0.06).

The Criterion and Predictive Validity of the Three Measures

As displayed in **Figure 2**, the results of the SEM analysis met the multiple criteria for adequate model fit, thus supporting the hypothesis that the set of instruments together predicted students' prospective intentions to use doping. Furthermore, students' moral disengagement and attitudes uniquely and significantly predicted their intentions to use doping substances. These latter estimates went in the expected directions, suggesting that a greater degree of moral disengagement and more positive attitudes would lead to stronger doping intentions. Although the observed scores of self-regulatory efficacy and intention to use doping are statistically significant and negatively correlated (**Table 1**), the path between these two variables seemed to be statistically insignificant in the SEM analyses. Therefore, in order to test whether this was a "suppressor effect" due to intercorrelations between the three predictors, a dominance analysis was computed assessing the relative importance of the three regressors in the linear model, using R (Grömping, 2006). The results of this analysis, which was conducted with the R package *relaimpo* using the metric

TABLE 2 | Measurement invariance across gender and country.

	χ^2	*df*	CFI	TLI	RMSEA	SRMR	χ^2/df	Models compared	ΔCFI
Gender									
Configural invariance	431.527	232	0.908	0.892	0.066	0.060	1.860		
Metric invariance	466.737	246	0.899	0.888	0.067	0.071	1.897	Metric against configural	0.009
Scalar invariance	490.595	260	0.894	0.889	0.067	0.073	1.886	Scalar against metric	0.005
Country									
Configural invariance	524.070	348	0.920	0.906	0.062	0.070	1.505		
Metric invariance	562.137	376	0.915	0.908	0.061	0.077	1.495	Metric against configural	0.005
Scalar invariance	720.767	404	0.856	0.855	0.077	0.093	1.784	Scalar against metric	0.059
Partial scalar invariance	602.178	398	0.907	0.905	0.062	0.079	1.513	Scalar against metric	0.008

CFI, comparative fit index; TLI, Tucker–Lewis index; RMSEA, root mean square error of approximation; SRMR, standardized root mean square residual; df, degrees of freedom.

TABLE 3 | Standardized factor loadings and internal reliability for the three instruments across genders and across the three countries participating in the study (i.e., Italy, Romania, and Turkey).

	Multigroup-CFA factor loadings				
	Gender		Countries		
	F	M	IT	RO	TU
(1) Attitudes toward doping					
The use of illegal substances to improve sporting performance or physical appearance would be for you:					
1. Useless/useful	0.64	0.57	0.49	0.76	0.74
2. Foolish/wise	0.57	0.53	0.47	0.68	0.73
3. Undesirable/desirable	0.73	0.81	0.75	0.67	0.74
4. Negative/positive	0.77	0.80	0.83	0.67	0.82
5. Harmful/beneficial	0.79	0.82	0.80	0.81	0.82
Cronbach's alpha	0.82	0.81	0.74	0.85	0.87
(2) Doping-specific self-regulatory efficacy					
You would be able to resist the temptation to use doping substances					
1. . . .even in the case you have a fall in performance	0.65	0.60	0.61	0.52	0.67
2. . . .to have a physique more appreciated by others, even if nobody will ever know it	0.60	0.67	0.76	0.61	0.62
3. . . .to make your body closer to how you would like it	0.65	0.76	0.75	0.64	0.70
4. . . .to achieve faster results, even if nobody will ever know it	0.76	0.77	0.82	0.65	0.79
5. . . .despite other people suggest me to do it	0.73	0.70	0.75	0.63	0.70
6. . . .to improve in the sport you practice, even if you know that wouldn't have any side effects	0.75	0.76	0.80	0.67	0.74
Cronbach's alpha	0.84	0.86	0.88	0.80	0.85
(3) Moral disengagement toward doping					
How much do you agree with each of these statements?					
1. Compared to the damaging effects of alcohol and tobacco, the use of illicit substances is not so bad	0.57	0.52	0.44	0.52	0.52
2. It is not right to condemn those who use illicit substances to improve their body, since many people do the same	0.50	0.48	0.51	0.49	0.49
3. Doping use is just another good way to "maximize its potential"	0.55	0.47	0.37	0.48	0.53
4. There is no reason to punish people who use illicit substances to improve their physical appearance, after all, no one gets hurt	0.56	0.51	0.49	0.54	0.55
5. People who use illicit substances in sport are not to blame, to blame are those who expect too much from him	0.54	0.53	0.51	0.50	0.55
6. To overcome their own limitation, it is reasonable to use also illicit substances	0.79	0.75	0.77	0.70	0.78
Cronbach's alpha	0.76	0.68	0.65	0.72	0.72

F, female; M, male; IT, Italy; RO, Romania; TU, Turkey; CFA, confirmatory factor analysis. All the factor loadings are statistically significant ($p < 0.001$). The items for which the intercepts were non-invariant between the countries are underlined.

lmg, confirmed the contribution of self-regulatory efficacy to the R^2 increase, therefore confirming the hypothesis of the suppressor effect. Results for the latter analysis are included in the **Supplementary Materials**.

TABLE 4 | Results of the latent factor mean difference tests.

Doping-related constructs	Difference tests		
	Gender	Countries	
	Females versus Males[a]	Italians versus Romanians[b]	Italians versus Turkish[b]
Attitudes	0.158	0.234	−0.110
Self-regulatory efficacy	−0.374*	−0.830***	−0.383*
Moral disengagement	0.041	0.553***	0.412***

*[a] Females are the reference group for the comparison (the latent mean for this group is fixed to be zero). [b] Italians are the reference group for the comparison (the latent mean for this group is fixed to be zero). *$p < 0.05$; ***$p < 0.001$*

DISCUSSION

Pro-doping attitudes, self-regulatory efficacy, and moral disengagement toward doping are self-reported measures that are widely used in the context of doping prevention (Lucidi et al., 2008; Ntoumanis et al., 2014; Kavussanu et al., 2016; Mallia et al., 2016). Those studies conducted hitherto have investigated the psychometric properties of these measures only in the context of sport, and none of them have either analyzed their factorial validity in a sample of non-athletes or tested their measurement invariance across different genders and cultures. The aim of the present study was therefore to test the three-factor structure of the measures in non-athletes, as well as their measurement invariance in males and females and in three different countries: Italy, Romania, and Turkey. Examining whether a set of instruments measuring the determinants of doping is invariant across cultures and genders will allow researchers to properly use and interpret their results. The findings of the CFA provided evidence for the factorial validity of the set of instruments.

Multigroup CFAs conducted on gender and on the three countries supported full configural, metric, and scalar invariance

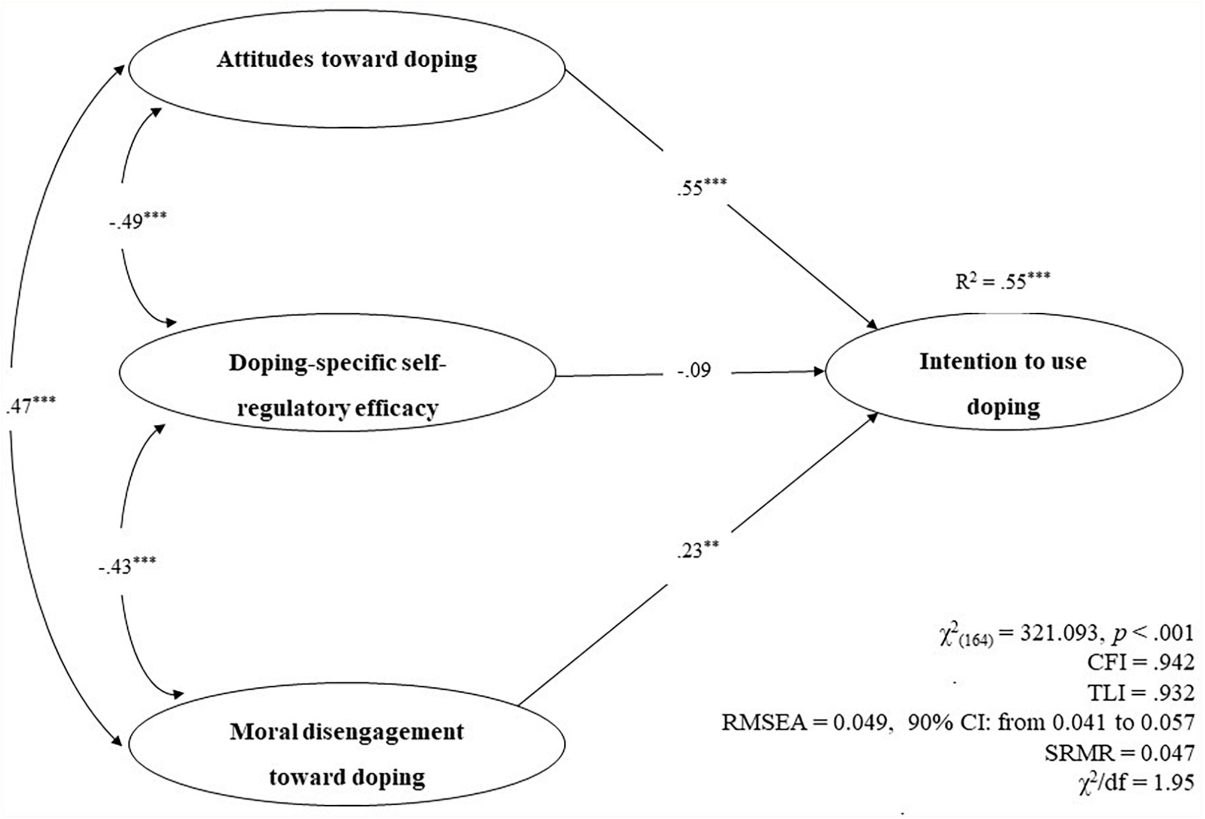

Attitudes toward doping
Doping-specific self-regulatory efficacy
Moral disengagement toward doping
Intention to use doping
R² = .55***
.55***
-.09
.23**
-.49***
-.43***
.47***
χ²(164) = 321.093, p < .001
CFI = .942
TLI = .932
RMSEA = 0.049, 90% CI: from 0.041 to 0.057
SRMR = 0.047
χ²/df = 1.95

FIGURE 2 | The results of the structural equation model (SEM) in which high-school student' attitudes, self-regulatory efficacy and moral disengagement predicted doping intention. ***$p < 0.001$.

between males and females and full configural, metric, and partial scalar invariance between Italian, Romanian, and Turkish high-school students. Although the full measurement invariance across countries was not achieved, findings may allow appropriate cross-group comparisons (Vandenberg and Lance, 2000). These results suggest that for males and females as well as for Italian, Romanian, and Turkish high-school students, the set of measures has the same structure, with most of the items being equally associated with pro-doping attitudes, self-regulatory efficacy, and moral disengagement toward doping. The scalar invariance across countries and genders allowed us to directly compare the latent means. The findings revealed that girls perceived themselves as significantly more efficacious in resisting social pressure to practice doping than did boys, whereas no significant gender differences were found in moral disengagement and pro-doping attitudes. However, boys expressed more positive attitudes toward doping use than girls did, even though this result did not reach a level of statistical significance. Differences between boys and girls are consistent with previous studies on doping conducted with high-school students (i.e., Lucidi et al., 2008) showing that girls are better able to deal with personal or interpersonal pressure than are boys. Regarding differences between countries, the results show that both Romanian and Turkish students had significantly higher levels of moral disengagement and significantly lower levels of self-regulatory efficacy than Italian students. When asked about their level of self-justification of doping conducts and their capability to resist the social pressure that encourage doping, Romanian and Turkish high-school students reported that they lowered their personal moral standards to a greater extent and that they felt higher levels of self-efficacy, as compared to their Italians peers. Mallia et al. (2016) found similar differences between countries, with German athletes considering themselves more able to resist social pressure with regard to doping use than did Greek and Italian athletes, and with Greek athletes reporting that they suspended their personal moral standards to a greater extent than did Italian and German peers. To our knowledge, no previous studies exist that compare Italian, Romanian, and Turkish high-school students concerning these factors. Differences in perceived self-efficacy and moral disengagement toward doping may simply reflect cultural differences and national customs or traditions. Further studies are needed in order to better explain cross-national differences regarding self-efficacy beliefs and moral disengagement.

Finally, the SEM findings of the present study supported the hypotheses that the measures had criterion and predictive validity. However, it is important to note that, contrary to our hypothesis, students' perceptions of their self-regulatory efficacy to resist external pressure toward doping did not uniquely predict adolescents' intentions to practice doping. This finding was probably due to the intercorrelations between self-regulatory efficacy and the other two predictors of intention as the dominance analysis confirmed that self-regulatory efficacy was a significant predictor of intention in a multiple regression model.

It is important to note some limitations of our results. First, the sample size was limited. Second, the study included only participants from Italy, Romania, and Turkey. It would be interesting to replicate this study in a larger sample size and in other countries. Despite these limitations, our results suggest that the instruments examined are reliable and can be used in high-school students to measure their positive attitudes, self-regulatory efficacy, and moral disengagement toward doping.

CONCLUSION

In recent years, interest has been growing in doping research. Various studies have shown that the use of illegal PAES is not limited to athletes (Mallia et al., 2013), especially with regard to adolescents (Barkoukis et al., 2016). Most of the research has identified positive attitudes toward doping, morality, and self-efficacy to resist doping as the strongest psychological predictors of doping intentions and behaviors (Ntoumanis et al., 2014). It is therefore fundamental to have a valid instrument for the screening and evaluation of these constructs in order to prevent doping among young people. Furthermore, doping use is influenced by a combination of factors, such as personal characteristics and social contexts (Ntoumanis et al., 2014), that may be associated with different countries or individual experiences. It is therefore important to adopt measures that can be utilized regardless of the national context or characteristics specific to the individual. Our findings suggest that the measures investigated are invariant across genders and partially invariant across countries, so that the reports of males or females, as well as of Italian, Romanian, and Turkish students on these constructs can be meaningfully compared. It seems safe to conclude that the instruments analyzed can be reliably used with high-school students to measure their positive attitudes toward doping, self-regulatory efficacy, and moral disengagement, thereby helping teachers and health practitioners to predict young people's use of doping or intentions to resort to doping in the future. Finally, the instruments may also be used to measure the effects of specific school-based interventions aimed at preventing the practice of doping (Lucidi et al., 2017).

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standard. All participants were in no risk of physical or emotional pressure. Written informed consent was obtained from all individual participants included in the study and the results were disseminated only anonymously. None of the participants were patients, or persons with disabilities.

AUTHOR CONTRIBUTIONS

LG made the greatest contribution to the manuscript, performing the statistical analyses, drafting the work, and contributing to all the steps of the work. EC revised the statistical analyses. SM, AC, FG, and MC revised the first draft of the manuscript. FL and FA revised the manuscript and monitored all the process providing scientific and theoretical contribution. All authors approved the final version of the manuscript and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

FUNDING

The present study was funded by the Italian Ministry of Health-Antidoping Surveillance Committee (2017-4; CUPB86G17000660005).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00663/full#supplementary-material

REFERENCES

Ajzen, I. (1991). The theory of planned behavior. *Organ. Behav. Hum. Decis. Process.* 50, 179–211. doi: 10.1016/0749-5978(91)90020-T

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organ. Behav. Hum. Decis. Process.* 50, 248–287. doi: 10.1016/0749-5978(91)90022-L

Bandura, A. (1997). *Self-Efficacy: The Exercise of Control*. New York, NY: W H Freeman.

Barkoukis, V., Kartali, K., Lazuras, L., and Tsorbatzoudis, H. (2016). Evaluation of an anti-doping intervention for adolescents: findings from a school-based study. *Sport Manag. Rev.* 19, 23–34. doi: 10.1016/j.smr.2015.12.003

Barkoukis, V., Lazuras, L., Tsorbatzoudis, H., and Rodafinos, A. (2013). Motivational and social cognitive predictors of doping intentions in elite sports: an integrated approach. *Scand. J. Med. Sci. Sport* 23, 330–340. doi: 10.1111/sms.12068

Bell, A., Dorsch, K. D., Mccreary, D. R., and Hovey, R. (2004). A look at nutritional supplement use in adolescents. *J. Adolesc. Heal.* 34, 508–516. doi: 10.1016/j.jadohealth.2003.07.024

Boardley, I. D., and Kavussanu, M. (2008). The moral disengagement in sport scale - Short. *J. Sports Sci.* 26, 1507–1517. doi: 10.1080/02640410802315054

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: a walk through the process. *Psicothema* 20, 872–882.

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Struct. Equ. Model. A Multidiscip. J.* 9, 233–255. doi: 10.1207/S15328007SEM0902

Dodge, T., and Hoagland, M. F. (2011). The use of anabolic androgenic steroids and polypharmacy: a review of the literature. *Drug Alcohol Depend.* 114, 100–109. doi: 10.1016/j.drugalcdep.2010.11.011

Dunn, M., and White, V. (2011). The epidemiology of anabolic-androgenic steroid use among Australian secondary school students. *J. Sci. Med. Sport* 14, 10–14. doi: 10.1016/j.jsams.2010.05.004

DuRant, R. H., Escobedo, L. G., and Heath, G. W. (1995). Anabolic-steroid use, strength training, and multiple drug use among adolescents in the United States. *Pediatrics* 96, 23–28.

Eisenberg, M. E., Wall, M., and Neumark-Sztainer, D. (2012). Muscle-enhancing behaviors among adolescent girls and boys. *Pediatrics* 130, 1019–1026. doi: 10.1542/peds.2012-2095

Goulet, C., Valois, P., Buist, A., and Côté, M. (2010). Predictors of the use of performance-enhancing substances by young athletes. *Clin. J. Sport Med.* 20, 243–248. doi: 10.1097/JSM.0b013e3181e0b935

Grömping, U. (2006). Relative importance for linear regression in R: the package relaimpo. *J. Stat. Softw.* 17, 1–27. doi: 10.18637/jss.v017.i01

Hambleton, R. K., and Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Soc. Indic. Res.* 45, 153–171. doi: 10.1023/A:1006941729637

Hoffman, J. R., Faigenbaum, A. D., Ratamess, N. A., Ross, R., Kang, J., and Tenenbaum, G. (2008). Nutritional supplementation and anabolic steroid use in adolescents. *Med. Sci. Sports Exerc.* 40, 15–24. doi: 10.1249/mss.0b013e31815a5181

Hu, L.-T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Kavussanu, M., Hatzigeorgiadis, A., Elbe, A. M., and Ring, C. (2016). The moral disengagement in doping scale. *Psychol. Sport Exerc.* 24, 188–198. doi: 10.1016/j.psychsport.2016.02.003

Kavussanu, M., Yukhymenko-Lescroart, M. A., Elbe, A. M., and Hatzigeorgiadis, A. (2019). Integrating moral and achievement variables to predict doping likelihood in football: a cross-cultural investigation*. *Psychol. Sport Exerc.* doi: 10.1016/j.psychsport.2019.04.008

Kindlundh, A. M. S., Isacson, D. G. L., Berglund, L., and Nyberg, F. (1999). Factors associated with adolescent use of doping agents: anabolic-androgenic steroids. *Addiction* 94, 543–553. doi: 10.1046/j.1360-0443.1999.9445439.x

Kline, R. B. (1998). *Principles and Practice of Structural Equation Modeling*. New York, NY: Guilford Press.

LaBotz, M., and Griesemer, B. A. (2016). Use of performance-enhancing substances. *Pediatrics* 115, 1103–1106. doi: 10.1542/peds.2016-1300

Lazuras, L., Barkoukis, V., Rodafinos, A., and Tzorbatzoudis, H. (2010). Predictors of doping intentions in elite-level athletes: a social cognition approach. *J. Sport Exerc. Psychol.* 32, 694–710. doi: 10.1123/jsep.32.5.694

Lucidi, F., Grano, C., Leone, L., Lombardo, C., and Pesce, C. (2004). Determinants of the intention to use doping substances: an empirical contribution in a sample of Italian adolescents. *Int. J. Sport Psychol.* 35, 133–148.

Lucidi, F., Mallia, L., Alivernini, F., Chirico, A., Manganelli, S., Galli, F., et al. (2017). The effectiveness of a new school-based media literacy intervention on adolescents' doping attitudes and supplements use. *Front. Psychol.* 8:1–9. doi: 10.3389/fpsyg.2017.00749

Lucidi, F., Zelli, A., Mallia, L., Grano, C., Russo, P. M., and Violani, C. (2008). The social-cognitive mechanisms regulating adolescents' use of doping substances. *J. Sports Sci.* 26, 447–456. doi: 10.1080/02640410701579370

Mallia, L., Lazuras, L., Barkoukis, V., Brand, R., Baumgarten, F., Tsorbatzoudis, H., et al. (2016). Doping use in sport teams: the development and validation of measures of team-based efficacy beliefs and moral disengagement from a cross-national perspective. *Psychol. Sport Exerc.* 25, 78–88. doi: 10.1016/j.psychsport.2016.04.005

Mallia, L., Lazuras, L., Violani, C., and Lucidi, F. (2015). Crash risk and aberrant driving behaviors among bus drivers: the role of personality and attitudes towards traffic safety. *Accid. Anal. Prev.* 79, 145–151. doi: 10.1016/j.aap.2015.03.034

Mallia, L., Lucidi, F., Zelli, A., and Violani, C. (2013). Doping attitudes and the use of legal and illegal performance-enhancing substances among Italian adolescents. *J. Child Adolesc. Subst. Abus.* 22, 179–190. doi: 10.1080/1067828X.2012.733579

Marsh, H. W., Hau, K.-T., and Wen, Z. (2004). Structural equation modeling: a modeling latent growth curves with incomplete data using different types of structural equation modeling and multilevel software. *Struct. Equ. Model.* 11, 452–483. doi: 10.1207/s15328007sem1103

Muthén, L. K., and Muthén, B. O. (2017). *Mplus: Statistical Analysis with Latent Variables: User's Guide (Version 8)*. Los Angeles, CA: Muthén, L. K. Muthén, B. O.

Nicholls, A. R., Cope, E., Bailey, R., Koenen, K., Dumon, D., Theodorou, N. C., et al. (2017). Children's first experience of taking anabolic-androgenic steroids can occur before their 10th birthday: a systematic review identifying 9 factors that predicted doping among young people. *Front. Psychol.* 8:1015. doi: 10.3389/fpsyg.2017.01015

Ntoumanis, N., Ng, J. Y. Y., Barkoukis, V., and Backhouse, S. (2014). Personal and psychosocial predictors of doping use in physical activity aettings: a meta-analysis. *Sport. Med.* 44, 1603–1624. doi: 10.1007/s40279-014-0240-4

Petróczi, A., and Strauss, B. (2015). Understanding the psychology behind performance-enhancement by doping. *Psychol. Sport Exerc.* 16, 137–139. doi: 10.1016/j.psychsport.2014.09.002

Smith, A. C. T., and Stewart, B. (2010). The special features of sport: a critical revisit. *Sport Manag. Rev.* 13, 1–13. doi: 10.1016/j.smr.2009.07.002

Tabachnick, B. G., and Fidell, L. S. (2006). *Using Multivariate Statistics*, 5th Edn. Needham Heights, MA: Allyn and Bacon, Inc.

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–69. doi: 10.1177/109442810031002

Wanjek, B., Rosendahl, J., Strauss, B., and Gabriel, H. H. (2007). Doping, drugs and drug abuse among adolescents in the state of Thuringia (Germany): prevalence, knowledge and attitudes. *Int. J. Sports Med.* 28, 346–353. doi: 10.1055/s-2006-924353

Wiefferink, C. H., Detmar, S. B., Coumans, B., Vogels, T., and Paulussen, T. G. W. (2008). Social psychological determinants of the use of performance-enhancing drugs by gym users. *Health Educ. Res.* 23, 70–80. doi: 10.1093/her/cym004

Zelli, A., Mallia, L., and Lucidi, F. (2010). The contribution of interpersonal appraisals to a social-cognitive analysis of adolescents' doping use. *Psychol. Sport Exerc.* 11, 304–311. doi: 10.1016/j.psychsport.2010.02.008

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 06 May 2020
doi: 10.3389/fpsyg.2020.00740

Investigating the Multidimensionality of the Work-Related Flow Inventory (WOLF): A Bifactor Exploratory Structural Equation Modeling Framework

Honglei Gu[1], *Zhonglin Wen*[2]* *and Xitao Fan*[3]

[1] *Cognition and Human Behavior Key Laboratory of Hunan Province, Department of Psychology, Hunan Normal University, Changsha, China,* [2] *Center for Studies of Psychological Application, School of Psychology, South China Normal University, Guangzhou, China,* [3] *School of Humanities & Social Science, The Chinese University of Hong Kong (Shenzhen), Shenzhen, China*

This study investigated the factor structure of the Work-Related Flow Inventory (WOLF) through the application of the bifactor exploratory structural equation modeling (B-ESEM) framework. Using a sample of 577 Chinese teachers, we contrasted a series of competing models, including CFA, ESEM, bifactor CFA, and B-ESEM models. The results suggested that the B-ESEM structure with three S-factors (absorption, work enjoyment, and intrinsic work motivation) and one G-factor (global flow) was the best representation of the WOLF ratings. The results also supported the composite reliability and the strict invariance of this measurement structure between male and female groups. Relative to males, female teachers showed a higher level of global work-related flow experience. Finally, the nomological validity of WOLF ratings was supported by the statistical relationships of the WOLF factors with job satisfaction and autonomy.

Keywords: work-related flow, WOLF, bifactor model, ESEM, B-ESEM, measurement invariance, nomological validity

OPEN ACCESS

Edited by:
Jason C. Immekus,
University of Louisville, United States

Reviewed by:
Daniela Traficante,
Catholic University of the Sacred Heart, Italy
Mengcheng Wang,
Guangzhou University, China

****Correspondence:***
Zhonglin Wen
wenzl@scnu.edu.cn

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *30 December 2019*
Accepted: *26 March 2020*
Published: *06 May 2020*

Citation:
Gu H, Wen Z and Fan X (2020) Investigating the Multidimensionality of the Work-Related Flow Inventory (WOLF): A Bifactor Exploratory Structural Equation Modeling Framework. Front. Psychol. 11:740. doi: 10.3389/fpsyg.2020.00740

INTRODUCTION

There has been an increasing interest in the construct of flow over the last forty years. Flow is a state of consciousness where people become totally immersed in an activity, and enjoy it intensely (Csikszentmihalyi, 1997). When a person is engaged in some activity of his or her preference, whether it be leisure (e.g., playing chess), sport (e.g., swimming), work, or study, it is more likely that the individual may experience flow. According to flow theory (Csikszentmihalyi, 1975), flow not only helps individuals to have pleasure and satisfaction in the activity, but also improves their self-efficacy and self-esteem and promotes their self-growth and subjective well-being.

Evidence shows that people may have more experience of flow during work as opposed to during their spare time (e.g., Delle Fave and Massimini, 2003). Bakker (2008) discussed that work-related flow experience could be conceptualized as three aspects: absorption, work enjoyment, and intrinsic work motivation. *Absorption* reflects a person's concentration on, and immersion in, the work. *Work enjoyment* reflects a person's happy feeling and positive view with regard to the quality of his work. *Intrinsic work motivation* reflects the tendency that a person does the work for pleasure and satisfaction in the work. Flow at work is most likely to occur when a balance is achieved between the demand of a job and a person's capacity and adequate organizational resources available for doing the job successfully (Bakker, 2008).

Psychometric Characteristics and Latent Structure of the WOLF

Despite the existence of many methods (e.g., experience sampling method, questionnaires, neuronal indicators, and psychophysiological measures) and psychometric instruments (e.g., the Swedish Flow Proneness Questionnaire, Ullén et al., 2012; the Flow State Scale, Jackson and Marsh, 1996) developed to assess flow, the work-related flow inventory (WOLF) developed by Bakker (2008) is the most widely administered measure in the work context. By design, the WOLF consists of three dimensions: *absorption* (4 items), *work enjoyment* (4 items), and *intrinsic work motivation* (5 items). As discussed in Csikszentmihalyi (1997), these three dimensions are the important components typically included in research for flow.

Research findings have generally supported the psychometric quality of the WOLF. For example, Bakker (2008) found that the WOLF showed good internal consistency reliability as well as test-retest reliability estimates, in addition to good evidence for its convergent, construct, and predictive validity. Other studies in different cultural settings also provided support for the WOLF's psychometric quality, such as reliability, predictive validity, and convergent validity, etc. The cultural settings in which such evidence came from included South Africa (Geyser et al., 2015), Norway (Christensen, 2009), Spain (Salanova et al., 2006), Pakistan (Zubair and Kamal, 2015), Italy (Colombo et al., 2013; Zito et al., 2015), and Turkey (Zekioğlu et al., 2017). Similarly, the WOLF has also been shown to have good psychometric characteristics (e.g., reliability and validity) when used in the Chinese cultural context (Zeng, 2013; Chen et al., 2016).

With regard to the latent structure of the WOLF, there are some unresolved issues. Using multiple samples of employees from different occupational groups in Netherlands, Bakker (2008) provided empirical support for the three-factor CFA model consistent with the original design of three components for the measure, over the one-factor model (i.e., only the general factor of flow), and a couple of competing two-factor models. In general, the three-factor structure of the WOLF has also been supported in other cultural settings, such as South Africa (Geyser et al., 2015), Brazil (Freitas et al., 2019), Italy (Colombo et al., 2013; Zito et al., 2015), Turkey (Zekioğlu et al., 2017), and China (Zeng, 2013; Chen et al., 2016). However, most research indicated that this three-factor model typically only showed borderline model fit at best (e.g., Bakker, 2008; Chen et al., 2016), especially with the consideration of the widely used criteria for model fit assessment (Hu and Bentler, 1999).

Furthermore, there was discussion that, of the three constructs proposed for WOLF, two of them – work enjoyment and intrinsic work motivation – overlap conceptually (Llorens et al., 2013; Happell et al., 2015), as enjoyment could already be covered by intrinsic motivation. Operationally, as discussed in Ryan and Deci (2000), self-report of enjoyment is often used for measuring intrinsic motivation. In a sample of Australian workers, Happell et al. (2015) found that the items representing the work enjoyment and intrinsic work motivation dimensions of the WOLF loaded on one dimension, and they argued that the two-factor solution (i.e., absorption and work enjoyment/intrinsic work motivation) should be retained instead of the conventional three-factor structure. Likewise, Bădoiu and Oprea (2018) showed that the two-factor model had a better fit for the sample data of a Romanian population. In fact, even in the original study (Bakker, 2008), work enjoyment was found to be considerably correlated with intrinsic motivation (e.g., ranging from 0.67 to 0.82), suggesting that these two factors conceptually overlap, which led to poor discriminant validity. Other studies (e.g., Geyser et al., 2015) also discussed these two issues (i.e., borderline model fit and two overlapping constructs).

On the practice side, how the WOLF score(s) is used is also inconsistent. Some researchers used the composited flow score as the measure of global flow (e.g., Fagerlind et al., 2013; Zubair and Kamal, 2015), while some others used three subscale scores to represent the three domain components of flow (e.g., Demerouti et al., 2012). Still others treated the global flow as a latent variable with the three subscale scores as its indicators (Salanova et al., 2006). These studies, however, did not provide any rationale or practical guidelines about why the WOLF score(s) should be used as shown in the respective studies. Ideally, the way in which the WOLF score(s) is used should be grounded in, and supported by, the latent structure of the measure, as how the score(s) of the WOLF is composited should be guided by the latent structure of the measure. When the latent structure of the WOLF is somewhat uncertain, we cannot be sure what scoring mechanism would be the best representation of the underlying structure of the WOLF.

Recent research (e.g., Stenling et al., 2015; Gu et al., 2017b; Tóth-Király et al., 2018) that examined different approaches for modeling the latent structure of some psychological measures indicated that a conventional confirmatory factor analysis approach may often fail to adequately capture the more complicated multidimensionality of the latent structure of some measures; more sophisticated modeling approaches may be needed to better model the multidimensionality of some measures. It is likely that we may develop a better understanding about the issues concerning the latent structure of the WOLF as discussed above by considering some more sophisticated modeling approaches that may better capture measurement multidimensionality. With all these considerations, it became necessary to revisit the issue of the latent structure of the WOLF, to develop a better understanding of the multiple issues discussed above.

Approaches for Modeling Multidimensionality

Confirmatory Factor Analysis (CFA)

The confirmatory factor analysis (CFA) is the most commonly used approach to model construct-relevant multidimensionality. CFA, however, is often criticized for its overly restrictive independent cluster model (ICM) assumption, which requires that each item is defined by one, and only one, content domain. This assumption is operationalized in a CFA (ICM-CFA) model by constraining all cross-loadings to zeros, which could lead to unintended consequences, such as inflated factor correlations, poor goodness-of-fit indices, and poor discriminant validity, etc. Indeed, research indicated that the ICM-CFA model, even when the model fit was satisfactory, could lead to inflated factor correlations (e.g., Morin et al., 2017).

Bifactor CFA

Bifactor CFA model assumes: (a) the existence of a general factor that accounts for the shared communality by all the items; and (b) the existence of several group factors, which contribute to a common variance shared within each cluster of items, beyond that of the general factor (Reise, 2012; Gu et al., 2015, 2017a). For model identification, orthogonality is assumed between the general factor and the specific factors. Such a bifactor CFA model better represents the multidimensionality of the underlying factor structure because of the coexistence of a general construct (e.g., flow at work) and some specific constructs (e.g., absorption, work enjoyment, and intrinsic work motivation).

Exploratory Structural Equation Modeling (ESEM)

Exploratory structural equation modeling provides an overarching framework which integrates CFA and exploratory factor analysis (EFA) into a single structural equation modeling (SEM) model. This model is more appropriate for investigating possible multidimensionality of a measure due to the associations between non-target constructs and imperfect items (Asparouhov and Muthén, 2009). ESEM relies on target rotation, which is a confirmatory form of rotation, to freely estimate cross-loadings. Compared with CFA, ESEM provides more accurate, typically lower, estimates of factor correlations, and these more accurate estimates of factor correlations result in better discriminant validity (Asparouhov and Muthén, 2009; Morin et al., 2016).

Bifactor Exploratory Structural Equation Modeling (B-ESEM)

Bifactor exploratory structural equation modeling was recently proposed by Morin et al. (2016) to examine the issue of construct-relevant multidimensionality. B-ESEM integrates both bifactor model and ESEM model into a single analytical framework. This new modeling approach not only allows the coexistence of the general construct and its subdomains (e.g., global flow, and absorption, work enjoyment, and intrinsic work motivation as specific components), but also takes the relations of non-target constructs and items into account. Theoretically, the B-ESEM is the most comprehensive and flexible model that can more accurately describe the complex psychological characteristics.

Compared with B-ESEM, ESEM ignores the possible presence of hierarchically higher order construct(s) (e.g., global flow at work), which can lead to inflated cross-loadings. By contrast, bifactor model, which is essentially a CFA model, neglects the possibility that items may have cross-loadings on the non-target specific factors. The consequence of fixing such cross-loadings to zero is to inflate the variance of the general factor (Morin et al., 2017; Sánchez-Oliva et al., 2017). B-ESEM, theoretically, overcomes these shortcomings as described above.

Nomological Validity of the WOLF

The nomological validity of WOLF could be supported by appropriate statistical relationships between work flow and external criterion variables such as autonomy and job satisfaction. As Morgeson et al. (2005) discussed, job autonomy reflects how much a job allows an employee to have discretion, freedom, and independence for work scheduling, or allows employees to make the necessary decisions to get the job done. Job satisfaction, on the other hand, is a person's agreeable or positive emotional state that is based on personal evaluation of one's occupation or job experiences (Locke, 1976). As Hackman and Oldham (1980) described in their job characteristics model, five important job characteristics (namely, task significance, skill variety, autonomy, feedback, and task identity) generate and enhance a person's flow experience. Of these five, autonomy seems to have the most beneficial effect on flow (Bakker, 2008; Mäkikangas et al., 2010; Lin and Joe, 2012). Empirical evidence also suggested that autonomy was significantly and positively associated with flow experience. For example, Fullagar and Kelloway (2009) revealed that autonomy was a significantly positive predictor for flow. In addition, many other studies showed that job satisfaction was closely related to work flow or its specific components (e.g., Maeran and Cangiano, 2013; Geyser et al., 2015; Zito et al., 2015).

The Present Study

We conducted this study with three specific aims. First, we intended to investigate WOLF's latent structure, by using both conventional and more recent modeling approaches, such as ESEM model, bifactor model, and B-ESEM model, for the purpose of resolving some issues related to WOLF's latent structure. Second, we intended to examine how invariant the WOLF structure was across gender groups. For this purpose, a series of progressively more stringent invariance conditions (e.g., ranging from configural, weak, strong, and to strict invariance) would be tested. Third, we intended to examine the nomological validity of the WOLF in relation to the relevant constructs of autonomy and job satisfaction, as suggested by the best model that emerged from the modeling analyses under the first aim.

MATERIALS AND METHODS

Participants

The participants were 577 teachers recruited in Zhengzhou, a metropolitan area in central China. The sample's average age was 36.80 years old (*SD* = 9.04), and their average work seniority was 12.20 years (*SD* = 9.95). The majority of participants were female (71.9%) and married (83.5%). Among the participating teachers, 21.0% were teaching in kindergartens, 40.0% in primary schools, and 39.0% in secondary schools.

Measures

Flow at Work

The *Work-Related Flow Inventory* (WOLF; Bakker, 2008) was used to measure *flow at work*. This measure had 13 items designed to assess three dimensions of flow experience: (a) *absorption* (four items; sample item: "I get carried away by my work"), (b) *work enjoyment* (four items; sample item: "I do my work with a lot of enjoyment"), and (c) *intrinsic work motivation* (five items; sample item: "I find that I also want to work in my free time"). The items had the response scale with 7-points ranging from 1 (never) to 7 (always). For using this measure in the sample of Chinese teachers, the standard procedure of translation and back-translation (Brislin, 1986) was used to translate the original WOLF into Chinese, and both the English and Chinese items of the WOLF were available in **Supplementary Table S1** of **Supplementary Appendix**. Cronbach's α was 0.92 for the total scale, and 0.85, 0.91, and 0.83 for the three subscales of *absorption*, *work enjoyment*, and *intrinsic work motivation*, respectively. The model-based reliability (i.e., omega coefficient, ω; Sijtsma, 2009) would be estimated and reported in section "Results."

Job Satisfaction

The *Job Satisfaction Scale* (Schriesheim and Tsui, 1980) was used to measure *job satisfaction*. The self-report scale contained six items, with each being rated on a 5-point Likert scale (1 = strongly disagree; 5 = strongly agree). A sample item is, "How satisfied are you with the nature of the work you perform?" Cronbach's α and the omega coefficient (ω) in the study sample were both 0.86.

Autonomy

Autonomy was measured by using the subscale of self-determination under the *Psychological Empowerment Scale* (Spreitzer, 1995). The subscale consisted of three items (e.g., "I have significant autonomy in determining how I do my job"). Participants responded to each item on a 5-point Likert scale (1 = strongly disagree; 5 = strongly agree). In this study, Cronbach's α and the omega coefficient (ω) were both 0.83.

Statistical Analysis

To achieve the aims of the study, statistical analyses were carried out in three phases. In the first phase of analyses, for understanding the measurement structure of the WOLF, a series of nine alternative models were examined to assess their respective goodness-of-fit, as follows.

Model of unitary dimension:

Model 1: One-factor CFA model (global flow).

Models with two sub-domains*:*

Model 2: Two-factor CFA model (absorption, work enjoyment/motivation).
Model 3: ESEM model (including absorption and work enjoyment/motivation).
Model 4: Bifactor CFA model (B-CFA) with two specific domains (absorption and work enjoyment/motivation).
Model 5: B-ESEM model, including two S-factors (absorption, work enjoyment/motivation), and one G-factor (global flow).

Models with three sub-domains:

Model 6: Three-factor CFA model (absorption, work enjoyment, and intrinsic work motivation).
Model 7: ESEM model (including absorption, work enjoyment, and intrinsic work motivation).
Model 8: Bifactor CFA model (B-CFA) with three specific domains (absorption, work enjoyment, and intrinsic work motivation).
Model 9: B-ESEM model, including three S-factors (absorption, work enjoyment, and intrinsic work motivation), and one G-factor (global flow).

Among the nine models above, Model 1 was the baseline model, which assumed one general factor of global flow without considerations for any sub-domains. Model 2 to Model 5 shared the general assumption of two sub-domains of work flow. Model 6 to Model 9 shared the general assumption of three sub-domains of work flow.

In the first-order CFA models (Model 2 and Model 6), each item was specified to load on the factor (i.e., the content domain) that the item was assumed to measure, and without cross-loadings on any other factors. In the first-order ESEM models (Model 3 and Model 7), all cross-loadings were specified to be freely estimated through oblique target rotation. The B-CFA models (Model 4 and Model 8) assumed that each item simultaneously loaded onto a global flow construct and one of the specific domains of flow, and that all factors were orthogonal (i.e., uncorrelated with each other). As for the B-ESEM models (Model 5 and Model 9), an item was not only defined by the G-factor and by a S-factor of its own, but it also reflected other conceptually adjacent subdomains (i.e., cross-loadings) through orthogonal bifactor-target rotation.

In the second phase of analyses, for the purpose of testing measurement invariance across gender groups, the best fitting model that emerged from the first phase of modeling analyses (i.e., Model 1 to Model 9; described above) was used, and measurement invariance analyses were conducted by using the sequence described in the literature (Millsap, 2011). The analyses tested progressively more stringent invariance assumptions: (a) configural invariance (invariance of factor structure), (b) weak invariance (#a satisfied, plus invariance of factor loadings), (c) strong invariance (#b satisfied, plus invariance of item intercepts), (d) strict invariance (#c satisfied, plus invariance of item uniquenesses), (e) latent variance-covariance invariance (#d satisfied, plus invariance of latent variance-covariance), and (f) latent means invariance (#e satisfied, plus invariance of latent factor means).

In the third phase of analyses, latent factors representing job satisfaction and autonomy were integrated to the retained measurement model to examine the nomological validity of the WOLF.

All modeling analyses were carried out by using the statistical modeling software M*plus* 7.0 (Muthén and Muthén, 2012). In the modeling analyses, the robust maximum likelihood (MLR) estimation method was used, which provides estimates of standard errors and fit indexes appropriate for conditions such as ordinal Likert-scale item responses and data non-normality. For model fit assessment, we considered the following model-fit indices: the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), and the root mean square error of approximation

(RMSEA) with its confidence intervals (CI). As suggested in the literature (Hu and Bentler, 1999; Marsh et al., 2004), adequate and excellent model fit may be indicated by values greater than 0.90 and 0.95, respectively, on CFI and TLI, and by values lower than 0.08 and 0.06, respectively, on RMSEA. For testing alternative models, as discussed in Chen (2007), ΔCFI and ΔTLI $\geq$ 0.01 and ΔRMSEA $\geq$ 0.015 could be considered to suggest a more restrictive model.

RESULTS

Descriptive Statistics

Table 1 presents the means, standard deviations, and Pearson correlations for the measured variables. As expected, the three components of the WOLF (absorption, work enjoyment, and intrinsic work motivation) were positively related to each other (r = 0.66–0.72, $p < 0.001$). Absorption, work enjoyment, and intrinsic motivation correlated with both autonomy and job satisfaction (r = 0.44–0.68, $p < 0.001$).

TABLE 1 | Means, standard deviations, and correlations of the study variables.

	1	2	3	4	5
1. Absorption	–				
2. Work enjoyment	0.66***	–			
3. Intrinsic work motivation	0.66***	0.72***	–		
4. Job satisfaction	0.44***	0.68***	0.62***	–	
5. Autonomy	0.49***	0.61***	0.60***	0.75***	–
M	4.99	5.05	5.06	3.59	3.68
SD	1.26	1.37	1.13	0.82	0.59

*n = 577. ***p < 0.001.*

Latent Structure of the WOLF

Results of model fit assessment for the nine alternative models, which represented different latent measurement structures of the WOLF as discussed previously, are displayed in the upper portion of **Table 2**. Based on comparison of the alternative two-subdomain (Models 2 to 5) and three-subdomain (Models 6 to 9) solutions, it is apparent that the three-subdomain solutions had a much better fit to the data than the two-subdomain counterparts. The parameter estimates for the two-subdomain models, which were reported in **Supplementary Tables S2, S3** of **Supplementary Appendix**, further supported the three-subdomain solutions.

With the superiority of the three-subdomain solutions clearly supported, we shifted our focus to the comparisons of different forms of three-subdomain solutions (i.e., comparisons among Models 6 to 9). As discussed in Morin et al. (2016), we first compared the CFA (Model 6) and ESEM (Model 7) models, and it was revealed that the ESEM model (Model 7) showed better model fit (ΔTLI = 0.04, ΔCFI = 0.06, and ΔRMSEA = –0.02) than the CFA model (Model 6).

Tables 3 and **4** present the standardized factor loadings and factor correlations of these two models, which also provided support for the ESEM solution. More specifically, most of the target loadings were statistically significant and practically acceptable in the three-factor CFA ($|\lambda|$ = 0.44–0.91; M = 0.74) and ESEM ($|\lambda|$ = 0.10–0.91; M = 0.58) models. CFA factor correlation estimates (r = 0.73–0.88; M = 0.81) were overall larger than the corresponding ESEM model factor correlations (r = 0.18–0.68; M = 0.39), which indicated that ESEM results showed better differentiation among the factors. More noticeably, the correlation of work enjoyment with intrinsic work motivation was reduced from 0.88 in the CFA model to 0.30 in the ESEM model. The findings based on this initial evaluation indicated that the ESEM should be the preferred model.

Morin et al. (2016) made the suggestion that the second comparison would compare the retained model from the CFA vs. ESEM comparison above with its bifactor counterparts (either B-CFA or B-ESEM). Here, as the ESEM model was retained, we now compare ESEM and B-ESEM models. Although most cross-loadings in ESEM remained small ($|\lambda|$ = 0.05–0.51; M = 0.20), some were large enough to indicate the possibility of an unmodelled G-factor. Out of 26 cross-loadings, seven were between 0.20 and 0.30 (Items a1, a2, w1, w3, i1, and i2), and six were over 0.30 (Items i1, i2, i3, i4, and i5).

As shown in **Table 2**, the B-ESEM solution (Model 9) had excellent model fit to the data as shown by all model fit indices (TLI = 0.97, CFI = 0.99, RMSEA = 0.05), which substantially exceeded (ΔTLI = 0.08, ΔCFI = 0.05, ΔRMSEA = –0.04) the model fit of the ESEM solution (Model 7). More importantly, the B-ESEM model's factor loadings reported in **Table 5** were indicative of a G-factor, as shown by the substantial and strong loadings from all the indicators ($|\lambda|$ = 0.40–0.86; M = 0.66). Beyond the global flow factor, the loadings on the target-specific factors ($|\lambda|$ = 0.01–0.59; M = 0.37) were substantially larger than the non-target loadings ($|\lambda|$ = 0.00–0.22; M = 0.09). The specific factors of absorption ($|\lambda|$ = 0.21–0.59; M = 0.42) and work enjoyment ($|\lambda|$ = 0.30 –0.59; M = 0.45) were well-defined with generally moderate to large target loadings. Loadings on the intrinsic work motivation S-factor ($|\lambda|$ = 0.01–0.53; M = 0.27) were lower in general than the loadings on the other group factors. In addition, two of five target loadings (Item i2, $|\lambda|$ = 0.10; Item i5, $|\lambda|$ = 0.01) on the intrinsic work motivation S-factor were statistically non-significant. These suggested that the intrinsic work motivation S-factor is less well-defined than the other two, but acceptable.

More importantly, the B-ESEM model's cross-loadings ($|\lambda|$ = 0.00–0.22; M = 0.09) were substantially lower than those of the ESEM model ($|\lambda|$ = 0.05–0.51; M = 0.20). Furthermore, in the B-ESEM solution, only two cross-loadings were between 0.20 and 0.30 (Items a2 and w3), and none were over 0.30. All these findings provided strong support for retaining the B-ESEM model as the best representation of the structure of the WOLF.

For further assessing the appropriateness of the B-ESEM model, we calculated model-based coefficients of composite reliability (Perreira et al., 2018) for both the G-factor and the S-factor, based on the standardized model estimates. The composite reliability of both the general flow factor (ω = 0.94) and the work enjoyment S-factor (ω = 0.82) were very good. The composite reliability of the S-factor for absorption (ω = 0.67)

TABLE 2 | Model fit statistics of alternative measurement models (upper) and measurement invariance tests of B-ESEM model (lower).

Model comparison analysis	**χ^2 (*df*)**	**RMSEA (90%CI)**	**CFI**	**TLI**					
Model 1: One-factor CFA	553.60 (65)	0.14 (0.13, 0.15)	0.79	0.74					
Model 2: Two-factor CFA	379.57 (64)	0.11 (0.10, 0.12)	0.86	0.83					
Model 3: Two-factor ESEM	503.70 (53)	0.12 (0.11, 0.13)	0.86	0.80					
Model 4: B-CFA: Two S-factors	239.81 (52)	0.09 (0.08, 0.10)	0.92	0.88					
Model 5: B-ESEM: Two S-factors	223.97 (42)	0.09 (0.08, 0.10)	0.94	0.89					
Model 6: Three-factor CFA	325.16 (62)	0.11 (0.10, 0.11)	0.88	0.85					
Model 7: Three-factor ESEM	223.97 (42)	0.09 (0.08, 0.10)	0.94	0.89					
Model 8: B-CFA: Three S-factors	245.15 (52)	0.09 (0.08, 0.10)	0.92	0.88					
Model 9: B-ESEM: Three S-factors	59.99 (32)	0.05 (0.04, 0.06)	0.99	0.97					
Measurement invariance analysis	**χ^2 (*df*)**	**RMSEA (90%CI)**	**CFI**	**TLI**	**CM**	**$\Delta\chi^2$ (Δ*df*)**	**ΔRMSEA**	**ΔCFI**	**ΔTLI**
Model A: Configural IN	116.98 (64)	0.05 (0.04, 0.07)	0.98	0.96		–	–	–	–
Model B: Weak IN	163.27 (100)	0.05 (0.03, 0.06)	0.98	0.97	Model A	46.29 (36)	0.00	0.00	0.01
Model C: Strong IN	172.26 (109)	0.05 (0.03, 0.06)	0.98	0.97	Model B	8.98 (9)	0.00	0.00	0.00
Model D: Strict IN	186.04 (122)	0.04 (0.03, 0.06)	0.98	0.98	Model C	13.78 (13)	–0.01	0.00	0.01
Model E: Latent v/c IN	246.62 (132)	0.06 (0.04, 0.07)	0.97	0.96	Model D	60.58 (10)	0.02	–0.01	–0.02
Model F: Latent means IN	273.91 (136)	0.06 (0.05, 0.07)	0.96	0.95	Model D	87.87 (14)	0.02	–0.02	–0.03

χ^2 = robust chi-square test of exact fit; df = degree of freedom; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis Index; CM = comparison model; B-CFA = bifactor CFA; IN = invariant; v/c = latent variance and covariance.

TABLE 3 | Standardized parameter estimates for three-factor CFA (Model 6) and three-factor ESEM (Model 7) models.

	Three-factor CFA		**Three-factor ESEM**			
	λ	**δ**	**AB (λ)**	**WE (λ)**	**IWM (λ)**	**δ**
***Absorption* (AB)**						
*a*1	0.58	0.66	**0.47**	0.05	0.26	0.63
*a*2	0.79	0.38	**0.70**	0.22	*–0.23*	0.28
*a*3	0.83	0.32	**0.87**	*–0.10*	*0.18*	0.27
*a*4	0.87	0.25	**0.91**	*–0.07*	*0.05*	0.24
***Work enjoyment* (WE)**						
*w*1	0.86	0.26	*–0.09*	**0.87**	0.20	0.21
*w*2	0.91	0.17	0.12	**0.87**	–0.13	0.13
*w*3	0.86	0.26	*–0.07*	**0.82**	0.30	0.17
*w*4	0.88	0.23	0.16	**0.78**	*–0.07*	0.22
***Intrinsic work motivation* (IWM)**						
*i*1	0.68	0.54	0.23	0.34	***0.33***	0.52
*i*2	0.47	0.78	0.36	0.21	***–0.19***	0.74
*i*3	0.44	0.81	0.41	–0.06	**0.30**	0.74
*i*4	0.63	0.60	*0.07*	*0.40*	**0.38**	0.56
*i*5	0.82	0.32	0.37	0.51	***–0.10***	0.37

Non-significant loadings (p > 0.05) are italicized. Target loadings of the ESEM model are shown in bold.

TABLE 4 | Inter-factor correlations for three-factor CFA (Model 6) and three-factor ESEM (Model 7) Solutions.

	Absorption	**Work Enjoyment**	**Intrinsic Work Motivation**
Absorption	–	0.73***	0.81***
Work enjoyment	0.68***	–	0.88***
Intrinsic work motivation	0.18**	0.30***	–

*ICM-CFA correlations are displayed above the diagonal and ESEM correlations are displayed below the diagonal. **p < 0.01; ***p < 0.001.*

was adequate, and that of the intrinsic work motivation S-factor (ω = 0.42) was marginal.

Assessment for Gender Group Measurement Invariance

The measurement invariance across gender for the B-ESEM model was assessed, and the findings were displayed in the lower portion of **Table 2**. The configural invariance model (Model A) showed very good model fit to the data (TLI = 0.96, CFI = 0.98, RMSEA = 0.05). Progressively more stringent invariance constraints were then successively imposed on factor loadings (Model B: weak invariance), item intercepts (Model C: strong invariance), and item uniquenesses (Model D: strict invariance). None of these more stringent invariance conditions caused model fit deterioration beyond the general guidelines (i.e. ΔCFI and ΔTLI $\geq$ 0.01 and ΔRMSEA $\geq$ 0.015). However, the model for the invariance of latent variance-covariance (Model E; ΔTLI = –0.02, ΔCFI = –0.01, ΔRMSEA = 0.02) and the model for latent means invariance (Model F; ΔTLI = –0.03, ΔCFI = –0.02, ΔRMSEA = 0.02) were not supported by the data. Further analysis showed that when males' factor means were fixed to zero for model identification purpose, female teachers' factor means were statistically higher on the work flow G-factor (M = 0.46, $p < 0.001$), but not statistically different on the S-factors ($p > 0.05$).

Nomological Validity

For the purpose of examining the nomological validity of the WOLF, external CFA factors for job satisfaction and autonomy

TABLE 5 | Standardized factor loadings for B-ESEM model with three S-factors and one G-factor (Model 9).

	GWF (λ)	S-AB (λ)	S-WE (λ)	S-IWM (λ)	δ
***Absorption* (AB)**					
*a*1	0.46	**0.34**	0.13	0.16	0.63
*a*2	0.80	**0.21**	–0.11	–0.20	0.27
*a*3	0.66	**0.59**	*0.03*	0.06	0.22
*a*4	0.70	**0.52**	*–0.05*	*–0.05*	0.23
ω		**0.67**			
***Work Enjoyment* (WE)**					
*w*1	0.68	*0.02*	**0.59**	0.13	0.17
*w*2	0.83	*0.01*	**0.40**	–0.15	0.12
*w*3	0.70	*0.00*	**0.52**	0.22	0.20
*w*4	0.83	*–0.02*	**0.30**	*–0.05*	0.23
ω			**0.82**		
***Intrinsic Work Motivation* (IWM)**					
*i*1	0.61	*0.05*	0.14	**0.37**	0.47
*i*2	0.52	*0.00*	–0.14	***–0.10***	0.70
*i*3	0.40	0.18	*–0.06*	**0.33**	0.70
*i*4	0.57	–0.10	0.13	**0.53**	0.37
*i*5	0.86	–0.07	*–0.03*	***0.01***	0.26
ω	**0.94**			**0.42**	

GWF = general work-related flow. Non-significant loadings (p > 0.05) are italicized. Target loadings on specific factors of the B-ESEM model are shown in bold.

TABLE 6 | Correlations between WOLF factors and two external factors (job satisfaction and autonomy) based on the B-ESEM model with three S-factors and one G-factor.

	Job satisfaction	Autonomy
General work-related Flow	0.66***	0.49***
Absorption	–0.09	0.03
Work Enjoyment	0.33***	0.32***
Intrinsic work motivation	0.44***	0.30***

****p < 0.001.*

were included into the B-ESEM model (Model 9), and this expanded model showed very good model fit (χ^2 = 282.10, *df* = 167, RMSEA = 0.04, RMSEA 90%CI = [0.04, 0.05], TLI = 0.96, CFI = 0.97). As displayed in **Table 6**, the flow G-factor and two S-factors (i.e. work enjoyment and intrinsic work motivation) were significantly and positively associated with job satisfaction and autonomy. By contrast, the absorption S-factor was not significantly associated with these two external factors.

DISCUSSION

This study is the first attempt to investigate the latent structure of the WOLF by using both CFA and ESEM approaches. Consistent with previous research (e.g., Bakker, 2008; Christensen, 2009; Happell et al., 2015), this study found that the one-factor CFA solution was far from being acceptable, indicating that work-related flow should be considered as consisting of multiple dimensions, rather than of a unitary dimension.

The WOLF was originally designed to assess three inter-related content domains (Bakker, 2008), and the three-factor structure was shown in various samples and in different cultures. However, Rodríguez-Sánchez et al. (2011) and Llorens et al. (2013) discussed that only enjoyment and absorption were the essence of the work-related flow experience. Enjoyment could be considered as some kind of motivation (Davis et al., 1992), and intrinsic motivation might be an antecedent, instead of a core component, of work-related flow (Deci and Ryan, 1985; Llorens et al., 2013). The study by Happell et al. (2015) in an Australian sample showed that the items for two domains, work enjoyment and intrinsic work motivation, loaded onto one dimension, providing support for the argument described above. For the purpose of understanding whether these two components might be combined into one factor, we compared the two-factor CFA (absorption and work enjoyment/motivation) and the three-factor CFA (absorption, work enjoyment, intrinsic work motivation) solutions, and found that the goodness-of-fit of the latter model substantially exceeded that of the former. More importantly, we found that the correlation between work enjoyment and intrinsic work motivation in the three-factor CFA model was indeed high (i.e., 0.88), which was in line with previous findings (e.g., Bakker, 2008; Geyser et al., 2015; Zito et al., 2015). With such findings based primarily on conventional CFA approaches, it is difficult to decide which model should be preferable for WOLF. Therefore, new modeling approaches (e.g., B-CFA, ESEM, and B-ESEM) could be needed to further examine the dimensionality of the WOLF structure.

In line with prior research on multidimensional data, the comparison between the ICM-CFA model and ESEM model in this study revealed that the ESEM model was preferable, as ESEM had better model fit, and the factors showed better differentiations between each other as indicated by the lower inter-factor correlations. The ESEM solution, similar to the three-factor CFA, only considered the subdomains as separate factors for absorption, work enjoyment, and intrinsic work motivation, without the consideration for a possible overarching global factor. The observation of multiple cross-loadings of sizable magnitude (|λ| > 0.20, or even 0.30) in the ESEM model suggested that a global work-related flow factor might be present in the data. The comparison of ESEM and B-ESEM solutions provided support for this possibility. First, B-ESEM had substantially better model fit to the data. Second, the general flow dimension in B-ESEM appeared to be well defined, with the items showing moderate to large loadings on this general flow factor. Third, the composite reliability of the flow G-factor (ω = 0.94) was excellent. Fourth, the specific factors of absorption and work enjoyment were well-defined, while the specific factor of intrinsic work motivation was less well-defined, but generally acceptable. Finally, cross-loadings in the B-ESEM solution were generally lower than those of the ESEM solution.

In general, if the composite reliability (ω) of a specific factor is sufficiently high (e.g., >0.5), it indicates that the subscale score accounts for a meaningful amount of variance beyond the G-factor (Perreira et al., 2018). The findings in this study showed that the specific factors of absorption (ω = 0.67) and work enjoyment (ω = 0.82) had a substantial amount of specificity of its own, over and above the global flow. On the other hand, the specific factor of intrinsic work motivation (ω = 0.42) was less well-defined and had relatively low composite reliability. But three of the five target loadings exceeded 0.3, indicating that this specific factor still had an acceptable degree of specificity beyond the G-factor. Therefore, it is suggested to report the total score and subscale scores of absorption and work enjoyment when using the WOLF in practice. The use of subscale score of intrinsic work motivation should be treated with caution.

As shown earlier, the B-ESEM solution showed the best model fit. In the B-ESEM model, however, Item w3 ("I feel happy during my work") not only reflected the global work-related flow and the subdomain of work enjoyment, but it also had a substantial cross-loading (λ = 0.22) on the non-target intrinsic work motivation S-factor. This, however, was reasonable, because employees who are happy at work are usually motivated intrinsically by their work (Geyser et al., 2015). Psychometrically, it may not be realistic to require that each item reflects one, and only one, content domain of multidimensional constructs (Asparouhov and Muthén, 2009).

In addition, the findings also provided support for strict measurement invariance of the B-ESEM solution across gender groups, suggesting that this model was well-replicated across subsamples of male and female teachers. For the latent mean differences, the results revealed that female teachers showed a higher level of global work-related flow experience than male teachers. These findings were consistent with previous research showing that female teachers reported greater engagement and satisfaction with the work and lower burnout (Okpara et al., 2005; Rey et al., 2012). This finding may be related to socially constructed gender roles. More specifically, as discussed in Motro and Ellis (2016), the society has a higher expectation for women to carry out communal roles and display the related traits (e.g., friendliness, sympathy, gentleness, caring, and kindness, etc.). On the other hand, society has a higher expectation for men to carry out agentic roles and display these associated traits (e.g., power, dominance, independence, aggression, and competence, etc.). The theory about role congruity suggests that, when a group's stereotype is not matched with the expected social roles, biased responses may occur (Diekman and Hirnisey, 2007). Due to the incongruity between the demands of teaching and the typically expected societal roles of males, male teachers may experience lower level of flow.

The relationships between the WOLF factors with external factors of autonomy and job satisfaction supported the nomological validity of the WOLF. The global flow experience was found to be positively associated with autonomy, and this makes sense, as previous research (e.g., Fried and Ferris, 1987; Saavedra and Kwun, 2000) indicated that when employees could schedule their work and determine some aspects of their job, this could contribute to the employees' positive affect and motivation. This finding is in line with the empirical findings in previous research related to the job characteristics model (Hackman and Oldham, 1980), and to the job demands-resources model (Bakker and Demerouti, 2007), in that high levels of job resources (e.g., autonomy and social support) lead to work-related flow (Zito et al., 2016).

The other finding that the overall work-related flow was positively related to job satisfaction is also in line with previous research, which indicated that flow experience had an important effect on job satisfaction (Geyser et al., 2015), and the psychological state of flow was considered critical in redesigning interventions in the workplace in order to promote job satisfaction (Maeran and Cangiano, 2013). Our results also revealed that only the specific factors of work enjoyment and intrinsic work motivation, but not the absorption S-factor, had positive relationship with job satisfaction and autonomy, confirming the notion that absorption might have some overlap with the holistic description of flow (Bakker, 2008).

Despite the strength of this study in using systematic modeling approaches to examine the latent structure of the WOLF, there are some limitations in this study. One limitation is that the study relied on a convenience sample of Chinese teachers, which may limit the generalizability of findings to a wider context. Future research could use samples from other cultures and from other types of employees. Another limitation is that our assessment of the underlying structure of the WOLF was based on cross-sectional data only. Future research may consider the longitudinal stability of the B-ESEM structure.

In summary, our results supported that the B-ESEM solution could best represent the underlying structure of WOLF scores, and this model incorporates two aspects of psychometric multidimensionality: one is the result of the conceptual adjacency of content domains of flow (e.g., work enjoyment and intrinsic work motivation), and the other is associated with the coexistence of the global work-related flow and the three specific components. Furthermore, the strict gender-group measurement invariance of the B-ESEM model was supported. Female teachers, however, showed a higher level of global work-related flow experience than the male teachers. Finally, the nomological validity of WOLF ratings was supported by the statistical relationships of the WOLF factors with job satisfaction and autonomy.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding authors.

ETHICS STATEMENT

The study protocol was approved by Hunan Normal University Research Ethnics Committee. All participants gave written informed consent.

AUTHOR CONTRIBUTIONS

HG and ZW designed the study. HG and XF wrote the first draft and revised the manuscript.

FUNDING

This research was funded by grants from the National Natural Science Foundation of China (31771245) and the Scientific Research Foundation of Hunan Provincial Education Department (19C1148).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00740/full#supplementary-material

REFERENCES

Asparouhov, T., and Muthén, B. O. (2009). Exploratory structural equation modeling. *Struct. Equ. Model.* 16, 397–438. doi: 10.1080/10705510903008204

Bădoiu, A., and Oprea, B. (2018). The work-related flow (WOLF) inventory: Romanian adaptation. *Hum. Resour. Psychol.* 16, 94–106.

Bakker, A. B. (2008). The work-related flow inventory: construction and initial validation of the WOLF. *J. Vocat. Behav.* 72, 400–414. doi: 10.1016/j.jvb.2007.11.007

Bakker, A. B., and Demerouti, E. (2007). The job demands-resources model: state of the art. *J. Manag. Psychol.* 22, 309–328. doi: 10.3390/ijerph17010069

Brislin, R. (1986). "The wording and translation of research instruments," in *Field Methods in Cross-Cultural Research*, eds W. J. Lonner and J. W. Berry (Beverly Hills, CA: Sage), 137–164.

Chen, F. F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Struct. Equ. Model.* 14, 464–504. doi: 10.1590/2237-6089-2017-0093

Chen, Y., Yu, X., and Huang, B. (2016). "The Chinese version of work-related flow inventory (WOLF): an examination of reliability and validity," in *Paper Presented at the International Conference on Humanities & Social Science*, London.

Christensen, M. (2009). *Validation and Test of Central Concepts in Positive Work and Organizational Psychology: The Second Report from the Nordic Project Positive Factors at Work*. Copenhagen: Renouf Publishing Company Limited.

Colombo, L., Zito, M., and Cortese, C. G. (2013). The Italian version of the work-related flow inventory (WOLF): first psychometric evaluations. *BPA Appl. Psychol. Bull.* 61, 37–42.

Csikszentmihalyi, M. (1975). *Beyond Boredom and Anxiety: Experiencing Flow in Work and Play*, 2nd Edn. San Francisco, CA: Jossey Bass.

Csikszentmihalyi, M. (1997). *Finding Flow: The Psychology of Engagement with Everyday Life*. New York, NY: HarperCollins.

Davis, F. D., Bagozzi, R. P., and Warshaw, P. R. (1992). Extrinsic and intrinsic motivation to use computers in the workplace. *J. Appl. Soc. Psychol.* 22, 1111–1132.

Deci, E. L., and Ryan, R. M. (1985). The general causality orientations scale: self-determination in personality. *J. Res. Pers.* 19, 109–134. doi: 10.1016/0092-6566(85)90023-6

Delle Fave, A., and Massimini, F. (2003). Optimal experience in work and leisure among teachers and physicians: individual and bio-cultural implications. *Leisure Stud.* 22, 323–342. doi: 10.1080/02614360310001594122

Demerouti, E., Bakker, A. B., Sonnentag, S., and Fullagar, C. J. (2012). Work-related flow and energy at work and at home: a study on the role of daily recovery. *J. Organ. Behav.* 33, 276–295. doi: 10.1002/job.760

Diekman, A. B., and Hirnisey, L. (2007). The effect of context on the silver ceiling: a role congruity perspective on prejudiced responses. *Pers. Soc. Psychol. Bull.* 33, 1353–1366. doi: 10.1177/0146167207303019

Fagerlind, A. C., Gustavsson, M., Johansson, G., and Ekberg, K. (2013). Experience of work-related flow: does high decision latitude enhance benefits gained from job resources? *J. Vocat. Behav.* 83, 161–170. doi: 10.1016/j.jvb.2013.03.010

Freitas, C. P. P., Damásio, B. F., Haddad, E. J., and Koller, S. H. (2019). Work-related flow inventory: evidence of validity of the Brazilian version. *Paidéia* 29:e2901.

Fried, Y., and Ferris, G. R. (1987). The validity of the job characteristics model: a review and meta-analysis. *Pers. Psychol.* 40, 287–322. doi: 10.1111/j.1744-6570.1987.tb00605.x

Fullagar, C., and Kelloway, E. K. (2009). "Flow" at work: an experience sampling approach. *J. Occup. Organ. Psychol.* 81, 595–615. doi: 10.1348/096317908x357903

Geyser, I., Geldenhuys, M., and Crous, F. (2015). The dimensionality of the work related flow inventory (WOLF): a South African study. *J. Psychol. Afr.* 25, 282–287. doi: 10.1080/14330237.2015.1078084

Gu, H., Wen, Z., and Fan, X. (2015). The impact of wording effect on reliability and validity of the core self-evaluation scale (CSES): a bi-factor perspective. *Pers. Individ. Dif.* 83, 142–147. doi: 10.1016/j.paid.2015.04.006

Gu, H., Wen, Z., and Fan, X. (2017a). Examining and controlling for wording effect in a self-report measure: a Monte Carlo simulation study. *Struct. Equ. Model.* 24, 545–555. doi: 10.1080/10705511.2017.1286228

Gu, H., Wen, Z., and Fan, X. (2017b). Structural validity of the Machiavellian personality scale: a bifactor exploratory structural equation modeling approach. *Pers. Individ. Dif.* 105, 116–123. doi: 10.1016/j.paid.2016.09.042

Hackman, J. R., and Oldham, G. R. (1980). *Work Redesign*. Reading, MA: Addison-Wesley.

Happell, B., Gaskin, C. J., and Platania-phung, C. (2015). The construct validity of the Work-related flow inventory in a sample of Australian workers. *J. Psychol.* 149, 42–62. doi: 10.1080/00223980.2013.838539

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Jackson, S. A., and Marsh, H. W. (1996). Development and validation of a scale to measure optimal experience: the flow state scale. *J. Sport Exerc. Psychol.* 18, 17–35. doi: 10.1123/jsep.18.1.17

Lin, C. P., and Joe, S. W. (2012). To share or not to share: assessing knowledge sharing, inter-employee helping, and their antecedents among online knowledge workers. *J. Bus. Ethics* 108, 439–449. doi: 10.1007/s10551-011-1100-x

Llorens, S., Salanova, M., and Rodríguez, A. M. (2013). How is flow experienced and by whom? Testing flow among occupations. *Stress Health* 29, 125–137. doi: 10.1002/smi.2436

Locke, E. A. (1976). "The nature and causes of job satisfaction," in *Handbook of Industrial and Organizational Psychology*, ed. M. D. Dunnette (Chicago, IL: Rand McNally), 1297–1343.

Maeran, R., and Cangiano, F. (2013). Flow experience and job characteristics: analyzing the role of flow in job satisfaction. *TPM Test. Psychom. Methodol. Appl. Psychol.* 20, 13–26.

Mäkikangas, A., Bakker, A. B., Aunola, K., and Demerouti, E. (2010). Job resources and flow at work: modeling the relationship via latent growth curve and mixture model methodology. *J. Occup. Health Psychol.* 83, 795–814. doi: 10.1348/096317909x476333

Marsh, H. W., Hau, K. T., and Wen, Z. (2004). In search of golden rules: comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralising Hu & Bentler's (1999) findings. *Struct. Equ. Model.* 11, 320–341. doi: 10.1207/s15328007sem1103_2

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. New York, NY: Routledge.

Morgeson, F. P., Delaney-Klinger, K., and Hemingway, M. A. (2005). The importance of job autonomy, cognitive ability, and job-related skills for predicting role breadth and job performance. *J. Appl. Psychol.* 90, 399–406. doi: 10.1037/0021-9010.90.2.399

Morin, A. J. S., Arens, A. K., and Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Struct. Equ. Model.* 23, 116–139. doi: 10.1080/10705511.2014.961800

Morin, A. J. S., Boudrias, J.-S., Marsh, H. W., McInerney, D. M., Dagenais-Desmarais, V., Madore, I., et al. (2017). Complementary variable- and person-centered approaches to the dimensionality of psychometric constructs: application to psychological wellbeing at work. *J. Bus. Psychol.* 32, 395–419. doi: 10.1007/s10869-016-9448-7

Motro, D., and Ellis, A. P. J. (2016). Boys, don't cry: gender and reactions to negative performance feedback. *J. Appl. Psychol.* 102, 227–235. doi: 10.1037/apl0000175

Muthén, L. K., and Muthén, B. O. (2012). *Mplus User's Guide*, 7th Edn. Los Angeles, CA: Muthén & Muthén.

Okpara, J. O., Squillace, M., and Erondu, E. A. (2005). Gender differences and job satisfaction: a study of university teachers in the United States. *Women Manag. Rev.* 20, 177–190. doi: 10.1108/09649420510591852

Perreira, T. A., Morin, A. J., Hebert, M., Gillet, N., Houle, S. A., and Berta, W. (2018). The short form of the workplace affective commitment multidimensional questionnaire (WACMQ-S): a bifactor-ESEM approach among healthcare professionals. *J. Vocat. Behav.* 106, 62–83. doi: 10.1016/j.jvb.2017.12.004

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behav. Res.* 47, 667–696. doi: 10.1080/00273171.2012.715555

Rey, L., Extremera, N., and Pena, M. (2012). Burnout and work engagement in teachers: are sex and level taught important? *Ansiedad Estrés* 18, 119–129.

Rodríguez-Sánchez, A. M., Schaufeli, W., Salanova, M., Cifre, E., and Sonnenschein, M. (2011). Enjoyment and absorption: an electronic diary study on daily flow patterns. *Work Stress* 25, 75–92. doi: 10.1080/02678373.2011.565619

Ryan, R. M., and Deci, E. D. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.* 55, 68–78. doi: 10.1037/0003-066x.55.1.68

Saavedra, R., and Kwun, S. K. (2000). Affective states in job characteristics theory. *J. Organ. Behav.* 2, 131–146. doi: 10.1002/(sici)1099-1379(200003)21:2<131::aid-job39>3.0.co;2-q

Salanova, M., Bakker, A. B., and Llorens, S. (2006). Flow at work: evidence for an upward spiral of personal and organizational resources. *J. Happiness Stud.* 7, 1–22. doi: 10.1007/s10902-005-8854-8

Sánchez-Oliva, D., Morin, A. J., Teixeira, P. J., Carraça, E. V., Palmeira, A. L., and Silva, M. N. (2017). A bifactor exploratory structural equation modeling representation of the structure of the basic psychological needs at work scale. *J. Vocat. Behav.* 98, 173–187. doi: 10.1016/j.jvb.2016.12.001

Schriesheim, C., and Tsui, A. N. (1980). "Development and validation of a short satisfaction instrument for use in survey feedback interventions," in *Paper Presented at the Western Academy of Management Meeting*, Salt Lake, UT.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* 74, 107–120. doi: 10.1007/s11336-008-9101-0

Spreitzer, G. M. (1995). Psychological empowerment in the workplace: dimensions, measurement, and validation. *Acad. Manag. J.* 38, 1442–1465. doi: 10.5465/256865

Stenling, A., Ivarsson, A., Hassmén, P., and Lindwall, M. (2015). Using bifactor exploratory structural equation modeling to examine global and specific factors in measures of sports coaches' interpersonal styles. *Front. Psychol.* 6:1303. doi: 10.3389/fpsyg.2015.01303

Tóth-Király, I., Morin, A. J. S., Bõthe, B., Orosz, G., and Rigó, A. (2018). Investigating the multidimensionality of need fulfillment: a bifactor exploratory structural equation modeling representation. *Struct. Equ. Model.* 25, 267–286. doi: 10.1080/10705511.2017.1374867

Ullén, F., Örjan, de Manzano, Almeida, R., Magnusson, P. K. E., Pedersen, N. L., et al. (2012). Proneness for psychological flow in everyday life: associations with personality and intelligence. *Pers. Individ. Dif.* 52, 167–172. doi: 10.1016/j.paid.2011.10.003

Zekioğlu, A., Tekingündüz, S., and Sünbül, Ö. (2017). The validity and reliability study of Turkish version of work-related flow inventory (WOLF). *Sanitas Magisterium* 3, 61–67.

Zeng, C. (2013). Reliability and validity of work-related flow inventory. *Chin. J. Clin. Psychol.* 25, 35–38.

Zito, M., Bakker, A. B., Colombo, L., and Cortese, C. G. (2015). A two-step study for the Italian adaptation of the work-related flow (WOLF) inventory: the I-WOLF. *Test. Psychom. Methodol. Appl. Psychol.* 22, 553–570.

Zito, M., Cortese, C. G., and Colombo, L. (2016). Nurses' exhaustion: the role of flow at work between job demands and job resources. *J. Nurs. Manag.* 24, E12–E22. doi: 10.1111/jonm.12284

Zubair, A., and Kamal, A. (2015). Authentic leadership and creativity: mediating role of work-related flow and psychological capital. *J. Behav. Sci.* 25, 150–171.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 12 May 2020
doi: 10.3389/fpsyg.2020.00947

bayes4psy—An Open Source R Package for Bayesian Statistics in Psychology

Jure Demšar[1,2]*, *Grega Repovš*[2] *and Erik Štrumbelj*[1]

[1] *Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia,* [2] *Mind & Brain Lab, Department of Psychology, Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia*

Research in psychology generates complex data and often requires unique statistical analyses. These tasks are often very specific, so appropriate statistical models and methods cannot be found in accessible Bayesian tools. As a result, the use of Bayesian methods is limited to researchers and students that have the technical and statistical fundamentals that are required for probabilistic programming. Such knowledge is not part of the typical psychology curriculum and is a difficult obstacle for psychology students and researchers to overcome. The goal of the bayes4psy package is to bridge this gap and offer a collection of models and methods to be used for analysing data that arises from psychological experiments and as a teaching tool for Bayesian statistics in psychology. The package contains the Bayesian *t*-test and bootstrapping along with models for analysing reaction times, success rates, and tasks utilizing colors as a response. It also provides the diagnostic, analytic and visualization tools for the modern Bayesian data analysis workflow.

Keywords: Bayesian statistics, R, psychology, reaction time, success rate, Bayesian *t*-test, color analysis, linear model

OPEN ACCESS

Edited by:
Jason C. Immekus,
University of Louisville, United States

Reviewed by:
Jose D. Perezgonzalez,
Massey University Business School,
New Zealand
Ratna Nandakumar,
University of Delaware, United States

****Correspondence:***
Jure Demšar
jure.demsar@fri.uni-lj.si

Specialty section:
This article was submitted to
Quantitative Psychology and
Measurement,
a section of the journal
Frontiers in Psychology

Received: *30 January 2020*
Accepted: *16 April 2020*
Published: *12 May 2020*

Citation:
Demšar J, Repovš G and Štrumbelj E (2020) bayes4psy—An Open Source R Package for Bayesian Statistics in Psychology. Front. Psychol. 11:947.
doi: 10.3389/fpsyg.2020.00947

1. INTRODUCTION

Bayesian data analysis with custom models offers a highly flexible, intuitive and transparent alternative to classical statistics. Throughout much of the modern era of science Bayesian approaches were on the sidelines of data analysis, mainly due to the fact that computations required for Bayesian analysis are usually quite complex. But computations that were only a decade or two ago too complex for specialized computers can now be executed on average desktop computers. In part also due to modern Markov chain Monte Carlo (MCMC) methods that make computations tractable for most parametric models. This, along with specialized probabilistic programming languages for Bayesian modeling, such as `Stan` (Carpenter et al., 2017) and `JAGS` (Plummer, 2003), drastically increased the accessibility and usability of Bayesian methodology for data analysis. Indeed, Bayesian data analysis is steadily gaining momentum in the twenty-first century (Gelman et al., 2014; Kruschke, 2014; McElreath, 2018), especially so in natural and technical sciences. Unfortunately, the use of Bayesian data analysis in social sciences remains scarce, most likely due to a steep learning curve associated with Bayesian analysis.

There are many advantages of Bayesian data analysis (Dunson, 2001; Gelman et al., 2014; Kruschke, 2014; McElreath, 2018), such as its ability to work with missing data and incorporating prior information about the data in a natural and principled way. Furthermore, Bayesian methods offer high flexibility through hierarchical modeling, while calculated posterior parameter values can be used as easily understandable alternatives to p-values. Bayesian methods provide very intuitive and interpretable answers, such as "the parameter μ has a probability of 0.95 of falling inside the $[-2, 2]$ interval."

One of the social sciences that can substantially benefit from Bayesian methodology is psychology. The majority of data that are acquired in psychological experiments, such as reaction times, success rates, and picked colors, can be analyzed in a Bayesian manner by using a small set of probabilistic models. To a certain degree Bayesian methodology could also alleviate the replication crisis that is pestering the field of psychology (Schooler, 2014; Open Science Collaboration, 2015; Stanley et al., 2018).

The ability to replicate scientific findings is of paramount importance to scientific progress (McNutt, 2014; Baker and Penny, 2016; Munafò et al., 2017). Unfortunately, more and more replications fail to reproduce original results and conclusions (Schooler, 2014; Open Science Collaboration, 2015; Amrhein et al., 2019). This so-called replication crisis is not only harmful to the authors of such studies but to science itself. A recent attempt to replicate 100 studies from three prominent psychology journals (Open Science Collaboration, 2015) showed that only approximately a third of studies that claimed statistical significance (p-value < 0.05) also showed statistical significance in replication. Another recent study (Camerer et al., 2018) tried to replicate systematically selected studies in the social sciences published in Nature and Science between 2010 and 2015, replication attempts were successful only in 13 out of 21 cases.

The main reasons behind the replication crisis seem to be poor quality control in journals, unclear writing and inadequate statistical analysis (Wasserstein and Lazar, 2016; Hurlbert et al., 2019; Wasserstein et al., 2019). One of the fundamental issues lies in the desire to claim statistical significance through p-values. Many manuscripts published today repeat the same mistakes even though prominent statisticians prepared extensive guidelines on what to do and mainly what not to do (Hubbard, 2015; Wasserstein and Lazar, 2016; Wasserstein et al., 2019; Ziliak, 2019). Reluctance to adhere to modern statistical practices has led scientist to believe that a more drastic shift in statistical thinking is needed, and some believe that it might come in the form of Bayesian statistics (Dunson, 2001; Gelman et al., 2014; Kruschke, 2014; McElreath, 2018).

Some software tools and packages already bring Bayesian statistics to broader audiences. JASP (Love et al., 2019) is a graphical statistical software that also implements Bayesian alternatives for some common statistical tests (e.g., t-test, ANOVA, ...). JASP allows execution of statistical analyses through its highly intuitive graphical user interface. Another great tool for executing elementary Bayesian analyses is Rasmus Bååth's `BayesianFirstAid` (Bååth, 2014). The goal of this `R` package is to replace the classic elementary statistical tests with their Bayesian counterparts. Since both JASP (Love et al., 2019) and `BayesianFirstAid` (Bååth, 2014) focus on the most elementary statistical tests, the tools they offer are often insufficient when working with more complex data sets. The development of a package that would cover all needs of modern science is impossible, but as a subset of specialized Bayesian models is sufficient to cover the majority of analyses in psychology, we developed the `bayes4psy` `R` package.

The `bayes4psy R` package provides a state-of-the art framework for Bayesian analysis of psychological data. It incorporates a set of probabilistic models for analysing data that arise during many types of psychological experiments. All models are pre-compiled, meaning that we do not need any specialized software or skills (e.g., knowledge of probabilistic programming languages). The only requirements are the `R` programming language and very basic programming skills (same skills as needed for classical statistical analysis in `R`). The package also incorporates the diagnostic, analytic and visualization tools required for modern Bayesian data analysis. The `bayes4psy` package represents a bridge into the exciting world of Bayesian statistics for students and researches in the field of psychology.

2. METHODS AND MODELS

For statistical computation (sampling from the posterior distributions) the `bayes4psy` package utilizes `Stan` (Carpenter et al., 2017). `Stan` is a state-of-the-art platform for statistical modeling and high-performance statistical computation and offers full Bayesian statistical inference with MCMC sampling. It also offers friendly interfaces with most programming languages used for statistical analysis, including `R`. `R` (R Core Team, 2017) is one of the most powerful and widespread programming languages for statistics and visualization. Visualizations in the `bayes4psy` package are based on the `ggplot2` package (Wickham, 2009).

Bayesian analysis requires three key pieces of information—the input data, the statistical model and the priors. By far the most complex of the three is the development of a statistical model, which requires extensive knowledge in probabilistic programming. To avoid this difficult step, the `bayes4psy` package includes an already prepared collection of models for analysing the most common types of data arising from psychological research.

2.1. The Input Data

In psychology and many other scientific fields data are typically gathered with experiments, surveys, questionnaires, observations, and other similar data collection methods. Gathering and preparing the data for use with the `bayes4psy` package is the same as for any other statistical analysis.

2.2. The Statistical Model

The `bayes4psy` package contains a collection of Bayesian models suitable for analysing common types of data that arise during psychological experiments. The packages includes the Bayesian t-test and bootstrap and models for analysing reaction times, success rates, and tasks utilizing colors as a response. Besides the models, we also prepared the diagnostic, analytic, and visualization tools for the modern Bayesian data analysis workflow.

Statistical models are defined through distributions and their parameters. For example, the Bayesian t-test utilizes a generalized t-distribution which has three parameters—degrees of freedom

ν, location/mean μ, and scale/variance σ. In the remainder of the paper, we describe and visualize all the models in the `bayes4psy` package.

2.3. Priors

In Bayesian statistics we use prior probability distributions (priors) to express our beliefs about the model's parameters before any evidence (data) is taken into account. Priors represent an elegant way of combining (pre)existing knowledge with new facts about the domain of analysis. Prior distributions are usually based on past research or domain expertise. If prior information is unavailable, we usually resort to weakly informative, vague priors. We can also leverage prior information to increase the power of small-sample studies.

In the `bayes4psy` package we can express prior knowledge with prior distributions on all of the model's parameters. The package supports uniform, normal, gamma and beta prior distributions. By default flat/improper priors are used for all of the model's parameters. For details, see the illustrative examples in section 3.

2.4. Outputs

The outputs of the MCMC-based Bayesian inference are samples. These samples represent credible values for parameters of the chosen statistical model. For example, the samples of the Bayesian *t*-test model contain values for the parameters of the underlying *t*-distribution—degrees of freedom ν, mean μ, and variance σ. Once we acquire these samples, typically hundreds or thousands of them, we can use them for statistical inference. The samples can be used in a number of ways, for example, we can use them to compare means of two or more groups, we can reconstruct the estimated distribution of the population, we can describe the group by calculating summary statistics (e.g., mean, confidence interval) of certain parameters.

2.5. A Simplified Example

Suppose we are interested in comparing the mean heights of Europe and US primary school pupils. First, we need to define our inputs—the input data, the statistical model and the priors. The input data are the actual height measurements of the pupils. Next, we have to pick an appropriate model. Since we are interested in comparison of the means, we can use the model for the Bayesian *t*-test (see the section 2.6 for a detailed explanation of this model). This model has three parameters—degrees of freedom ν, mean μ, and variance σ. We can specify priors for these parameters or use the default non-informative priors. An example of a weakly informative or vague prior in this example would be a uniform distribution $\mathcal{U}(0, 200)$ for the μ parameter. With this prior on μ we are postulating that mean height of primary school pupils lies strictly somewhere between 0 and 200 cm. Priors can be based on previous studies or expert knowledge. For example, since mean height of primary school pupils is around 120 ± 20 cm a reasonable informative prior for the μ parameter could be $\mathcal{N}(120, 20)$. In a similar way we can define priors for ν and σ.

Once we have selected the priors, we are ready to infer the distributions underlying the chosen model (fit the model) to our data for each of the two groups (height of pupils in Europe and height of pupils in USA). The output of the inference process are the generated samples of the model's parameters. Suppose that the generated samples are μ_{EU} = [123, 128, 121, 137, 110 cm] and μ_{US} = [118, 126, 119, 110, 122 cm]. We can compare the mean height of these two groups by executing a pair-wise comparison of the μ samples. In this example we can claim with 80% certainty that European pupils are higher than their US counterparts (in four out of five samples, the μ parameter of European pupils is higher—123 > 118 cm, 128 > 126, 121 > 119 cm, 137 > 110 cm, 110 < 122 cm). Note that in practice we would typically have hundreds or thousands of samples.

We can also check if means of two groups is equal. One way of doing this is by defining the ROPE (Region Of Practical Equivalence) interval. For example, if our measuring equipment had a tolerance of 0.2 cm, then it would make sense to set the ROPE interval to [−0.2, 0.2]. Samples from both groups that differ for <0.2 cm would be interpreted as equal and we would be able to compute the probability that the means are (practically) equal.

2.6. Bayesian *t*-Test

The *t*-test is one of the most popular statistical tests. In `bayes4psy` it is based on Kruschke's model (Kruschke, 2013, 2014) which uses a scaled and shifted Student's *t*-distribution (**Figure 1**). This distribution has three parameters—degrees of freedom (ν), mean (μ), and variance (σ).

There are some minor differences between our implementation and Kruschke's. Instead of pre-defined vague priors for all parameters, we can define custom priors for the ν, μ, and σ. Kruschke's implementation models two data sets simultaneously, while in `bayes4psy` we can model several data sets individually and then make pairwise comparisons or

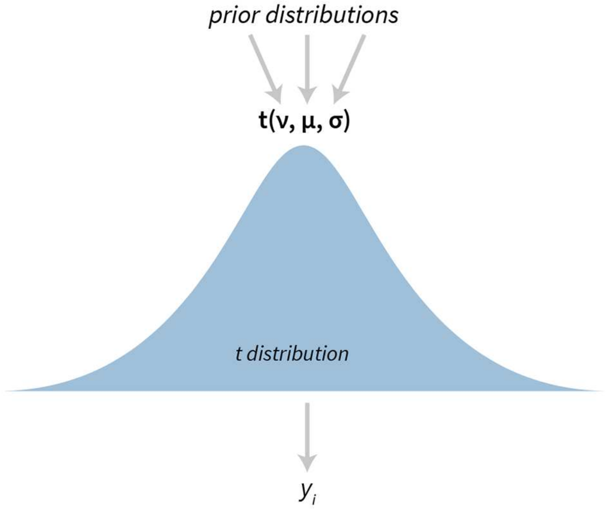


FIGURE 1 | The visualization of the Bayesian *t*-test. The model has three parameters—degrees of freedom ν, mean μ, and variance σ. y_i denotes *i*-th datum in the provided data set.

a simultaneous cross comparison between multiple fits. We illustrate the use of the t-test in section 3.3.

2.7. Model for Analysing Reaction Times

Psychological experiments typically have a hierarchical structure—each subject (participant) performs the same test for a number of times, several subjects are then grouped together by their characteristics (e.g., by age, sex, health) and the final statistical analysis is conducted at the group level. Such structure is ideal for Bayesian hierarchical modeling (Kruschke, 2014).

Our subject-level reaction time model is based on the exponentially modified normal distribution. This distribution has proven to be a suitable interpretation for the long tailed data that arise from reaction time measurements Lindeløv (2019). Note here, that the exponentially modified normal distribution is flexible and can also accommodate the cases in which data are distributed normally. To model the data at the group level we put hierarchical normal priors on all parameters of the subject-level exponentially modified normal distribution.

The subject level parameters are thus μ_i, σ_i, and λ_i, where i is the subject index. And hierarchical normal priors on these parameters are $\mathcal{N}(\mu_\mu, \sigma_\mu)$ for the μ parameter, $\mathcal{N}(\mu_\sigma, \sigma_\sigma)$ for the σ parameter and $\mathcal{N}(\mu_\lambda, \sigma_\lambda)$ for the λ parameter. See **Figure 2** for a graphical representation of the Bayesian reaction time model. For a practical application of this model see section 3.1.

In the case of an exponentially modified normal distribution means are calculated using the μ and λ parameters. By default, `bayes4psy` reports means on the group level, calculated as $E = \mu_\mu + 1/\mu_\lambda$.

2.8. Model for Analyzing Success Rates

The success rate model is based on the Bernoulli-Beta model that can be found in most Bayesian statistics textbooks (Gelman et al., 2014; Kruschke, 2014; McElreath, 2018). This model is used for modeling binary data. In our case this binary output represents whether a subject successfully solved the given task or not.

The success rates model also has a hierarchical structure. The success rate of individual subjects is modeled using Bernoulli distributions, where the p_i is the success rate of subject i. A reparameterized Beta distribution, Beta$(p\tau, (1-p)\tau)$, is used as a hierarchical prior on subject-level parameters, where p is the group level success rate and τ is the scale parameter. A graphical representation of our hierarchical success rate model can be seen in **Figure 3**. For a practical application of this model see section 3.1.

2.9. Model for Analysis of Sequential Tasks

In some psychological experiments data have a time component or some other ordering. For example, when subjects are asked to perform a sequence of tasks. To model how a subject's performance changes over time, we implemented a hierarchical linear normal model.

The sequence for a subject is modeled using a simple linear model with subject-specific slope and intercept. To model the data at the group level we put hierarchical normal priors on all parameters of the subject-level linear models. The parameters of subject i are α_i for the intercept, β_i for the slope and σ_i for modeling errors of the fit (residuals). The hierarchical normal priors on these parameters are $\mathcal{N}(\mu_\alpha, \sigma_\alpha)$ for the intercept α, $\mathcal{N}(\mu_\beta, \sigma_\beta)$ for the slope β and $\mathcal{N}(\mu_\sigma, \sigma_\sigma)$ for the residuals (σ).

A graphical representation of the model is shown in **Figure 4**. For a practical application of this model see section 3.2.

2.10. Model for Analysis of Tasks Utilizing Colors as a Response

This model is designed for experiments in which subject's response comes in the form of a color (e.g., subjects have to pick a color that describes their mood, subject have to remember a color and then pick it from a color palette after a certain time interval ...). Color stimuli and subject responses in psychological experiments are most commonly defined through the RGB color model. The name of the model comes from the initials of the three additive primary colors, red, green, and blue. These colors are also the three components of the model, where each component has a value ranging from 0 to 255 which defines the presence of a particular color. Since defining and analysing colors through the RGB model is not very user friendly and intuitive, our Bayesian model is capable of working with both the RGB and HSV color models. HSV (hue, saturation and value) is an alternative representation of the RGB model that is usually easier to read and interpret for most human beings.

The Bayesian color model works in a component-wise fashion. Six distributions (three for the RGB components and three for the HSV components) are inferred from the data for each component individually. For RGB components we use normal distributions (truncated to the [0, 255] interval). In the HSV case, we used [0, 1]-truncated normal distributions for saturation and value components and the von Mises distribution for the hue component. The von Mises distribution (also known as the circular normal distribution) is a close approximation to the normal distribution wrapped on the $[0, 2\pi]$ interval. A visualization of our Bayesian model for colors can be seen in **Figure 5** and its practical application in section 3.4.

2.11. The Bayesian Bootstrap

The bootstrap is a resampling technique for computing standard deviations, confidence intervals and other estimates for quantifying uncertainty. It uses sampling with replacement to approximate the sampling distribution of an estimator and is applicable in a uniform way to a wide range of scenarios.

The Bayesian bootstrap in `bayes4psy` is the analog of the classical bootstrap (Efron, 1979). It is based on Rasmus Bååth's implementation (Bååth, 2015), which in turn is based on methods developed by Rubin (1981). The Bayesian bootstrap does not simulate the sampling distribution of a statistic estimating a parameter, but instead simulates the posterior distribution of the parameter. The statistical model underlying the Bayesian bootstrap can be characterized by drawing weights from a uniform Dirichlet distribution with the same dimension as the number of data points. These draws are then used for calculating the statistic in question and weighing the data (Bååth, 2015). For more details about the implementation see Bååth (2015) and Rubin (1981).

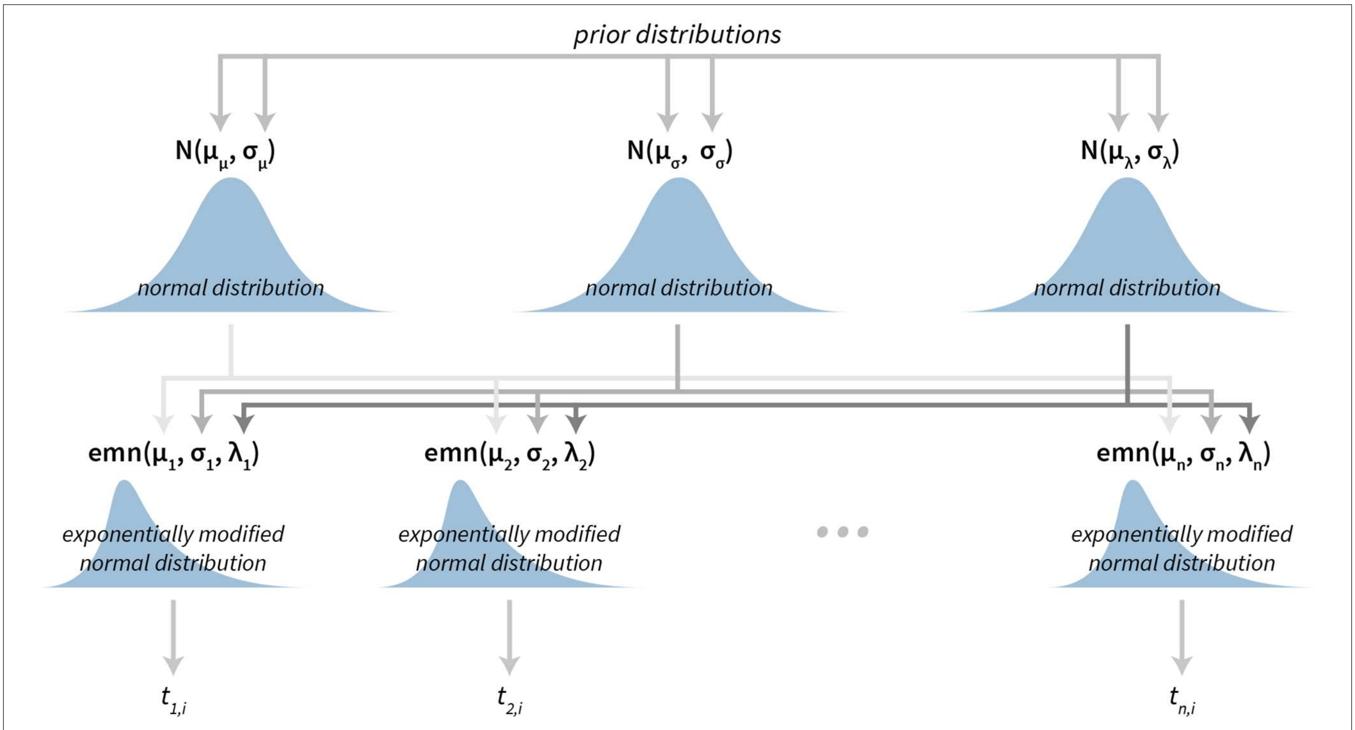

prior distributions
N(μ_μ, σ_μ)
normal distribution
N(μ_σ, σ_σ)
normal distribution
N(μ_λ, σ_λ)
normal distribution
emn(μ_1, σ_1, λ_1)
exponentially modified normal distribution
emn(μ_2, σ_2, λ_2)
exponentially modified normal distribution
• • •
emn(μ_n, σ_n, λ_n)
exponentially modified normal distribution
$t_{1,i}$
$t_{2,i}$
$t_{n,i}$

FIGURE 2 | The visualization of the Bayesian reaction time model. The model has a hierarchical structure. Reaction times belonging to each individual subject ($t_{n,i}$ depicts i-th reaction time of the subject n) are used to construct exponentially modified normal distributions at the subject level. Parameters of subject level distributions are then connected at the group level by using normal distributions, which can then be used for group level analysis.

2.12. Methods for Fitting and Analysing Bayesian Fits

This section provides a quick overview of all the methods for fitting and analysing the models described in previous sections. For a more detailed description of each function we invite the reader to consult the `bayes4psy` package documentation and examples.

The first set of functions infers the parameters of model's distributions from the input data, in other words these functions fit the model to the data. We can also use these functions to define priors (for an example, see the second part of section 3.1) or configure the fitting parameters. This way we can set the number of generated samples (number of MCMC iterations) along with several other parameters of the MCMC algorithm. Some basic MCMC settings are described in this manuscript and the documentation of this package, for more advanced settings consult the official `Stan` documentation (Carpenter et al., 2017).

- `b_ttest` is used for fitting the Bayesian *t*-test model. The input data comes in the form of a vector of normally distributed real numbers.
- `b_linear` is used for fitting the hierarchical linear model, suitable for analysing sequential tasks. The input data are three vectors—x a vector containing values of the independent variable (time, question index ...), y a vector containing values of the dependent variable (subject's responses) and s a vector containing IDs of subjects, these IDs are used for denoting that x_i/y_i pair belongs to a particular subject.
- `b_reaction_time` is used for the Bayesian reaction time model. Its input data are two vectors—vector t includes reaction times while vector s is used for linking reaction times with subjects.
- `b_success_rate` is used for fitting the Bayesian success rate model. Its input data are two vectors, the first vector r contains results of an experiment with binary outcomes (e.g., success/fail, hit/miss ...) and the second vector s is used to link these results to subjects.
- `b_color` is used for fitting the color model. The input data to this model is a three column matrix or a `data.frame` where each column represents one of the components of the chosen color model (RGB or HSV). If the input data are provided in the HSV format then we also have to set the *hsv* parameter to `TRUE`.
- `b_bootstrap` function can be used for Bayesian bootstraping. The input data can be in the form of a vector, matrix or a `data.frame`. The Bayesian bootstrap also requires the specification of the statistics function.

Before interpreting the results, we can use the following functions to check if the model fits are a credible representation of the input data:

- `plot_trace` draws the Markov chain trace plot for main parameters of the model, providing a visual way to

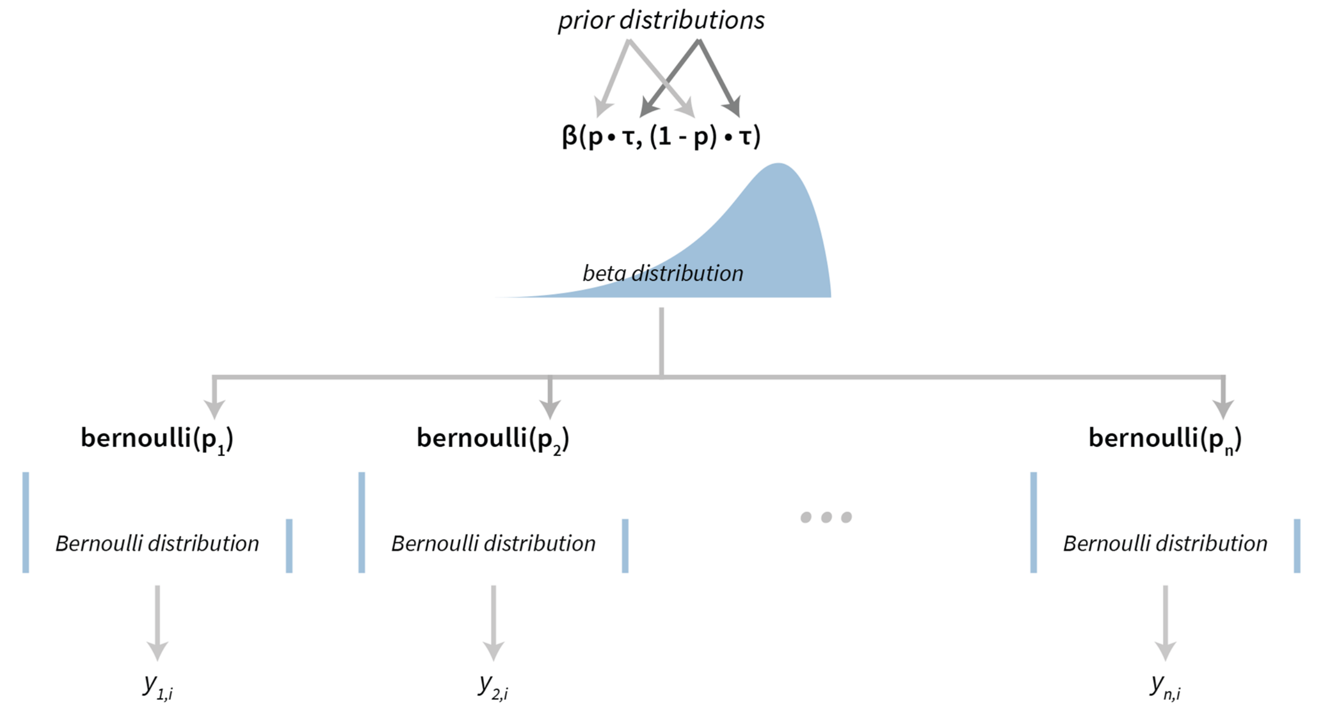


FIGURE 3 | The visualization of the Bayesian success rate model. The model has a hierarchical structure. Data about success of individual subjects ($y_{n,i}$ depicts success on the i-th attempt of the subject n) are used for inferring Bernoulli distributions on the subject level. Parameters of subject level distributions are then connected at the group level with a Beta distribution.

inspect sampling behavior and assess mixing across chains and convergence.

- `plot` or `plot_fit` draws the inferred distributions against the input data. With hierarchical models we can use the subjects parameter to draw fits on the subject level.
- `plot_hsv` or `plot_fit_hsv` are special functions for inspecting color model fits by using a color wheel visualization of HSV components.

For a summary of the posterior with Monte Carlo standard errors and confidence intervals we can use the `summary` or `print`/`show` functions:

- `summary` prints summary statistics of the main model's parameters.
- `print`, `show` prints a more detailed summary of the model's parameters. It includes estimated means, Monte Carlo standard errors (`se_mean`), confidence intervals, effective sample size (`n_eff`, a crude measure of effective sample size), and the R-hat statistic for measuring auto-correlation. R-hat measures the potential scale reduction factor on split chains and equals 1 at convergence (Gelman and Rubin, 1992; Brooks and Gelman, 1998).

The `compare_means` function can be used for comparison of parameters that represent means of the fitted models. To visualize these means one can use the `plot_means` function and for visualizing the difference between means the `plot_means_difference` function. All comparison functions (functions that print or visualize the difference between fitted models) also offer the option of defining the ROPE interval by setting the `rope` parameter.

- `compare_means` prints and returns a `data.frame` containing the comparison. It can be used for comparing two or multiple models at the same time.
- `plot_means_difference` visualizes the difference of means between two or multiple models at the same time.
- `plot_means` plots the distribution of parameters that depict means. It can be used on a single or multiple models at the same time.
- `plot_means_hsv` is a special function for the Bayesian color model that plots means of HSV components by using a color wheel visualization.

The following set of functions works in a similar fashion as the one for comparing means, the difference is that this one compares entire distributions and not just the means. This analysis is based on the comparison of a large amount of samples drawn from the distributions.

- `compare_distributions` prints and returns a `data.frame` containing the comparison results. It can be used for comparing two or multiple models at the same time.
- `plot_distributions_difference` visualizes the difference of distributions underlying two or multiple fits at the same time.

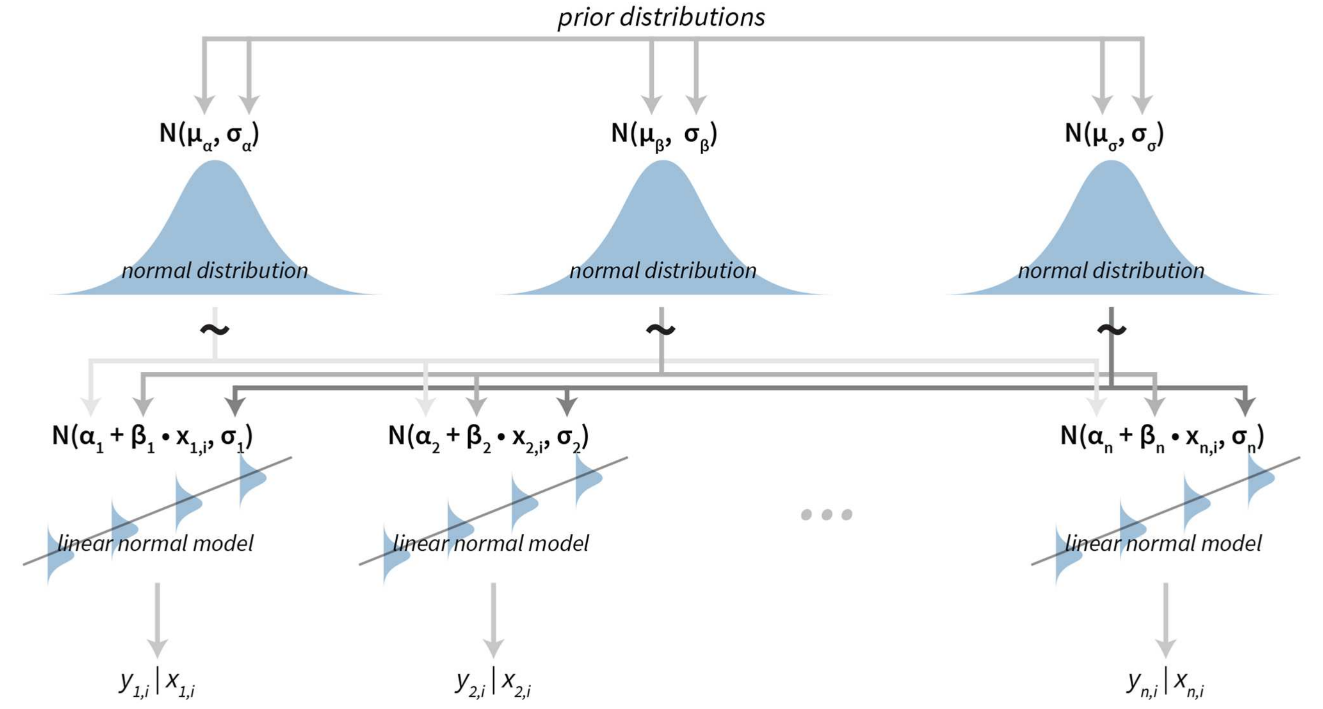


FIGURE 4 | The visualization of the hierarchical linear model. The model has a hierarchical structure, linear normal models are fitted on the subject level from data belonging to each particular subject. Since the ordering is important input data come in pairs of dependent (e.g., result or answer) and independent variables (e.g., time or the question index). The term $y_{n,i}|x_{n,i}$ is the value of the *i*-th dependent variable given the value of the independent variable *i* for the subject *n*. Parameters of subject level distributions are joined on the group level by using normal distributions. These group level distributions can then be used for group level analysis of the data.

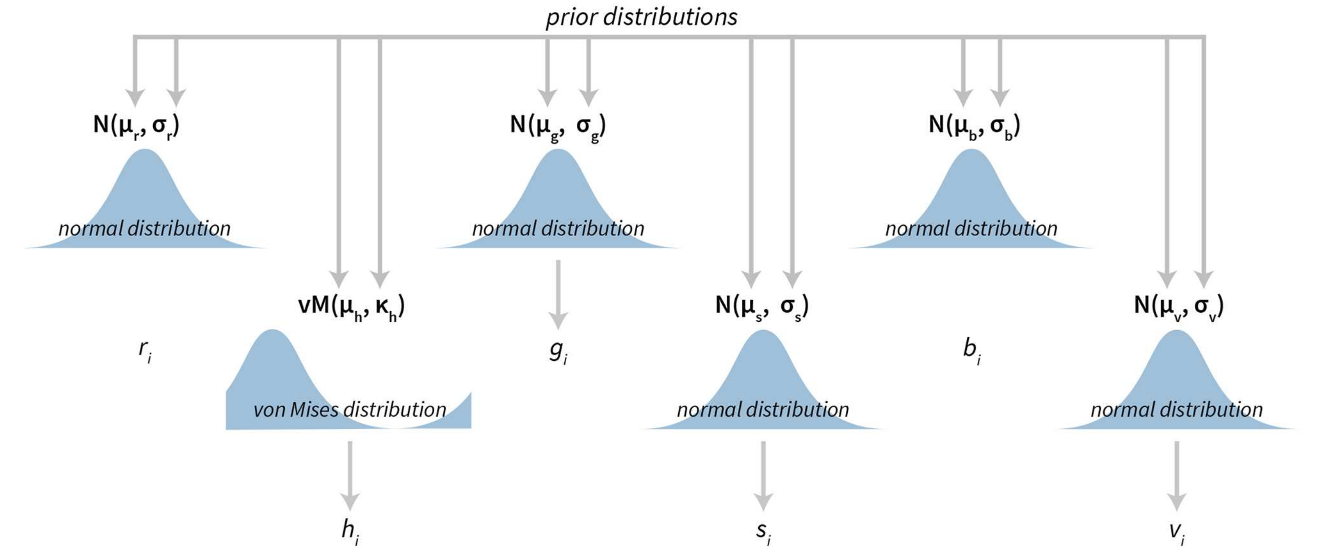


FIGURE 5 | The visualization of the Bayesian color model. The model is composed of six parts. Three parts are used to describe the RGB (red, green, blue) color model components and three parts are used to describe the HSV (hue, saturation, value) color model components. All components, except hue, are modeled with normal distributions, while hue is modeled with the von Mises distribution—a circular normal distribution.

- `plot_distribution` plots the distributions underlying the fitted models, can be used on a single or multiple models at the same time.
- `plot_distributions_hsv` is a special function for the Bayesian color model that plots the distribution behind HSV components by using a color wheel like visualization.

We can also extract samples from the posterior for further custom analyses:

- `get_parameters` returns a `data.frame` of model's parameters. In hierarchical models this returns a `data.frame` of group level parameters.
- `get_subject_parameters` can be used to extract subject level parameters from hierarchical models.

3. ILLUSTRATIVE EXAMPLES

For the sake of brevity, we are presenting diagnostic visualizations and outputs only the first time they appear and omit them in later examples. The datasets used in the examples are based on the experiments conducted by the Mind & Brain Lab at the Faculty of Arts, University of Ljubljana. All datasets are included in the `bayes4psy` package.

3.1. The Flanker Task

In the Eriksen flanker task (Eriksen and Eriksen, 1974) participants are shown an image of an odd number of arrows (usually five or seven). Their task is to indicate the orientation (left or right) of the middle arrow as quickly as possible. There are two types of stimuli: in the *congruent* condition (e.g., "<<<<<<<") both the middle arrow and the flanking arrows point in the same direction and in the *incongruent* condition (e.g., "<<<><<<") where the middle arrow points in the opposite direction.

The participants have to consciously ignore and inhibit the misleading information provided by the flanking arrows in the incongruent condition, which leads to robustly longer reaction times and a higher proportion of errors. The difference between reaction times and error rates in congruent and incongruent conditions is a measure of the subject's ability to focus and to inhibit distracting stimuli.

In the illustration below we compare reaction times and error rates when performing the flanker task between the control group (healthy subjects) and the test group (subjects suffering from a certain medical condition).

First, we load `bayes4psy` and `dplyr` (Wickham et al., 2018) for data wrangling. Second, we load the data and split them into control and test groups. For reaction time analysis we use only data where the response to the stimuli was correct:

```
R> library(bayes4psy)
R> library(dplyr)

R> data <- flanker

R> control_rt <- data %>%
          filter(result == "correct" &
                 group == "control")

R> test_rt <- data %>%
       filter(result == "correct" &
              group == "test")
```

The model requires subjects to be indexed from 1 to n. Control group subject indexes range from 22 to 45, so we have to cast them to an interval that ranges from 1 to 23. Note here, that even though this way both control and test subject have some indexes, they will be still treated as separate individuals because the models for test and control subjects will be fitted separately.

```
R> control_rt$subject <- control_rt$subject - 21
```

Now we are ready to fit the Bayesian reaction time model to data from both groups. The modeling function (`b_reaction_time`) requires two parameters—a vector of reaction times t and the vector of subject indexes s.

```
R> rt_control_fit <- b_reaction_time(t=control_rt$
     rt,
                                     s=control_rt$
                                        subject)

R> rt_test_fit <- b_reaction_time(t=test_rt$rt,
                                  s=test_rt$
                                     subject)
```

Before we interpret the results, we check MCMC diagnostics (such as the traceplot on **Figure 6**, the `Rhat` metric and the effective sample size) and inspect model's fit.

```
R> plot_trace(rt_control_fit)
R> plot_trace(rt_test_fit)

R> print(rt_control_fit)

Inference for Stan model: reaction_time.
4 chains, each with iter=2000; warmup=1000; thin
     =1;
post-warmup draws per chain=1000, total post-
     warmup draws=4000.

             mean se_mean sd 2.5% 97.5% n_eff Rhat
mu[1]        0.46   0.00 0.01 0.44  0.47  4789  1
mu[2]        0.36   0.00 0.01 0.35  0.38  4661  1
...
sigma[1]     0.04   0.00 0.01 0.03  0.05  5406  1
sigma[2]     0.03   0.00 0.01 0.02  0.04  5165  1
...
lambda[1]   14.41   0.02 1.62 1.59 17.87  4441  1
lambda[2]   11.59   0.02 1.15 9.53 14.01  5271  1
...
mu_m         0.51   0.00 0.01 0.48  0.54  5589  1
mu_l         6.86   0.01 0.91 5.12  8.75  5299  1
mu_s         0.07   0.00 0.01 0.06  0.08  4115  1
sigma_m      0.06   0.00 0.01 0.05  0.09  6078  1
sigma_l      4.24   0.01 0.78 3.02  5.99  3940  1
sigma_s      0.02   0.00 0.00 0.01  0.03  3862  1
rt           0.66   0.00 0.02 0.61  0.71  5112  1
rt_subjects  0.53   0.00 0.01 0.51  0.54  4261  1
[1]
rt_subjects  0.45   0.00 0.01 0.44  0.47  5654  1
[2]
...

R> print(rt_test_fit)
```

The output above is truncated and shows only values for 2 of the 24 subjects on the subject level of the hierarchical model. The output provides further MCMC diagnostics, which again do not give us any cause for concern. The convergence diagnostic `Rhat` is practically 1 for all parameters and there is little auto-correlation—effective sample sizes (`n_eff`) are of the order of

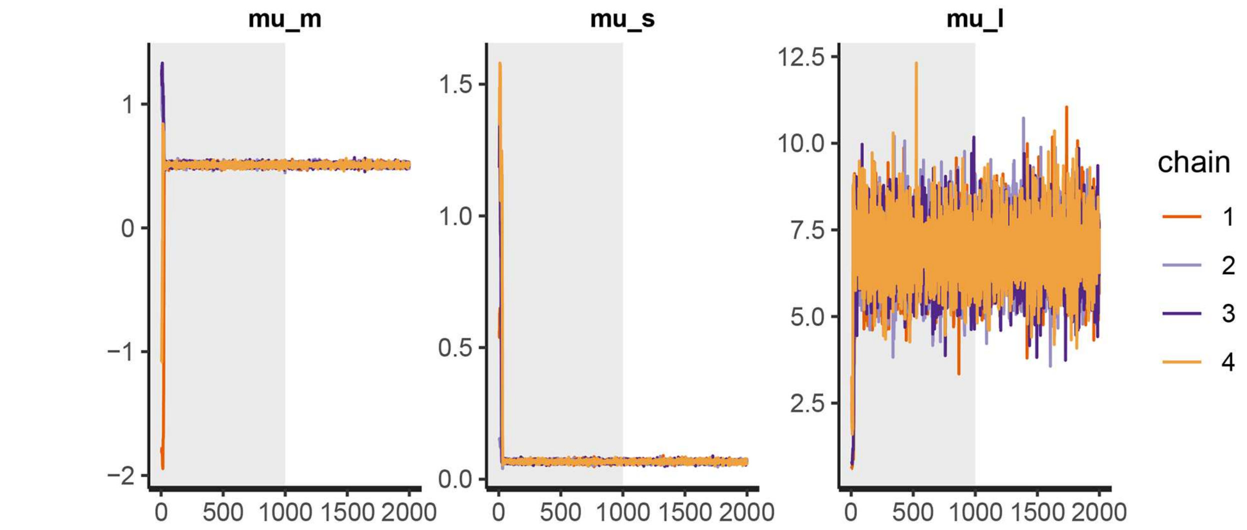


FIGURE 6 | The trace plot for `rt_control_fit`. The traceplot gives us no cause for concern regarding MCMC convergence and mixing. The trace plot for `rt_test_fit` is similar. Note that the first 1,000 iterations (shaded gray) are used for warmup (tuning of the MCMC algorithm) and are discarded. The next 1,000 iterations are used for sampling. Informally speaking, if trace plots after the warmup period look like "hairy caterpillars" there is no reason for concern.

samples taken and Monte Carlo standard errors (`se_mean`) are relatively small.

What is a good-enough effective sample sizes depends on our goal. If we are only interested in estimating the mean, 100 effective samples is in most cases enough for a practically negligible Monte Carlo error. On the other hand if we are interested in posterior quantities, such as extreme percentiles for example, the effective sample sizes might have to be 10,000 or higher.

We can increase the effective sample size by increasing the amount of MCMC iterations with the `iter` parameter. In our case we can achieve an effective sample size of 10,000 by setting `iter` to 4,000. Because the MCMC diagnostics give us no cause for concern, we can leave the `warmup` parameter at its default value of 1,000.

```
R> rt_control_fit <-
   b_reaction_time(t=control_rt$rt,
                   s=control_rt$subject,
                   iter=4000)

R> rt_test_fit <-
   b_reaction_time(t=test_rt$rt,
                   s=test_rt$subject,
                   iter=4000)
```

Because we did not explicitly define priors, default flat (improper) priors were used. In some cases, flat priors are a statement that we have no prior knowledge about the experiment results (in some sense). In general, even flat priors can express a preference for a certain region of parameter space. In practice, we will almost always have some prior information and we should incorporate it into the modeling process.

Next, we should check whether the model fits the data well by using the `plot` function (see **Figure 7**). If we set the `subjects` parameter to `FALSE`, we will get a less detailed group level fit.

```
R> plot(rt_control_fit)

R> plot(rt_test_fit)
```

Since the model fits the data well we can move on with our analysis and use the `compare_means` function to compare reaction times between healthy (control) and unhealthy (test) subjects. In the example below we use a ROPE interval of 0.01 s, meaning that differences smaller that 0.01 of a second are treated as equal. The `compare_means` function provides us with a friendly output of the comparison and the results in the form of a `data.frame`.

```
R> rt_control_test <-
   compare_means(rt_control_fit,
                 fit2=rt_test_fit,
                 rope=0.01)

---------- Group 1 vs. Group 2 ----------
Probabilities:
  - Group 1 < Group 2: 0.98 +/- 0.00409
  - Group 1 > Group 2: 0.01 +/- 0.00304
  - Equal: 0.01 +/- 0.00239
95% HDI:
  - Group 1 - Group 2: [-0.17, -0.01]
```

The `compare_means` function outputs probabilities that one group has shorter reaction times than the other, the probability that both groups are equal (if ROPE interval is provided) and the 95% HDI [highest density interval, Kruschke (2014)] for the difference between groups. Based on the output (Group 1 < Group 2) we can confidently claim (98% ± 0.4%) that the healthy group's (`rt_control_fit`, Group 1) expected reaction times are lower than those from the unhealthy group (`rt_test_fit`, Group 2).

We can also visualize this difference with the `plot_means_difference` function (**Figure 8**), `plot_means` provides an alternative and visualizes the parameters that define the means of each model (**Figure 9**).

```
R> plot_means_difference(rt_control_fit,
```

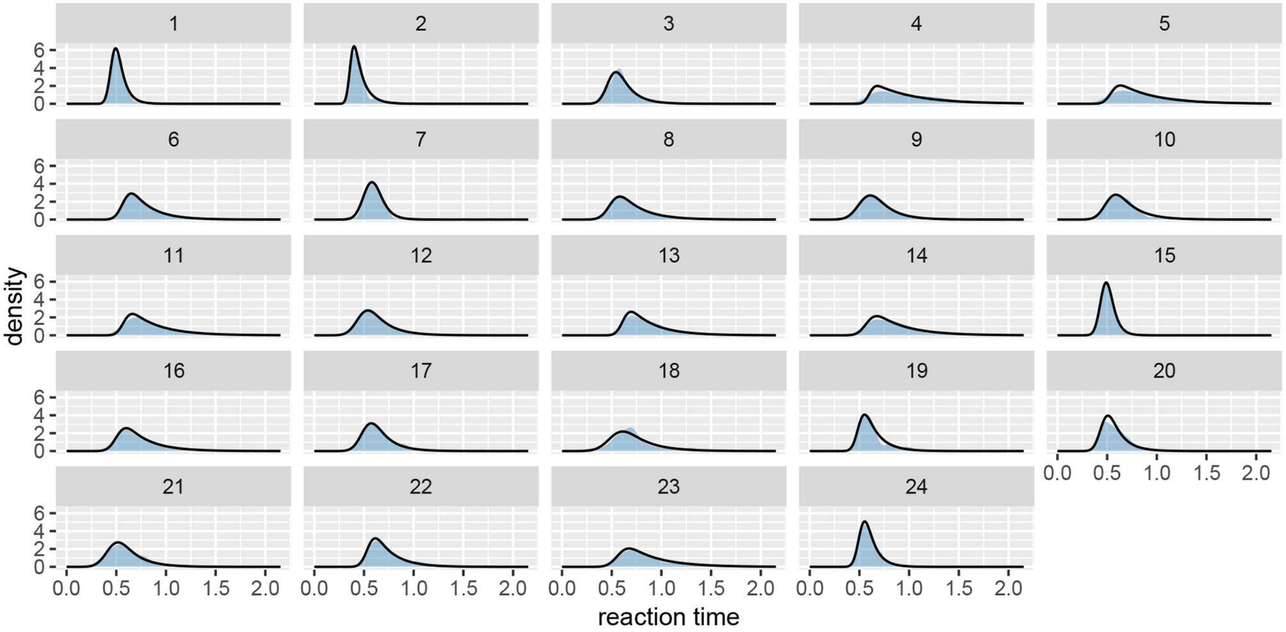


FIGURE 7 | The fit plot for the `rt_control_fit`. The data are visualized as a blue region while the fit is visualized with a black line. In this case the model fits the underlying data well, similar conclusions can be reached for the test group (`rt_test_fit`).

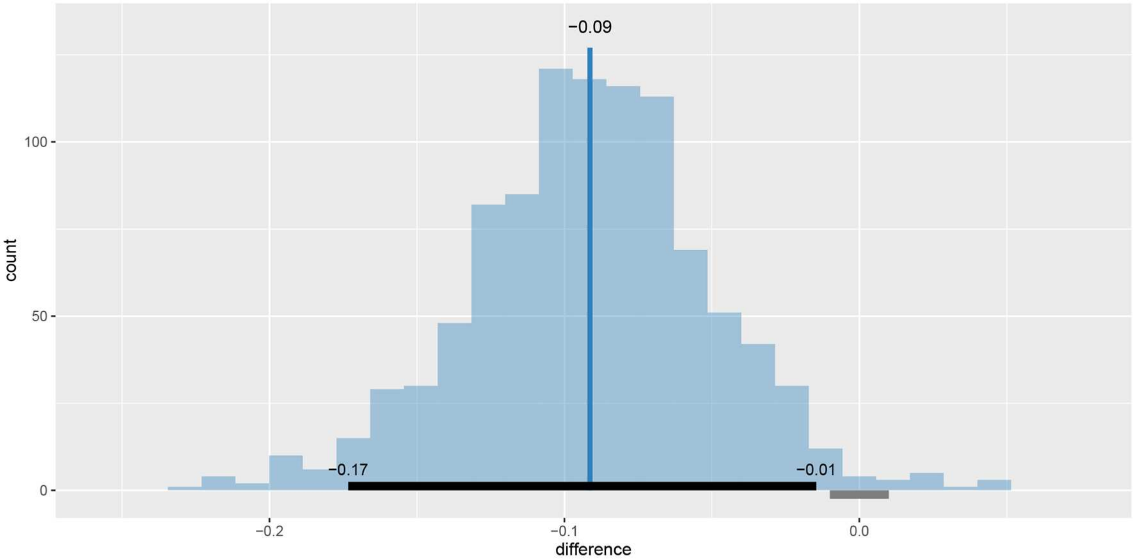


FIGURE 8 | The visualization of the difference in mean reaction times between `rt_control_fit` and `rt_test_fit`. The histogram visualizes the distribution of the difference, vertical blue line denotes the mean, the black band at the bottom marks the 95% HDI interval and the gray band marks the ROPE interval. Since the entire 95% HDI of difference is negative and lies outside of the ROPE interval, we can confidently conclude that healthy subjects are faster on average.

```
                    fit2=rt_test_fit,
                    rope=0.01)

R> plot_means(rt_control_fit,
              fit2=rt_test_fit)
```

To summarize, based on our analysis we can confidently claim that healthy subjects have a lower mean reaction time when solving the flanker task than unhealthy subjects. Next, we analyse if the same applies to success rates.

The information about success of subject's is stored as correct/incorrect. However, the Bayesian success rate model requires binary (0-1) inputs so we first have to transform the data. Also, just like in the reaction time example, we have to correct the indexes of control group subjects.

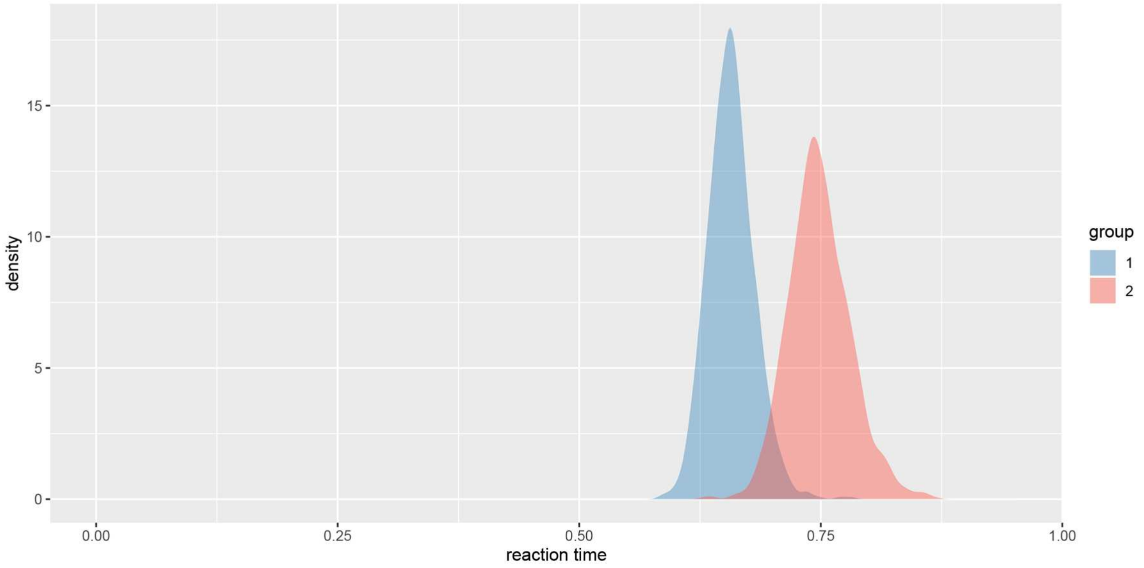


FIGURE 9 | The visualization of means for `rt_control_fit` and `rt_test_fit`. Group 1 visualizes means for the healthy subjects and group 2 for the unhealthy subjects.

```
R> data <- data %>%
     mutate(result_numeric=
         ifelse(result == "correct", 1, 0))

R> control_sr <- data %>%
           filter(group == "control")
R> test_sr <- data %>% filter(group == "test")

R> control_sr$subject <- control_sr$subject - 21
```

Since the only prior information we have about the success rate of participants is that it is between 0 and 1, we used a beta distribution to put a uniform prior on the [0, 1] interval (we put a Beta(1, 1) prior on the p parameter). We fit the model by running the `b_success_rate` function with appropriate input data.

```
R> p_prior <- b_prior(family="beta", pars=c(1, 1))

R> priors <- list(c("p", p_prior))

R> sr_control_fit <-
     b_success_rate(r=control_sr $result_numeric,
                s=control_sr$subject,
                priors=priors,
                iter=4000)

R> sr_test_fit <-
     b_success_rate(r=test_sr$result_numeric,
                s=test_sr$subject,
                priors=priors,
                iter=4000)
```

The process for inspecting Bayesian fits (through `plot_trace` and `print` functions) is the same and since the results are similar as above we omitted them here. When visually inspecting the quality of the fit (the `plot` function) we can set the `subjects` parameter to `FALSE`, which visualizes the fit on the group level. This offers a quicker, but less detailed method of inspection.

```
R> plot_trace(sr_control_fit)
R> plot_trace(sr_test_fit)

R> print(sr_control_fit)
R> print(sr_test_fit)

R> plot(sr_control_fit, subjects=FALSE)
R> plot(sr_test_fit, subjects=FALSE)
```

Since diagnostic functions show no cause for concern and the fits look good we can proceed with the actual comparison between the two fitted models. We will again estimate the difference between two groups with `compare_means`.

```
R> sr_control_test <-
     compare_means(sr_control_fit,
             fit2=sr_test_fit)

---------- Group 1 vs. Group 2 ----------
Probabilities:
  - Group 1 < Group 2: 0.53 +/- 0.01052
  - Group 1 > Group 2: 0.47 +/- 0.01052
95% HDI:
  - Group 1 - Group 2: [-0.02, 0.02]
```

As we can see the success rate between the two groups is not that different. Since the probability that healthy group is more successful is only 53% (± 1%) and the 95% HDI of the difference ([−0.02, 0.02]) includes the 0 we cannot claim inequality (Kruschke, 2014). We can visualize this result by using the `plot_means_difference` function (**Figure 10**).

```
R> plot_means_difference  (sr_control_fit,
              fit2=sr_test_fit)
```

3.2. Adaptation Level

In the adaptation level experiment participants had to assess weights of the objects placed in their hands by using a verbal scale: very very light, very light, light, medium light, medium, medium heavy, heavy, very heavy, and very very heavy. The task was to assess the weight of an object that was placed on the palm of their hand. To standardize the procedure the participants had to place the elbow on the desk, extend the palm and assess the weight of the object after it was placed on their palm by slight up and down movements of their arm. During the experiment participants were blinded by using non-transparent fabric. In total there were 15 objects of the same shape and size but different mass (photo film canisters filled with metallic balls). Objects were grouped into three sets:

- the light set: 45, 55, 65, 75, 85 g (weights 1–5),
- the medium set: 95, 105, 115, 125, 135 g (weights 6–10),
- the heavy set: 145, 155, 165, 175, 185 g (weights 11–15).

The experimenter sequentially placed weights in the palm of the participant and recorded the trial index, the weight of the object and participant's response. The participants were divided into two groups, in group 1 the participants first assessed the weights of the light set in ten rounds within which the five weights in the set were weighted in a random order. After completing the 10 rounds with the light set, the experimenter switched to the medium set. The participant then weighted the medium set across another 10 rounds of weighting the five weights in the medium set in a random order. In group 2 the overall procedure was the same, the only difference being that they started with the 10 rounds of the heavy set and then performed another 10 rounds of weighting on the medium set. Importantly, the weights within each set were given in random order and the experimenter switched between sets seamlessly without any break or other indication to the participant.

We will use the `bayes4psy` package to show that the two groups provide different assessment of the weights in the second part of the experiment even though both groups are responding to weights from the same (medium) set. We will use Bayesian analysis to test the hypothesis that in the second part of the experiment the difference is very pronounced at first but then fades away with subsequent assessments of weights from the medium set. This is congruent with the hypothesis that each group formed a different adaptation level during the initial phase of the task, the formed adaptation level then determined the perceptual experience of the same set of weights at the beginning of the second part of the task.

We will conduct the analysis by using the hierarchical linear model. First we have to construct fits for the second part of the experiment for each group independently. The code below loads and prepares the data, just like in the previous example, subject indexes have to be mapped to a [1, n] interval. We will use the `ggplot2` package to fine-tune graph axes and properly annotate graphs returned by the `bayes4psy` package.

```
R> library(bayes4psy)
R> library(dplyr)
R> library(ggplot2)

R> data <- adaptation_level

R> group1 <- data %>% filter(group == 1)
R> group2 <- data %>% filter(group == 2)

R> n1 <- length(unique(group1$subject))
R> n2 <- length(unique(group2$subject))
```

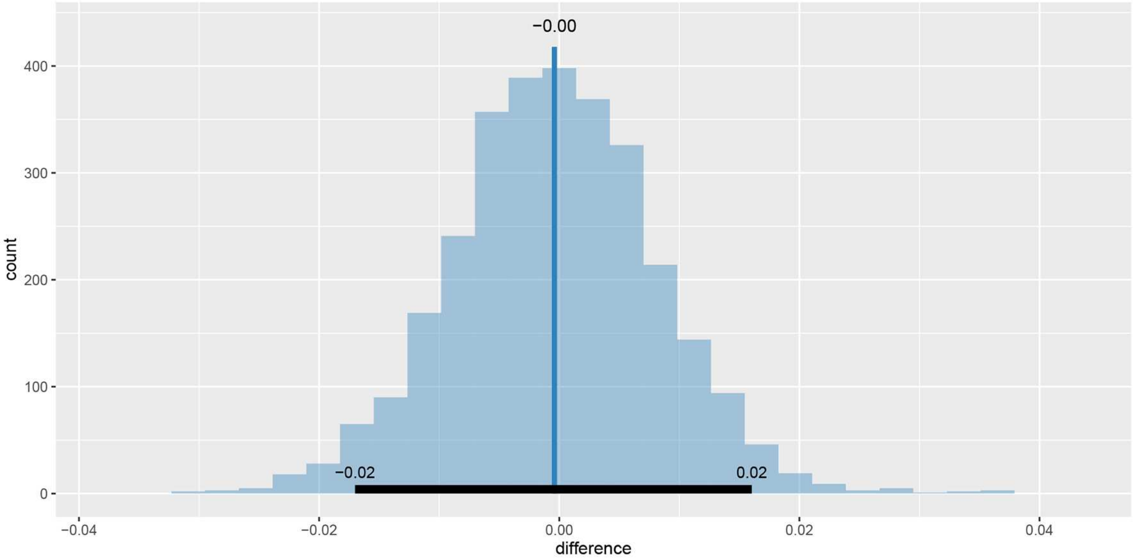


FIGURE 10 | The visualization of the difference between `sr_control_fit` and `sr_test_fit`. The histogram visualizes the distribution of the difference, vertical blue line denotes the mean difference and the black band at the bottom marks the 95% HDI interval. Since the 95% HDI of difference includes 0 we cannot claim inequality. If we used a ROPE interval and the whole ROPE interval lied in the 95% HDI interval we could claim equality.

```
R> group1$subject <-
   plyr::mapvalues(group1$subject,
            from=unique(group1$subject),
            to=1:n1)

R> group2$subject <-
   plyr::mapvalues(group1$subject,
            from=unique(group1$subject),
            to=1:n2)

R> group1_part2 <- group1 %>% filter(part == 2)
R> group2_part2 <- group2 %>% filter(part == 2)
```

Once the data is prepared we can start fitting the Bayesian models, the input data comes in the form of three vectors, *x* stores indexes of the measurements, *y* the subject's responses and *s* indexes of the subjects. The `warmup` and `iter` parameters are set in order to achieve an effective sample size of 10,000.

```
R> fit1 <- b_linear(x=group1_part2$sequence,
                    y=group1_part2$response,
                    s=group1_part2$subject,
                    iter=10000, warmup=500)

R> fit2 <- b_linear(x=group2_part2$sequence,
                    y=group2_part2$response,
                    s=group2_part2$subject,
                    iter=10000, warmup=500)
```

The fitting process is always followed by the quality analysis.

```
R> plot_trace(fit1)
R> plot_trace(fit2)

R> print(fit1)
Inference for Stan model: linear.
4 chains, each with iter=10000; warmup=500; thin
    =1;
post-warmup draws per chain=9500, total post-
    warmup draws=38000.

             mean se_mean   sd    2.5%   97.5% n_
                 eff Rhat
alpha[1]     7.66    0.00 0.31    7.07    8.28
    25452    1
alpha[2]     8.63    0.00 0.23    8.19    9.08
    23074    1
...
beta[1]     -0.14    0.00 0.04   -0.24   -0.06
    20097    1
beta[2]     -0.12    0.00 0.03   -0.19   -0.05
    30442    1
...
sigma[1]     1.67    0.00 0.15    1.41    2.00
    45998    1
sigma[2]     0.99    0.00 0.10    0.82    1.21
    44379    1
...
mu_a         8.05    0.00 0.18    7.68    8.41
    25983    1
mu_b        -0.11    0.00 0.02   -0.15   -0.07
    20126    1
mu_s         1.10    0.00 0.09    0.92    1.29
    33871    1
sigma_a      0.61    0.00 0.16    0.38    0.98
    24984    1
sigma_b      0.05    0.00 0.02    0.01    0.09
    6726    1
sigma_s      0.34    0.00 0.08    0.21    0.54
    30901    1
lp__      -374.28    0.09 6.47 -387.21 -361.12
    5372    1

R> print(fit2)

R> plot(fit1)
R> plot(fit2)
```

The trace plot showed no MCMC related issues (for an example of trace plot see **Figure 6**), effective sample sizes of parameters relevant for our analysis (μ_a, μ_b, and μ_s) are large enough. Since the visual inspection of the fit also looks good we can continue with our analysis. To get a quick description of fits we can take a look at the summary statistics of the model's parameters.

```
R> summary(fit1)
intercept (alpha):
   8.05 +/- 0.00266, 95% HDI: [7.69, 8.39]
slope (beta):
  -0.11 +/- 0.00033, 95% HDI: [-0.15, -0.07]
sigma:
   1.10 +/- 0.00094, 95% HDI: [0.91, 1.28]

R> summary(fit2)
intercept (alpha):
   5.81 +/- 0.00461, 95% HDI: [5.20, 6.43]
slope (beta):
   0.12 +/- 0.00036, 95% HDI: [0.08, 0.16]
sigma:
   1.40 +/- 0.00165, 95% HDI: [1.13, 1.66]
```

Values of intercept (95% HDI intercept equals [7.69, 8.39] for the first group and [5.20, 6.43] for the second group) suggest that our initial hypothesis about adaptation level is true. Subject's that weighted lighter object in the first part of the experiment (`fit1`) find medium objects at the beginning of experiment's second part heavier than subjects that weighted heavier objects in the first part (`fit2`). We can confirm this assumption by using functions that perform a more detailed analysis (e.g., `compare_means` and `plot_means_difference`, see the output below and **Figure 11**).

```
R> comparison_results <-
     compare_means(fit1, fit2=fit2)
---------- Intercept ----------
Probabilities:
  - Group 1 < Group 2: 0.00 +/- 0.00000
  - Group 1 > Group 2: 1.00 +/- 0.00000
95% HDI:
  - Group 1 - Group 2: [1.54, 2.91]

---------- Slope ----------
Probabilities:
  - Group 1 < Group 2: 1.00 +/- 0.00000
  - Group 1 > Group 2: 0.00 +/- 0.00000
95% HDI:
  - Group 1 - Group 2: [-0.29, -0.18]

R> plot_means_difference(fit1,
              fit2=fit2,
              par="intercept")
```

The fact that we are confident in the claims that the slope for the first group is negative (95% HDI for the first group's slope equals [−0.15, −0.07] and lies entirely below 0) and positive for the second group (95% HDI for the second group's slope equals [0.08, 0.16] and lies entirely above 0) suggests that the adaptation

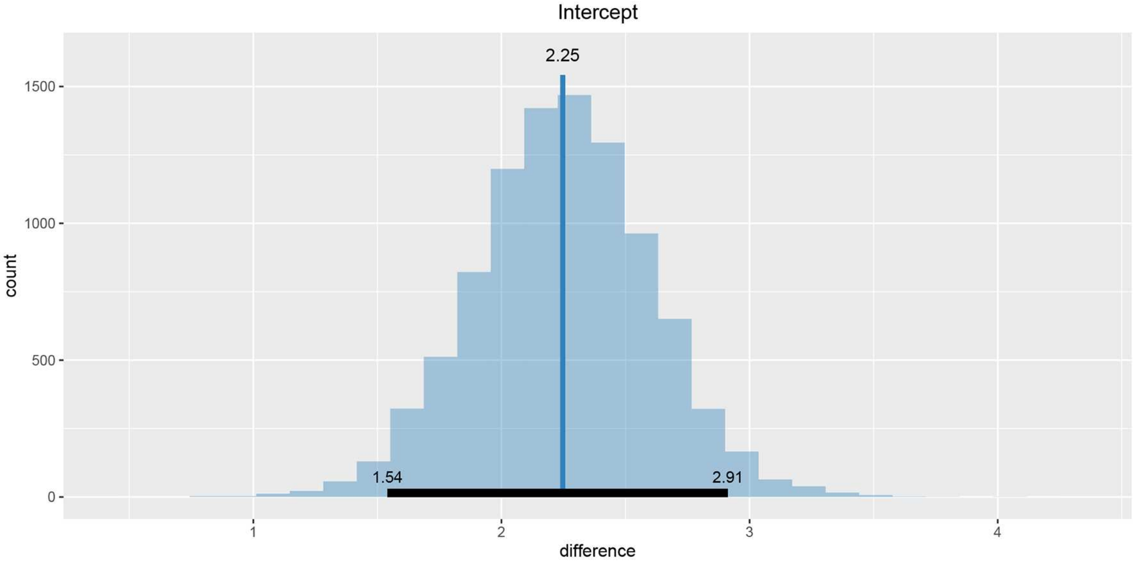


FIGURE 11 | The difference between the intercept of the two fits. Since the entire 95% HDI is positive we are confident that the subject's that weighted lighter object in the first part of the experiment (`fit1`) find medium objects heavier than subjects that initially weighted heavier objects (`fit2`).

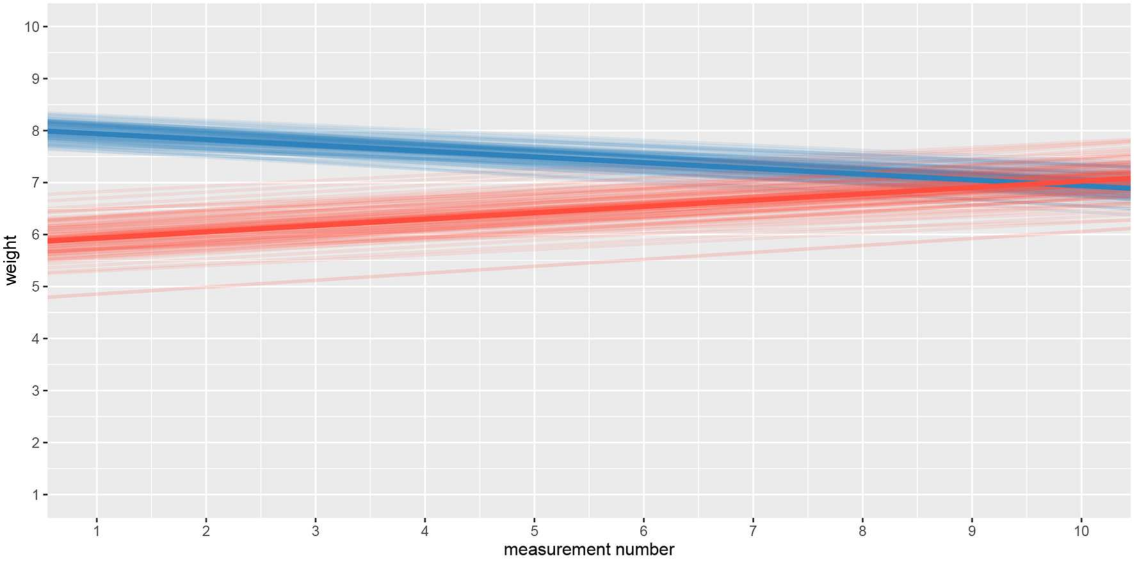


FIGURE 12 | Comparison of distributions underlying `fit1` and `fit2`. The hypothesis that each group formed a different adaptation level during the initial phase of the task seems to be true. The group that switches from heavy to medium weights assesses weights as lighter than they really are, while for the group that switches from light to medium the weights appear heavier. These adaptation levels fade with time and assessments converge to similar estimates of weights.

level phenomenon fades away with time. We can visualize this by plotting means and distributions underlying both fits. The plotting functions in the `bayes4psy` package return regular `ggplot2` plot objects, so we can use the same techniques to annotate or change the look and feel of graphs as we would with the usual `ggplot2` visualizations (see the code below and **Figure 12**).

```
R> plot_distributions(fit1, fit2) +
     labs(title="Part␣II",
          x="measurement␣number",
          y="") +
     theme(legend.position="none") +
     scale_x_continuous(limits=c(1, 10),
                   breaks=seq(1:10)) +
     ylim(0, 10)
```

3.3. The Stroop Color-Word Test

The Stroop test (Stroop, 1935) showed that when the stimuli are incongruent—the name of a color is printed in different ink than the one denoted by its name (for example, red)—naming the color takes longer and is more error-prone than naming the color of a rectangle or a set of characters that does not form a word (for example, XXXXX).

In our version of the Stroop test participants were faced with four types of conditions:

- Reading neutral—the name of the color was printed in black ink, the participant had to read the color's name.
- Naming neutral—string XXXXX was written in colored ink (red, green or blue), the participant had to name the ink color.
- Reading incongruent—name of the color was printed in incongruent ink, the participant had to read the written name of the color.
- Naming incongruent—name of the color was printed in incongruent ink, the participant had to name the ink color.

In each of the listed conditions the participants had to name or read 100 stimuli presented on an A4 sheet of paper organized in 5 columns of 20 stimuli as quickly as possible. The specific order of the stimuli was pseudo-random and balanced across the sheet. We recorded the time to complete each sheet.

In our example analysis, we are primarily interested in expected task completion times. Since our data is composed from average times needed to complete the task we can use the Bayesian *t*-test. The nature of the Stroop test requires the use of *t*-test for dependent samples. This example first shows how to execute the Bayesian *t*-test for dependent samples and in the second part, for illustrative purposes only, also how to execute the Bayesian *t*-test for independent samples. The example for independent samples also shows how to use bayes4psy to compare multiple groups simultaneously.

To execute the Bayesian *t*-test for dependent samples we first have to calculate the difference between the samples and then perform Bayesian modeling on those differences. The example below compares reading times between neutral and incongruent conditions.

```
R> library(bayes4psy)
R> library(dplyr)
R> library(ggplot2)

R> data <- stroop_simple

R> ri_vs_rn <- data$reading_incongruent -
            data$reading_neutral

R> fit_ri_vs_rn <- b_ttest(ri_vs_rn,
                   iter=4000,
                   warmup=500)
```

Once we fit the Bayesian *t*-test model to the differences between the reading neutral and reading incongruent conditions, we can compare whether the means differ from 0.

```
R> comparison <- compare_means(fit_ri_vs_rn, mu=0)
---------- Group 1 vs. Group 2 ----------
Probabilities:
  - Group 1 < Group 2: 0.00 +/- 0.00000
  - Group 1 > Group 2: 1.00 +/- 0.00000
95% HDI:
  - Group 1 - Group 2: [2.03, 3.94]
```

Since the 95% HDI of means ([2.03, 3.94]) lies above 0 we can confidently claim that subject's read neutral stimuli faster than incongruent stimuli. In a similar fashion we can also execute a comparison between other conditions.

The examples that follow are for illustrative purposes only, they analyse the Stroop data under the wrongful assumption that the samples are independent. These examples are in the manuscript mainly to explain how we can use bayes4psy to compare multiple groups simultaneously. The examples also include priors, we based them on our previous experience with similar tasks—participants finish the task in ∼1 min and the typical standard deviation for a participant is <2 min.

```
R> mu_prior <- b_prior(family="normal",
               pars=c(60, 30))

R> sigma_prior <- b_prior(family="uniform",
                  pars=c(0, 120))

R> priors <- list(c("mu", mu_prior),
            c("sigma", sigma_prior))

R> fit_reading_neutral <-
     b_ttest(data$reading_neutral,
             priors=priors,
             iter=4000,
             warmup=500)

R> fit_reading_incongruent <-
     b_ttest(data$reading_incongruent,
             priors=priors,
             iter=4000,
             warmup=500)

R> fit_naming_neutral <-
     b_ttest(data$naming_neutral,
            priors=priors,
            iter=4000,
            warmup=500)

R> fit_naming_incongruent <-
     b_ttest(data$naming_incongruent,
            priors=priors,
            iter=4000,
            warmup=500)
```

There were no causes for concern in the MCMC diagnostics and model fits, so we omit them for brevity. In practice, we should of course always perform these steps. We proceed by cross-comparing several fits with a single line of code.

```
R> fit_list <- c(fit_reading_incongruent,
                 fit_naming_neutral,
                 fit_naming_incongruent)

R> multiple_comparison <-
    compare_means(fit_reading_neutral,
            fits=fit_list)
---------- Group 1 vs. Group 2 ----------
Probabilities:
  - Group 1 < Group 2: 1.00 +/- 0.00054
  - Group 1 > Group 2: 0.00 +/- 0.00054
95% HDI:
  - Group 1 - Group 2: [-4.66, -0.96]

---------- Group 1 vs. Group 3 ----------
Probabilities:
  - Group 1 < Group 3: 1.00 +/- 0.00000
  - Group 1 > Group 3: 0.00 +/- 0.00000
95% HDI:
  - Group 1 - Group 2: [-15.34, -10.19]

---------- Group 1 vs. Group 4 ----------
Probabilities:
  - Group 1 < Group 4: 1.00 +/- 0.00000
  - Group 1 > Group 4: 0.00 +/- 0.00000
95% HDI:
  - Group 1 - Group 2: [-36.72, -28.44]

---------- Group 2 vs. Group 3 ----------
Probabilities:
  - Group 2 < Group 3: 1.00 +/- 0.00000
  - Group 2 > Group 3: 0.00 +/- 0.00000
```

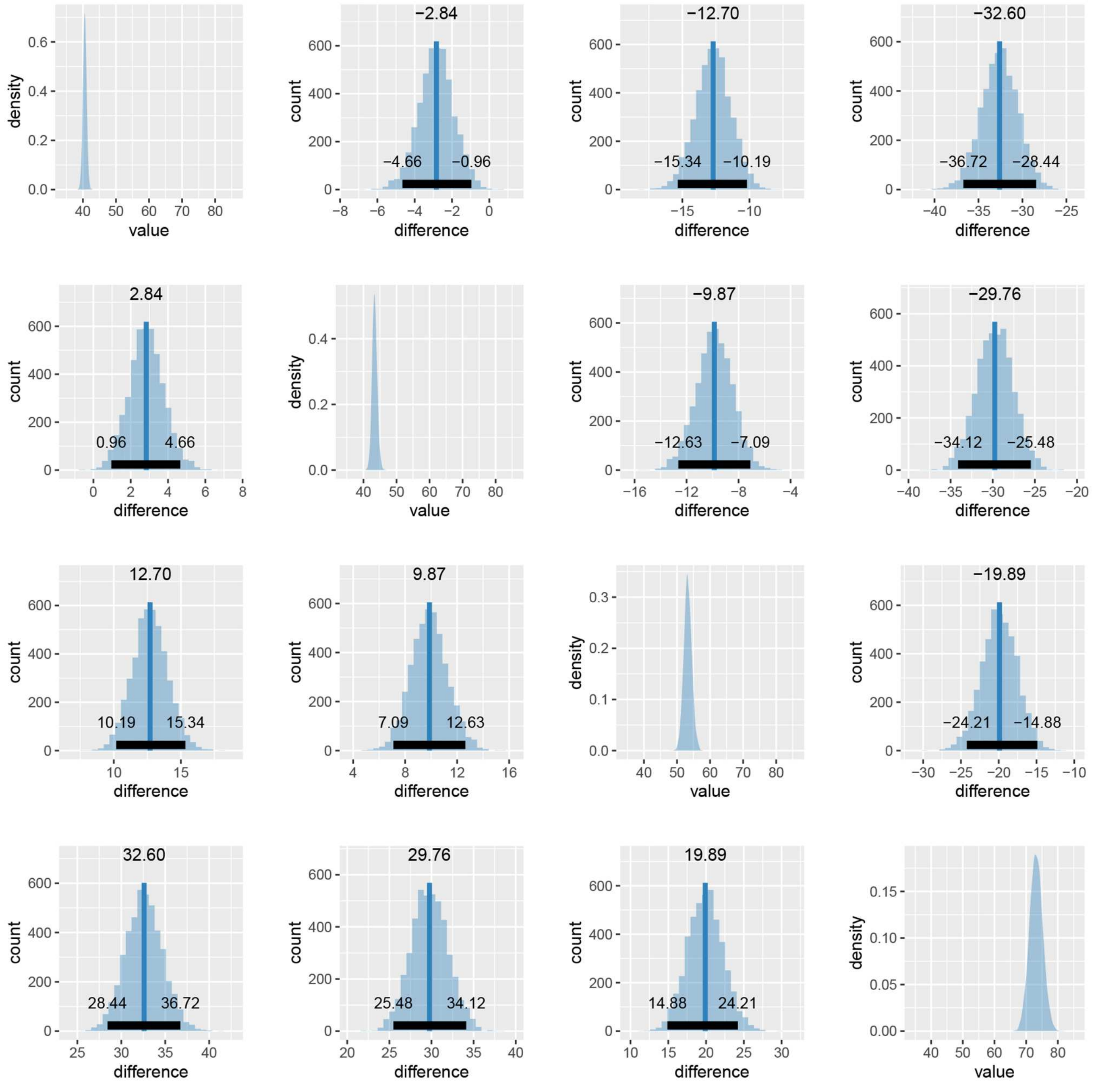


FIGURE 13 | Differences in the mean task completion times for the four conditions. Row and column 1 represent the reading neutral task, row and column 2 the reading incongruent task, row and column 3 the naming neutral task and row and column 4 the naming incongruent task. Since 95% HDI intervals (black bands at the bottom of graphs) in all cases exclude 0 we are confident that the task completion times between conditions are different.

```
95% HDI:
  - Group 1 - Group 2: [-12.63, -7.09]

---------- Group 2 vs. Group 4 ----------
Probabilities:
  - Group 2 < Group 4: 1.00 +/- 0.00000
  - Group 2 > Group 4: 0.00 +/- 0.00000
95% HDI:
  - Group 1 - Group 2: [-34.12, -25.48]

---------- Group 3 vs. Group 4 ----------
Probabilities:
  - Group 3 < Group 4: 1.00 +/- 0.00000
  - Group 3 > Group 4: 0.00 +/- 0.00000
95% HDI:
  - Group 1 - Group 2: [-24.21, -14.88]

----------------------------------------
Probabilities that a certain group is
smallest/largest or equal to all others:

  largest     smallest equal
1       0 0.9991111111     0
2       0 0.0008888889     0
3       0 0.0000000000     0
4       1 0.0000000000     0
```

When we compare more than two fits, we also get an estimate of the probabilities that a group has the largest or the smallest expected value. Based on the above output, the participants are best at the reading neutral task (Group 1), followed by the reading incongruent task (Group 2) and the naming neutral task (Group 3). They are the worst at the naming incongruent task (Group 4). We are very confident that this ordering is correct (the probabilities distinguishing the groups are extremely high), so we can conclude that both naming and incongruency of stimuli increase the response times of subjects, with naming having a bigger effect. We can also visualize this in various ways, either as distributions of mean times needed to solve the given tasks or as a difference between these means (**Figure 13**).

```
R> plot_means(fit_reading_neutral,
        fits=fit_list) +
scale_fill_hue(labels=c("Reading_neutral",
                 "Reading_incongruent",
                 "Naming_neutral",
                 "Naming_incongruent")) +
theme(legend.title=element_blank())

R> plot_means_difference(fit_reading_neutral,
            fits=fit_list)
```

3.4. Afterimages

In the afterimages task participants were asked to fix their gaze on a fixation point in the middle of the computer screen. Stimulus—a colored rectangle—was then shown above the fixation point. After 20 s the rectangle disappeared and a color palette was shown on the right-hand side of the screen. Participants were asked to keep their gaze on the fixation point while using the mouse to select the color that best matched the color of the afterimage that appeared above the fixation point. To help select the correct color, a rectangle of the same size as the adapting stimuli was shown below the fixation point in the color currently under the mouse cursor. Participants confirmed their selection by pressing a mouse button when they were satisfied that color of the rectangle below the fixation point matched the color of the afterimage experienced above the fixation point. For each trial the color of the stimulus rectangle, the subject's response in RGB and the subject's response time were recorded. The goal of this study was to determine which of the two color coding mechanisms (trichromatic or opponent-process) better explains the perceived color of the afterimages. We used six differently colored rectangles: red, green, blue, cyan, magenta, yellow.

We start our analysis by loading the experiment and stimuli data. The experiment data include subject index, reaction time, response in RGB format, stimuli name (e.g., blue) and stimuli values in RGB and HSV. The stimuli data include the information about stimuli (stimuli names and their RGB/HSV values).

```
R> library(bayes4psy)
R> library(dplyr)
R> library(ggplot2)

R> data_all <- after_images

R> stimuli <- after_images_stimuli
```

Once we load required libraries and data we can start fitting Bayesian color models. Below is a detailed example of fitting the Bayesian color model for the red color stimuli. For a visual inspection of the fit (see **Figure 14**).

```
R> data_red <- data_all %>%
         filter(stimuli == "red")
R> data_red <- data.frame(r=data_red$r,
                          g=data_red$g,
                          b=data_red$b)

R> fit_red <- b_color(colors=data_red)

R> plot_trace(fit_red)
R> print(fit_red)

R> plot_hsv(fit_red)
```

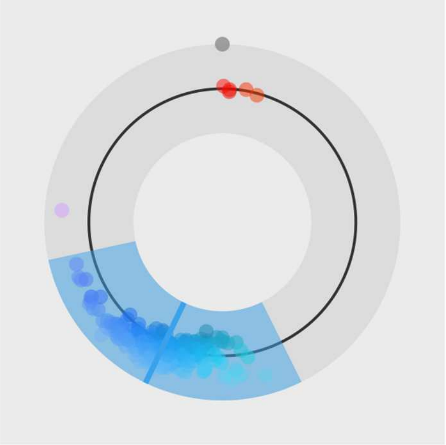


FIGURE 14 | The special `plot_hsv` function developed for the color model. Input data points are visualized with circles, mean of the fit is visualized with a solid line and the 95% HDI of the underlying distribution is visualized as a colored band.

We repeat the same process five more times for the remaining five colors of stimuli. We start the analysis by loading data about the colors predicted by the trichromatic and the opponent-process theory.

```
R> trichromatic <- after_images_trichromatic

R> opponent_process <-
     after_images_opponent_process
```

We can then use the `plot_distributions_hsv` function of the Bayesian color model to produce a visualization of the accuracy of both color coding mechanism predictions for each stimuli independently. Each graph visualizes the inferred distribution, displayed stimuli, and responses predicted by the trichromatic and opponent-process coding. This additional information can be added to the visualization via annotation points and lines. Below is an example for the red stimulus, visualizations for other five stimuli are practically the same.

```
R> stimulus <- "red"
R> lines <- list()
R> lines[[1]] <-
c(
trichromatic[trichromatic$stimuli == stimulus, ]$h,
trichromatic[trichromatic$stimuli == stimulus, ]$s,
trichromatic[trichromatic$stimuli == stimulus, ]$v
)
```

```
R> lines[[2]] <-
c(
opponent_process[
  opponent_process$stimuli == stimulus, ]$h,
opponent_process[
  opponent_process$stimuli == stimulus, ]$s,
opponent_process[
  opponent_process$stimuli == stimulus, ]$v
)

R> points <- list()

R> points[[1]] <-
c(
stimuli[stimuli$stimuli == stimulus, ]$h_s,
stimuli[stimuli$stimuli == stimulus, ]$s_s,
stimuli[stimuli$stimuli == stimulus, ]$v_s
)

R> plot_red <-
   plot_distributions_hsv(fit_red,
                          points=points,
                          lines=lines,
                          hsv=TRUE)

R> plot_red <- plot_red +
   ggtitle("Red") +
   theme(plot.title = element_text(hjust = 0.5))
```

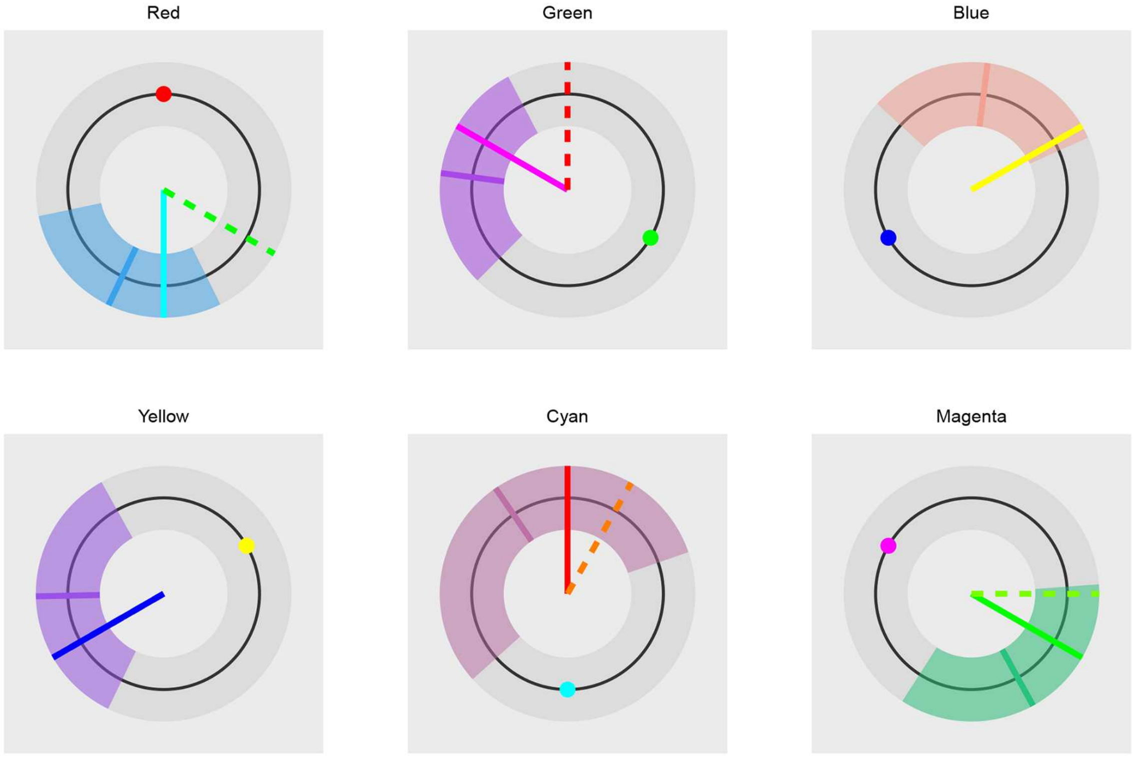


FIGURE 15 | The comparison of trichromatic and opponent-process color coding prediction. The long solid line visualizes the trichromatic color coding prediction. The dashed line visualizes the opponent-process color coding prediction. Short solid line represents the mean hue of the fit. The colored band the 95% HDI of the distribution underlying the fit. The small colored circle visualizes the color of the presented stimuli. In the case of blue and yellow stimuli the dashed line is not visible because both color codings predict the same outcome. The prediction based on the trichromatic color coding seems more accurate as its prediction is always inside the 95% of the most probable subject's responses and is always closer to the mean predicted hue than the opponent-process prediction. The opponent-process prediction is outside of the 95% of the most probable subject's responses in cases of red and green stimuli.

We can use the cowplot (Wilke, 2019) package to combine the plots into a single figure (see **Figure 15**).

```
R> cowplot::plot_grid(plot_red,
                      plot_green,
                      plot_blue,
                      plot_yellow,
                      plot_cyan,
                      plot_magenta,
                      ncol=3,
                      nrow=2,
                      scale=0.9)
```

4. DISCUSSION

The bayes4psy package helps psychology students and researchers with little or no experience in Bayesian statistics or probabilistic programming to do modern Bayesian analysis in R. The package includes several Bayesian models that cover a wide range of tasks that arise in psychological experiments. We can perform a Bayesian *t*-test or Bayesian bootstrap, analyse reaction times, success rates, colors, or sequential tasks. The package covers all parts of Bayesian data analysis, from fitting and diagnosing fitted models to visualizations and comparisons.

We plan to continuously upgrade the package with new tools and Bayesian statistics even closer to non-technical researchers. For example, we will implement probability distribution elicitation tools, which will ease the extraction of prior knowledge from domain experts and the prior construction process (Morris et al., 2014). Over the last couple of years neuroimaging techniques (e.g., fMRI and EEG) have become very popular for tracking brain activity during psychological experiments. The implementation of Bayesian models for analysing such data is also one of our future goals.

DATA AVAILABILITY STATEMENT

The results in this paper were obtained using R 3.5.3. Core R and all packages used are available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/.

The source code of bayes4psy can be found at https://github.com/bstatcomp/bayes4psy and the illustrative examples from section 3 are included in the package as vignettes. The bayes4psy package is also published on the CRAN repository (https://cran.r-project.org/package=bayes4psy).

AUTHOR CONTRIBUTIONS

JD, GR, and EŠ designed the study. GR determined which models should be implemented and gathered and prepared example data for these models. JD with supervision and guidance from EŠ developed the package and Bayesian models. JD prepared the illustrative examples. All the authors wrote the paper.

FUNDING

The research behind this manuscript was partially funded by the Slovenian Research Agency (ARRS) through grants L1-7542 (Advancement of computationally intensive methods for efficient modern general-purpose statistical analysis and inference), P3-0338 (Physiological mechanisms of neurological disorders and diseases), J3-9264 (Decomposing cognition: working memory mechanism and representations), P5-0410 (Digitalization as driving force for sustainability of individuals, organizations, and society), and P5-0110 (Psychological and neuroscientific aspects of cognition).

REFERENCES

Amrhein, V., Greenland, S., and McShane, B. (2019). Scientists rise up against statistical significance. *Nature* 567, 305–307. doi: 10.1038/d41586-019-00857-9

Bååth, R. (2014). Bayesian first aid: a package that implements bayesian alternatives to the classical *. test functions in R. *Proc. UseR* 33:2. Available online at: https://github.com/rasmusab/bayesian_first_aid

Bååth, R. (2015). *Bayesboot: An Implementation of Rubin's (1981) Bayesian Bootstrap*. Lund.

Baker, M., and Penny, D. (2016). Is there a reproducibility crisis? *Nature* 533, 452–454. doi: 10.1038/533452a

Brooks, S. P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* 7, 434–455. doi: 10.1080/10618600.1998.10474787

Camerer, C. F., Paulson, J. A., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., et al. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nat. Hum. Behav.* 2, 637–644. doi: 10.1038/s41562-018-0399-z

Carpenter, B., Lee, D., Brubaker, M. A., Riddell, A., Gelman, A., Goodrich, B., et al. (2017). Stan: a probabilistic programming language. *J. Stat. Softw.* 76, 1–32. doi: 10.18637/jss.v076.i01

Dunson, D. B. (2001). Commentary: practical advantages of Bayesian analysis of epidemiologic data. *Am. J. Epidemiol.* 153, 1222–1226. doi: 10.1093/aje/153.12.1222

Efron, B. (1979). Bootstrap methods: another look at the Jackknife. *Ann. Stat.* 7, 1–26. doi: 10.1214/aos/1176344552

Eriksen, B. A., and Eriksen, C. W. (1974). Effects of noise letters upon identification of a target letter in a nonsearch task. *J. Child Psychol. Psychiatry Allied Discipl.* 16, 143–149. doi: 10.3758/BF03203267

Gelman, A., Carlin, J. B. B., Stern, H. S. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. B. (2014). *Bayesian Data Analysis, 3rd Edn*. New York, NY: Chapman and Hall/CRC. doi: 10.1201/b16018

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–511. doi: 10.1214/ss/1177011136

Hubbard, R. (2015). *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science*. Des Moines, IA: Sage Publications. doi: 10.4135/9781506305332

Hurlbert, S. H., Levine, R. A., and Utts, J. (2019). Coup de Grâce for a tough old bull: "statistically significant" expires. *Am. Stat.* 73, 352–357. doi: 10.1080/00031305.2018.1543616

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t*-test. *J. Exp. Psychol.* 142, 573–603. doi: 10.1037/a0029146

Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial With R, JAGS, and Stan, 2nd Edn*. Bloomington, IN: Academic Press. doi: 10.1016/B978-0-12-405888-0.00008-8

Lindeløv, J. K. (2019). *Reaction Time Distributions: An Interactive Overview*. Aalborg.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., et al. (2019). JASP: Graphical statistical software for common statistical designs. *J. Stat. Softw*. 88, 1–17. doi: 10.18637/jss.v088.i02

McElreath, R. (2018). *Statistical Rethinking: A Bayesian Course With Examples in R and Stan*. Leipzig: CRC Press. doi: 10.1201/9781315372495

McNutt, M. (2014). Reproducibility. *Science* 343:229. doi: 10.1126/science.1250475

Morris, D. E., Oakley, J. E., and Crowe, J. A. (2014). A web-based tool for eliciting probability distributions from experts. *Environ. Model. Softw*. 52, 1–4. doi: 10.1016/j.envsoft.2013.10.010

Munafó, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie Du Sert, N., et al. (2017). A manifesto for reproducible science. *Nat. Hum. Behav*. 1, 1–9. doi: 10.1038/s41562-016-0021

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *J. Am. Stat. Assoc*. 349:aac4716. doi: 10.1126/science.aac4716

Plummer, M. (2003). "JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling," in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (Vienna), 1–40.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Auckland.

Rubin, D. B. (1981). The Bayesian bootstrap. *Ann. Stat*. 9, 130–134. doi: 10.1214/aos/1176345338

Schooler, J. W. (2014). Metascience could rescue the 'replication crisis'. *Nature* 515:9. doi: 10.1038/515009a

Stanley, T. D., Carter, E. C., and Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychol. Bull*. 144, 1325–1346. doi: 10.1037/bul0000169

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *J. Exp. Psychol*. 18, 643–661. doi: 10.1037/h0054651

Wasserstein, R. L., and Lazar, N. A. (2016). The ASA's statement on *p*-values: context, process, and purpose. *Am. Stat*. 70, 129–133. doi: 10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$". *Am. Stat*. 73, 1–19. doi: 10.1080/00031305.2019.1583913

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Auckland: CRAN. doi: 10.1007/978-0-387-98141-3

Wickham, H., François, R., Henry, L., and Müller, K. (2018). *dplyr: A Grammar of Data Manipulation*. Auckland: CRAN.

Wilke, C. O. (2019). *cowplot: Streamlined Plot Theme and Plot Annotations for ggplot2*. Austin, TX: CRAN.

Ziliak, S. T. (2019). How large are your G-values? Try Gosset's guinnessometrics when a little "p" is not enough. *Am. Stat*. 73, 281–290. doi: 10.1080/00031305.2018.1514325

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

METHODS
published: 13 May 2020
doi: 10.3389/fpsyg.2020.00815

Farewell to Bright-Line: A Guide to Reporting Quantitative Results Without the S-Word

Kevin M. Cummins[1,2,3]* *and Charles Marks*[1]

[1] *Division of Infectious Disease and Global Public Health, SDSU-UCSD Joint Doctoral Program in Interdisciplinary Research on Substance Use, San Diego, CA, United States,* [2] *Department of Psychology, University of California, San Diego, San Diego, CA, United States,* [3] *Department of Psychiatry, University of California, San Diego, San Diego, CA, United States*

Recent calls to end the practice of categorizing findings based on statistical significance have focused on what not to do. Practitioners who subscribe to the conceptual basis behind these calls may be unaccustomed to presenting results in the nuanced and integrative manner that has been recommended as an alternative. This alternative is often presented as a vague proposal. Here, we provide practical guidance and examples for adopting a research evaluation posture and communication style that operates without bright-line significance testing. Characteristics of the structure of results communications that are based on conventional significance testing are presented. Guidelines for writing results without the use of bright-line significance testing are then provided. Examples of conventional styles for communicating results are presented. These examples are then modified to conform to recent recommendations. These examples demonstrate that basic modifications to written scientific communications can increase the information content of scientific reports without a loss of rigor. The adoption of alternative approaches to results presentations can help researchers comply with multiple recommendations and standards for the communication and reporting of statistics in the psychological sciences.

Keywords: scientific communication, statistical significance, null hypothesis significance testing, confidence intervals, bright-line testing

OPEN ACCESS

Edited by:
Pietro Cipresso,
Italian Auxological Institute (IRCCS),
Italy

Reviewed by:
Jose D. Perezgonzalez,
Massey University Business School,
New Zealand
Julio Sánchez-Meca,
University of Murcia, Spain

****Correspondence:***
Kevin M. Cummins
kcummins@ucsd.edu

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *07 January 2020*
Accepted: *02 April 2020*
Published: *13 May 2020*

Citation:
Cummins KM and Marks C (2020) Farewell to Bright-Line: A Guide to Reporting Quantitative Results Without the S-Word. Front. Psychol. 11:815. doi: 10.3389/fpsyg.2020.00815

INTRODUCTION

The abandonment of significance testing has been proposed by some researchers for several decades (Hunter, 1997; Krantz, 1999; Kline, 2004; Armstrong, 2007). In place of heavy reliance on significance testing, a thorough interrogation of data and replication of findings can be relied upon to build scientific knowledge (Carver, 1978) and has been recommended in particular for exploratory research (Gigerenzer, 2018). The replication crisis has demonstrated that significance testing alone does not ensure that reported findings are adequately reliable (Nosek et al., 2015). Indeed, the practice of focusing on significance testing during analysis is a motivator for "P-hacking" and expeditions into the "garden of forking paths" (Gelman and Loken, 2014; Szucs and Ioannidis, 2017). These problems are compounded by the outright misunderstanding and misuse of *p*-values (Goodman, 2008; Wasserstein, 2016). The practice of bright-line significance testing has been considered a generator of scientific confusion (Gelman, 2013), though appropriately interpreted *p-value*s can provide guidance in results interpretation.

A recent issue of *The American Statistician* and a commentary in *Nature* suggest a seemingly simple conciliatory solution to the problems associated with current statistical practices (Amrhein et al., 2019; Wasserstein et al., 2019). The authors of these articles advocate ending the use of bright-line statistical testing in favor of a thoughtful, open, and modest approach to results reporting and evaluation (Wasserstein et al., 2019). This call to action opens the door to the widespread adoption of various alternative practices. A transition will require authors to adopt new customs for analysis and communication. Reviewers and editors will also need to recognize and accept communication styles that are congruent with these recommendations.

Many researchers may subscribe to the ideas behind the criticisms of traditional significance testing based on bright-line decision rules but are unaccustomed to communicating findings without them. This is a surmountable barrier. While recent calls to action espouse principles that researchers should follow, tangible examples, both of traditional approaches to statistical reporting, as well as the newly recommended ones, may serve as a needed resource for many researchers. The aim of this paper is to provide researchers guidance in ensuring their repertoire of approaches and communication styles include approaches consistent with these newly reinforced recommendations. Guidance on crafting results may facilitate some researchers' transition toward the execution of the recent recommendations.

What Is the Dominant Approach?

Null hypothesis significance testing (NHST) dominates the contemporary application of statistics in psychological sciences. A common approach to structuring a research report based on NHST (and we note that there are many variations) follows these steps: *first*, a substantive question is articulated; *then*, an appropriately matched statistical null hypothesis is constructed and evaluated. The statistical question is often distinct from the substantive question of interest. Next, upon execution of the given approach, the appropriate metrics, including the p-value, are extracted, and, *finally*, if this p-value (or equivalently the test statistic) is more extreme than a pre-determined bright-line α, typically 0.05, a declaration that the result is significant is issued.

The dominant style of research communication that has arisen from this approach has emphasized the dissemination of findings that meet the significance threshold, often disregarding the potential for non-significant findings to provide some utility in addressing substantive scientific questions at hand. The concern with dichotomizing findings was distilled by Altman (1990) when he wrote, "It is ridiculous to interpret the results of a study differently according to whether or not the P-value obtained was, say, 0.055 or 0.045. These should lead to very similar conclusions, not diametrically opposed ones." Further, this orientation has facilitated the de-emphasis of the functional associations between variables under investigation. In the simplest case, researchers have failed to focus on the association magnitude (Kirk, 1996). Whereas Kelley and Preacher (2012) have described the differences between effect size and effect magnitude, we propose a more general focus on the functional associations between our variables of interest, which are often complex, contingent, and curvilinear, and so often cannot be adequately distilled into a single number. Although we will refer to effect sizes using the conventional definition, we want the reader to recognize that this usage is not consistently tied to causal inference, in practice. Adapting Kelley and Preacher's (2012) definition, we treat effect size as the quantitative reflection(s) of some feature(s) of a phenomenon that is under investigation. In other words, it is the quantitative features of the functional association between variables in a system under study; this tells us how much our outcome variable is expected to change based on differences in the predictors. If the outcome variable displays such small changes as a result of changes in a predictor that the variance is of little practical value, a finding of statistical significance may be irrelevant to the field. The shape, features, and magnitude of functional associations in studied phenomena should be the focus of researchers' description of findings. To this end, the reader is encouraged to consult several treatments of effect size indices to assist in the identification of appropriate statistics (Ellis, 2010; Grissom and Kim, 2014; Cumming, 2017).

Herein, we present a generalized version of this significance orientation communication style (SOCS), steps that can be taken to transition to a post-significance communication style (POCS) that will facilitate researchers' focus on the structure of the associations they are studying rather than just evidence of an association. Examples of how SOCS results write-ups may be updated to meet the standards of this new style follow.

Significance Orientation Communications Structure

The structure for a passage in a results section written in the SOCS frequently includes:

1. A reference to a table or figure,
2. A declaration of significance, and,
3. A declaration of the direction of the association (positive or negative).

The ordering is not consistent but often begins with the reference. The authors write a statement such as, "Table 1 contains the results of the regression models," where Table 1 holds the statistics from a series of models. There may be no further verbal description of the pattern of findings. The second sentence is commonly a declaration of the result of a significance test, such as, "In adjusted models, depression scores were significantly associated with the frequency of binge drinking episodes ($p < 0.05$)." If the direction of association was not incorporated into the second sentence, a third sentence might follow; for instance, "After adjustment for covariates, depression scores were positively associated with binge episodes." Variation in this structure occurs, and in many instances, some information regarding the magnitude of associations (i.e., effect sizes) is presented. However, because this approach focuses on the results of a significance test, the description of the effect is often treated as supplemental or perfunctory. This disposition explains many misunderstandings and misuses of standardized effect sizes (Baguley, 2009). Findings that do not meet the significance threshold are often only available to the reader in the tables and frequently not considered when answering the substantive question at hand. However, interval estimates can and should be leveraged even when the null hypothesis is not rejected (Kelley and Preacher, 2012).

Post-significance Communications Structure

Here, we present an overarching structure for what text in the results could look like when using post-significance communications structure (POCS). The emphasis shifts from identifying significant results to applying all findings toward the purpose of answering the substantive question under study. The **first sentence** can be considered a direct answer to this question, which the authors proposed in the introduction – the findings of the statistical tests should be placed in the context of the scientific hypotheses they are addressing. **Next,** the quantitative results of the statistical analyses should be described, and, as a part of this description, a directional reference to supporting tables and figures can be noted. Emphasis should be placed on making sure the results are presented in a form that allows the reader to confirm if the author's assessment in the first sentence is appropriate. This will often include a parenthetical notation of the p-value associated with the presented parameter estimates. The significance is not an isolated focus and its presentation is not contingent on the p-value reaching a threshold. Instead, p-values are part of the support and context for the answer statement (Schneider, 2015). This is reflected in their position within the paper. They can be placed in tables, presented parenthetically, or set off from the rest of the text through the use of commas when parentheses would add an additional level of enclosure. P-values should always be presented as continuous statistics and recognized as providing graded levels of evidence (Murtaugh, 2014; Wasserstein et al., 2019).

Even where p-values are large, the authors should focus on describing patterns relevant to the question at hand. Assuming a good study design, the best estimate, based on the data being presented, are the point estimates, regardless of the p-value. Considering the context of the interval estimates is also critical in all circumstances because we do not want to conflate random noise with effects. The **remaining sentences** should be descriptions of the auxiliary patterns in the data that are pertinent to the scientific questions at hand. In many cases, these descriptions function as annotations of the key patterns found in the tables and figures.

To help clarify how we can transition from the SOCS style to the POCS style, we provide two examples from our own research.

Example: Factors Related to Injection Drug Use Initiation Assistance

Using data from a multi-site prospective cohort study, we investigated factors that were associated with providing injection assistance to previously injection-naïve individuals the first time they injected (Marks et al., 2019). Most initiations (i.e., the first time an injection-naïve person injects drugs) are facilitated by other people who inject drugs (PWID). There is evidence that PWID receiving opioid agonist treatment have a reduced likelihood of providing assistance to someone initiating injection drug use (Mittal et al., 2017). We are interested in understanding the extent to which opioid agonist treatment enrollment and other factors are associated with assisting injection drug use initiation. The following describes part of what we recently found, using a conventional SOCS approach (Marks et al., 2019):

Conventional example 1

> *As shown in Table B*[1]*, the likelihood of recently (past 6 months) assisting injection drug use initiation was significantly related to recent enrollment in opioid agonist treatment ($z = -2.52$, $p = 0.011$), and methamphetamine injecting ($z = 2.38$, $p = 0.017$), in Vancouver. Enrollment in the opioid agonist treatment arm was associated with a lower likelihood of assisting injection initiation. The relative risk was significantly elevated for those injecting methamphetamine, whereas speedball injecting was not significantly associated with initiation assistance ($z = 1.84$, $p = 0.064$).*

This example starts with a reference to a table. It then indicates the patterns of significance and the direction of the effects. The parameter estimates that describe the magnitude and functional form can be extracted from the table (see Marks et al., 2019); however, significance tests are the focus of what is being communicated. The parameter estimates are absent from the text. No information about the non-significant association is developed. Abandoning significance tests and broadening the focus to include parameter estimates increases both the total information content and information density of the text. Now we will rewrite this paragraph in the POCS style.

The first sentence can be a direct answer to the research question proposed. Based on prior evidence, we had hypothesized opioid agonist treatment enrollment would decrease the likelihood of assisting an initiation; thus, for our new first sentence, we propose:

> *Results of our multivariable model are consistent with our hypothesis that recent enrollment in opioid agonist treatment was associated with a decreased likelihood of recently assisting injection initiation in Vancouver.*

We have begun by directly addressing how our findings answer our research question. Next, we want to present the details of the quantitative patterns. This can also be the first sentence when the functional association is simple. We also want to make sure to present the results in a way that increases the value of information available to the reader – in this case, instead of presenting regression point estimates, we present the relative risk and proportional effects. As such, we propose:

> *Recent opioid agonist treatment enrollment was associated with a 12 to 63% reduction in likelihood of assisting initiation (RR: 0.58 95% CI: 0.37–0.88, $p = 0.011$, Table B).*

This lets the reader know not only that we have a high degree of confidence in the direction of the effect (both indicated by the confidence interval and p-value), but also that the magnitude of the effect warrants further consideration that opioid agonist treatment should be considered as a tool for addressing injection

[1] This reference is to Table B in the supplement to Marks et al. (2019).

initiation. If, for example, our confidence interval had been 0.96–0.98, even though we feel confident in the direction of the effect, we may deem it inappropriate to suggest changes to treatment implementation as a result based on such a small potential return on investment. Relying solely upon significance testing to determine the value of findings could result in the glossing over this critical piece of information (i.e., the effect size). In addition, we note that we have now included a reference to the table where further details and context can be inspected.

Finally, we want to examine additional patterns in the data. In the SOCS style paragraph, we reflected on the significance of both the effect of recent methamphetamine injection and recent speedball (heroin and cocaine) injection. While our primary research question focused on the impact of opioid agonist treatment, we can still also present results for related secondary questions regarding methamphetamine and speedball injection, so we write:

> *Recent methamphetamine injection was associated with a 12% to 227% increase in likelihood of assisting initiation (RR: 1.91 95% CI: 1.12–3.27, p-value = 0.017). Similarly, recent speedball injection was associated with an effect ranging from a 3% decrease to a 193% increase in likelihood of assisting initiation (RR: 1.68 95% CI: 0.97–2.93, p = 0.064).*

Here, we find that methamphetamine and speedball injection had similar confidence interval estimates. Instead of saying that the impact of speedball injection was "not significant" where the *p*-values exceed.05, we present the confidence interval of methamphetamine and speedball injection relative risks. From this, the reader can evaluate if our conclusion that the findings preclude the possibility that speedball may have at most a small protective effect on assisting initiation. We can gain some knowledge from non-significant findings. Relevant stakeholders may determine that a 3% reduction to a 193% increase in risk is strong enough evidence to allocate resources to further study and/or intervene on speedball injection. We note that assessing the acceptability of characterizing an effect in this way that did not meet traditional standards of significance is a complex task and that it will be dependent on the consensus of the authors, reviewers, and editors. This subjectivity of assessment exemplifies the importance of the POCS style, as it requires all stakeholders in the peer-review process to engage critically with the interpretation of "not significant" findings.

Our new POCS paragraph reads:

> *Results of our multivariable model were consistent with our hypothesis that recent enrollment in opioid agonist treatment was associated with a decreased likelihood of recently assisting injection initiation in Vancouver. Recent opioid agonist treatment enrollment was associated with a 12 to 63% reduction in likelihood of assisting initiation (RR: 0.58 95% CI: 0.37–0.88, p = 0.011, Table B). Recent methamphetamine injection was associated with a 12% to 227% increase in likelihood of assisting initiation (RR: 1.91 95% CI: 1.12–3.27, p-value = 0.017). Similarly, recent speedball injection was associated with an effect ranging from a 3% decrease to a 193% increase in likelihood of assisting initiation (RR: 1.68 95% CI: 0.97–2.93, p = 0.064).*

Example 2: Associations Among Adolescent Alcohol Use, Expectancies, and School Connectedness

Using data from a community survey of high school students, we investigated the relationship among drinking expectancies, school connectedness and heavy episodic drinking (Cummins et al., 2019). Student perceptions of acceptance, respect, and support at their schools are reported to be protective against various risky health behaviors, including drinking. We wanted to know if the association was contingent on alcohol expectancies. Alcohol expectancies are cognitions related to the expected outcomes that a person attributes to drinking (Brown et al., 1987). In this study, higher expectancies indicate the respondent expects the outcomes of consuming alcohol to be more rewarding.

Conventional example 2

> ***Figure 1*** *and Table 2*[2] *depict the associations among recent (past 30 days) binge drinking, school connectedness, and alcohol expectancies. The model for recent (past 30 days) binge drinking with school connectedness, party-related alcohol expectancies, and their interaction as independent variables was statistically significant (Likelihood Ratio χ2 (3) = 171, p < 0.0001). Significant moderation was observed (OR = 9.89, SC X pAE interaction: z = 2.64, p = 0.008). The prevalence of binging was significantly higher for students reporting the highest expectancies as compared to those reporting the lowest expectancies when students also reported the highest school connectedness (z = 9.39, p < 0.001). The predicted prevalence of binge drinking was 17.9 times higher among students with the highest expectancies, as compared to those with the lowest expectancies. This same comparison was non-significant, where school connectedness was at its lowest (z = 1.84, p = 0.066).*

While this is not an archetypal version of the SOCS style, its primary focus is on the patterns of significance. Some information is presented on the magnitudes of the associations

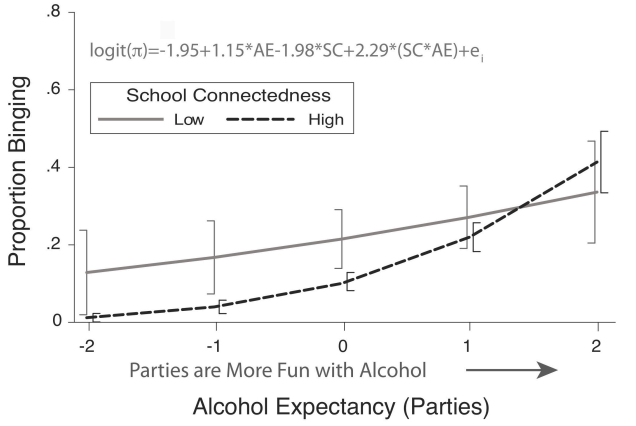


FIGURE 1 | Modeled proportion of high school students engaging in heavy episodic drinking (binging) as a function of school connectedness (SC) and party-related alcohol expectancies (AE). Error bars represent 95% confidence intervals. Low and high school connectedness are at the minimum and maximum observed school connectedness, respectively. Modified from Cummins et al. (2019).

[2] This reference is to Table 2 in the supplement to Cummins et al. (2019).

in the text; however, metrics of estimation uncertainty are absent, as is information on the non-significant patterns. Much of the text is redundant with the tables or is unneeded (Cummins, 2009). For example, the first sentence is a simple reference to a figure with no indication of what the authors extracted from their inspection of the figure. The text itself does not help answer the scientific question. The functional associations and their magnitudes are not described, so there is no value added by the presence of the sentence. The latter sentences primarily function to identify statistically significant patterns. The measure of association strength is presented for one of the features of the model, which was statistically significant. The uncertainty of the estimate and the features were not described. Finally, it remains unclear how the results answer the substantive question under study. Now, we will rewrite this paragraph in the POCS style.

The first sentence needs to be a direct answer to the substantive question under study. We expected students who reported higher party-related expectancies (i.e., had more positive views of attending parties) would report higher odds of recent binge drinking. Further, we wanted to assess if school connectedness moderated this relationship. For our new first sentence, we propose:

> *We found that higher levels of alcohol expectancies were associated with greater odds of binge drinking and that this relationship was attenuated among those reporting the lowest level of school connectedness.*

Here, we have directly answered our research question. Results of moderation analyses can be challenging to parse, exemplifying the need to clearly articulate how the results reflect upon the question under study is particularly important. This first sentence helps the reader navigate the more complex statements about contingent effects. We do this by presenting a data visualization and highlighting the relationship of party-related expectancies and binge drinking at the highest and lowest levels of school connectedness in the text, as such:

> *For students reporting the highest level of school connectedness, the modeled odds of binge drinking was 30.5 (95% CI: 15.4–60.2) times higher for students with the highest level of alcohol expectancies as compared to the those with the lowest. This pattern was attenuated for students reporting the lowest levels of school connectedness (OR = 3.18), such that the confidence interval for the modeled odds ratio ranged from 1.18 to 10.6 (95% CI) (**Figure 1**).*

Here, we have provided the reader two key pieces of information that quantifies our initial qualitative statement: first, for both students with the lowest and highest levels of school connectedness, we are confident there is a positive relationship between alcohol expectancies and binge drinking; and, second, that this relationship is attenuated amongst those with the lowest levels of school connectedness, as indicated by their non-overlapping confidence intervals. We have also provided the reference to **Figure 1**, reducing the initial examples' 20-word directional sentence, to two words.

We also focus upon estimation uncertainty by presenting the confidence intervals in the body of the text. We direct the reader to recognize the lower bound of the for the odds ratio was near 1.18 for students with the lowest school connectedness. This can be returned to in the discussion. It could be pointed out that it is plausible alcohol expectancies are not strongly associated with binge drinking for these students. Deploying interventions targeting expectancies among these students could be an inefficient use of resources. Thus, getting an improved estimate of the effect size could be valuable to practitioners before committing to a rigid plan for deploying intervention resources. Not only should authors present measures of uncertainty (e.g., confidence intervals, credibility intervals, prediction intervals), they should base their interpretations on those intervals.

Finally, we want to reflect on additional patterns in the data. In the SOCS example, we presented the significance of the model fit, as well as the significance of the interaction effect. We provide additional information for the reader to assess the magnitude of the interaction. We give the reader a way to gauge this by contrasting the association at the extremes of school connectedness. We note that, in cases where the models are complex, word limits and a disposition toward being concise will force authors to be selective about which features are to be verbalized. Patterns of lower importance may not be described in the text but should be accessible to the reader through tables and figures.

Here, we also note that the reader should be able to evaluate the authors' descriptive choices in the text and ensure those are faithful to the overall patterns. On the flip side of the coin, the author's selections also initially guide the reader through the answers to the study's questions that are supported by the content within the tables and figures. For reviewers and editors assessing works in the POCS style, it will be important to assess if the authors' descriptive choices are faithful to the overall patterns of the results. This requires that authors provide adequate information in their tables and figures for reviewers to make such an assessment.

As a result, our new POCS paragraph reads as follows:

> *We found that higher levels of alcohol expectancies were associated with greater odds of binge drinking and that there was evidence that the strength of this relationship was contingent on school connectedness, such that it was attenuated among students reporting the lowest level of school connectedness. For students reporting the highest level of school connectedness, the modelled odds of binge drinking was 30.5 (95% CI: 15.4–60.2, **Figure 1** and Table 2) times higher for students with the highest level of alcohol expectancies as compared to the those with the lowest. This association was attenuated for students reporting the lowest levels of school connectedness (OR = 3.18), such that the interval estimate of modelled odds ratio ranged from 1.18 to 10.6 (95% CI).*

DISCUSSION

We present a communication style that abandons the use of bright-line significance testing. By introducing the POCS style as a formal structure for presenting results, we seek to reduce barriers faced by researchers in their efforts to follow recommendations for abandoning the practice of declaring results statistically significant (Amrhein et al., 2019; Wasserstein et al., 2019). The examples provided demonstrate how the adoption of this general approach could help improve the field by shifting its focus during results generation to the simultaneous and integrated consideration of measures of effect and inferential statistics. Reviewers should also recognize that the use of POCS is not an indicator of statistical naivety, but rather one of a differing view on traditional approaches–this paper can be a useful resource for explaining POCS to unfamiliar reviewers. Writing results without the word "significant" is completely counter to the training and experience for most researchers. We hope that these examples will motivate researchers to attempt to draft their results without using or reporting significance tests. Although some researchers may fear that they will be left with a diminished ability to publish, this need not be the case. If the research findings do not stand up when described in terms of the functional associations, perhaps that research is not ready to be published. Indeed, with greater recognition of the replication crisis in the psychological sciences, we should pay more attention to the design features and basic details of the patterns of effects.

Significance testing should not be used to reify a conclusion. Fisher (1935) warned that an "isolated record" of a significant result does not warrant its consideration as a genuine effect. Although we want our individual works to be presented as providing a strong benefit to the field, our confidence that individual reports will hold is often unwarranted. We may benefit from cautiously reserving our conclusions until a strong and multi-faceted body of confirmatory evidence is available. This evidence can be compiled without bright-line significance testing. Improved reporting, that presents a full characterization of the functional relationships under study, can help to facilitate the synthesis of research generated knowledge into reviews and metanalyses. It is also consistent with American Psychological Association reporting standards, which promotes the reporting of exact *p*-values along with point and interval estimates of the effect-size (Appelbaum et al., 2018).

The strongest support for some of our research conclusions have been obtained from Bayesian probabilities based on informative priors (e.g., Cummins et al., 2019). This point serves to highlight the general limitations of focusing on frequentist based NHST in scientific research and the benefit of gauging evidence other substantive features, such as the design, explanatory breadth, predictive power, assumptions, and competing alternative models (de Schoot et al., 2011; Trafimow et al., 2018). The POCS is compatible with a more integrated approach to the valuation of research reports, whereas the continued use of bright-line significance testing is not (Trafimow et al., 2018).

We suspect that the quality of many papers will increase through the application of POCS. In part, this will be driven by a change in orientation toward the aims of research reports where the emphasis on the establishment of the presence of an association is substituted with an emphasis on estimating the functional form (magnitude, shape, and contingencies) of those relationships. The examples from our own work demonstrate that there should be no barrier to drafting papers with POCS. Research-based on an integrated examination of all statistical metrics (effect sizes, *p*-values, error estimates, etc.) shall lead to more meaningful and transparent communication and robust development of our knowledge base. Research findings should not be simply dichotomized – the quantitative principle that the categorization of a continuous variable will always lead to a loss of information also applies to *p*-values (Altman and Royston, 2006). In this paper, we provide examples of different ways to apply that principle.

AUTHOR CONTRIBUTIONS

Both authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

Altman, D. G. (1990). *Practical Statistics for Medical Research*. Boca Raton: CRC Press.

Altman, D. G., and Royston, P. (2006). The cost of dichotomising continuous variables. *BMJ* 332:1080. doi: 10.1136/bmj.332.7549.1080

Amrhein, V., Greenland, S., and McShane, B. (2019). Scientists rise up against statistical significance. *Nature* 567, 305–307.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., and Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: the APA publications and communications board task force report. *Am. Psychol.* 73, 3–25. doi: 10.1037/amp0000191

Armstrong, J. S. (2007). Significance tests harm progress in forecasting. *Intern. J. Forecast.* 23, 321–327.

Baguley, T. (2009). Standardized or simple effect size: what should be reported. *Br. J. Psychol.* 100, 603–617. doi: 10.1348/000712608X377117/pdf

Brown, S. A., Christiansen, B. A., and Goldman, M. S. (1987). The alcohol expectancy questionnaire: an instrument for the assessment of adolescent and adult alcohol expectancies. *J. Stud. Alcohol* 48, 483–491.

Carver, R. (1978). The case against statistical significance testing. *Harvard Educ. Rev.* 48, 378–399.

Cumming, G. (2017). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta Analysis*. Abingdon: Routledge.

Cummins, K. (2009). *Tips on Writing Results For A Scientific Paper*. Alexandria: American Statistical Association.

Cummins, K. M., Diep, S. A., and Brown, S. A. (2019). Alcohol expectancies moderate the association between school connectedness and alcohol consumption. *J. Sch. Health* 89, 865–873. doi: 10.1111/josh.12829

de Schoot, R. V., Hoijtink, H., and Jan-Willem, R. (2011). Moving beyond traditional null hypothesis testing: evaluating expectations directly. *Front. Psychol.* 2:24. doi: 10.3389/fpsyg.2011.00024

Ellis, P. D. (2010). *The Essential Guide to Effect Sizes*. Cambridge: Cambridge University Press.

Fisher, S. R. A. (1935). *The Design of Experiments.* Edinburgh: Oliver and Boyd.

Gelman, A. (2013). P values and statistical practice. *Epidemiology* 24, 69–72. doi: 10.1097/EDE.0b013e31827886f7

Gelman, A., and Loken, E. (2014). The statistical crisis in science data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *Am. Sci.* 102, 459–462.

Gigerenzer, G. (2018). Statistical rituals: the replication delusion and how we got there. *Adv. Methods Pract. Psychol. Sci.* 1, 198–218.

Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. *Semin. Hematol.* 45, 135–140. doi: 10.1053/j.seminhematol.2008.04.003

Grissom, R. J., and Kim, J. J. (2014). *Effect Sizes for Research*. Abingdon: Routledge.

Hunter, J. E. (1997). Needed: a ban on the significance test. *Psychol. Sci.* 8, 3–7.

Kelley, K., and Preacher, K. J. (2012). On effect size. *Psychol. Methods* 17, 137–152. doi: 10.1037/a0028086

Kirk, R. E. (1996). Practical significance: a concept whose time has come. *Educ. Psychol. Measur.* 56, 746–759.

Kline, R. B. (2004). Beyond significance testing: reforming data analysis methods in behavioral research. *Am. Psychol. Assn.* 2004:325.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *J. Am. Statist. Assoc.* 94, 1372–1381.

Marks, C., Borquez, A., Jain, S., Sun, X., Strathdee, S. A., Garfein, R. S., et al. (2019). Opioid agonist treatment scale-up and the initiation of injection drug use: a dynamic modeling analysis. *PLoS Med.* 16:2973. doi: 10.1371/journal.pmed.1002973

Mittal, M. L., Vashishtha, D., Sun, S., Jain, S., Cuevas-Mota, J., Garfein, R., et al. (2017). History of medication-assisted treatment and its association with initiating others into injection drug use in San Diego. *CA Subst. Abuse Treat. Prev. Policy* 12, 42. doi: 10.1186/s13011-017-0126-1

Murtaugh, P. A. (2014). In defense of P values. *Ecology* 95, 611–617.

Nosek, B., Aarts, A., Anderson, C., Anderson, J., Kappes, H., Baranski, E., et al. (2015). Estimating the reproducibility of psychological science. *Science* 349:aac4716.

Schneider, J. W. (2015). Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations. *Scientometrics* 102, 411–432. doi: 10.1007/s11192-014-1251-5

Szucs, D., and Ioannidis, J. P. A. (2017). When null hypothesis significance testing is unsuitable for research: a reassessment. *Front. Hum. Neurosci.* 11:390. doi: 10.3389/fnhum.2017.00390

Trafimow, D., Amrhein, V., Areshenkoff, C. N., Barrera-Causil, C. J., Beh, E. J., Bilgiç, Y. K., et al. (2018). Manipulating the alpha level cannot cure significance testing. *Front. Psychol.* 9:699. doi: 10.3389/fpsyg.2018.00699

Wasserstein, R. (2016). *American Statistical Association Releases Statement On Statistical Significance And P-Values: Provides Principles To Improve The Conduct And Interpretation Of Quantitative Science*. Washington, DC: American Statistical Association.

Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019). Moving to a World Beyond "p<*0.05*". *Am. Statist.* 73, 1–19. doi: 10.1080/00031305.2019.1583913

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 18 May 2020
doi: 10.3389/fpsyg.2020.00827

Mind and Body: Italian Validation of the Postural Awareness Scale

Eleonora Topino[1], *Alessio Gori*[2]* *and Holger Cramer*[3]

[1] Department of Human Science, LUMSA University, Rome, Italy, [2] Department of Health Sciences, University of Florence, Florence, Italy, [3] Department of Internal and Integrative Medicine, Evang. Kliniken Essen-Mitte, Faculty of Medicine, University of Duisburg-Essen, Essen, Germany

Postural awareness (PA) refers to a subjective conscious awareness of body posture and falls within the framework of mind–body integration. The aim of this research was to validate and evaluate psychometric properties of the Postural Awareness Scale (PAS) in an Italian population sample (n = 928; 45.04% men and 54.96% women; mean age = 29.96 years, standard deviation = 11.44). The results obtained with Velicer's Minimum Average Partial Test, Horn's Parallel Analysis, and exploratory factor analysis showed a two-factor solution, as supported by the confirmatory factor analysis: ease/familiarity with postural awareness and need for attention regulation with postural awareness. Furthermore, the findings highlighted both a good internal consistency (α = 0.76 for the total scale and α = 0.80, α = 0.79 for the two subscales) and a satisfactory construct validity. Furthermore, multivariate analysis of variance was carried out to assess differences in PA between specific subgroup. In particular, the positive effects of physical activity and healthy body weight were confirmed, whereas no significant differences related to gender or age were found. All these findings suggest that the Italian version of the PAS is a rapid instrument with good psychometric properties, which can be useful both for research and clinical practice.

Keywords: posture, awareness, mind–body integration, Italian validation, self-report questionnaire

OPEN ACCESS

Edited by:
Pietro Cipresso,
Italian Auxological Institute (IRCCS),
Italy

Reviewed by:
Carlo Chiorri,
University of Genoa, Italy
Alen Hajnal,
University of Southern Mississippi,
United States

****Correspondence:***
Alessio Gori
alessio.gori@unifi.it

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: *26 November 2019*
Accepted: *03 April 2020*
Published: *18 May 2020*

Citation:
Topino E, Gori A and Cramer H
(2020) Mind and Body: Italian
Validation of the Postural Awareness
Scale. Front. Psychol. 11:827.
doi: 10.3389/fpsyg.2020.00827

INTRODUCTION

The postural awareness is "the subjective conscious awareness of body posture that is mainly based on proprioceptive feedback from the body periphery to the central nervous system" (Cramer et al., 2018b, paragraph 1). It is a fundamental element for controlling posture in a process of adaptation based on sensory, motor, and cognitive aspects (Balasubramaniam and Wing, 2002). The body posture, in fact, can be influenced by a certain number of conditioning factors: mechanical aspects, heredity, race, flexibility, muscular strength, vision, and habits, but it is also involved in relationships of mutual interdependence with emotional and psychological factors (Brito, 1995; Wright et al., 2000). The scientific literature confirms the close relationship between posture and psychological dimensions, as demonstrated in several studies concerning assertiveness levels and action trends (e.g., Maner et al., 2010; Huang et al., 2011; Arnette and Pettijohn, 2012; Van der Toorn et al., 2015), self-efficacy (Anderson and Galinsky, 2006), self-esteem (Stepper and Strack, 1993; Dijkstra et al., 2007; Kwon and Kim, 2015), and mood (Hackford et al., 2019). These findings fall within a framework of mind–body interaction supported by different lines of research. The field of trauma studies, for example, increasingly focuses on this reciprocal influence, describing somatically driven individuals, with strong emotions accompanied by dysregulated physical sensations: these activations derive from reminders to previous adverse experiences that were not elaborated and repeated in the body (Van Der Kolk, 2014; Fisher, 2017). Furthermore, trauma affects self-awareness, specifically the sensory one (Bluhm et al., 2009), and alters the "postural body scheme" (Gurfinkel, 1994), involved in the perception of self with respect to the external world and in the actions directed to it. All this seriously alters the psychological perception of being able to manage one's life, which is closely linked to the possibility of experiencing control of the physical sphere (Van Der Kolk, 2008). Therefore, maintaining or increasing one's postural awareness levels allows the management of one's "postural body scheme," developing more adaptive attitudes through reflection and intention (Massion, 1994). Thus, the vision of James Grotstein (1997) who spoke of the mind and the body as a "strangely coupled unity" appears pertinent. He depicts them in a single entity with two inseparable aspects, like two sides of the same coin: they are considered like two different categories dependent on the perspective of the observer (Solano, 2010). In this theoretical framework, body awareness and mindfulness are parallel to the construct of postural awareness and strongly associated with it.

Body awareness concerns attention to bodily sensations and implies access to consciousness of proprioceptive (including posture) and interoceptive aspects (Mehling et al., 2009). It allows the participation of bodily sensations in everyday life and the observation of changes and physical responses to emotions and environment. It finds a good application and positive feedback in many contexts of clinical care such as, for example, those for recovery from physical and/or psychological traumas (Herman, 1992; Bishop et al., 2004), for substance abuse (Marlatt and Ostafin, 2005), for eating disorders (Zerbe, 1995), and for personality disorders (Friis et al., 1989). On the contrary, the concept of bodily dissociation is characterized by the avoidance of inner experience (Price and Thompson, 2007); it could represent a protective strategy against painful memories, thoughts, or feelings and is a mechanism commonly used for defense against physical suffering (Bakal, 2001) and trauma (Van der Kolk, 2006).

Closely associated, mindfulness is an awareness of the present moment with total acceptance of it (Brown and Ryan, 2003). Mindfulness intertwines focused attention with meta-awareness, allowing deep insight and clarifying the nature of the elements that constitute the experience (Wallace, 2006). This presence disposition is closely connected to higher levels of physical and mental health, better postural control through the conscious management of attentional focus (Kee et al., 2012), and more likely to maintain healthy habits such as sufficient sleep (Roberts and Danoff-Burg, 2010), physical activity (Murphy et al., 2012), and healthy eating (Gilbert and Waltz, 2010).

In this frame of mind–body relationship, postural awareness has still been little investigated, even though it is related to other important constructs explored above. Some studies support its effectiveness in chronic pain situations: in this field, considering the impact of this condition on people's lives and the psychological difficulties that ensue, new treatment models have been developed based on the association of physical experiences to states of greater awareness and mindfulness (Mattsson, 1998; Rosberg, 2000). Specifically, postural awareness training has proved particularly effective for chronic low back pain conditions (Moseley et al., 2004; Ahmed et al., 2016). These favor the control of one's physical disposition and the maintenance of healthy postural patterns in everyday life, important elements to avoid chronicization and further deterioration (Cramer et al., 2018b); a faulty posture, in fact, increases stress on muscles, tendons, ligaments, and bones (Yamak et al., 2018).

The Postural Awareness Scale: A Measure of Body Posture Awareness

The Postural Awareness Scale (PAS) is a German self-report measure designed by Cramer et al. (2018b), which allowed them to grasp the increases of this variable on subjects with chronic pain following the implementation of a multimedia mind–body training program. In particular, they found that improvements in body posture awareness were longitudinally associated with reduced pain in patients with spinal/shoulder pain, in line with other research on this topic (e.g., Lauche et al., 2017). The scale consists of 12 items, grouped into two factors (explaining the 58.8% of the variance in the original study); the first one is "ease/familiarity with postural awareness," which refers to an effortless awareness and connectedness; the second factor is "need for attention regulation with postural awareness" and indicates a forced awareness. The original scale and both its factors demonstrated satisfactory internal consistency and good validity converging with other measures related to body awareness and mindfulness. Specifically, the subscale *ease/familiarity with postural awareness* showed important associations with the measures related to the connection with one's body (Cramer et al., 2018b), significantly correlating with the scores of the Body Awareness Questionnaire (BAQ; Shields et al., 1989), of the *trust in bodily sensations* subscale [Body Responsiveness Questionnaire (BRS); Daubenmier, 2005; Cramer et al., 2018a], and of the Conscious Presence and Self-Control scale (Büssing et al., 2013). The *need for attention regulation with postural awareness* subscale, on the other hand, did not significantly correlate with the BAQ, but showed a relevant association with the BRS Perceived Connection between Mental and Physical Process subscale, reflecting the need to strive for achieving or maintaining a link between cognitive process and bodily needs. To conclude, both factors were also significantly correlated to the subscales of the Dresden Body Image Inventory (Pöhlmann, 2014), indicating the association between high levels of posturalawareness and a more positive attitude toward one's body and appearance.

Rationale for the Study

Further studies on postural awareness would add useful contributions to the mind–body integration perspective, with possible positive repercussions in the field of psychological and psychotherapeutic intervention. The scientific literature shows the efficacy of some interventions for the improvement of posture aspects such as balance (Wayne et al., 2004; Kee et al., 2012), coordination (Jay et al., 2013), control (Pluchino et al., 2012), and awareness (Roll et al., 2002). However, up to now, particularly complex (Barrus, 1996; Aminian and Najafi, 2004) and/or hardly usable outside the laboratory setting (Lanningham-Foster et al., 2005; Wong et al., 2007) tools have been used to measure these outcomes. With the exception of PAS, no self-report tools have been found to allow a more agile assessment of subjective postural awareness (Cramer et al., 2018b). The simplicity of this self-administered scale would enable a measurement of postural awareness in the absence of technical devices and within a psychological setting.

The aim of the present research is the validation (and evaluation of psychometric properties) of the Italian PAS, originally created in German by Cramer et al. (2018b), to allow its use in research and clinical practice. In light of the excellent psychometric characteristics of the original instrument, we hypothesize to obtain an Italian version with a good internal coherence and a similar and equally good factor structure.

MATERIALS AND METHODS

Participants

The study involved 928 individuals (45.04% men and 54.96% women) with an age ranging from 18 to 77 years (mean = 29.96, standard deviation = 11.44). The sample included participants from Northern (37.50%), Central (32.54%), and Southern (29.96%) Italy. Most individuals were unmarried (71.55% single). Of the 928 participants, 456 (49.14%) were students, and 255 (27.48%) were employed; 44.61% of them held a secondary school diploma, 27.37% a bachelor's, and 19.83% a master's degree; 48.28% of the sample was Catholic Christian, and 45.37% was atheist. Three hundred seventy-four participants (40.30%) did not practice any type of sports, whereas 260 (28.02%) trained in the gym (**Table 1**).

TABLE 1 | Demographics variables of the sample (*N* = 928).

Age		
		Mean = 29.96, Standard deviation = 11.44
	n	**%**
Sex		
Male	418	45.04
Female	510	54.96
Provenance		
Northern Italy	348	37.50
Central Italy	302	32.54
Southern Italy	278	29.96
Marital status		
Single	664	71.55
Married	111	11.96
Separated	34	3.66
Divorced	22	2.37
Widowed	11	1.19
Cohabitant	86	9.27
Professional condition		
Unemployed	64	6.90
Student	456	49.14
Housewife	12	1.23
Freelance	123	13.25
Employee	255	27.48
Retired	10	1.08
Other	8	0.86
Study degree		
Middle school diploma	52	5.60
High school diploma	414	44.61
University degree	254	27.37
Master's degree	184	19.83
Postlaurea specialization	24	2.59
Religion		
Catholic Christian	448	48.28
Muslim	2	0.22
Buddhist	11	1.19
Atheist	421	45.37
Jehovah's Witness	3	0.32
Agnostic	30	3.23
Other	13	1.40
Sport		
Gym	260	28.02
Water sports	46	4.96
Football/soccer	34	3.66
Cycling and running	31	3.34
Walk and trekking	27	2.91
Bodyweight exercises, free exercises, yoga, fitness	35	3.77
Dance and skating	23	2.48
Volley	20	2.16
Basket and rugby	16	1.72
Martial arts and combat sports	33	3.56
Other	29	3.13
No sport	374	40.30

Procedures

Items of the original version of the PAS have been translated into Italian by a native German speaker living in Italy. Then, the Italian version was back-translated by a bilingual Italian German teacher, and the outcome was submitted to the author of the original measure, with the help of which the remaining inaccuracies were corrected. The researchers compared the translated version with the original text until a consensus on cross-language equivalence was reached. The participants were recruited on the internet with an anonymous link spread through a snowball-like procedure, and the presence of psychological or orthopedic issues was adopted as criteria for exclusion from the sample. All the subjects were informed about the aim of the research and gave written informed consent in accordance with the Declaration of Helsinki. The self-report measures together with a demographic questionnaire (i.e., age, sex, weight, height) were administered to participants, who did not take any compensation for their involvement in the study. The subjects were guaranteed privacy and anonymity.

Measures

Postural Awareness Scale

The PAS is a brief self-report measure designed to assess awareness of body posture (Cramer et al., 2018b), and it consists of 12 items scored on a 7-point scale anchored by 1 (not at all true for me) and 7 (very true for me). Results supported the internal consistency of the original German PAS, with a Cronbach α of 0.80 for the total scale and 0.81 and 0.77 for the two subscales (*ease/familiarity with postural awareness* e *need for attention regulation with postural awareness,* respectively). The scale scores range from 12 to 84, with higher scores being indicative of greater postural awareness. The scores were computed by adding up the answers to all the items, after reversing the values of items 1, 2, 3, 4, 5, and 12. In this study, an Italian version obtained by a back-translation process was used.

Body Image Concern Inventory

The Body Image Concern Inventory (BICI) is a self-report measure for assessing experiences related to dysmorphic concern (Littleton et al., 2005). In this study, the Italian version of the BICI (I-BICI; Luca et al., 2011) was used. It consists of 19 items divided into two subscales: dysmorphic symptoms and symptom interference. Response categories ranged from 1 (never) to 5 (always), and the scale scores range from 19 to 95. The aspects investigated were dissatisfaction and concern about appearance, checking and camouflaging behavior, reassurance seeking, social concerns, and avoidance related to appearance. In this sample, the I-BICI possesses good internal consistency, with a Cronbach α of 0.92 and 0.76 for the two subscales and α = 0.93 for the total scale.

Rosenberg Self-Esteem Scale

The Rosenberg Self-Esteem Scale (RSES) is a 10-item self-report questionnaire designed for assessing global self-esteem with items answered on a 4-point scale from *strongly agree* to *strongly disagree* (Rosenberg, 1965). The scale scores range from 0 to 30, in which scores between 15 and 25 are within normal range, whereas scores less than 15 suggest low self-esteem. In this study, the Italian version of the RSES (Prezza et al., 1997), showing good internal consistency (α = 0.90), was used.

General Self-Efficacy Scale

The General Self-Efficacy Scale (GSE) is a self-report measure of self-beliefs to cope with a variety of difficult demands in life (Schwarzer and Jerusalem, 1995). It consists of 10 items scored on a 4-point scale anchored by 1 (not at all true for me) and 4 (very true for me). The scale scores range from 10 to 40, with higher scores being indicative of a sense of personal competence in stressful situations. In this sample, the Italian versions of the GSE (Sibilia et al., 1995) showed a high internal consistency (α = 0.90).

Body Awareness Questionnaire

The BAQ is an 18-item self-report questionnaire designed to assess the sensitivity to normal and non-emotional body processes (Shields et al., 1989). Each item on the measure is rated on a 7-point scale ranging from 1 (not at all true for me) to 7 (very true for me). In this study, the Italian translation of BAQ (Shields et al., 1989; for the Italian version Cardinali, unpublished manuscript) possesses good internal consistency with a Cronbach α of 0.88.

West Haven–Yale Multidimensional Pain Inventory – Short Version

The West Haven–Yale Multidimensional Pain Inventory (WHYMPI-S) is a self-report measure designed to examine the impact of chronic pain on patients' lives, quality of social support, and general activities (Kerns et al., 1985). In the present study, a short version of this measure was used: five items (2, 8, 9, 12, 19) of the 52 taken from the Italian version (Ferrari et al., 2000), showing a good internal consistency (α = 0.87), were readapted. The selected items evaluated interference in daily life, changes in the ability to participate in recreational and social activities, in the level of satisfaction deriving from involvement in family activities, in the level of suffering, and in friendship. Responses were on a 5-point Likert scale, and higher scores indicated higher levels of suffering and impact of chronic pain.

20-Item Toronto Alexithymia Scale

The 20-item Toronto Alexithymia Scale (TAS-20) is a well-known 20-item questionnaire, scored on a 1- to 5-point Likert scale, which assesses the level of alexithymia (Bagby et al., 1994). The scale measures three main dimensions: (1) difficulty in identifying feelings and distinguishing between feelings and bodily sensations in emotional activation, (2) difficulty in the verbal expression of emotions, and (3) externally oriented thinking. In this sample, the Italian version of the TAS-20 (Bressi et al., 1996), showing a good internal consistency with a Cronbach α of 0.86 for the total score (α = 0.84, 0.79, 0.65 for the subscales), was used.

Beck Depression Inventory II

The Beck Depression Inventory II (BDI-II) is a 21-item self-report multiple-choice inventory designed to assess the intensity of depression (Beck et al., 1996). Response categories range from 1 to 3, and the scale scores range from 0 to 63. It is composed of two subscales: a cognitive–affective and a somatic–performance subscale. In this study, the Italian translation of BDI-II (Ghisi et al., 2006) possesses high internal consistency with a Cronbach α of 0.91 for the total score (α = 0.84 and 0.88 for the subscales).

Mindfulness Attention Awareness Scale

The Mindfulness Attention Awareness Scale (MAAS) is a self-report measure designed to assess present attention and awareness (Brown and Ryan, 2003). In this study, the Italian version of the MAAS (Veneziani and Voci, 2015) was used. It includes 15 items to be rated on a 7-point Likert scale from 1 (almost always) to 7 (almost never), with higher scores being indicative of greater mindfulness. In this sample, the Italian version possesses a good internal consistency (α = 0.87).

Modified Somatic Perception Questionnaire

The Modified Somatic Perception Questionnaire (MSPQ) is a self-report measure of somatic and autonomic perceptions (Main, 1983). In this study, the Italian translation of MSPQ (Conti, 1999) was used. It consists of 22 items scored on a 0- to 4-point Likert scale, 13 of which are used for the final score (the others have a masking function). In the present sample, the Italian version possesses a good internal consistency ($\alpha = 0.85$).

Data Analysis

All the statistical analyses were performed using the software SPSS 25.0 for Windows (IBM Corp, 2017, Armonk, NY, United States) and MPlus Version 8.1 (Muthén and Muthén, 1998–2017). Descriptive statistics were examined. To test the factor structure of the Italian PAS, the sample was randomly split. On the first subsample, Velicer's Minimum Average Partial Test (MAP), Horn's Parallel Analysis (HPA), and an exploratory factor analysis (EFA) with principal axis factoring extraction method (Promax rotation) were performed. Then, the factor structure was verified with a confirmatory factor analysis (CFA) on the second subsample, using the following fit indices: (1) the model χ^2, which indicates a good model fit when $p > 0.05$ (Hooper et al., 2008); (2) the goodness-of-fit statistic (GFI), with recommended values ≥ 0.95 (Hooper et al., 2008); (3) the non-normed fit index (NNFI) with recommended values ≥ 0.95 (Hu and Bentler, 1999); (4) the comparative fit index (CFI), for which the recommended values are ≥ 0.95, although values between 0.90 and 0.95 indicate reasonable fit (Kline, 2005); (5) the root mean square error of approximation (RMSEA), with recommended values ≤ 0.05, although values up to 0.08 represent reasonable errors of approximation (Marsh et al., 2004); (6) the standardized root mean square residual, with recommended values ≤ 0.08 (Hooper et al., 2008; Hu and Bentler, 1999). After that, the reliability of the scale was calculated both with the Cronbach α coefficient and item-total correlation indices. In order to assess some aspects of construct validity, Pearson correlation was calculated between PAS, I-BICI, RSES, GSE, BAQ, WHYMPI-S, TAS-20, BDI-II, MAAS, and MSPQ. The choice of these measures was driven by the observation that there are no other self-report questionnaires for the assessment of postural awareness: measures evaluating aspects of awareness and somatic perceptions were therefore included. Moreover, as for large samples even low correlations could be significant, greater precision was searched in the evaluation of the discriminating validity of the two subscales of the PAS, by implementing a correlation coefficients comparison according to Meng et al. (1992). Finally, to assess the differences between specific subgroups, the multivariate analysis of variance (MANOVA) was carried out, by simultaneously entering all the background variables [gender, age, practice of sport, body mass index (BMI)] as fixed factors in a multivariate general linear model. Separate follow-up ANOVAs were conducted for the dependent variables when it was necessary, and *post hoc* analyses using Scheffé test were performed to support the interpretation of the differences between averages where needed.

RESULTS

Descriptive Statistics

The descriptive statistics of the sample were reported in **Table 1**. The mean values of the PAS items ranged from 2.69 to 5.61 (**Table 2**).

Factor Structure of the Italian PAS

First, in accordance with O'connor (2000), the MAP and HPA were carried out (**Table 3**). Both the original MAP (Velicer, 1976) and the revised MAP (Velicer et al., 2000) suggested the retention of two factors, as well as the HPA (Horn, 1965).

Furthermore, an EFA with principal axis factoring extraction method (Promax rotation) yielded two interpretable factors, which explained 51.00% of the total variance (**Table 4** and **Figure 1**). The first factor (ease/familiarity with postural awareness) was made up of six items related to high postural awareness without effort and accounted for 27.82% of the total variance. The second factor (need for attention regulation with postural awareness) consisted of six items related to high efforts required to be aware of their own body posture; it accounted for 23.18% of the total variance.

Concerning the CFA, although the χ^2 was significant with $\chi^2(36, n = 463) = 134.877$, $p < 0.001$, the other indices showed satisfactory values and supported the two-factor solution of the Italian PAS: GFI = 0.954, NNFI = 0.921, CFI = 0.940, RMSEA = 0.077, SRMR = 0.066.

Reliability of the Scale

A Cronbach α coefficient ($\alpha = 0.76$ for the total scale and $\alpha = 0.80$, 0.79 for the two subscales) suggested satisfactory reliability. Item-total correlations (**Table 2**) showed values ranging from 0.19 (Item 7) to 0.57 (Item 8).

Construct Validity

Intercorrelations between PAS subscale scores were $r = 0.11$, $p < 0.01$, and they significantly and positively correlated with the PAS total score (F1, $r = 0.73$, $p < 0.01$; F2, $r = 0.76$, $p < 0.01$).

The Italian PAS showed significant correlations with most measures used to assess construct validity (**Table 5**). More specifically, correlations of particular importance for the convergent validity were those shown with BAQ ($r = 0.23$, $p < 0.01$, for the total PAS scale; and $r = 0.32$, $p < 0.01$, for the first PAS subscale, but there was no significant correlation with the second PAS factor), RSES ($r = 0.19$, $p < 0.01$; $r = 0.07$, $p < 0.05$; $r = 0.22$, $p < 0.01$ for total PAS score, the first and the second PAS subscales, respectively), GSE ($r = 0.25$, $p < 0.01$; $r = 0.24$, $p < 0.01$; $r = 0.14$, $p < 0.01$ for total PAS score, the first and the second PAS subscales, respectively), and MAAS ($r = 0.19$, $p < 0.01$; $r = 0.13$, $p < 0.01$; $r = 0.15$, $p < 0.01$ for total PAS score, the first and the second PAS subscales, respectively). Regarding discriminant validity, specific relevance has been given to I-BICI and TAS-20 measurements. The PAS total score and its second subscale were significantly and negatively correlated with the I-BICI total scale ($r = -0.28$, $p < 0.01$; $r = -0.37$, $p < 0.01$, respectively), the first I-BICI subscale ($r = -0.28$, $p < 0.01$; $r = -\ 0.37$,

TABLE 2 | Descriptive statistics and item-total correlations of each of the Italian PAS items.

Item	*N*	Minimum	Maximum	Mean	Standard deviation	Skewness	Kurtosis	Item-total correlation
1[a]	928	1.00	7.00	4.44	1.76	−0.13	−0.91	0.49
2[a]	928	1.00	7.00	3.88	1.95	0.10	−1.23	0.42
3[a]	928	1.00	7.00	3.27	2.05	0.56	−0.98	0.36
4[a]	928	1.00	7.00	2.69	1.69	1.01	0.25	0.34
5[a]	928	1.00	7.00	4.30	2.08	−0.15	−1.33	0.50
6	928	1.00	7.00	3.71	1.84	0.12	−1.08	0.38
7	928	1.00	7.00	4.61	1.93	−0.46	−0.90	0.19
8	928	1.00	7.00	3.47	1.75	0.32	−0.91	0.57
9	928	1.00	7.00	4.54	1.72	−0.35	−0.79	0.41
10	928	1.00	7.00	3.08	1.69	0.53	−0.65	0.56
11	928	1.00	7.00	3.58	1.88	0.21	−1.05	0.29
12[a]	928	1.00	7.00	3.90	1.87	0.17	−1.01	0.26
Valid *N* (listwise)	928							

[a] *Reversed scoring.*

TABLE 3 | MAP test and parallel analysis results for the number of components.

	Average partial correlations				Random data eigenvalues	
N	Squared	Power 4	*N*	Eigenvalues	Means	95% Percentile
0	0.08	0.01	1	3.34	1.27	1.32
1	0.07	0.01	**2**	**2.78**	**1.20**	**1.25**
2	**0.03**	**0.00**	3	0.96	1.14	1.20
3	0.04	0.00	4	0.83	1.10	1.13
4	0.05	0.01	5	0.66	1.05	1.08
5	0.07	0.01	6	0.64	1.02	1.04
6	0.10	0.03	7	0.61	0.98	1.01
7	0.14	0.06	8	0.56	0.93	0.97
8	0.22	0.11	9	0.54	0.89	0.93
9	0.31	0.20	10	0.42	0.86	0.89
10	0.60	0.47	11	0.39	0.81	0.84
11	1	1	12	0.28	0.76	0.81

Bold values show the number of components, according to the tests.

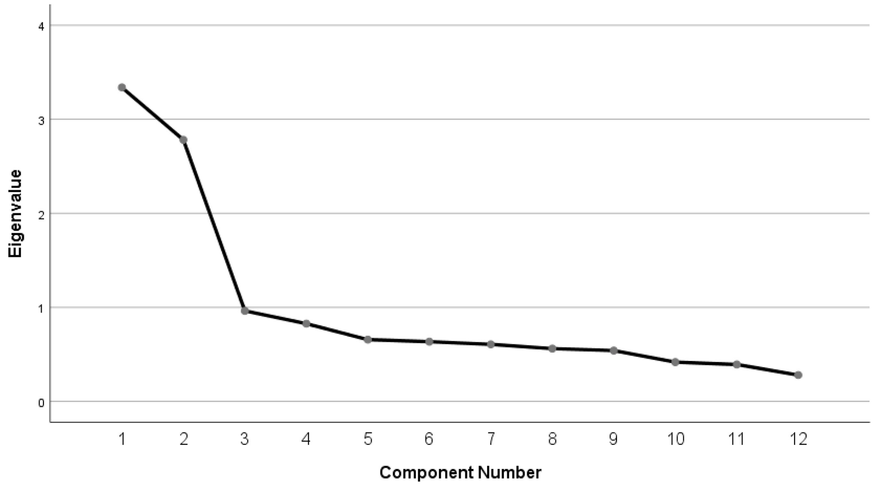


FIGURE 1 | Scree plot.

TABLE 4 | Factor structure of the Italian PAS.

Item	F1	F2
1. Ich muss mich sehr konzentrieren, um meine Körperhaltung wahrzunehmen.[b] (Needs to concentrate for being aware of posture) *Devo concentrarmi molto per percepire la mia postura*[a,c]	0.24	**0.63**
2. Wenn ich eine ungünstige Körperhaltung einnehme, bemerke ich dies oft erst, wenn ich Schmerzen bekomme.[b] (Awareness of bad posture only by pain) *Spesso, mi accorgo di assumere posture scorrette solo quando provo dolore* [a,c]	0.07	**0.56**
3. Im Sitzen sacke ich oft unbewusst in mich zusammen.[b] (Slumps down when sitting) *Quando sono seduto/a, spesso mi "accascio" inconsapevolmente*[a,c]	0.02	**0.62**
4. Wenn ich mich auf eine Tätigkeit konzentriere, nehme ich oft unbewusst eine bestimmte Körperhaltung ein.[b] (Unaware of posture when focused) *Mi capita spesso di assumere inconsapevolmente una determinata postura quando sono concentrato/a su un'attività*[a,c]	−0.09	**0.65**
5. Es fällt mir schwer, bewusst eine bestimmte Körperhaltung einzunehmen.[b] (Difficulties to consciously adopt a posture) *Ho difficoltà ad adottare consapevolmente una certa postura*[a,c]	0.25	**0.66**
6. Während der Arbeit überprüfe ich immer wieder meine Körperhaltung.[b] (Often checks posture when working) *Controllo spesso la mia postura mentre lavoro*[c]	**0.60**	0.08
7. Über meine Körperhaltung kann ich beeinflussen, wie ich auf andere Menschen wirke.[b] (Influences her/his own appeal by posture) *Attraverso la mia postura sono in grado di influenzare l'impressione che do alle altre persone*[c]	**0.53**	−0.23
8. Mir ist im Alltag immer bewusst, wie ich im Moment sitze oder stehe.[b] (Always aware of sitting or standing posture) *Nella vita di tutti i giorni sono sempre consapevole di com'è la mia postura quando sono seduto/a o in piedi*[c]	**0.72**	0.27
9. Ich rufe mir oft aktiv ins Bewusstsein, wie ich im Moment sitze oder stehe.[b] (Often makes her/himself aware of her/his posture) *Spesso cerco di essere consapevole della mia postura da seduto/a o in piedi*[c]	**0.73**	0.04
10. Selbst bei konzentrierten Arbeiten bin ich mir meiner Körperhaltung stets bewusst.[b] (Aware of posture even when focused) *Sono sempre consapevole della mia postura anche quando sto svolgendo attività che richiedono concentrazione*[c]	**0.68**	0.29
11. Über meine Körperhaltung kann ich bewusst steuern, wie es mir geht.[b] (Regulates how she/he feels through posture) *Riesco a influenzare consapevolmente come mi sento attraverso la mia postura*[c]	**0.59**	−0.04
12. Ob eine Körperhaltung mir gut tut oder nicht merke ich meist erst, wenn ich mich darauf konzentriere.[b] (Needs to concentrate to feel whether a posture benefits her/him or not) *Il più delle volte, noto se una postura va bene o meno per me solo se mi concentro su di essa*[a,c]	−0.03	**0.49**
Factor correlation matrix		
Factor 1	1	
Factor 2	0.14	1

Factor 1, ease/familiarity with postural awareness (α = 0.80); Factor 2, need for attention regulation with postural awareness (α = 0.79). [a]*Reverse item.* [b]*Original version of the PAS.* [c]*Italian version of the PAS. Bold values indicate strong factor loadings.*

$p < 0.01$, respectively), and the second I-BICI subscale ($r = -0.20$, $p < 0.01$; $r = -\ 0.28$, $p < 0.01$, respectively). Furthermore, the PAS total score and its subscales were significantly and negatively correlated with the TAS-20 total scale ($r = -0.25$, $p < 0.01$; $r = -\ 0.09$, $p < 0.01$; $r = -0.28$, $p < 0.01$, respectively), the first TAS-20 subscale ($r = -0.22$, $p < 0.01$; $r = -0.31$, $p < 0.01$, only total PAS and the second PAS factor, respectively), the second TAS-20 subscale ($r = -0.18$, $p < 0.01$; $r = -0.21$, $p < 0.01$, only total PAS and the second PAS factor, respectively), and the third TAS-20 subscale ($r = -0.18$, $p < 0.01$; $r = -\ 0.14$, $p < 0.01$; $r = -0.13$, $p < 0.01$, respectively).

Then, a correlation coefficients comparison (Meng et al., 1992) was used to assess the discriminant validity of the PAS subscales (**Table 6**). The analysis showed that the subscales correlations with total PAS ($z = -1.54$, $p = 0.124$), the third factor of the TAS20 ($z = -0.23$, $p = 0.817$), and MAAS ($z = -0.46$, $p = 0.644$) were not significantly different.

General Linear Model

The results of the MANOVA revealed no significant differences regarding gender or age on level of postural awareness (**Tables 7**, **8**).

TABLE 5 | Correlations of the measures used to assess construct validity.

	1	1a	1b	2	2a	2b	3	4	5	6	7	7a	7b	7c	8	8a	8b	9	10
1) PAS	1	0.73** [0.69, 0.78]	0.76** [0.72, 0.80]	−0.28** [−0.34, −0.21]	−0.28** [−0.34, −0.21]	−0.20** [−0.27, −0.14]	0.19** [0.13, 0.26]	0.25** [0.19, 0.32]	0.23** [0.16, 0.29]	−0.14** [−0.20, −0.07]	−0.25** [−0.31, −0.19]	−0.22** [−0.29, −0.16]	−0.18** [−0.24, −0.12]	−0.18** [−0.24, −0.11]	−0.23** [−0.29, −0.17]	−0.24** [−0.30, −0.17]	−0.18** [−0.25, −0.12]	0.19** [0.12, 0.25]	−0.15** [−0.21, −0.09]
1a) PAS (F1)		1	0.11** [0.05, 0.18]	−0.04 [−0.10, 0.03]	−0.04 [−0.11, 0.02]	−0.02 [−0.09, 0.04]	0.07* [0.01, 0.13]	0.24** [0.18, 0.30]	0.32** [0.26, 0.38]	−0.01 [−0.08, 0.05]	−0.09** [−0.15, −0.02]	−02 [−0.09, 0.04]	−0.06 [−0.12, 0.01]	−0.14** [−0.20, −0.07]	−0.08* [−0.14, −0.01]	−0.07* [−0.14, −0.01]	−0.06 [−0.12, 0.01]	0.13** [0.06, 0.19]	−0.01 [−0.08, 0.05]
1b) PAS (F2)			1	−0.37** [−0.43, −0.31]	−0.37** [−0.43, −0.31]	−0.28** [−0.34, −0.21]	0.22** [0.15, 0.28]	0.14** [0.08, 0.20]	0.02 [−0.04, 0.09]	−0.19** [−0.25, −0.13]	−0.28** [−0.34, −0.22]	−31** [−0.37, −0.25]	−0.21** [−0.27, −0.15]	−0.13** [−0.19, −0.06]	−0.27** [−0.33, −0.20]	−0.27** [−0.34, −0.21]	−0.21** [−0.28, −0.15]	0.15** [0.09, 0.21]	−0.21** [−0.27, −0.15]
2) BICI				1	0.98** [0.98, 1.00]	0.80** [0.77, 0.84]	−0.44** [−0.50, −0.38]	−0.33** [−0.39, −0.27]	0.00 [−0.06, 0.06]	0.23** [0.17, 0.29]	0.31** [0.25, 0.37]	0.39** [0.33, 0.45]	0.26** [0.19, 0.31]	0.04 [−0.02, 0.11]	0.49** [0.43, 0.55]	0.45** [0.39, 0.50]	0.46** [0.40, 0.51]	−0.18** [−0.24, −0.11]	0.34** [0.28, 0.40]
2a) BICI (F1)					1	0.69** [0.64, 0.74]	−44** [−0.50, −0.83]	−0.33** [−0.39, −0.27]	−0.00 [−0.07, 0.06]	0.21** [0.15, 0.28]	0.30** [0.24, 0.36]	0.39** [0.33, 0.45]	0.26** [0.19, 0.32]	0.03 [−0.04, 0.09]	0.48** [0.42, 0.53]	0.44** [0.38, 0.49]	0.44** [0.38, 0.50]	−0.17** [−0.24, −0.11]	0.33** [0.27, 0.39]
2b) BICI (F2)						1	−0.34** [−0.40, −0.28]	−0.23** [−0.30, −0.17]	0.01 [−0.06, 0.07]	0.23** [0.17, 0.30]	0.25** [0.19, 0.31]	0.31** [0.24, 0.37]	0.18** [0.12, 0.25]	0.08* [0.01, 0.14]	0.42** [0.36, 0.48]	0.37** [0.31, 0.43]	0.41** [0.35, 0.47]	−0.15** [−0.21, −0.08]	0.29** [0.23, 0.35]
3) RSES							1	0.47** [0.42, 0.53]	0.16** [0.10, 0.23]	−0.13** [−0.20, −0.07]	−0.40** [−0.46, −. 34]	−0.42** [−0.47, −0.36]	−0.32** [−0.39, −0.26]	−0.17** [−0.23, −0.11]	−0.54** [−0.60, −0.49]	−0.42** [−0.48, −0.36]	−0.59** [−0.64, −0.54]	0.22** [0.16, 0.29]	−0.23** [−0.29, −0.17]
4) GSE								1	0.26** [0.19, 0.32]	−0.14** [−0.20, −0.08]	−0.37** [−0.43, −0.31]	−0.33** [−0.39, −0.27]	−0.27** [−0.33, −0.21]	−0.26** [−0.32, −0.19]	−0.44** [−0.50, −0.38]	−0.34** [−0.40, −0.28]	−0.47** [−0.53, −0.41]	0.20** [0.13, 0.26]	−0.16** [−0.22, −0.10]
5) BAQ									1	0.08* [0.01, 0.14]	−0.16** [−0.22, −0.09]	−0.06 [−0.13, 0.00]	−0.11** [−0.17, −0.04]	−0.22** [−0.28, −0.16]	−0.08* [−0.14, −0.01]	−0.04 [−0.10, 0.03]	−0.11** [−0.17, −0.04]	0.11** [0.04, 0.17]	0.04 [−0.02, 0.11]
6) WHYMPI-S										1	0.20** [0.14, 0.27]	0.27** [0.21, 0.33]	0.13** [0.07, 0.19]	0.05 [−0.01, 0.12]	0.30** [0.24, 0.37]	0.34** [0.27, 0.40]	0.23** [0.17, 0.29]	−0.10** [−0.16, −0.03]	0.37** [0.31, 0.43]
7) TAS20											1	0.83** [0.80, 0.87]	0.83** [0.79, 0.87]	0.69** [0.65, 0.74]	0.49** [0.43, 0.55]	0.44** [0.39, 0.50]	0.46** [0.40, 0.52]	−0.29** [−0.35, −0.23]	0.30** [0.24, 0.36]
7a) TAS20 (F1)												1	0.59** [0.54, 0.64]	0.30** [0.24, 0.36]	0.59** [0.43, 0.64]	0.55** [0.50, 0.61]	0.52** [0.47, 0.58]	−0.27** [−0.34, −0.21]	0.40** [0.34, 0.46]
7b) TAS20 (F2)													1	0.40** [0.34, 0.46]	0.37** [0.31, 0.43]	0.31** [0.25, 0.38]	0.36** [0.30, 0.42]	−0.20** [−0.16, −0.13]	0.20** [0.14, 0.26]
7c) TAS20 (F3)														1	0.16** [0.10, 0.22]	0.13** [0.06, 0.19]	0.17** [0.10, 0.23]	−0.20** [−0.26, −0.14]	0.07* [0.05, 0.13]
8) BDI-II															1	0.93** [0.90, 0.95]	0.92** [0.89, 0.94]	−0.28** [−0.34, −0.22]	0.49** [0.44, 0.55]
8a) BDI-II (F1)																1	0.70** [0.66, 0.75]	−0.27** [−0.34, −0.21]	0.55** [0.49, 0.60]
8b) BDI-II (F2)																	1	−0.24** [−0.30, −0.18]	0.35** [0.29, 0.41]
9) MAAS																		1	−0.18** [−0.24, −0.12]
10) MSPQ																			1

***Correlation is significant at the 0.01 level (2-tailed). *Correlation is significant at the 0.05 level (2-tailed). PAS, Postural Awareness Scale; PAS (F1), Factor 1 "need for attention regulation with postural awareness"; PAS (F2), Factor 2, "ease/familiarity with postural awareness"; BICI, Body Image Concern Inventory; BICI (F1), Factor 1 "dysmorphic symptoms"; BICI (F2), Factor 2 "symptom interference"; RSES, Rosenberg Self-Esteem Scale; GSE, General Self-Esteem; BAQ, Body Awareness Questionnaire; WHYMPI-S, West Haven–Yale Multidimensional Pain Inventory – Short Version; TAS20, 20-item Toronto Alexithymia Scale; TAS20 (F1), Factor 1 "difficulty identifying feelings and distinguishing between feelings and bodily sensations in emotional activation"; TAS20 (F2), Factor 2 "difficulty in the verbal expression of emotions"; TAS20 (F3), Factor 3 "externally oriented thinking"; BDI-II, Beck Depression Inventory II; BDI-II (F1): Factor 1 "cognitive–affective subscale"; BDI-II (F2): Factor 2 "somatic–performance subscale"; MAAS, Mindfulness Attention Awareness Scale; MSPQ, Modified Somatic Perception Questionnaire. Bold values indicate significant correlations.*

TABLE 6 | Comparison of correlation coefficients between PAS subscales and the other variables.

		95% Confidence interval				
	r Diff.	Lower limit	Upper limit	*z*	*p*	Effect size
1) PAS	−0.03	−0.15	0.02	−1.54	0.124	0.05
2) BICI	0.33	0.26	0.41	7.79	< 0.001	0.26
2a) BICI (F1)	0.33	0.26	0.41	7.79	< 0.001	0.26
2b) BICI (F2)	0.26	0.18	0.34	6.04	< 0.001	0.20
3) RSES	−0.15	−0.24	−0.07	−3.47	< 0.001	0.11
4) GSE	0.10	−0.02	0.19	2.34	0.019	0.08
5) BAQ	0.30	0.22	0.38	7.00	< 0.001	0.23
6) WHYMPI-S	0.18	0.10	0.26	4.14	< 0.001	0.14
7) TAS	0.19	0.11	0.28	4.45	< 0.001	0.15
7a) TAS (F1)	0.29	0.21	0.37	6.76	< 0.001	0.22
7b) TAS (F2)	0.15	0.07	0.24	3.47	< 0.001	0.11
7c) TAS (F3)	−0.01	−0.10	0.08	−0.23	0.817	0.01
8) BDI	0.19	0.11	0.28	4.43	< 0.001	0.15
8a) BDI (F1)	0.20	0.12	0.29	4.66	< 0.001	0.15
8b) BDI (F2)	0.15	0.07	0.24	3.47	< 0.001	0.11
9) MAAS	−0.02	−0.11	0.07	−0.46	0.644	0.02
10) MSPQ	0.20	−0.12	0.28	4.60	< 0.001	0.15

BICI, Body Image Concern Inventory; BICI (F1), Factor 1 "dysmorphic symptoms"; BICI (F2), Factor 2 "symptom interference"; RSES, Rosenberg Self-Esteem Scale; GSE, General Self-Esteem; BAQ, Body Awareness Questionnaire; WHYMPI-S, West Haven–Yale Multidimensional Pain Inventory – Short Version; TAS20, 20-item Toronto Alexithymia Scale; TAS20 (F1), Factor 1 "difficulty identifying feelings and distinguishing between feelings and bodily sensations in emotional activation"; TAS20 (F2), Factor 2 "difficulty in the verbal expression of emotions"; TAS20 (F3): Factor 3 "externally oriented thinking"; BDI-II: Beck Depression Inventory II; BDI-II (F1): factor 1 "cognitive–affective subscale"; BDI-II (F2): factor 2 "somatic–performance subscale"; MAAS, Mindfulness Attention Awareness Scale; MSPQ, Modified Somatic Perception Questionnaire.

TABLE 7 | Summary of PAS total scale and PAS subscales scores by men and women.

				95% Confidence interval		
Dependent variable	Sex	Mean	Standard error	Lower	Upper	Partial η^2
PAS	Male	44.97[a]	0.97	43.06	46.87	0.000
	Female	44.35[a]	1.00	42.39	46.31	
PAS (F1)	Male	22.99[a]	0.66	21.70	24.29	0.000
	Female	22.98[a]	0.68	21.65	24.31	
PAS (F2)	Male	21.98[a]	0.67	20.65	23.30	0.000
	Female	21.36[a]	0.69	20.01	22.72	

[a] *Based on modified population marginal mean. PAS, Postural Awareness Scale; PAS (F1), Factor 1 "need for attention regulation with postural awareness"; PAS (F2), Factor 2 "ease/familiarity with postural awareness."*

There was a significant difference between those who practice sport and those who do not (**Table 9**) when considered jointly on the variables total PAS, PAS (F1) and PAS (F2), Wilk's Λ = 0.991, $F(2,843) = 3.93$, $p = 0.020$, partial $\eta^2 = 0.01$. A separate ANOVA was conducted for each dependent variable, with each ANOVA evaluated at an α level of 0.025. There were significantly higher scores in those who practice sport than those who do not on both total PAS score and first PAS subscale, but not on the second one: $F(1,844) = 7.80$, $p = 0.005$, partial $\eta^2 = 0.01$; $F(1,844) = 5.87$, $p = 0.028$, partial $\eta^2 = 0.01$, respectively.

Indeed, significant differences related to BMI (**Table 10**) were found when considered jointly on the variables total PAS, PAS (F1) and PAS (F2), Wilk's Λ = 0.980, $F(8,1686) = 1.12$, $p = 0.031$, partial $\eta^2 = 0.01$. A separate ANOVA was conducted for each dependent variable, with each ANOVA evaluated at an α level of 0.025. There was a significant difference among the different BMI range only on total PAS score: $F(4,844) = 2.38$, $p = 0.050$, partial $\eta^2 = 0.01$. More specifically, *post hoc* analysis (Scheffé) showed that the group "normal weight" had a higher level of postural awareness.

DISCUSSION

The aim of this research was to analyze the psychometric characteristics of the Italian version of the PAS (Cramer et al., 2018b), a measure of body posture awareness. This tool fits into a perspective that connects posture to well-being (Lauche et al., 2017) and which, in turn, falls within a broader theoretical frame including a growing literature supporting the close link between

TABLE 8 | Summary of PAS total scale and PAS subscales scores in different age range.

				95% Confidence interval		
Dependent variable	**Age range**	**Mean**	**Standard error**	**Lower**	**Upper**	**Partial η^2**
PAS	18–24	42.942[a]	1.099	40.786	45.099	0.008
	25–34	42.873[a]	1.220	40.478	45.268	
	35–44	44.350[a]	1.637	41.138	47.563	
	45–54	45.327[a]	1.885	41.626	49.027	
	> 54	48.658[a]	2.090	44.555	52.761	
PAS (F1)	18–24	22.354[a]	0.745	20.891	23.817	0.006
	25–34	23.279[a]	0.828	21.654	24.904	
	35–44	23.385[a]	1.111	21.205	25.564	
	45–54	21.597[a]	1.279	19.085	24.108	
	> 54	24.383[a]	1.418	21.599	27.167	
PAS (F2)	18–24	20.588[a]	0.761	19.095	22.081	0.010
	25–34	19.594[a]	0.845	17.935	21.252	
	35–44	20.966[a]	1.133	18.741	23.190	
	45–54	23.730[a]	1.306	21.167	26.293	
	> 54	24.275[a]	1.448	21.434	27.117	

[a]*Based on modified population marginal mean. PAS, Postural Awareness Scale; PAS (F1), Factor 1 "need for attention regulation with postural awareness"; PAS (F2), Factor 2 "ease/familiarity with postural awareness."*

TABLE 9 | Summary of PAS total scale and PAS subscales scores by people who practice or not sport.

				95% Confidence interval		
Dependent variable	**Sport activity**	**Mean**	**Standard error**	**Lower**	**Upper**	**Partial η^2**
PAS	No	42.725[a]	1.003	40.757	44.693	0.009
	Yes	46.557[a]	0.973	44.648	48.466	
PAS (F1)	No	21.987[a]	0.680	20.652	23.322	0.006
	Yes	23.986[a]	0.660	22.691	25.282	
PAS (F2)	No	20.738[a]	0.694	19.376	22.101	0.004
	Yes	22.571[a]	0.673	21.249	23.893	

[a]*Based on modified population marginal mean. PAS, Postural Awareness Scale; PAS (F1), Factor 1 "need for attention regulation with postural awareness"; PAS (F2), Factor 2 "ease/familiarity with postural awareness."*

physical and mental aspects (e.g., Ogden et al., 2012; Van Der Kolk, 2014; Fisher, 2017).

The Italian version of the PAS showed satisfactory psychometric properties with good indications of internal consistency and construct validity. The results obtained with MAP, HPA, and EFA supported a two-factor solution, as confirmed by the CFA and in line with the original version: the first regards the ability to have a high postural awareness in a natural and effortless way (Factor 1 "ease/familiarity with postural awareness"); the second refers the need of high efforts to be aware of their own posture (Factor 2 "need for attention regulation with postural awareness"). In line with what the authors of the original instrument indicated, the two subscales (both with good internal consistency) would seem to indicate the extremes of a continuum concerning the effort employed to be aware of one's posture (Cramer et al., 2018b).

Positive and significant correlations were found with the mindfulness (MAAS) and the body awareness (BAQ) measurements, although in the relationship with the latter the second factor of the PAS (need for attention regulation with postural awareness) is an exception (in line with the results of the original version, in which there was a low association). The absence of association of this subscale could be interpreted looking at the need of efforts to be aware of his own posture as a difficulty and a lower spontaneity to have mental representation of body aspects. More specifically, the Multiple Code Theory (Bucci, 1999) considers the visceral and physical sensations as subsymbolic processes that, through a referential process, can be depicted within the symbolic register provided by language and images. A lack of integration of these elements does not allow having a full bodily processes awareness, which is a fundamental element for the distinction between emotive or physiological physical activations.

This condition causes tensions and dysregulated states of emotional arousal that could lead to psychosomatic problems (Ruesch, 1948; MacLean, 1949): all this could result in greater attention to somatic aspects, which, however, do not lead to awareness, but only to excessive worry and anxiety. The above is confirmed by the negative associations of postural awareness with alexithymia (especially externally oriented thinking) and

TABLE 10 | Summary of PAS total scale and PAS subscales scores in different BMI range.

				95% Confidence interval			
Dependent variable	**BMI range**	**Mean**	**Standard Error**	**Lower**	**Upper**	**Partial η²**	**Scheffé *post hoc***
PAS	Underweight	45.757[a]	2.085	41.664	49.849	0.011	G2 > G1 > G3 > G4 > G5
	Normal weight	47.482	0.877	45.762	49.203		
	Overweight	44.603	1.047	42.548	46.657		
	Class I obesity	43.058[a]	1.589	39.938	46.177		
	Classes II and III obesity	41.718[a]	2.527	36.758	46.678		
PAS (F1)	Underweight	24.237[a]	1.415	21.461	27.014	0.010	–
	Normal weight	24.010	0.595	22.842	25.177		
	Overweight	21.690	0.710	20.296	23.084		
	Class I obesity	21.671[a]	1.079	19.554	23.788		
	Classes II and III obesity	23.996[a]	1.715	20.631	27.362		
PAS (F2)	Underweight	21.519[a]	1.444	18.685	24.353	0.010	–
	Normal weight	23.472	0.607	22.281	24.664		
	Overweight	22.913	0.725	21.490	24.335		
	Class I obesity	21.386[a]	1.101	19.226	23.547		
	Classes II and III obesity	17.721[a]	1.750	14.287	21.156		

[a] *Based on modified population marginal mean. PAS, Postural Awareness Scale; PAS (F1), Factor 1 "need for attention regulation with postural awareness"; PAS (F2), Factor 2 "ease/familiarity with postural awareness"; G1, underweight; G2, normal weight; G3, overweight; G4, Class I obesity; G5, Classes II and III obesity.*

the perception of physiological functions linked to states of anxiety and malaise (respectively, TAS-20 and MSPQ, which are instead positively correlated to each other). A lack of integration between symbolic and subsymbolic processes, therefore, does not allow to understand, express, and elaborate the somatic activations. In fact, the data show that both natural focus and active attention aimed at achieving and maintaining high levels of postural awareness are linked with a decrease of negative effect perception from the pain experiences (WHYMPI-S), which is in line with the scientific literature that shows that a higher non-judgmental bodily consciousness is associated with lower physical pain (Zeidan and Vago, 2016; Anheyer et al., 2017) and with a decrease in the anxiety that this condition determines (Flink et al., 2009). Furthermore, regarding the attention to aesthetic features, a negative and worried attitude toward one's appearance would seem to be associated with a sense of detachment from the body and a complete unwillingness to make efforts to be aware of the posture assumed. Indeed, negative correlations were found between PAS and dysmorphic concern scores (BICI), except for the first factor (in line with the original study).

Positive correlations with self-esteem (RSES) and self-efficacy (GSE) and negative associations with depression (BDI-II) are also identified. Scientific literature supports evidence that certain bodily attitudes can influence self-confidence, the perception of being able to cope with difficulties, and emotional state (Keltner et al., 2003; Michalak et al., 2014; Nair et al., 2015; Cuddy, 2016); on the one hand, it could deduce that a greater posture awareness allows greater control over it and over the states it influences, favoring a more positive self-image; on the other hand, this could also be interpreted taking into account that higher levels of self-esteem and self-efficacy are associated to higher insight (Gori et al., 2015), also allowing a greater sense of mastery in one's environment and a greater awareness of how body fits and interacts with it, facilitating a state of well-being.

Besides, to have more accurate interpretations about the differences between PAS subscales correlations, an inferential test was used to determine whether relevant pairs of correlations were statistically different in magnitude. The findings support the construct validity: significant associations were found between the positive correlations that the PAS subscales have with the PAS total score, between those with the MAAS and between the negative correlations that they have with the TAS20 "external oriented thinking" subscale.

Other important results obtained from the present research confirm the positive effects of physical activity and healthy body weight. Indeed, previous studies suggest that repeated exposure to bodily functions related to physical activity (e.g., increased breathing and heart rate) may lead to better body awareness in the various aspects that characterize it (Skrinar et al., 1986; Mehling et al., 2009), which in turn can be associated with greater body satisfaction and a decrease in disordered eating attitudes (Daubenmier, 2005). On the contrary, no significant differences were found regarding gender and age, in line with other research (Price and Thompson, 2007; Cramer et al., 2018b). This study has some limitations that need to be identified and discussed. First, several statistical comparisons have been carried out without any control procedure for false discovery rate, and this should be considered in the interpretation of the results: future research could overcome this limit, also correcting the *p*-values for multiple comparisons. Besides, as self-report tool, the PAS requires a self-assessment of aspects for which there could be a low level of consciousness; by definition, it is not possible to understand the actual association between self-report and the real postural awareness. Future research could use a multimodal approach (e.g., adding laboratory measurements and in-depth interviews) to have more complete assessments and overcome this limit, albeit with a greater expenditure of resources. Furthermore, the sample is composed only of Italian subjects, and this impacts the generalizability of the results in other cultures. Specifically, it could be interesting to study and analyze the differences in postural awareness levels in Eastern countries, considering, for example, the positive impact that different martial arts having their origin and diffusion have on this aspect (e.g., Lauche et al., 2017). Thus, future research could expand the sample by including employees from different geographical areas, to test the cross-cultural invariance of the results too.

Despite these limitations, the results of this validation study suggest that the Italian version of the PAS is a rapid tool, simple in its administration and evaluation, and with good psychometric properties; these data imply the possibility of using this self-report easily both for research and clinical practice, elaborating interventions within the psychotherapeutic process that can act on the two dimensions of the postural awareness construct ("need for attention regulation with postural awareness" and "ease/familiarity with postural awareness").

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors conceptualized and designed the study, collaborated for the realization of the back-translation, contributed to manuscript revision, read, and approved the submitted version. ET recruited the subjects, administered them the protocol, and wrote the first draft of the manuscript. AG and ET performed the data analysis. AG was the scientific supervisor of the study.

REFERENCES

Ahmed, H., Shaphe, M. A., Iqbal, A., Khan, A. R., and Anwer, S. (2016). Effect of trunk stabilization exercises using a gym ball with or without electromyography-biofeedback in patients with chronic low back pain: an experimental study. *Physikalische Med. Rehabil. Kurortmed.* 26, 79–83. doi: 10.1055/s-0042-102537

Aminian, K., and Najafi, B. (2004). Capturing human motion using body-fixed sensors: outdoor measurement and clinical applications. *Comput. Anim. Virtual Worlds* 15, 79–94. doi: 10.1002/cav.2

Anderson, C., and Galinsky, A. D. (2006). Power, optimism, and risk- taking. *Eur. J. Soc. Psychol.* 36, 511–536. doi: 10.1002/ejsp.324

Anheyer, D., Haller, H., Barth, J., Lauche, R., Dobos, G., and Cramer, H. (2017). Mindfulness-based stress reduction for treating low back pain: a systematic review and meta-analysis. *Ann. Intern. Med.* 166, 799–807. doi: 10.7326/m16-1997

Arnette, S. L., and Pettijohn, T. F. II (2012). The effects of posture on self- perceived leadership. *Int. J. Bus. Soc. Sci.* 3, 8–13.

Bagby, R. M., Parker, J. D., and Taylor, G. J. (1994). The twenty-item Toronto Alexithymia Scale—I. Item selection and cross-validation of the factor structure. *J. Psychosom. Res.* 38, 23–32. doi: 10.1016/0022-3999(94)90005-1

Bakal, D. A. (2001). *Minding the Body: Clinical Uses of Somatic Awareness*. New York, NY: Guilford Press.

Balasubramaniam, R., and Wing, A. M. (2002). The dynamics of standing balance. *Trends Cogn. Sci.* 6, 531–536. doi: 10.1016/s1364-6613(02)02021-1

Barrus, J. W. (1996). *U.S. Patent No. 5,570,301*. Washington, DC: U.S. Patent and Trademark Office.

Beck, A. T., Steer, R. A., and Brown, G. K. (1996). Beck depression inventory-II. *San Antonio* 78, 490–498.

Bishop, S. R., Lau, M., Shapiro, S., Carlson, L., Anderson, N. D., Carmody, J., et al. (2004). Mindfulness: a proposed operational definition. *Clin. Psychol.* 11, 230–241. doi: 10.1093/clipsy.bph077

Bluhm, R. L., Williamson, P. C., Osuch, E. A., Frewen, P. A., Stevens, T. K., Boksman, K., et al. (2009). Alterations in default network connectivity in posttraumatic stress disorder related to early-life trauma. *J. Psychiatry Neurosci.* 34, 187.

Bressi, C., Taylor, G., Parker, J., Bressi, S., Brambilla, V., Aguglia, E., et al. (1996). Cross validation of the factor structure of the 20-item Toronto Alexithymia Scale: an Italian multicenter study. *J. Psychosom. Res.* 41, 551–559. doi: 10.1016/S0022-3999(96)00228-0

Brito, C. A. Jr. (1995). *Alterações Posturais. Lianza S. Medicina de Reabilitação*. Rio de Janeiro: Guanabara Koogan.

Brown, K. W., and Ryan, R. M. (2003). The benefits of being present: mindfulness and its role in psychological well-being. *J. Pers. Soc. Psychol.* 84:822. doi: 10.1037/0022-3514.84.4.822

Bucci, W. (1999). *Psicoanalisi e Scienza Cognitiva: Una Teoria Del Codice Multiplo*. Italy: Fioriti.

Büssing, A., Walach, H., Kohls, N., Zimmermann, F., and Trousselard, M. (2013). Conscious Presence and Self Control as a measure of situational awareness in soldiers–A validation study. *Int. J. Mental Health Syst.* 7:1. doi: 10.1186/1752-4458-7-1

Conti, L. (1999). *Repertorio Delle Scale Di Valutazione in Psichiatria*. Firenze: SEE Editrice.

Cramer, H., Lauche, R., Daubenmier, J., Mehling, W., Büssing, A., Saha, F. J., et al. (2018a). Being aware of the painful body: validation of the German Body Awareness Questionnaire and Body Responsiveness Questionnaire in patients with chronic pain. *PLoS One* 13:e193000. doi: 10.1371/journal.pone.0193000

Cramer, H., Mehling, W. E., Saha, F. J., Dobos, G., and Lauche, R. (2018b). Postural awareness and its relation to pain: validation of an innovative instrument measuring awareness of body posture in patients with chronic pain. *BMC Musculoskelet. Disord.* 19:109. doi: 10.1186/s12891-018-2031-9

Cuddy, A. (2016). *Il Potere Emotivo Dei Gesti*. Italy: Sperling & kupfer.

Daubenmier, J. J. (2005). The relationship of yoga, body awareness, and body responsiveness to self-objectification and disordered eating. *Psychol. Women Q.* 29, 207–219. doi: 10.1111/j.1471-6402.2005.00183.x

Dijkstra, K., Kaschak, M. P., and Zwaan, R. A. (2007). Body posture facilitates retrieval of autobiographical memories. *Cognition* 102, 139–149. doi: 10.1016/j.cognition.2005.12.009

Ferrari, R., Novara, C., Sanavio, E., and Zerbini, F. (2000). Internal structure and validity of the multidimensional pain inventory, Italian language version. *Pain. Med.* 1, 123–130. doi: 10.1046/j.1526-4637.2000.00020.x

Fisher, J. (2017). *Healing the Fragmented Selves Of Trauma Survivors: Overcoming Internal Self-Alienation*. Abingdon: Routledge.

Flink, I. K., Nicholas, M. K., Boersma, K., and Linton, S. J. (2009). Reducing the threat value of chronic pain: a preliminary replicated single-case study of interoceptive exposure versus distraction in six individuals with chronic back pain. *Behav. Res. Ther.* 47, 721–728. doi: 10.1016/j.brat.2009.05.003

Friis, S., Skatteboe, U. B., Hope, M. K., and Vaglum, P. (1989). Body awareness group therapy for patients with personality disorders. *Psychother. Psychosom.* 51, 18–24. doi: 10.1159/000288128

Ghisi, M., Flebus, G. B., Montano, A., Sanavio, E., and Sica, C. (2006). *Beck Depression Inventory II Italian Version*. Firenze: Editore Giunti OS.

Gilbert, D., and Waltz, J. (2010). Mindfulness and health behaviors. *Mindfulness* 1, 227–234. doi: 10.1007/s12671-010-0032-3

Gori, A., Craparo, G., Giannini, M., Loscalzo, Y., Caretti, V., La Barbera, D., et al. (2015). Development of a new measure for assessing insight: psychometric properties of the insight orientation scale (IOS). *Schizophr. Res.* 169, 298–302. doi: 10.1016/j.schres.2015.10.014

Grotstein, J. S. (1997). "Mens sane in Corpore Sano": the mind and body as an "odd couple" and as an oddly coupled unity. *Psychoanal. Inq.* 17, 204–222. doi: 10.1080/07351699709534120

Gurfinkel, V. S. (1994). The mechanisms of postural regulation in man. *Sov. Sci. Rev. F Phys. Gen. Biol.* 7, 59–89.

Hackford, J., Mackey, A., and Broadbent, E. (2019). The effects of walking posture on affective and physiological states during stress. *J. Behav. Ther. Exp. Psychiatry* 62, 80–87. doi: 10.1016/j.jbtep.2018.09.004

Herman, J. L. (1992). Complex PTSD: a syndrome in survivors of prolonged and repeated trauma. *J. Trauma Stress* 5, 377–391. doi: 10.1016/j.janxdis.2005.04.003

Hooper, D., Coughlan, J., and Mullen, M. R. (2008). Structural equation modelling: guidelines for determining model fit. *Electr. J. Bus. Res. Methods* 6, 53–60.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika* 30, 179–185. doi: 10.1007/BF02289447

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equa. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Huang, L., Galinsky, A. D., Gruenfeld, D. H., and Guillory, L. E. (2011). Powerful postures versus powerful roles: which is the proximate correlate of thought and behavior? *Psychol. Sci.* 22, 95–102. doi: 10.1177/0956797610391912

IBM Corp (2017). *IBM SPSS Statistics for Windows, Version 25. 0*. Armonk, NY: IBM Corp.

Jay, K., Jakobsen, M. D., Sundstrup, E., Skotte, J. H., Jørgensen, M. B., Andersen, C. H., et al. (2013). Effects of kettlebell training on postural coordination and jump performance: a randomized controlled trial. *J. Strength Condition. Res.* 27, 1202–1209. doi: 10.1519/JSC.0b013e318267a1aa

Kee, Y. H., Chatzisarantis, N. N., Kong, P. W., Chow, J. Y., and Chen, L. H. (2012). Mindfulness, movement control, and attentional focus strategies: effects of mindfulness on a postural balance task. *J. Sport Exerc. Psychol.* 34, 561–579. doi: 10.1123/jsep.34.5.561

Keltner, D., Gruenfeld, D. H., and Anderson, C. (2003). Power, approach, and inhibition. *Psychol. Rev.* 110:265. doi: 10.1037/0033-295x.110.2.265

Kerns, R. D., Turk, D. C., and Rudy, T. E. (1985). The west haven-yale multidimensional pain inventory (WHYMPI). *Pain* 23, 345–356. doi: 10.1016/0304-3959(85)90004-1

Kline, R. B. (2005). *Principles and Practice of Structural Equation Modeling*, 2nd Edn. New York, NY: Guilford.

Kwon, J., and Kim, S. Y. (2015). "The effect of posture on stress and self-esteem: comparing contractive and neutral postures," in *Proceedings of International Academic Conferences (No. 2705176)*, London: International Institute of Social and Economic Sciences, doi: 10.20472/IAC.2015.018.067

Lanningham-Foster, L. M., Jensen, T. B., McCrady, S. K., Nysse, L. J., Foster, R. C., and Levine, J. A. (2005). Laboratory measurement of posture allocation and physical activity in children. *Med. Sci. Sports Exerc.* 37, 1800–1805. doi: 10.1249/01.mss.0000175050.03506.bf

Lauche, R., Wayne, P. M., Fehr, J., Stumpe, C., Dobos, G., and Cramer, H. (2017). Does postural awareness contribute to exercise-induced improvements in neck pain intensity? A secondary analysis of a randomized controlled trial evaluating Tai Chi and neck exercises. *Spine* 42, 1195–1200. doi: 10.1097/BRS.0000000000002078

Littleton, H. L., Axsom, D., and Pury, C. L. (2005). Development of the body image concern inventory. *Behav. Res. Ther.* 43, 229–241. doi: 10.1016/j.brat.2003.12.006

Luca, M., Giannini, M., Gori, A., and Littleton, H. (2011). Measuring dysmorphic concern in Italy: psychometric properties of the Italian Body Image Concern Inventory (I-BICI). *Body Image* 8, 301–305. doi: 10.1016/j.bodyim.2011.04.007

MacLean, P. D. (1949). Psychosomatic disease and the" visceral brain"; recent developments bearing on the Papez theory of emotion. *Psychosom. Med.* 11, 338–353. doi: 10.1097/00006842-194911000-00003

Main, C. J. (1983). The modified somatic perception questionnaire (MSPQ). *J. Psychosom. Res.* 27, 503–514. doi: 10.1016/0022-3999(83)90040-5

Maner, J. K., Kaschak, M. P., and Jones, J. L. (2010). Social power and the advent of action. *Soc. Cogn.* 28, 122–132. doi: 10.1521/soco.2010.28.1.122

Marlatt, G. A., and Ostafin, B. D. (2005). "Being mindful of automaticity in addiction: a clinical perspective," in *Handbook of Implicit Cognition and Addiction*, eds R. W. Wiers and A. W. Stacy (Thousand Oaks, CA: Sage), 489–493.

Marsh, H. W., Hau, K. T., and Wen, Z. (2004). In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Struct. Equa. Model.* 11, 320–341. doi: 10.1207/s15328007sem1103_2

Massion, J. (1994). Postural control system. *Curr. Opin. Neurobiol.* 4, 877–887. doi: 10.1016/0959-4388(94)90137-6

Mattsson, M. (1998). *Body Awareness: Applications in Physiotherapy*. Sweden: Umeå Universitet.

Mehling, W. E., Gopisetty, V., Daubenmier, J., Price, C. J., Hecht, F. M., and Stewart, A. (2009). Body awareness: construct and self-report measures. *PLoS One* 4:e5614. doi: 10.1371/journal.pone.0005614

Meng, X. L., Rosenthal, R., and Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychol. Bull.* 111:172. doi: 10.1037/0033-2909.111.1.172

Michalak, J., Mischnat, J., and Teismann, T. (2014). Sitting posture makes a difference—embodiment effects on depressive memory bias. *Clin. Psychol. Psychother.* 21, 519–524. doi: 10.1002/cpp.1890

Moseley, G. L., Nicholas, M. K., and Hodges, P. W. (2004). A randomized controlled trial of intensive neurophysiology education in chronic low back pain. *Clin. J. Pain* 20, 324–330. doi: 10.1097/00002508-200409000-00007

Murphy, M. J., Mermelstein, L. C., Edwards, K. M., and Gidycz, C. A. (2012). The benefits of dispositional mindfulness in physical health: a longitudinal study of female college students. *J. Am. Coll. Health* 60, 341–348. doi: 10.1080/07448481.2011.629260

Nair, S., Sagar, M., Sollers, J. III, Consedine, N., and Broadbent, E. (2015). Do slumped and upright postures affect stress responses? A randomized trial. *Health Psychol.* 34:632. doi: 10.1037/hea0000146

O'connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behav. Res. Methods Instr. Comput.* 32, 396–402. doi: 10.3758/bf03200807

Ogden, P., Minton, K., and Pain, C. (2012). *Il Trauma E Il Corpo: Manuale Di Psicoterapia Sensomotoria*. Italy: Istituto di scienze cognitive.

Pluchino, A., Lee, S. Y., Asfour, S., Roos, B. A., and Signorile, J. F. (2012). Pilot study comparing changes in postural control after training using a video game balance board program and 2 standard activity-based balance intervention programs. *Arch. Phys. Med. Rehabil.* 93, 1138–1146. doi: 10.1016/j.apmr.2012.01.023

Pöhlmann, K., et al. (2014). The Dresden Body Image Inventory (DKB-35): validity in a clinical sample. *Psychother. Psychosom. Med. Psychol.* 64, 93–100. doi: 10.1055/s-0033-1351276

Prezza, M., Trombaccia, F. R., and Armento, L. (1997). La scala dell'autostima di Rosenberg: traduzione e validazione italiana. *Boll. Psicol. Appl.* 223, 35–44.

Price, C. J., and Thompson, E. A. (2007). Measuring dimensions of body connection: body awareness and bodily dissociation. *J. Alternat. Complement. Med.* 13, 945–953. doi: 10.1089/acm.2007.0537

Roberts, K. C., and Danoff-Burg, S. (2010). Mindfulness and health behaviors: is paying attention good for you? *J. Am. Coll. Health* 59, 165–173. doi: 10.1080/07448481.2010.484452

Roll, R., Kavounoudias, A., and Roll, J. P. (2002). Cutaneous afferents from human plantar sole contribute to body posture awareness. *Neuroreport* 13, 1957–1961. doi: 10.1097/00001756-200210280-00025

Rosberg, S. (2000). *Kropp, Varande Och Mening i Ett Sjukgymnastiskt Perspektiv*. Göteborg: Department of Social Work Institutionen för socialt arbete.

Rosenberg, M. (1965). Rosenberg self-esteem scale (RSE). Acceptance and commitment therapy. *Measures Package* 61:52.

Ruesch, J. (1948). The infantile personality; the core problem of psychosomatic medicine. *Psychosom. Med.* 10, 134–144. doi: 10.1097/00006842-194805000-00002

Schwarzer, R., and Jerusalem, M. (1995). Generalized self-efficacy scale. Measures in health psychology: a user's portfolio. *Causal Control Beliefs* 1, 35–37.

Shields, S. A., Mallory, M. E., and Simon, A. (1989). The body awareness questionnaire: reliability and validity. *J. Pers. Assess.* 53, 802–815. doi: 10.1207/s15327752jpa5304_16

Sibilia, L., Schwarzer, R., and Jerusalem, M. (1995). *Italian Adaptation of the General Self-Efficacy Scale. Resource Document*. Available online at: http://userpage.fu-berlin.de/~{}health/italian.htm (accessed October 17, 2019).

Skrinar, G. S., Bullen, B. A., Cheek, J. M., McArthur, J. W., and Vaughan, L. K. (1986). Effects of endurance training on body-consciousness in women. *Percept. Mot. Skills* 62, 483–490. doi: 10.2466/pms.1986.62.2.483

Solano, L. (2010). Some thoughts between body and mind in the light of Wilma Bucci's multiple code theory. *Int. J. Psychoanal.* 91, 1445–1464. doi: 10.1111/j.1745-8315.2010.00359.x

Stepper, S., and Strack, F. (1993). Proprioceptive determinants of emotional and nonemotional feelings. *J. Pers. Soc. Psychol.* 64:211. doi: 10.1037/0022-3514.64.2.211

Van Der Kolk, B. (2008). "Il disturbo traumatico dello sviluppo: verso una diagnosi razionale per bambini cronicamente traumatizzati," in *Trauma e Psicopatologia: un Approcio Evolutivo-Relazionale*, eds V. Caretti and G. Craparo (Rome: Astrolabio Ubaldini), 81–93.

Van Der Kolk, B. (2014). *Il Corpo Accusa il colpo. Mente, Corpo e Cervello Nell'elaborazione Delle Memorie Traumatiche*. Milano: Raffaello Cortina, 2015.

Van der Kolk, B. A. (2006). "Clinical implications of neuroscience research in PTSD," in *Neuroscience and Psychoanalysis*, eds G. Leo, D. Mann, G. Northoff, A. N. Schore, R. Skgold, B. A. Van Der Kolk, et al. (Italy: Frenis Zero Press), 159–196. doi: 10.1196/annals.1364.022

Van der Toorn, J., Feinberg, M., Jost, J. T., Kay, A. C., Tyler, T. R., Willer, R., et al. (2015). A sense of powerlessness fosters system justification: implications for the legitimation of authority, hierarchy, and government. *Politic. Psychol.* 36, 93–110. doi: 10.1111/pops.12183

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika* 41, 321–327. doi: 10.1007/BF02293557

Velicer, W. F., Eaton, C. A., and Fava, J. L. (2000). "Construct explication through factor or component analysis: a review and evaluation of alternative procedures for determining the number of factors or components," in *Problems and Solutions In Human Assessment: Honoring Douglas N. Jackson at Seventy*, eds R. D. Goffin and E. Helmes (Norwell, MA: Kluwer Academic), 41–77. doi: 10.1007/978-1-4615-4397-8_3

Veneziani, C. A., and Voci, A. (2015). The Italian adaptation of the Mindful Awareness Attention Scale and its relation with individual differences and quality of life indexes. *Mindfulness* 6, 373–381. doi: 10.1007/s12671-013-0270-2

Wallace, A. B. (2006). *The Nature of Mindfulness and Its Role in Buddhist Meditation: A Correspondence Between B. Alan Wallace and the Venerable Bhikkhu Bodhi*. Santa Barbara, CA: Santa Barbara Institute for Consciousness Studies.

Wayne, P. M., Krebs, D. E., Wolf, S. L., Gill-Body, K. M., Scarborough, D. M., McGibbon, C. A., et al. (2004). Can Tai Chi improve vestibulopathic postural control? *Arch. Phys. Med. Rehabil.* 85, 142–152. doi: 10.1016/S0003-9993(03)00652-X

Wong, W. Y., Wong, M. S., and Lo, K. H. (2007). Clinical applications of sensors for human posture and movement analysis: a review. *Prosthetic. Orthotics Int.* 31, 62–75. doi: 10.1080/03093640600983949

Wright, E. F., Domenech, M. A., and Fischer, J. R. (2000). Usefulness of posture training for patients with temporomandibular disorders. *J. Am. Dent. Assoc.* 131, 202–210. doi: 10.14219/jada.archive.2000.0148

Yamak, B., ımamoğlu, O., ımamoğlu, ı, and Çebi, M. (2018). The effects of exercise on body posture. *Electr. Turkish Stud.* 13, 1377–1388. doi: 10.7827/TurkishStudies.13911

Zeidan, F., and Vago, D. R. (2016). Mindfulness meditation–based pain relief: a mechanistic account. *Ann. N. Y. Acad. Sci.* 1373, 114–127. doi: 10.1111/nyas.13153

Zerbe, K. J. (1995). *The Body Betrayed: A Deeper Understanding of Women, Eating Disorders, and Treatment*. Carlsbad, CA: Gürze Books.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 20 May 2020
doi: 10.3389/fpsyg.2020.00911

Time Series Analysis in Forecasting Mental Addition and Summation Performance

*Anmar Abdul-Rahman**

Department of Ophthalmology, Counties Manukau DHB, Auckland, New Zealand

An ideal performance evaluation metric would be predictive, objective, easy to administer, estimate the variance in performance, and provide a confidence interval for the level of uncertainty. Time series forecasting may provide objective metrics for predictive performance in mental arithmetic. Addition and summation (addition combined with subtraction) using the Japanese Soroban computation system was undertaken over 60 days. The median calculation time in seconds for adding 10 sequential six digit numbers [CT_{Add}) was 63 s (interquartile range (IQR) = 12, range 48–127 s], while that for summation (CT_{Sum}) was 70 s (IQR = 14, range 53–108 s), and the difference between these times was statistically significant $p < 0.0001$. Using the mean absolute percentage error (MAPE) to measure forecast accuracy, the autoregressive integrated moving average (ARIMA) model predicted a further reduction in both CT_{Add} to a mean of 51.51 ± 13.21 s (AIC = 5403.13) with an error of 6.32%, and CT_{Sum} to a mean of 54.57 ± 15.37 s (AIC = 3852.61) with an error of 8.02% over an additional 100 forecasted trials. When the testing was repeated, the actual mean performance differed by 1.35 and 4.41 s for each of the tasks, respectively, from the ARIMA point forecast value. There was no difference between the ARIMA model and actual performance values (p-value $CT_{Add} = 1.0$, CT_{Sum}=0.054). This is in contrast to both Wright's model and linear regression (p-value < 0.0001). By accounting for both variability in performance over time and task difficulty, forecasting mental arithmetic performance may be possible using an ARIMA model, with an accuracy exceeding that of both Wright's model and univariate linear regression.

Keywords: ARIMA model, time series, mathematics, forecasting–methodology, cognitive performance

1. INTRODUCTION

Learning curves aim to model the gain in efficiency (increase in productivity, decrease in activity time, or both) of a repetitive task with increasing experience. The mathematical representation of the learning process is of particular interest across several disciplines including psychology (Mazur and Hastie, 1978; Balkenius and Morén, 1998; Glautier, 2013), medicine (Sutton et al., 1998; Ramsay et al., 2000; Dinçler et al., 2003; Hopper et al., 2007; Harrysson et al., 2014; Blehar et al., 2015), economics/industry (Cunningham, 1980; Lieberman, 1984; Badiru, 1991; Smunt and Watts, 2003) and more recently, artificial intelligence (Schmajuk and Zanutto, 1997; Perlich et al., 2003; Li et al., 2015).

OPEN ACCESS

Edited by:
Jason C. Immekus,
University of Louisville, United States

Reviewed by:
Raydonal Ospina,
Federal University of Pernambuco, Brazil
Alen Hajnal,
University of Southern Mississippi, United States

***Correspondence:**
Anmar Abdul-Rahman
anmar_rahman@hotmail.com

Specialty section:
This article was submitted to Quantitative Psychology and Measurement,
a section of the journal Frontiers in Psychology

Received: *16 January 2020*
Accepted: *14 April 2020*
Published: *20 May 2020*

Citation:
Abdul-Rahman A (2020) Time Series Analysis in Forecasting Mental Addition and Summation Performance. Front. Psychol. 11:911.
doi: 10.3389/fpsyg.2020.00911

Learning occurs most rapidly early in training, with equal increments in performance requiring a longer practice time in the later stages of the learning process. The classical understanding is that these diminishing returns result in learning curves that are smooth, decelerating functions (Mazur and Hastie, 1978; Jaber and Maurice, 2016). In 1880 Hermann Ebbinghaus first described the learning curve as a forgetting function; in a series of rigorous experiments he approximated the parameter as a negative exponential equation (Murre and Dros, 2015). In 1936 TP Wright investigated direct labor costs of assembling a particular aircraft and noted that the cost decreased with worker experience, a theory subsequently confirmed by other aircraft manufacturers (Wright, 1936). Analogous to Ebbinghaus's forgetting curve, he predicted the acquisition of skill followed a negative power function currently referred to as Wright's Model:

$$y_t = a \cdot x^b \tag{1}$$

Where y_t = the cumulative average time per unit, x = the cumulative number of units produced, a = the time to produce the first unit and b = learning coefficient (the slope of the function) ranging from −1 to 0; values close to −1 indicate a high learning rate and fast adaptation to task execution. Subsequently, JR Crawford described an incremental unit time model aimed at improving time representation in the algorithm, by substituting (x) in Wrights' model with the algebraic midpoint of the time required to produce a batch of units; this modification was a consequence of an observation that the time to complete a task decreased by a constant percentage, whenever the sum of the units doubled (Crawford, 1944; Jones, 2018). Three-parameter, two-parameter and the constant time exponential models were described to improve longterm predictions (Knecht, 1974; Mazur and Hastie, 1978). These algorithms were outperformed by multi-parameter hyperbolic models, where neutral, positive and negative learning episodes are represented through corresponding variable slope smooth curve profiles (Mazur and Hastie, 1978; Nembhard and Uzumeri, 2000; Shafer et al., 2001; Anzanello and Fogliatto, 2007). While the conventional univariate learning curves express a quantitative dependent variable in terms of an independent variable, multivariate models were eventually formulated to encode both qualitative and quantitative factors influencing the learning process (Badiru, 1992).

The smooth curves generated by these formal models provide an estimate at the average level of a set of observations. However, variation in performance demands a more rigorous representation of the learning process. This variation can be represented in a time series through two stochastic terms. an autoregressive term (AR), calculated as a weighted value from another point in the series, and a moving average (MA), which is estimated from the error terms in the series (Hyndman and Athanasopoulos, 2018). Characterization of time series using either an AR, MA, or combined ARMA processes was suggested independently in the 1920s by the Russian statistician and economist Eugen Slutsky (Slutzky, 1937), and the British statistician George Yule (Yule, 1921, 1926, 1927). It was not until the 1970s when Box and Jenkins described the autoregressive integrated moving average (ARIMA) model, which uses differencing of successive observations to render the series stationary, which is an essential property of the series for statistical validity (Milgate and Newman, 1990; Manuca and Savit, 1996). This study aims to investigate whether accounting for variance in the mental arithmetic using an ARIMA model is more accurate at forecasting performance, compared to Wright's model and univariate linear regression.

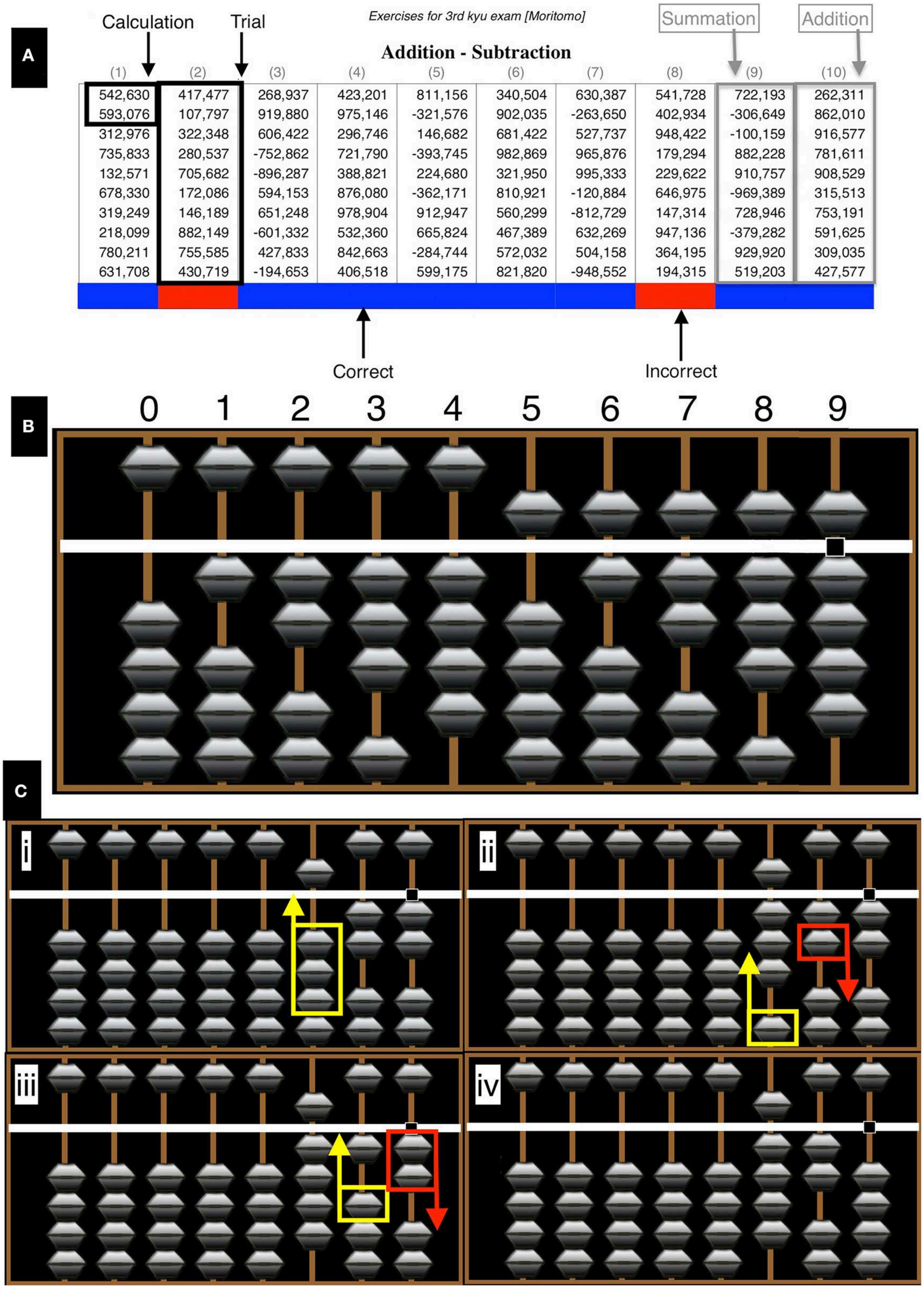
2. MATERIALS AND METHODS

2.1. Test Description

The learning period duration was 60 days, followed by 8 test days to assess the model forecasts. Tests were conducted between 7:00 and 7:30 a.m. Test sheets were randomly generated from the Soroban exam website (www.sorobanexam.org). Each sheet lays out both the questions and answers of a set of 10 columns of numbers (called a trial in this study). A sheet was composed of a mixture of six addition and four summation trials. The test difficulty was set to what is known in the Japanese Soroban exam system as difficulty level 3rd kyu, which consists of numbers ranging between 100,000 and 999,999. At the end of the test, a trial outcome was compared with the printed result and recorded. The last cell of the trial column was color coded with either a blue or red color, to indicate a successful or unsuccessful trial, respectively. Also, the time to complete a set of additions or summations was recorded in seconds. An example test sheet is provided in **Figure 1**. The in-built iOS voice over app (High Sierra 10.13.6) was used to vocalize the list of numbers from a .pdf list from the test sheet. Cumulative calculation time was defined as calculation time in seconds for adding 10 sequential 6-digit numbers, which either represented the addition task only (CT_{Add}) or a combination of addition and subtraction (CT_{Sum}). The author had limited prior experience with the Soroban (self-taught in 2017). Refer to the **Supplementary Material** section for the learning and test phase of the dataset.

2.2. The Soroban

The Soroban is a mechanical calculator, of which the origins are traced back to Mesopotamia, 2,500 years BC. Basic mathematical operations (addition, subtraction, multiplication, and division) can be performed using the device. There are two principles of operation: all calculations are performed as a number is pronounced, i.e., from the left to right. In addition, it reduces the complex mental mathematics to a simpler task, by using an algebraic principle of the method of complements, being in this case, either five or ten (Association, 1989; Schumer, 1999). Number representation and an example calculation is demonstrated in **Figure 1**. For clarification colored arrows are the next move in an operation (yellow = up, red = down). All computations are performed from left to right. Beads in contact with the central horizontal beam are considered in the final calculation. In this example of an addition operation (522+398), computation is started by representing the number 522 on the Soroban (**Figure 1C**i). The number (300) is then added to the hundreds rod (**Figure 1C**ii). Direct representation can take place with this step as there are an adequate number of beads not



Exercises for 3rd kyu exam [Moritomo]

Addition - Subtraction

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
542,630	417,477	268,937	423,201	811,156	340,504	630,387	541,728	722,193	262,311
593,076	107,797	919,880	975,146	-321,576	902,035	-263,650	402,934	-306,649	862,010
312,976	322,348	606,422	296,746	146,682	681,422	527,737	948,422	-100,159	916,577
735,833	280,537	-752,862	721,790	-393,745	982,869	965,876	179,294	882,228	781,611
132,571	705,682	-896,287	388,821	224,680	321,950	995,333	229,622	910,757	908,529
678,330	172,086	594,153	876,080	-362,171	810,921	-120,884	646,975	-969,389	315,513
319,249	146,189	651,248	978,904	912,947	560,299	-812,729	147,314	728,946	753,191
218,099	882,149	-601,332	532,360	665,824	467,389	632,269	947,136	-379,282	591,625
780,211	755,585	427,833	842,663	-284,744	572,032	504,158	364,195	929,920	309,035
631,708	430,719	-194,653	406,518	599,175	821,820	-948,552	194,315	519,203	427,577
🟦	🟥	🟦	🟦	🟦	🟦	🟦	🟥	🟦	🟦

FIGURE 1 | Test example. **(A)** The test consisted of 10 columns (trials) of 6-digit numbers labeled 1–10. There were six addition and four summation trials per test sheet. To provide a visual indicator of performance in each sheet, color coding at the last cell of each column, where a blue or red color was used to indicate a correct or incorrect result, respectively. **(B)** Digits from 0 to 9 are represented on each rod by adding the numerical value of all beads contacting the central horizontal beam. The lower beads have a numerical value of 1, whereas the single upper bead has a value of 5. **(C)** An example of an addition operation (522+398) showing the principles of number representation and complementary number calculations (details provided in the text). Images generated with abacus software http://www.komodousa.com.

contacting the central beam. Adding 90 to the tens rod is not possible directly therefore, the complementary technique is used, in this method $100 - 10 = 90$, a bead is added to the hundreds rod and another subtracted from the tens rod (**Figure 1C**iii). To add 8 to the ones rod the complementary technique again, where $10 - 2 = 8$, a bead is added to the tens rod and 2 subtracted from the ones rod giving a result of 920 (**Figure 1C**iv).

2.2.1. Time Series Model Description

An ARIMA time series model is mainly defined by three terms (p,d,q), which represent the autoregressive (p), integrative (d), and the moving average (q) parameters of the model, respectively. The general mathematical description of the model is provided below (Box et al., 2015):

$$\varphi(B)z_t = \phi(B)\nabla^d z_t = \theta_0 + \theta(B)a_t \tag{2}$$

where

$$\phi(B) = 1 - \phi_1(B) - \phi_2(B)^2 .... - \phi_p(B)^p \tag{3}$$

$$\theta(B) = 1 - \theta_1(B) - \theta_2(B)^2 .... - \theta_q(B)^q \tag{4}$$

1. (B) is the backward shift (lag) operator, which is defined by $B^k z_t = z_{t-k}$. This operator is convenient for describing the process of shifting between successive points in the series. That is to say B, operating on z_t, has the effect of shifting the data back one period.
2. $\phi(B)$ is the autoregressive polynomial operator in B of degree (p); it is assumed to be stationary, that is, the roots of $\phi(B) = 0$ lie outside the unit circle.
3. $\varphi(B) = \theta(B)\nabla^d$ is the generalized autoregressive (backward difference ∇z_t) operator; which is a non-stationary operator with d of the roots of $\varphi(B) = 0$ equal to unity, that is, d unit roots. The backwards difference operator is defined as $\nabla z_t = z_t - z_{(t-1)} = (1 - B)z_t$. Differencing is used to stabilize the series when the stationarity assumption is not met.
4. $\theta(B)$ is the moving average polynomial operator in B of degree (q); it is assumed to be invertible, that is, the roots of $\theta(B) = 0$ lie outside the unit circle.
5. The error term (a_t), which is assumed to have a Gaussian distribution, with a mean (μ) of zero and a constant variance of (σ_ϵ^2).

In practical terms, fitting the ARIMA model requires defining the model order (p,d,q). The autoregressive (ar) term, determines the value of (p), which is a datapoint in the series weighted by the value of proceeding data points. The term is given a number (ar_n); this represents the lag value in the series from where the correlation was calculated. The moving average (ma_n) corrects future forecasts based on errors made on recent forecasts; this term determines the (q) of the model order calculated from the partial autocorrelation function, which is a summary of the relationship between an observation in a time series with observations at prior time steps with the relationships of intervening observations removed. The integrated (d) portion of ARIMA models does not add predictors to the forecasting equation, rather, it indicates the order of differencing that has been applied to the time series to remove any trend in the data and render it stationary.

2.2.2. Statistical Analysis

Data was analyzed in R. The distributions of CT_Add and CT_Sum were modeled using the fitdistplus() package (Delignette-Muller and Dutang, 2015). Results are expressed as the median, range and interquartile range (IQR). Pearson's Chi-squared test (χ^2) with Yates' continuity correction was used to assess the differences in the accuracy of the calculated result between the addition and summation tasks. The Wilcoxon ranked sum test was used to assess the differences in CT_Add and CT_Sum.

Parameters of Wright's model were estimated using the learningCurve() package in R. This package uses Equation (1) in its calculations (Boehmke and Freels, 2017). The learning natural slope estimate (b) was calculated using the equation:

$$b < -\frac{Log_{10}T - Log_{10}t}{Log_{10}n - 1} \tag{5}$$

where T = total time (or cost) required to produce the first n units, t = time of all trials, n = total trials. The learning rate estimate (s) is calculated from the natural slope estimate by applying the following equation:

$$s = \frac{10^{b*log10(2)+2}}{100} \tag{6}$$

To forecast the 100th additional trial direct substitution in Equation (1) of (x) was done, where x = time for the 947th and 663rd attempt for the addition and summation tasks, respectively (a = the time for the first attempt in each of these tasks).

Univariate linear regression was utilized to assess the correlation between the time to perform the task and the number of trials and the equation of the best line fit was derived. The adjusted correlation coefficient (R^2) was used to represent the proportion of the variance explained by the model fits. The residual standard error (RSE) was used to assess model fit to the residuals.

The autoregressive integrated moving average model (ARIMA) was used for forecasting. The time series was plotted together with autocorrelation (acf) and partial autocorrelation functions (pacf). Although automated fitting of the time series (auto.arima) from the forecast package was initially used, acf and pacf graphs were used to confirm the order (p,q,d) of the series. After visual inspection of the time series plot suggested stationarity (mean, variance, and auto-covariance being independent from time), this assumption was confirmed by applying two statistical tests: the augmented Dickey-Fuller test (ADF), which is unit root test for stationarity, and The Kwiatkowski–Phillips–Schmidt–Shin test (KPSS). Unit roots (difference stationary process, i.e., a stochastic trend in a time series, sometimes called a "random walk with drift"), which exist in a time series if the value of α=1 in the general time series equation:

$$Y_t = \alpha Y_{t-1} + \beta X_e + \epsilon \tag{7}$$

The lag length (k) was chosen for this test (CT_{Add} k = 6, and CT_{Sum} k = 5) to avoid serial correlation of the residuals by choosing the last statistically significant lag, as determined by the partial autocorrelation function (pacf). The KPSS test was then applied, which is used for testing the null hypothesis that an observable time series is stationary around a deterministic trend (mean) or is non-stationary due to a unit root. Selection of the ARIMA model order (p,d,q) was chosen using the automated R auto.arima() command, which combines unit root tests, minimization of the corrected Akaike's Information Criterion (AICc) and Maximum likelihood estimation (MLE) to obtain an ARIMA model (Hyndman and Athanasopoulos, 2018). Validity of the model parameter choice was confirmed by plotting the autocorrelation (ACF) and partial autocorrelation (PACF) plots of the stationary data to determine a possible model candidate as suggested by the minimal AICc. Model fitting diagnostics also considered the lowest root mean square error (RMSE) and mean absolute percentage error (MAPE). A plot of the ACF of the residuals was done to confirm if the residuals appeared to be white noise. Once these criteria were met the forecast equations were calculated. The characteristic roots of both time series equations were plotted to assess whether the model is close to invertibility or stationarity in relation to the complex unit circle. Any roots close to the unit circle may be numerically unstable, and the corresponding model will not be suitable for forecasting. This possibility is mitigated through the auto.arima() function, which avoids selecting a model with roots close to the unit circle (Hyndman and Khandakar, 2008). Plotting the fitted model against the time series plot was performed. The models were tested for autoregressive conditional heteroscedasticity using the McLeod-Li test. Plotting the acf of the residuals and the Ljung-Box test were performed to assess for autocorrelations. In order to assess the model performance, the mean point forecast was reported from each model. In addition, forecasted data was generated from the model parameters and compared with actual test performance for an additional 100 trials using a pairwise-Wilcoxon test with Bonferroni correction. A p-value of <0.05 was considered statistically significant for all tests.

3. RESULTS

Over 60 days a total of 1,410 trials were conducted. The actual test time was 26.28 h during which a total of 847 addition and 563 summation trials were conducted. A variable number of trials, ranging from 0 to 70 trials per day were carried out. The distribution of both CT_{Add} and CT_{Sum} was non-normal and best fit a skewed exponential power type 4 distribution (model coefficients fit $p < 0.0001$). The skew and kurtosis of CT_{Add} were 1.38 and 7.46, respectively, whereas the corresponding values for CT_{Sum} were 0.80 and 3.40. The density distribution plot is demonstrated in **Figure 2**.

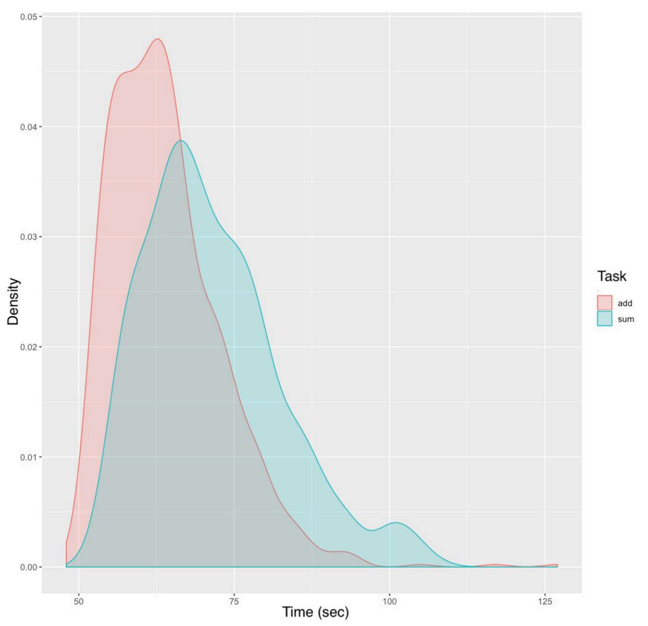


FIGURE 2 | Density plot of performance subset by task. The density plot uses kernel smoothing to reduce noise in the data, which generates a more defined distribution. The peaks of a density plot display represents values that are concentrated over the test interval. Noted in the plot are the narrower range of values and the shift to the left of the peak density for CT_{Add} compared to CT_{Sum}.

Addition tasks, being the simpler of the two, were more likely to yield an accurate result, and this difference compared to the outcome of the summation task was statistically significant ($\chi^2 = 9.33$, df = 1, $p < 0.002$). There was an increasing trend of total successful trials, as demonstrated in **Figure 3**. As expected, there was an improved performance with training, there were correct outcomes were recorded for 449/660 (68%) of trials in the first half compared to 598/750 (79.7%) in the second half of the learning period. The outcome of all tests classified by task type are summarized in **Table 1**.

Table 2 Summarizes the performance timing characteristics. From this table it can be noted that the median CT_{Add} was shorter compared to the median CT_{Sum}, this difference persisted throughout the learning period duration. Trials that concluded with an accurate calculated result (median time = 64 s, IQR = 13, range 48–117 s) were quicker compared to those where the result was incorrect (median time = 70 s, IQR = 15, range 51–127 s). These differences between addition, summation and performance times at the mid-learning interval were statistically significant ($p < 0.0001$). These results are shown in **Figure 3**.

3.1. Mathematical Models

3.1.1. Wright's Model

The time for the first event for CT_{Add} was 76 s, the total time 54,302 s and the number of trials 847, whereas for CT_{Sum} the first event was 74 s, the total time 40,297 s and the number of trials was 563. The learning rate was 0.98 and 0.99 for the addition and the summation tasks, respectively, this was calculated as a ratio of learning time at each doubling of the event i.e., time to event 1/time for event 2, time for event 2/time for event 4, time for event 4/time for event 8......etc. The natural slope is calculated by dividing the log of the learning rate by log2, this was further refined by calculating the natural slope estimate when the total number of trials, total time of all trials and the time for the first trial is known. Substituting in Equation (5), the natural slope

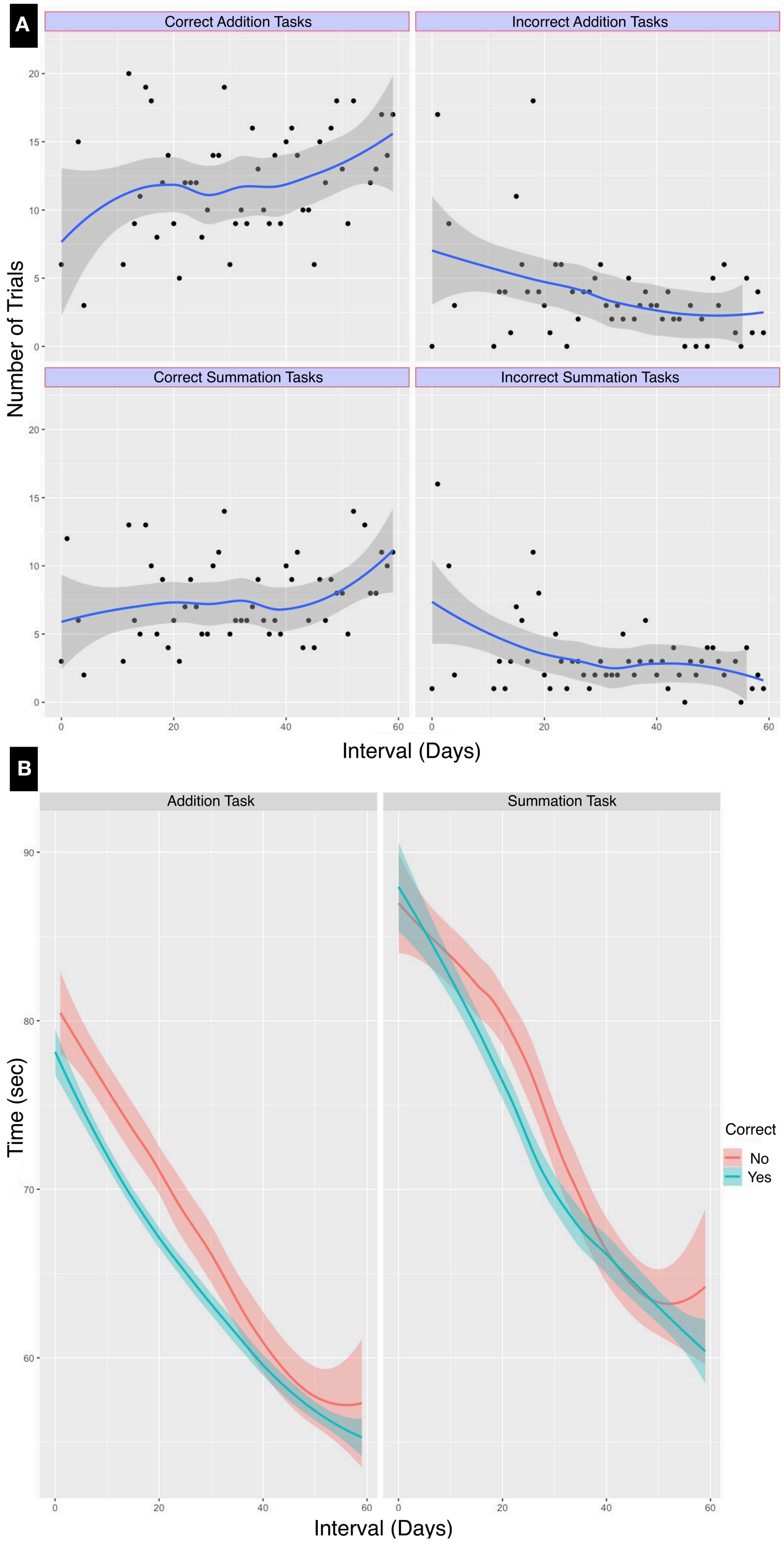


FIGURE 3 | Line graph of performance classified by task type and duration. Loess smoothed line graphs with 95% confidence intervals show the progress in test performance classified by the calculation result accuracy over the test interval of 60 days. There was both **(A)** an increase in the number of correct calculations and **(B)** a reduction of test time for addition and summation tasks over the learning period.

TABLE 1 | Learning outcome performance.

Trial	Correct	Incorrect	Total
LEARNING PERIOD (DAY 0–59)			
Addition	654 (r = 77%, c = 62%)	193 (r = 23%, c = 53%)	847 (60%)
Summation	393 (r = 70%, c = 38%)	170(r = 30%, c = 47%)	563 (40%)
Total	1047 (74%)	363 (26%)	1410
TESTING PERIOD (DAY 0–7)			
Addition	82 (r = 82%, c = 51.9%)	18 (r = 18.0%, c = 42.9%)	100 (50%)
Summation	76 (r = 76%, c = 48.1%)	24 (r = 24%, c = 57.1%)	100 (50%)
Total	158 (79%)	42 (21%)	

A comparison of the outcome of the trials during the learning and test periods. The learning period (60 days). The testing period (8 days) was to verify the accuracy of the model predictions, r = percentage by row, c = percentage by column.

TABLE 2 | The time to complete the tasks showed an expected reduction with learning.

			Range	
	Median	IQR	Min	Max
TOTAL LEARNING PERIOD				
CT_{Add}	63	12	48	127
CT_{Sum}	70	14	53	108
INITIAL LEARNING PERIOD (DAY 0–29)				
CT_{Add}	68	10	57	127
CT_{Sum}	78	14	59	108
LATE LEARNING PERIOD (DAY 30–59)				
CT_{Add}	58	7	48	88
CT_{Sum}	64	9	53	87
TEST PERIOD				
CT_{Add}	53	4	45	69
CT_{Sum}	58	6.25	47	75

The learning period was 60 days during which 1,410 trials were conducted, the performance modeled, and forecasting equations were used to generated expected timing data. The test period consisted of 100 additional trials.

estimate was −0.025 and −0.005 for the addition and summation tasks, respectively. The learning rate was further refined by taking into account the natural slope estimate by applying Equation (7), therefore the learning rate was estimated at 0.983 for the addition task and 0.996 for the summation task.

Substituting in Equation (1) where (x) is the forecasted performance time at the end of the 100th trial, CT_{Add} would be 62.24 s and CT_{Sum} would be 67.34 s. The plot of the model parameters is outlined in **Figure 4**.

3.1.2. Univariate Linear Regression

In the simplest case, when the variance in performance was disregarded, univariate linear regression was used to estimate the correlation between the cumulative time to perform both tasks and number of trials conducted. Consequently, the following equations were derived, where (x) is a variable representing the number of trials:

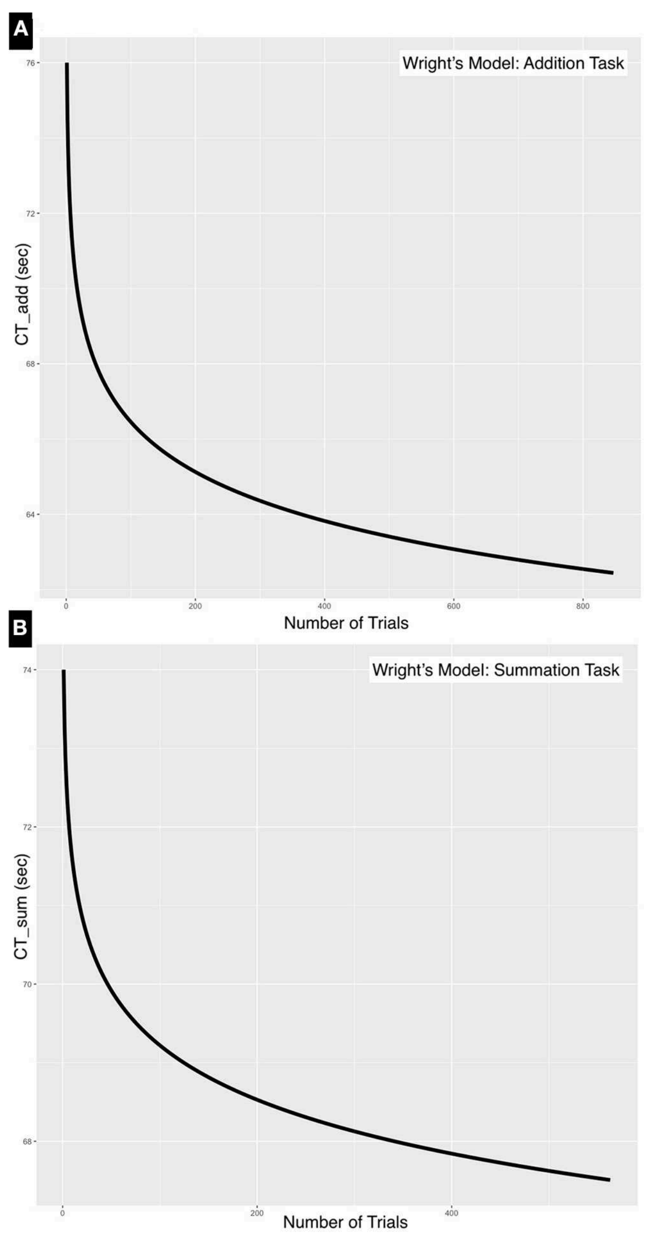


FIGURE 4 | Negative exponential curves for **(A)** addition and **(B)** summation tasks. These learning curves were generated using Wright's model. The natural slope estimate is steeper for CT_{Add} (−0.025) compared to that of CT_{Sum} (−0.005). This is a consequence of the simpler addition task reaching a plateau quicker than the more complex summation task.

$$CT_{Add} = -0.027(\pm 0.00085) \cdot x + 75.63(\pm 0.41) \tag{8}$$

$$CT_{Sum} = -0.051(\pm 0.0021) \cdot x + 85.63(\pm 0.64) \tag{9}$$

The intercepts of Equations (8) and (9), +75.63(±0.41) and +85.63(±0.64), respectively, indicate the values of CT_{Add} and CT_{Sum} at baseline (day 0) when commencing the test, and therefore represent the level of prior expertise with the task.

Both the negative slope of the regression line (**Figure 5**) and the negative (x) variable coefficient demonstrate a reduction of performance time with learning. The summation task resulted in a higher RSE (7.55, df = 561) compared to the addition task (6.02, df = 845) due to the lower deviation from the regression line as shown by the median of the residuals for summation (−1.46) compared to (−0.6) for the addition task. The model's predictor (number of trials) explained about half of the variance in the dependent variable (CT_{Add} and CT_{Sum}) as indicated by an adjusted R^2 of 0.55 and 0.54 for Equations (8) and (9), respectively. Statistical significance was achieved for all model coefficients ($p < 0.0001$). Substituting in these equations, the forecast for the for the 100th additional forecasted trial yields a mean time of $CT_{Add} = 50.06$ s and $CT_{Sum} = 51.82$ s.

3.1.3. Autoregressive Integrated Moving Average Model (ARIMA)

The time series was of sub-daily frequency ranging from 10 to 70 trials (median 30) per day. There were three discontinuities in testing: interval 2, intervals 5–10, and interval 53. The interval was calculated from day 0 at the commencement of the test. As shown in **Figure 6** there was an overall declining average for both the addition and the summation trials. The ADF test confirmed stationarity of the series for both CT_{Add} (ADF value = −6.86, $p < 0.01$) and CT_{Sum} (ADF value = −6.86, $p < 0.01$). However the KPSS test was statistically significant for both CT_{Add} (KPSS value = 10.23, $p < 0.01$) and CT_{Sum} (KPSS value = 6.81, $p < 0.01$), this result indicated that the time series had stationary autoregressive terms (ar) and non-stationary moving average terms (ma), which was consistent with the declining trend in the performance time for both tasks as shown in **Figure 3**, this analysis confirmed that the series was weakly stationary and that differencing using the ARIMA model was required to render the series stationary for further analysis.

The correlograms in **Figure 6** show, for both CT_{Add} and CT_{Sum} that the acf is highly correlated at all lag values up to lag 30; therefore the suggested q would be of order 1. The pacf plot is used to select the order of the p term. For the addition task the highest significant value was at lag 6, whereas the value for the summation task was at lag 5. Therefore, a custom ARIMA model would be (6, 1, 1), AIC = 5409.09 for addition and (5, 1, 1), AIC = 3863.65 for the summation task.

Software packages (like R) provide the option of an automated ARIMA model order approximation, when this was trialed for the series, an 18 and a 23-order permutation was tested for addition and summation, from both these approaches the model order (2, 1, 3) with drift for addition, where the AIC was 5403.13 and the order for summation was (5, 1, 4) with drift, where the AIC was 3852.61. Hence the automated approximation provided more favorable model parameters. The coefficients and accuracy criterion of the model are listed in **Table 3**.

Autoregressive conditional heteroscedasticity (ARCH) among the lags was considered, the Mcleod-Li test for the addition model ARCH effects are were absent, for the summation task, from a total of 30 lags there was minimal (13%) ARCH effects in lags 3–7.

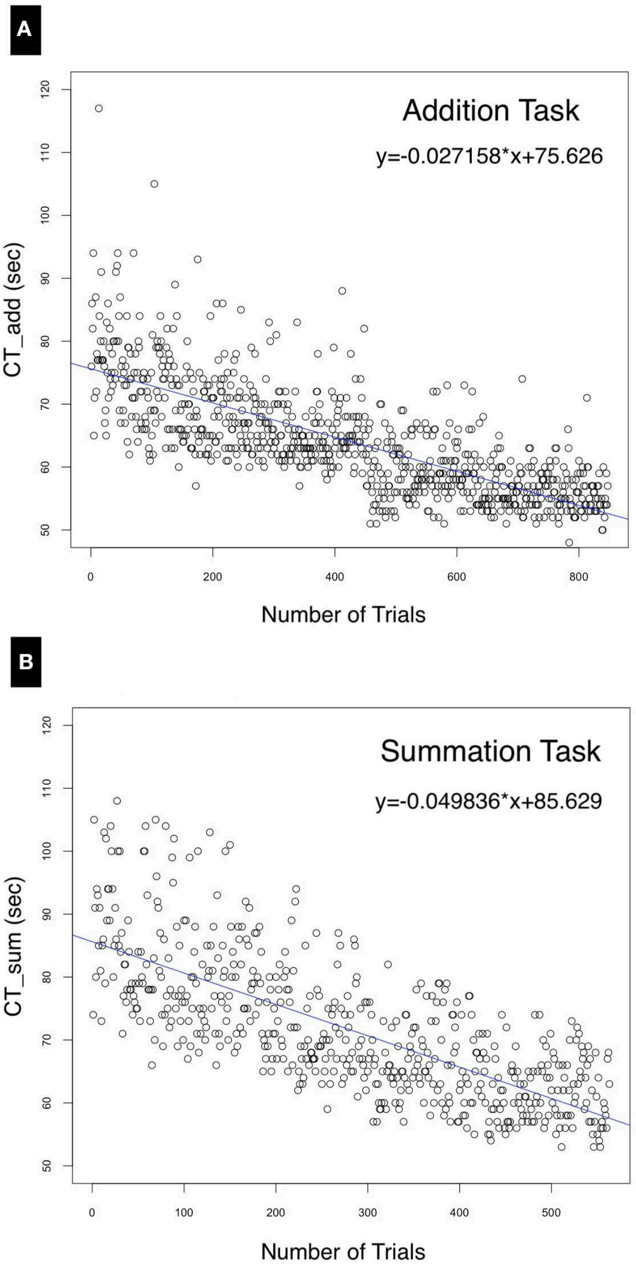


FIGURE 5 | Univariate linear regression model. correlating the performance time (seconds) with the number of trials for **(A)** the addition and **(B)** the summation trials. The distribution of residuals from the regression line for **(B)** is further from the regression line compared to **(A)** due to the wider variance in performance time for the summation task.

The following equations can be used to describe the time series fitted in **Figure 6** derived in standard notation:

$$CT_{Add}(Y_t) = -1.39Y_{(t-1)} - 0.86Y_{(t-2)} + 0.46e_{(t-1)} - 0.49e_{(t-2)} - 0.79e_{(t-3)} - 0.032e_t \quad (10)$$

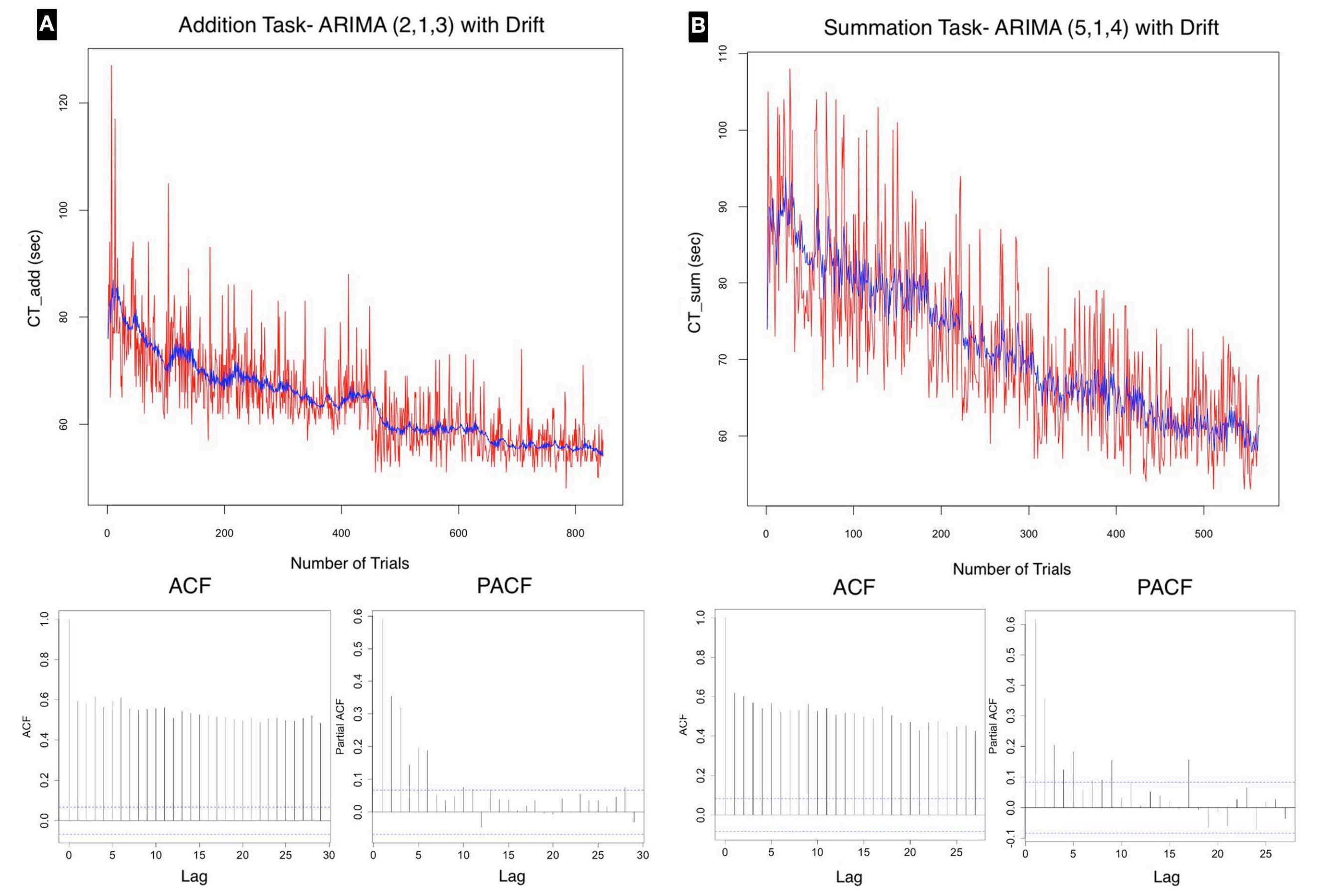


FIGURE 6 | Time series plot for **(A)** addition and **(B)** summation tasks with autocorrelation (acf) and partial autocorrelation (pacf) correlograms, which allow visual interpretation of the correlation of present data points with past points in the series. The time series for both tasks is demonstrating a declining trend. The acf plots for both tasks show a geometric pattern, with strong correlation of the sequential points up to lag 30. The pacf showed significant correlation at the 95% confidence interval up to lag 6 for addition, and lag 5 for the summation task. These tests confirm a gradual improvement in performance over the test time with a <20% correlation of test scores at each 6th trial in the series for addition and 5th trial for summation at the 95% confidence interval. A favorable model fit in the time series is demonstrated as a blue line in each of the series plots.

$$\begin{aligned} CT_{Sum}(Y_t) = &-1.81Y_{(t-1)} - 1.53Y_{(t-2)} - 0.41Y_{(t-3)} \\ &+ 0.31Y_t(t-4) + 0.092Y_{(t-5)} + 0.97e_{(t-1)} \\ &- 0.01e_{(t-2)} - 0.97e_{(t-3)} - 0.87e_{(t-4)} - 0.051e_t \quad (11) \end{aligned}$$

Where (Y) is the autoregressive term, (e) is the moving average term, (e_t) is the error term and (t-n) is the lag (time interval between two data points).

As shown in **Figure 7**. The mean forecasted performance for the 100th trial for CT_{Add} was 51.50 ± 13.21 and for CT_{Sum} was 54.57 ± 15.37. From **Table 3** using (MAPE) to measure forecast accuracy, the model was able to forecast with an error of 6.42% for the addition task and 8.02% for the summation task.

Independence of the residuals for both ARIMA models was evaluated using the acf plot of the residuals, which showed the absence of autocorrelation. This was confirmed by the Ljung-Box test. Parameters for the addition task ($\chi^2 = 759.99$, df = 800, p-value = 0.84) and the summation task ($\chi^2 = 480.93$, df = 500, p-value = 0.72) failed to achieve statistical significance, therefore an absence of serial autocorrelation in both series, thereby confirming an appropriate model fit. The three model comparisons on predicting the actual means on repeating the tests for a further 100 trials for each of the addition and the summation tasks are listed in **Table 4**.

The forecasted mean ARIMA model values offered a closer match with actual test performance (p-value $CT_{Add} = 1.0$, $CT_{Sum} = 0.054$), this in contrast to both Wright's model and univariate linear regression, for which mean values differed from these test (p-value < 0.0001) for both tasks. Simulated data for the three models for the forecasted period were compared using the paired Wilcoxon rank sum test, which showed no difference for the ARIMA model values from the actual test values (p-value CT_{Add} 1.0, CT_{Sum}=0.054), this is in contrast to both Wright's and the linear regression models which forecasted statistically significant (<0.0001) values for both tasks.

TABLE 3 | ARIMA model parameters.

	ar1	ar2	ma1	ma2	ma3	Drift
ADDITION TRIALS ARIMA (2,1,3) MODEL PARAMETERS						
Coefficients	−1.3976	−0.864	0.4623	−0.4916	−0.7876	−0.0321
se	0.0922	0.0763	0.1077	0.0936	0.0766	0.0115
p-value	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.005
RMSE	5.84					
MAPE	6.32					
AIC = 5403.13, AICc = 5403.27, BIC = 5436.32						

	ar1	ar2	ar3	ar4	ar5	ma1	ma2	ma3	ma4	Drift
SUMMATION TRIALS ARIMA (5,1,4) MODEL PARAMETERS										
Coefficients	−1.8077	−1.5259	−0.4061	0.2994	0.0916	0.9703	−0.0115	−0.9706	−0.8742	−0.0495
se	0.0859	0.1229	0.1278	0.099	0.0483	0.0749	0.0426	0.0518	0.0591	0.0089
p-value	< 0.0001	< 0.0001	< 0.001	< 0.002	< 0.06	< 0.0001	< 0.8	< 0.0001	< 0.0001	< 0.0001
RMSE	7.28									
MAPE	8.02									
AIC = 3852.61, AICc = 3853.09, BIC = 3900.26										

ARIMA model parameters for the addition and the summation tasks. ar, autoregressive coefficients; ma, moving average coefficients; se, standard error. RMSE, root mean square error; MAPE, mean absolute percentage error; AIC, akaike information criterion; AICc, corrected akaike information criterion; BIC, Bayseian information criterion.

4. DISCUSSION

The ARIMA model provided a more accurate approximation to actual performance after 100 additional trials, compared to both univariate linear regression and Wright's model. Considering the model means, in their predictions, the former overestimated and the latter underestimated the actual performance (**Table 1**). Many formal models of learning generate smooth learning curves, which are seldom observed except at the level of average data (Glautier, 2013). In this example both Wright's model and simple linear regression hide important information regarding performance variance. The ARIMA model predicted $\mathrm{CT}_{\mathrm{Add}}$ more accurately than that of $\mathrm{CT}_{\mathrm{Sum}}$, and this may be accounted for by the larger number of addition trials of which the test trials constituted 60% compared to the more complex summation task, where the test trials constituted approximately 40% of the total learning dataset. As shown in **Figure 2** the distribution of calculation times for both tasks were non-linear. In addition, the decline in both $\mathrm{CT}_{\mathrm{Add}}$ and $\mathrm{CT}_{\mathrm{Sum}}$ followed a non-linear trajectory over time (**Figure 3**). These patterns are consistent with the three phases of learning theory, which predict a three phase life cycle: the incipient phase during which a familiarization with the task occurs, which is characterized by a slow improvement; the learning phase, is where most of the improvement takes place; and the final phase, where the performance levels off (Carlson and Rowe, 1976). Whereas, prior knowledge of the task would have masked the incipient phase, the limitations of the univariate linear model become apparent by concealing the different performance phases altogether due to the constant slope of its regression line.

The neurophysiological basis of the Soroban remains unclear, however, it is known that computations using the Soroban involves a higher level of visual imagery (Tanaka et al., 2012). A longitudinal functional magnetic resonance imaging study of a patient with abacus-based acalculia suggested an important role in the parietal cortex and the dorsal premotor cortex in arithmetic ability of abacus users (Tanaka et al., 2012). Several cognitive processes required for mental arithmetic take place in these regions including retrieval, computation, reasoning, and decision making about arithmetic relations in addition to resolving interference between multiple competing solutions (interference resolution) (Menon, 2010). These factors may have played a role in the differences in variance in the performance of tasks as shown in **Figures 4–6**.

Calculation time for both addition and summation tasks as shown in **Figure 6** demonstrate a predictable downward trend and a slightly higher learning rate. The slope in the univariate linear regression was more negative for $\mathrm{CT}_{\mathrm{Sum}}$ than $\mathrm{CT}_{\mathrm{Add}}$, although the former was a more complex task. There may be some influence of the difference in the scale of comparison, as the number of trials for the summation tasks were less than the addition tasks by about 20%. A comparison of the ARIMA model pacf plot in **Figure 6** also suggests a slightly higher learning rate with summation compared to the addition task as indicated by the loss of correlation over shorter lags with the former task. In Wright's model the steeper natural slope estimate was approximately twice that for addition compared to summation, perhaps reflecting the higher complexity of the latter and a more gradual departure from the learning phase. As multiple neural systems and pathways involved in mathematical information processing mainly the parietal cortex, prefrontal cortex with several functional dissociations between brain regions differentially involved in specific operations such as addition, subtraction, and multiplication have been suggested in literature. Menon (2010), it is therefore difficult to speculate on the underlying structural reason behind the detected difference

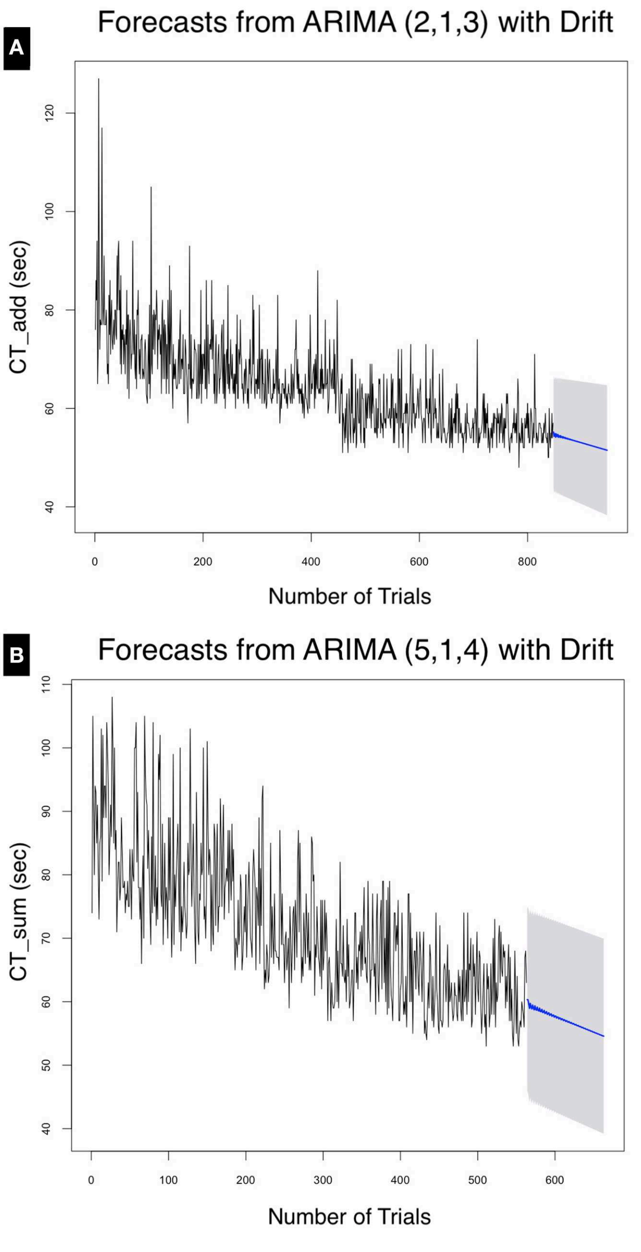


FIGURE 7 | Forecasts for **(A)** CT_{Add} and **(B)** CT_{Sum} at the 95% confidence interval for the next 100 trials. The point forecast for the 947th trial was calculated at 51.51 ± 13.21 s, whereas that for CT_{Sum} at the 663rd trial was calculated at 54.57 ± 15.36 s.

in learning speed. The other possibility behind the large variance seen in summation tasks is the measured difference in the number of complementary calculations per computation in each trial which was not considered in this experiment.

Inherent to the mathematical property of a times series analysis, is the capability of the model to capturing both linear and nonlinear relationships of the variables in the model. This property distinguishes it from other analysis methodologies which are either linear or non-linear (Yanovitzky and VanLear, 2008). In addition to describing the learning process in rigorous mathematical terms at an individual level, a psychological benefit is conferred to the test subject through accurate feedback of the improvement in performance. The limitations of this technique include the amount of data required to perform the analysis and the mathematical skill required to interpret the results. The protracted nature of the data collection requires a commitment in the testing process and may hinder some practicality as a routine test of learning performance. Although with the current experiment, the model fit was appropriate and delivered a high level of forecasting accuracy, most time series model predictions falter with extended forecast times due to non-stationarity, cohort effects, time-in-sample bias, and other challenges of longitudinal analyses (Taris, 2000; Yanovitzky and VanLear, 2008). Therefore, it is not clear from the current analysis how far into the future the forecast would be able to extend and retain its predictive accuracy. While the Soroban is still widely taught in Asian schools and therefore time series modeling may be beneficial for a more directed approach to teaching this skill, it may not apply to a wider population in other regions of the world where the use of the Soroban is less common. Future studies involving simultaneously recording an encephalogram may uncover wave activity associated with performance and the neurological basis of calculation errors in this task.

TABLE 4 | Mean point forecast.

	CT_{Add}	CT_{Sum}
POINT FORECAST COMPARISON (SEC)*		
Wright's model	62.24 (+9.38)	67.34 (+8.36)
Linear regression	50.06 (−2.74)	51.82 (−7.16)
ARIMA	51.50 (−1.36)	54.57 (−4.41)
Actual mean	52.86	58.98

*Comparison of the mean point forecasted value at the 100th trial for CT_{Add} and CT_{Sum}. *The number in brackets is the difference from the actual mean in seconds. The ARIMA model provided the closest prediction to actual performance.*

5. CONCLUSION

Time series analysis, by capturing the variance in performance may offer a more accurate mathematical representation of the learning process than classical learning theory models. The additional advantage of the ARIMA model to accurately forecast cognitive performance, with an accuracy exceeding that of both Wright's model and univariate linear regression, offers a potential for a wider applications for evaluation of cognitive function.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

AA-R designed the study, analyzed the data, and created the current version of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00911/full#supplementary-material

Data Sheet 1 | Learning period addition and summation times.

Data Sheet 2 | Testing period addition and summation times. For both **Supplementary Data Sheets**: rate, iOS accessibility program voice over speed; kyu, Calculation difficulty rate; Correct, correct answer calculated; time, test time in seconds; interval, Duration in days from start to the end of the learning period.

REFERENCES

Anzanello, M. J., and Fogliatto, F. S. (2007). Learning curve modelling of work assignment in mass customized assembly lines. *Int. J. Prod. Res.* 45, 2919–2938. doi: 10.1080/00207540600725010

Badiru, A. B. (1991). Manufacturing cost estimation: a multivariate learning curve approach. *J. Manuf. Syst.* 10, 431–441. doi: 10.1016/0278-6125(91)90001-I

Badiru, A. B. (1992). Computational survey of univariate and multivariate learning curve models. *IEEE Trans. Eng. Manage.* 39, 176–188. doi: 10.1109/17.141275

Balkenius, C., and Morén, J. (1998). "A computational model of emotional conditioning in the brain," in *Proceedings of Workshop on Grounding Emotions in Adaptive Systems* (Zurich).

Blehar, D. J., Barton, B., and Gaspari, R. J. (2015). Learning curves in emergency ultrasound education. *Acad. Emerg. Med.* 22, 574–582. doi: 10.1111/acem.12653

Boehmke, B., and Freels, J. (2017). learningcurve: an implementation of crawford's and wright's learning curve production functions. *J. Open Source Softw.* 2:202. doi: 10.21105/joss.00202

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control.* 5th Edn. Hoboken, NJ: John Wiley & Sons.

Carlson, J. G., and Rowe, A. J. (1976). How much does forgetting cost. *Ind. Eng.* 8, 40–47.

Crawford, J. R. (1944). *Learning Curve, Ship Curve, Ratios, Related Data.* Fort Worth, TX: Lockheed Aircraft Corporation.

Cunningham, J. A. (1980). Management: using the learning curve as a management tool: the learning curve can help in preparing cost reduction programs, pricing forecasts, and product development goals. *IEEE Spectr.* 17, 45–48. doi: 10.1109/MSPEC.1980.6330359

Delignette-Muller, M. L., and Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *J. Stat. Softw.* 64, 1–34. doi: 10.18637/jss.v064.i04

Dinçler, S., Koller, M. T., Steurer, J., Bachmann, L. M., Christen, D., and Buchmann, P. (2003). Multidimensional analysis of learning curves in laparoscopic sigmoid resection. *Dis. Colon Rectum* 46, 1371–1378. doi: 10.1007/s10350-004-6752-5

Glautier, S. (2013). Revisiting the learning curve (once again). *Front. Psychol.* 4:982. doi: 10.3389/fpsyg.2013.00982

Harrysson, I. J., Cook, J., Sirimanna, P., Feldman, L. S., Darzi, A., and Aggarwal, R. (2014). Systematic review of learning curves for minimally invasive abdominal surgery: a review of the methodology of data collection, depiction of outcomes, and statistical analysis. *Ann. Surg.* 260, 37–45. doi: 10.1097/SLA.0000000000000596

Hopper, A., Jamison, M., and Lewis, W. (2007). Learning curves in surgical practice. *Postgrad. Med. J.* 83, 777–779. doi: 10.1136/pgmj.2007.057190

Hyndman, R., and Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *J. Stat. Softw.* 27, 1–22. doi: 10.18637/jss.v027.i03

Hyndman, R. J., and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. Melbourne, VIC: OTexts.

Japan Abacus Association (1989). *Soroban*: *The Japan Chamber of Commerce and Industry, 2nd Edn*. Tokyo: Japan Abacus Association.

Jones, A. R. (2018). *Learning, Unlearning and Re-learning Curves*. 1st Edn. London: Routledge; Taylor and Francis Group.

Jaber, M. Y., and Maurice B. (2016). "Chapter 14: The lot sizing problem and the learning curve: a review," in *Learning Curves: Theory, Models, and Applications*, ed M. Y. Jaber (London: CRC Press), 265–291. doi: 10.1201/b10957-17

Knecht, G. (1974). Costing, technological growth and generalized learning curves. *J. Oper. Res. Soc.* 25, 487–491. doi: 10.1057/jors.1974.82

Li, N., Matsuda, N., Cohen, W. W., and Koedinger, K. R. (2015). Integrating representation learning and skill learning in a human-like intelligent agent. *Artif. Intell.* 219, 67–91. doi: 10.1016/j.artint.2014.11.002

Lieberman, M. B. (1984). The learning curve and pricing in the chemical processing industries. *RAND J. Econ.* 15, 213–228. doi: 10.2307/2555676

Manuca, R., and Savit, R. (1996). Stationarity and nonstationarity in time series analysis. *Phys. D* 99, 134–161. doi: 10.1016/S0167-2789(96)00139-X

Mazur, J. E., and Hastie, R. (1978). Learning as accumulation: a reexamination of the learning curve. *Psychol. Bull.* 85:1256. doi: 10.1037/0033-2909.85.6.1256

Menon, V. (2010). Developmental cognitive neuroscience of arithmetic: implications for learning and education. *ZDM: Int. J. Math. Educ.* 42, 515–525. doi: 10.1007/s11858-010-0242-0

Milgate, M., and Newman, P. (1990). "Chapter 3: Arima models," in *Time Series and Statistics*, ed J. Eatwell (London: Palgrave Macmillan), 22–35.

Murre, J. M., and Dros, J. (2015). Replication and analysis of ebbinghaus forgetting curve. *PLoS ONE* 10:e0120644. doi: 10.1371/journal.pone.0120644

Nembhard, D. A., and Uzumeri, M. V. (2000). Experiential learning and forgetting for manual and cognitive tasks. *Int. J. Ind. Ergon.* 25, 315–326. doi: 10.1016/S0169-8141(99)00021-9

Perlich, C., Provost, F., and Simonoff, J. S. (2003). Tree induction vs. logistic regression: a learning-curve analysis. *J. Mach. Learn. Res.* 4, 211–255. doi: 10.1162/153244304322972694

Ramsay, C. R., Grant, A. M., Wallace, S. A., Garthwaite, P. H., Monk, A. F., and Russell, I. T. (2000). Assessment of the learning curve in health technologies: a systematic review. *Int. J. Technol. Assess. Health Care* 16, 1095–1108. doi: 10.1017/S0266462300103149

Schmajuk, N. A., and Zanutto, B. S. (1997). Escape, avoidance, and imitation: a neural network approach. *Adapt. Behav.* 6, 63–129. doi: 10.1177/105971239700600103

Schumer, G. (1999). Mathematics education in japan. *J. Curric. Stud.* 31, 399–427. doi: 10.1080/002202799183061

Shafer, S. M., Nembhard, D. A., and Uzumeri, M. V. (2001). The effects of worker learning, forgetting, and heterogeneity on assembly line productivity. *Manage. Sci.* 47, 1639–1653. doi: 10.1287/mnsc.47.12.1639.10236

Slutzky, E. (1937). The summation of random causes as the source of cyclic processes. *Econometrica* 5, 105–146. doi: 10.2307/1907241

Smunt, T. L., and Watts, C. A. (2003). Improving operations planning with learning curves: overcoming the pitfalls of messyshop floor data. *J. Oper. Manage.* 21, 93–107. doi: 10.1016/S0272-6963(02)00088-8

Sutton, D., Wayman, J., and Griffin, S. (1998). Learning curve for oesophageal cancer surgery. *Br. J. Surg.* 85, 1399–1402. doi: 10.1046/j.1365-2168.1998.00962.x

Tanaka, S., Seki, K., Hanakawa, T., Harada, M., Sugawara, S. K., Sadato, N., et al.(2012). Abacus in the brain: a longitudinal functional mri study of a skilled abacus user with a right hemispheric lesion. *Front. Psychol.* 3:315. doi: 10.3389/fpsyg.2012.00315

Taris, T. W. (2000). *A Primer in Longitudinal Data Analysis*. 1st Edn. London: SAGE Publications Ltd.

Wright, T. P. (1936). Factors affecting the cost of airplanes. *J. Aeronaut. Sci.* 3, 122–128. doi: 10.2514/8.155

Yanovitzky, I., and VanLear, A. (2008). "Time series analysis: traditional and contemporary approaches," in *The Sage Sourcebook of Advanced Data Analysis Methods for Communication Research,* eds A. Hayes, M. Slater, and L. Snyder (London: SAGE Publications Ltd) 89–124. doi: 10.4135/9781452272054.n4

Yule, G. U. (1921). On the time-correlation problem, with especial reference to the variate-difference correlation method. *J. R. Stat. Soc.* 84, 497–537. doi: 10.2307/2341101

Yule, G. U. (1926). Why do we sometimes get nonsense-correlations between time-series?–a study in sampling and the nature of time-series. *J. R. Stat. Soc.* 89, 1–63. doi: 10.2307/2341482

Yule, G. U. (1927). VII. on a method of investigating periodicities disturbed series, with special reference to wolfer's sunspot numbers. *Philos. Trans. R. Soc. Lond. Ser. A* 226, 267–298. doi: 10.1098/rsta.1927.0007

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REVIEW
published: 27 May 2020
doi: 10.3389/fpsyg.2020.00917

The Drives for Driving Simulation: A Scientometric Analysis and a Selective Review of Reviews on Simulated Driving Research

Alessandro Oronzo Caffò[1], Luigi Tinella[1], Antonella Lopez[1], Giuseppina Spano[2], Ylenia Massaro[1], Andrea Lisi[1], Fabrizio Stasolla[3], Roberto Catanesi[4], Francesco Nardulli[5], Ignazio Grattagliano[1] and Andrea Bosco[1]*

[1] *Dipartimento di Scienze della Formazione, Psicologia, Comunicazione, Università degli Studi di Bari Aldo Moro, Bari, Italy,* [2] *Department of Agricultural and Environmental Science, Faculty of Agricultural Science, University of Bari Aldo Moro, Bari, Italy,* [3] *Giustino Fortunato University, Benevento, Italy,* [4] *Department of Interdisciplinary Medicine, School of Medicine, University of Bari Aldo Moro, Bari, Italy,* [5] *Commissione Medica Locale Patenti Speciali, Azienda Sanitaria Locale, Bari, Italy*

OPEN ACCESS

Edited by:
*Pietro Cipresso,
Italian Auxological Institute (IRCCS),
Italy*

Reviewed by:
*Leonardo Carlucci,
Università degli Studi "G. d'Annunzio"
Chieti – Pescara, Italy
Zaida Chinchilla-Rodríguez,
Consejo Superior de Investigaciones
Científicas (CSIC), Spain*

***Correspondence:**
*Alessandro Oronzo Caffò
alessandro.caffo@uniba.it*

Specialty section:
*This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology*

Received: *30 November 2019*
Accepted: *14 April 2020*
Published: *27 May 2020*

Citation:
*Caffò AO, Tinella L, Lopez A,
Spano G, Massaro Y, Lisi A,
Stasolla F, Catanesi R, Nardulli F,
Grattagliano I and Bosco A (2020)
The Drives for Driving Simulation:
A Scientometric Analysis
and a Selective Review of Reviews on
Simulated Driving Research.
Front. Psychol. 11:917.
doi: 10.3389/fpsyg.2020.00917*

Driving behaviors and fitness to drive have been assessed over time using different tools: standardized neuropsychological, on-road and driving simulation testing. Nowadays, the great variability of topics related to driving simulation has elicited a high number of reviews. The present work aims to perform a scientometric analysis on driving simulation reviews and to propose a selective review of reviews focusing on relevant aspects related to validity and fidelity. A scientometric analysis of driving simulation reviews published from 1988 to 2019 was conducted. Bibliographic data from 298 reviews were extracted from Scopus and WoS. Performance analysis was conducted to investigate most prolific Countries, Journals, Institutes and Authors. A cluster analysis on authors' keywords was performed to identify relevant associations between different research topics. Based on the reviews extracted from cluster analysis, a selective review of reviews was conducted to answer questions regarding validity, fidelity and critical issues. United States and Germany are the first two Countries for number of driving simulation reviews. United States is the leading Country with 5 Institutes in the top-ten. Top Authors wrote from 3 to 7 reviews each and belong to Institutes located in North America and Europe. Cluster analysis identified three clusters and eight keywords. The selective review of reviews showed a substantial agreement for supporting validity of driving simulation with respect to neuropsychological and on-road testing, while for fidelity with respect to real-world driving experience a blurred representation emerged. The most relevant critical issues were the a) lack of a common set of standards, b) phenomenon of simulation sickness, c) need for psychometric properties, lack of studies investigating d) predictive validity with respect to collision rates and e) ecological validity. Driving simulation represents a cross-cutting topic in scientific literature on driving, and there are several evidences for considering it as a valid alternative to neuropsychological and on-road testing. Further research efforts could be aimed at establishing a consensus statement for protocols assessing fitness to drive, in order to (a) use standardized systems, (b) compare systematically driving simulators with regard to their validity and fidelity, and (c) employ shared criteria for conducting studies in a given sub-topic.

Keywords: driving simulation, driving simulator, fitness to drive, scientometric analysis, review

INTRODUCTION

Driving is a multifaceted activity involving cognitive and physical tasks. It requires the integration of visual and perceptual stimuli, information processing, decision making, vehicle control responses, motor programming and execution, and the capability of carefully responding to a dynamic environment (Heikkilä et al., 1998; Anstey et al., 2005; Classen et al., 2006; Ranchet et al., 2011).

In order to measure driving behavior and to assess fitness to drive, researchers have been using different assessment tools over time. The gold standard seems to be the on-road assessment of actual driving performance. This kind of evaluation is considered costly, stressful, and time-consuming; furthermore, it is very difficult to evaluate the driving performance in different conditions, such as in heavy traffic, at night, in various types of weather, or in dangerous circumstances (i.e., collision avoidance, obstacles on the road). Moreover, testers often experience anxiety and stress, and experimenters do not completely manage to control all variables, such as errors and violations (e.g., Brown and Ott, 2004; Kraft et al., 2010).

Neuropsychological evaluation by means of psychometric tests is also used to evaluate driving behavior as well as fitness to drive. The underlying assumption is that significant cognitive impairments should prevent safe operation of a motor vehicle (e.g., Kraft et al., 2010). The most widely appraised cognitive domains are visual perception (e.g., contrast sensitivity; Uc et al., 2006a; Worringham et al., 2006), visual attention (e.g., Uc et al., 2006a, 2007), visual and verbal memory (e.g., Heikkilä et al., 1998; Radford et al., 2004; Uc et al., 2007), information processing (e.g., Heikkilä et al., 1998; Worringham et al., 2006), motor dexterity (e.g., Radford et al., 2004; Grace et al., 2005), executive functioning (Stolwyk et al., 2006; Uc et al., 2006b), and visuospatial organization and planning (e.g., Grace et al., 2005; Uc et al., 2007). Neuropsychological tests' performance can predict driving ability, but initial evidences suggested that neuropsychological screening batteries explained less than 70% of the variance in driving ability and correctly classified about 70% of participants (e.g., see Heikkilä et al., 1998; Radford et al., 2004; Worringham et al., 2006; Amick et al., 2007; Devos et al., 2007). More recently, Verster and Roth (2012) showed that psychometric test batteries predicted on-road test performance at only 33.4%, showing that combinations of basic neuropsychological/psychometric tests are not always good predictors of driving performance. The screening batteries considered most reliable, with sensitivity and specificity ranging between 61 and 94%, included the Trail Making Test (TMT), the Useful Field of View (UFOV), the Pelli–Robson contrast sensitivity test, and the Symbol Digit Modalities Test (SDMT) (e.g., Jacobs et al., 2017). All together, these findings compel researchers to shed further light on the role of neuropsychological tests in predicting fitness to drive.

The use of a driving simulator is another widespread method for assessing fitness to drive (Shechtman, 2010). It provides the opportunity to test many challenging/hazardous conditions or events that may not be presented during on-road testing in a standardized setting. Moreover, a lot of advantages contribute to make this approach a promising alternative to both neuropsychological and on-road testing for a safe assessment procedure as well as for cost cutting, time efficiency, and reliability (Lew et al., 2005; de Winter et al., 2009; Shechtman et al., 2009; Mayhew et al., 2011). Additionally, a large amount of data could be collected, capturing several variables and measures. On the other hand, the main limitations of driving simulation seem to be: (a) the difficulty to compare research findings adopting different driving simulators because of how parameters are collected and how driving simulator performance is quantified (e.g., Jacobs et al., 2017) and (b) sickness, dizziness, nausea, vomiting, and sweating associated with simulations (Brooks et al., 2010; Domeyer et al., 2013).

A considerable amount of literature on driving simulation has been produced since 1970. The rapid and continuous advancements of technology in the last 50 years have allowed for a massive development and employment of driving simulators. A recent bibliometric analysis (Guo et al., 2019) has explored the paths through which literature on simulated driving has evolved in the last 20 years. Authors filtered out 3,766 documents published from 1997 to 2016 and performed several bibliometric computations. The Countries which contributed and collaborated most in publishing studies on simulated driving were the United States followed by Germany and China. The most productive institutes were located in Netherlands and in the United States. The most recognized journals were in transportation and ergonomics, and the most productive authors were "J. D. Lee," "D. L. Fisher," "J. H. Kim," and "K. A. Brookhuis." A co-citation analysis was also performed showing different trends in topic over time—from early works on task-induced stress, drivers with neurological disorders, alertness and sleepiness, driving assistance systems, driver distraction in the first 10 years to the effect of drug use on driving behavior, the validity of driving simulators, and automated driving in more recent years.

Regarding the latter point highlighted by Guo et al. (2019), in a recent literature review, Wynne et al. (2019) pointed out the poor consistency among measures employed to assess the simulated performance and on-road driving. Several studies do not report all the employed measures to assess simulated driving and/or do not provide a direct comparison with measures assessing driving performance in the real world. Authors claimed that these results suggest the lack of a common research practice. Indeed, evidences of validity on one measure in one simulator do not mean that other measures may be equally valid in the same simulator, or that the same measures can be considered valid in other simulators. Furthermore, each setup is unique even when modeled on previously validated simulators and may be validated in light of those uniqueness (Pinto et al., 2008). Thus, simulated driving cannot be considered a universally valid measure of on-road driving performance (George, 2003; Wynne et al., 2019).

Similarly, a lot of studies have been devoted to investigating the predictive validity of cognitive performance measured with paper-and-pencil neuropsychological tests on simulated driving performance. Several reviews in the last 20 years aimed to summarize the results provided from primary studies putting together specific cognitive tasks or tests able to predict both simulated and real driving measures (Reger et al., 2004; Mathias and Lucas, 2009). Despite the above, there seems to be no clear evidence supporting the validity of driving simulation measures compared to neuropsychological testing ones in the assessment of fitness to drive (Marcotte and Scott, 2004; Mathias and Lucas, 2009).

The driving simulators are widespread employed in research in several disciplines and for different aims. Moreover, they are widely used to assess driving performance and driving behavior in several populations (Marcotte and Scott, 2004; Wynne et al., 2019). Evidences of validity on simulated driving measures observed in specific populations are not representative of all populations. This issue increases the controversy in literature regarding driving simulators' validation due to differences between studies also in special populations (Shechtman, 2010), and thus concerns remain regarding their employability. Taken together, the evidences from primary studies provide a framework of puzzling and blurred results which may prevent generalizable conclusions about the validity of driving simulators with respect to neuropsychological testing and on-road performance.

This variability among primary studies has elicited a high number of secondary studies. In order to gain a comprehensive picture of secondary studies and a "state-of-the-art" snapshot of the domain, a scientometric analysis was conducted exclusively on driving simulation reviews. The choice to filter only secondary studies will allow to have a sort of *second-order analysis* on the topic as well as an overview on the different uses of driving simulators across research fields and several academic disciplines.

The present work has two aims: 1) to perform a scientometric analysis on the corpus of reviews on driving simulation studies conducted in the last 30 years, i.e., from January 1, 1988, to July 1, 2019, and 2) to propose a selective review of reviews of the main clusters emerged from the scientometric analysis, with a special focus on psychometric issues related to validity of driving simulators compared to standardized neuropsychological and on-road testing as well as to fidelity with respect to real-world driving experience. Reviews may provide an overview of primary studies on a certain topic, thus highlighting similarities and differences among the findings of the studies included. While contemplating the extensive variability of the results provided in primary studies for each review, the review of reviews is aimed at better understanding the effectiveness of driving simulator in predicting measures of fitness to drive, with respect to both neuropsychological and on-road testing.

The scientometric analysis and the review focused on secondary studies could summarize more clearly whether the driving simulator is a useful and effective tool for the assessment of the fitness to drive, specifying in which discipline or population this happens, in a reliable manner. A comprehensive overview given by secondary studies could be also useful in order to highlight critical issues related to the effectiveness of driving simulators.

SCIENTOMETRIC ANALYSIS

Materials and Methods

Data Collection

The great variety of disciplines interested in the topic of driving simulation and the not perfect overlap in search results on scientific databases has required to proceed with a search on two databases, thus improving the likelihood to carry out a fully exhaustive work (e.g., Meneghini et al., 2006; Pollack and Adler, 2015). Consequently, a literature search was conducted on July 1, 2019, on two databases, Scopus and Web of Science (WoS). The former is the largest abstract and citation database of peer-reviewed research literature in the fields of science, technology, medicine, social sciences, and arts and humanities. The latter is composed of several citation indexes for different disciplines, from social sciences to engineering to chemical sciences *et cetera*.

The search expression used for data collection was "Driv* Simulat*" OR "Simulat* Driv*" in the "title, abstract, keywords" search in Scopus database and in "Topic" search in WoS, which comprises title, abstract, author, keywords, and Keywords Plus. Scopus search returned 15,518 records, and WoS search returned 10,379 records. Such results were refined selecting "Review" in the field "Document type" of each database. There were 228 documents classified as reviews in Scopus and 151 in WoS. Two datasets containing several information for each record, such as abstract and keywords, bibliographical information, citation information, funding details, and the list of references, were exported in BibTeX format. Subsequently, they were converted into *dataframes* using bibliometrix R package (Aria and Cuccurullo, 2017) and merged together. After deleting duplicates, the final sample was composed of 298 records.

Data Analysis

Bibliometrics, scientometrics, and infometrics are methodological and quantitative approaches in which the scientific literature itself becomes the subject of analysis. Although their historical origins differ and they are not necessarily synonymous (Hood and Wilson, 2001), nonetheless, they share theories, methods, technologies, and applications. Their main aim is to measure the evolution of a scientific domain, the impact of scholarly publications, and the process of scientific knowledge production, and they often comprehend the monitoring of research in a given field, the assessment of the scientific contribution of authors, journals, or specific articles, as well as the analysis of the dissemination process of scientific knowledge (Mao et al., 2015).

Several tools and software have been developed and proposed in order to perform scientometric analysis, among the most known there are BibExcel, Bibliometrix R Package, CiteSpace, VOSviewer, *et cetera*. For the present work, two of them were used, namely, Bibliometrix R Package (Aria and Cuccurullo, 2017) and VOSviewer (Van Eck et al., 2010). Bibliometrix R Package is an open-source tool for quantitative research in scientometrics and bibliometrics that includes all the main bibliometric methods of analysis. VOSviewer is an open-source software tool for constructing and visualizing, among other functionalities, bibliometric networks of relevant information extracted from a body of scientific literature.

For the present study, a particular focus has been given to performance analysis, i.e., the statistical analysis based on publication outputs and received citation to gauge the research performance and also the leadership of Institutes, Departments, Journal or Persons (Noyons et al., 1999; Van Raan, 2004; Cantos-Mateos et al., 2012; Muñoz-Écija et al., 2017; Vargas-Quesada et al., 2017). Performance of Countries, Journals, Institutes, and Authors which published reviews on simulated driving was analyzed in order to highlight research contents and trends associated with such topic. Cluster analysis based on authors Keywords Co-occurrence Network (KCN) was employed in order to conceptualize the deep structure of the research field and its trends throughout different disciplines and methodologies.

Results

Performance of Countries

Table 1 shows the number of reviews on driving simulation studies by Country, the number of single and multiple Country publications (SCP and MCP, respectively), and the Relative International Collaboration Rate (RICR; Elango et al., 2015) for the 10 most productive Countries. It can be noted that those Countries have produced together almost the 70% of all the reviews, with a high prominence of the United States, followed by the most industrialized Countries all over the world. The number of SCP and MCP together with the RICR may provide a measure of the degree of collaboration between different Countries. Australia, Netherlands, Canada, United Kingdom, and China showed an international collaboration rate equal or greater than the global rate (= 1), while the other five Countries showed an international collaboration rate lesser than the global rate. **Supplementary Table S1** contains the number of reviews, the number of SCP and MCP, and the RICR for all the Countries present in the dataframe.

Figure 1 shows the number of reviews by Country and by year, from 1988 to 2019, for the first four most productive Countries, and the total number of reviews in the same years range. A visual inspection of the graph shows that the total trend is mimicked by that of the United States and only in part by that of Germany, which are the first two most productive Countries. It also emerges that after the year 2000, there has been a strong increase in the number of reviews, followed by a substantial drop in the year 2007 and a constant recovery in subsequent years, with high levels of interest in the last 5 years. As Guo et al. (2019) stated in their recent scientometric analysis on primary studies, the rapid advancement of technological tools applied to driving simulation in the last 20 years promoted thousands of studies and conversely a high interest for summarizing their findings.

TABLE 1 | Number of reviews on driving simulation studies, single and multiple Country publications, and Relative International Collaboration Rate for the 10 most productive Countries.

Country	Number of reviews	SCP	MCP	RICR
United States	87	55	32	0.85
Germany	31	23	8	0.60
Canada	20	10	10	1.16
France	15	10	5	0.77
United Kingdom	14	7	7	1.16
Australia	12	1	11	2.13
Netherlands	8	3	5	1.45
China	7	4	3	0.99
Poland	6	4	2	0.77
Switzerland	6	4	2	0.77

SCP, Single Country Publications; MCP, Multiple Country Publications; RICR, Relative International Collaboration Rate.

Performance of Journals

Table 2 shows the number of reviews on driving simulation studies by the top 10 journals, as well as the total global citation score (TGCS), which refers to the number of times the document has been cited in the scientific databases used for retrieval. The number of citations was a piece of information present in the bibliographic record for each review, no matter where it came from (Scopus or WoS). The software we used for obtaining the performance of journals simply summed up the citations of the reviews published on each journal. The 10 most productive journals on a total of 223 journals accounted for 18.79% of the total 298 reviews. Surprisingly, it can be noted that only two journals in the top 10 belong to the transportation field (i.e., *Transportation Research Record* and *Traffic Injury Prevention*) with a relatively low citation score, while journals in the ergonomics and human factors field have a high citation score. In the first place *ex aequo* with another journal, there is a journal devoted to sleep medicine (i.e., *Sleep Medicine Reviews*) with a high number of global citations, as a demonstration of the strong interest for the relationship between sleep-related disorders and simulated driving. Two national journals (i.e., *VDI Berichte* and *Medycyna Pracy*) are also present, with a scarce number of citations. A decision was made not to exclude them in the retrieval phase in order to have a broader representation of the topic. Indeed, 258 reviews were published entirely in English language, and 40 were published in a double language, i.e., abstract in English and text in another language, or entirely in another language, nonetheless, all of them were indexed in Scopus or in WoS. Moreover, several journals with two (seven journals) or even only one (16 journals) review published obtained a good or very good performance, having more than 100 global citations. **Supplementary Table S2** contains the number of reviews and the global number of citations for all the Countries present in the dataframe.

Performance of Institutes

Table 3 shows the number of reviews on driving simulation studies by the top 10 Institutes of the first author, by Country, as well as total global citation scores, and total citations per year. The most productive Institutes are located in the United States; University of Florida, Yale University, University of Iowa, University of Massachusetts, and University of Michigan have the highest global citations, as well as the highest total citations per year. Other productive Institutes are distributed worldwide

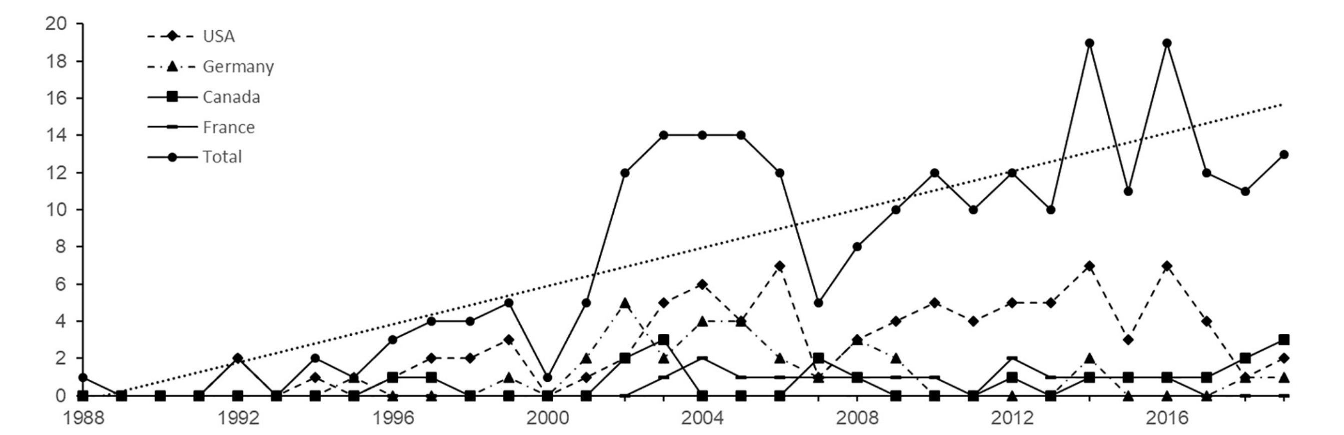


FIGURE 1 | Number of reviews by Country and by year for the first four most productive Countries and total number of reviews by year from 1988 to 2019.

TABLE 2 | Number of reviews on driving simulation studies by Source and total global citation scores.

Source	Number of reviews	TGCS
Sleep Medicine Reviews	12	637
VDI Berichte	12	9
Transportation Research Record	6	27
Human Factors	4	897
American Journal of Occupational Therapy	4	103
Traffic Injury Prevention	4	47
Medycyna Pracy	4	14
International Journal of RF and Microwave Computer-Aided Engineering	4	7
Ergonomics	3	214
Frontiers in Psychology	3	46

TGCS, Total Global Citation Score.

TABLE 3 | Number of reviews on driving simulation studies by Institute, Country, total global citation scores, and total citations per year.

Institute	Country	Number of reviews	TGCS	TCpY
University of Florida	United States	8	200	23.95
Yale University	United States	5	307	13.69
University of Iowa	United States	5	147	13.80
Nofer Institute of Occupational Medicine	Poland	5	23	4.11
Utrecht University	Netherlands	4	165	11.97
Reykjavik University	Iceland	4	7	1.71
University of Massachusetts	United States	3	129	9.84
University of Toronto	Canada	3	112	8.58
University of Western Ontario	Canada	3	80	5.00
University of Michigan	United States	3	43	4.50

TGCS, Total Global Citation Score; TCpY, Total Citations per Year.

between Europe, i.e., in Poland, Iceland and in the Netherlands, and Canada. Utrecht University and University of Toronto obtained a comparable high number of total citations and citations per year. Nofer Institute of Occupational Medicine and Reykjavik University obtained a lower number of total citations and citations per year. It is noteworthy that there are a number of Institutes which published one or two reviews that have reached a high or very high number of total citations and citations per year, such as Harvard University (United States, Number of reviews: 2, TGCS: 799, TCpY: 73.40), CNRS-Collège de France (France, Number of reviews: 2, TGCS: 546, TCpY: 41.27), University of Maryland (United States, Number of reviews: 2, TGCS: 395, TCpY: 59.48), Max Planck Institute (Germany, Number of reviews: 1, TGCS: 1,238, TCpY: 77.38), and University of Illinois (United States, Number of reviews: 1, TGCS: 583, TCpY: 53.00). **Supplementary Table S3** contains the number of reviews by Institutes, by Country, as well as total global citation scores and total citations per year for all the Institutes present in the dataframe.

Performance of Authors

Table 4 shows the number of reviews on driving simulation studies by the top 10 Authors and the number of single-, multi-, and first-authored reviews for each Author. These Authors wrote or co-wrote 37 out of 298 (i.e., about 12.4%) reviews on driving simulation. "S. Classen" dominates the ranking with seven reviews, followed by "D.L. Fisher," "S. Koziel," and "M. Rizzo" with four reviews each, and all other Authors wrote three reviews each. Only five reviews were single-authored, and about half of them (19) were first-authored by one of the top 10 Authors.

Cluster Analysis

To identify and understand possible ensembles of semantic knowledge in this scientific area, a cluster analysis based on the KCN was performed. Cluster analysis is a multivariate technique that allows to minimize the distance between items belonging to the same group and to maximize the distance between items from different groups (Irani et al., 2016). VOSviewer software perform a cluster analysis throughout the "VOS mapping technique,"

TABLE 4 | Number of reviews and number of single-, multi-, and first-authored reviews on driving simulation studies by Author.

Author	Number of reviews	Single-Authored	Multi-Authored	First-Authored
Classen S.	7	0	7	4
Fisher D.L.	4	0	4	1
Koziel S.	4	1	3	2
Rizzo M.	4	1	3	2
Andysz A.	3	0	3	2
Bekasiewicz A.	3	0	3	1
George C.F.P.	3	3	0	3
Pearlson G.D.	3	0	3	0
Uc E.Y.	3	0	3	1
Verster J.C.	3	0	3	3

which is based on a weighted and parameterized variant of modularity-based clustering (for a detailed explanation, see Waltman et al., 2010). Keyword co-occurrences refer to the common presence, frequency, and proximity of keywords that are similar to others, i.e., based on the same topic, but not exactly the same. In other words, keyword co-occurrence is an association or combination of terms that marks the presence of a keyword in several papers (more than one) of a bibliographic database. Since the keywords of a paper are supposed to indicate the core concept of the study, this method is useful to systematically explore the knowledge-components and the knowledge-structure constructed by the keywords of papers in a specific research field. The KCN's modularity is the network ability to decompose into separated modules or clusters. Each link between keywords in the network has a strength represented by a positive numerical value; the higher this strength value, the stronger the linkage (Radhakrishnan et al., 2017). The total link strength represents the number of publications in which two keywords occur together. In other words, link strength refers to the strength of semantics association between keywords. Highly cited keywords were analyzed and visualized with VOSviewer (Van Eck and Waltman, 2014). The type of analysis was selected by choosing "Co-occurrence" among the alternatives offered by the software. Subsequently, the analysis' unit was chosen selecting only the "author's keywords" and excluding "keywords plus" in which there were general and non-specific terms such as "human," "review," and "computer." Furthermore, the counting method employed in this analysis was the "Fractional counting," in which the weight of a link is fractionalized. For example, if a keyword co-occurs in a document with five other keywords, each of the five co-occurrences has a weight of 1/5. Considering the minimum and the maximum number of possible co-occurrences in the database (respectively 1 and 13), the co-occurrences threshold (i.e., the minimum number of occurrences of a keyword to enter the network) was based on the median value and set as 7. In this way, only eight of the 763 keywords in the database met the threshold and were brought into visualization (**Figure 2**). The purpose of this choice was to extract and visualize only the most relevant keywords. According to VOSviewer manual, the nodes represent the keywords, and the co-occurring frequency of a keyword is represented by the circle size; the larger a circle, the more a keyword has been co-selected in the driving simulation reviews. The analysis clearly defined three clusters: cluster 1 includes "driving simulator," "driving simulation," "sleepiness," and "attention" grouping together ergonomic, anthropic, environmental, and psychophysiological factors; cluster 2 includes "driving," "simulation," and "alcohol" grouping very different human, environmental, and technical subtopics linked to driving simulation research; finally, cluster 3 includes only "dementia" which refers to a wide range of subtopics (i.e., assessment; treatment; assistive driving systems, etc.). **Table 5** lists the main clusters identified, as well as the associated keywords, the occurrences, the links, and total link strength. Keyword "Sleepiness" was the most cited of cluster 1 and had the highest number of occurrences (13), links (five), and a total link strength equal to 8. Keyword "Attention" had seven occurrences, five links, and a total link strength of 7. Keyword "Dementia" had the same number of occurrences of "Attention," but a fewer number of links (three) and a lower total link strength (5). Keyword "Driving simulation" also had the same number of occurrences, and a number of links equal to 3 and total link strength equal to 3. Keyword "Driving" was the most cited of cluster 2 and of the whole network, had the highest number of occurrences (35) and links (six) with a total link strength equal to 14, suggesting for a key role in the network. Keyword "Simulation" had 22 occurrences, three links, and achieved a total link strength of 6. Keyword "Alcohol" had seven occurrences, two links, and a total link strength of 3. Keyword "Driving simulator"

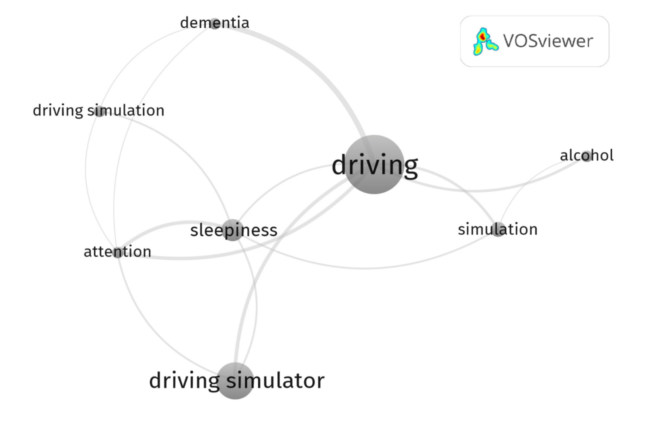


FIGURE 2 | Author's keywords network as a result of cluster analysis.

TABLE 5 | Keywords and related occurences, links, and total link strength for the three major clusters.

# Cluster	Keyword	Occurrences	Links	Total link strength
1	*Sleepiness*	13	5	8
1	*Attention*	7	5	7
1	*Dementia*	7	3	5
1	*Driving simulation*	7	3	3
2	*Driving*	35	6	14
2	*Simulation*	9	3	6
2	*Alcohol*	7	2	3
3	*Driving simulator*	22	3	5

was the only one present in cluster 3, with 22 occurrences, three links, and a total link strength of 5.

Discussion

The first purpose of the present study was to perform a scientometric analysis on driving simulation reviews and to provide a comprehensive picture of secondary studies on this topic based on 298 reviews obtained by Scopus and WoS Core Collection in the last 30 years.

Performance analysis was conducted on Countries, Journals, Institutions, and Authors. The United States and Germany are the first two Countries for the number of driving simulation reviews, and their production has increased constantly in the last 20 years. Surprisingly, journals which contributed to the highest number of driving simulation reviews comprise only two titles belonging to the transportation field. This could be taken as a cue of the wide interest in the topic from different disciplines, such as medicine, engineering, psychology, *et cetera*. Regarding performance of institutes, again, the United States is the leading Country with five Institutes in the top 10. Top authors wrote from three to seven reviews each and belong to institutes located in North America (United States and Canada) and Europe (Netherlands, Germany, Poland, and Iceland).

The comparison between the results of the present study on the reviews of simulated driving and those obtained by Guo et al. (2019) on primary studies allows some considerations. Regarding performance of Countries, it is noteworthy that the United States and Germany dominate both ranks. Although Canada did not appear in the rank of the first four most productive Countries for primary studies, it appears as the third most productive Country in the rank of the reviews. Several European Countries, namely, France, United Kingdom, Netherlands, Poland, and Switzerland, are present in the top 10 of the reviews. Finally, even though in 2016, China was the second most prolific Country of primary studies in the world, in the reviews' rank, it is located at the seventh position. This is an interesting result because it shows an increase in the gap between the United States and China switching from primary studies to reviews or a clear preference of China research centers for empirical studies.

The comparison between the performance of journals shows that those present in both ranks refer to the fields of transportation and engineering on one hand and to the field of human factors and ergonomics on the other. More specifically, Transportation Research Record, Human Factors, and Traffic Injury Prevention are classified, respectively, at the third, fourth, and sixth positions in both ranks. Two national journals are in the top 10 of the reviews, namely, *VDI Berichte* (Germany) and *Medycyna Pracy* (Czechia). This seems to be a clue of the attention European Countries devotes in reviewing and summarizing studies on driving simulation. Furthermore, Guo et al. (2019) emphasized that several journals published primary studies on the topic of sleepiness, but none of them appeared in the top 10 list. Regarding the reviews, it is noticeable that *Sleep Medicine Reviews* is in the first position. This highlights that sleepiness is one of the most recurrent topics in both primary and secondary studies on driving simulation. Indeed, sleepiness leads to physical conditions which may increase the rate of accidents and reduce the safety during driving (Guo et al., 2019). A comparison of citation scores among journals belonging to different subjects and disciplines is not desirable in this case, since differences in such scores may be partly due to differences in status and spread of journals themselves (e.g., impact factor and other bibliographic indexes related to the journals). Nonetheless, it is reasonable to think that all the reviews as well as the most part of the citations refer to the same topic, and that driving simulation is a topic which cannot be ascribed to one subject in particular, but belongs to several disciplines and research fields.

Regarding the performance of institutes, three Universities in the United States (i.e., University of Iowa, University of Massachusetts, and University of Michigan) are present in both the top 10 ranks. European institutes are almost equally present in the primary studies rank than in the reviews' rank. Utrecht University is currently in the top 10 reviews' rank, while Delft University of Technology and University of Groningen are among the most productive institutes regarding primary studies. Further, the Nofer Institute is located at the fourth position in the reviews' rank. Iceland and Canada are represented by Reykjavik University and University of Toronto and University of Western Ontario, respectively. These institutes were not present in the top 10 rank of primary studies.

Concerning the most prolific authors, "D.L. Fisher" is the only author who appears to be present in both top 10 ranks. No other correspondences emerge by the comparison of authors' performances among the two top 10 ranks.

The conceptual structure of driving simulation reviews was outlined using a co-occurrence network analysis to map and cluster high-frequency author keywords. Cluster analysis makes clear the interdisciplinary nature of this research topic. Three main clusters were identified together with eight relevant keywords. It is noteworthy that the eight keywords represent two distinct areas of interest, namely, an area devoted to the investigation of technical factors of driving simulation and another area devoted to the investigation of human factors, taking into account participants coming from special populations (i.e., persons with dementia, sleep-related disorders, alcohol-related disorders, and attention deficit). The choice to set a co-occurrences threshold using the median value among those available probably led to the best trade-off between the high heterogeneity of research topics and the need to summarize the main trends within the driving simulation framework.

A direct comparison between the results of cluster analysis proposed in this study and those from the cluster analysis conducted by Guo et al. (2019) is not possible. Indeed, in the present study, the cluster analysis was conducted on the Co-occurrence Network between authors' keywords. In the study conducted by Guo et al. (2019), the cluster analysis was based on the Co-citation Network. The different nature of the data allows only a tentative comparison of semantic labels emerged by the respective cluster analyses. Labels associated with the human factor were predominant in both primary (10/13; 76.2%) and secondary studies (5/8; 62.5%). These labels were in the top positions in terms of productivity (number of documents) in the analysis on primary studies and were also present as authors' keywords in a large amount of the reviews. This may be an indirect clue that simulated driving is a topic still strongly related to the human component. For example, driving simulation methods are employed to assess human driving abilities in different medical conditions, under the effect of various substances and medications, and to study safety behaviors and cognitive functioning related to driving activity.

SELECTIVE REVIEW OF REVIEWS

Methods

Cluster analysis based on authors' keywords identified a total of 61 reviews. In order to identify the most relevant results about validity and fidelity of driving simulation, it was decided to conduct a further investigation within this subgroup. A checklist was created in order to answer for each review to the following questions: (1) Was driving simulation performance associated with or predictive of on-road testing performance? (2) Was driving simulation performance associated with or predictive of standard neuropsychological testing performance? (3) Did driving simulation exhibit the same or similar features of real driving? (4) Was a formal meta-analysis feasible? (5) Was there any critical issue highlighted regarding driving simulation?

The first two questions were aimed at investigating validity of driving simulations with respect to the other two methods currently used for assessing fitness to drive, i.e., on-road testing and standardized neuropsychological testing. The third question was aimed at investigating fidelity about the experience of driving simulation with respect to the experience of real driving. The fourth question was aimed at investigating the possibility of summarizing throughout meta-analytic techniques quantitative results for a given research topic. The fifth question was aimed at investigating critical issues regarding the use of driving simulation in research and clinical practice.

The answers to the first two questions were coded as following: "Yes" if in the text of the review Authors clearly stated that there was an association or a prediction between driving simulation and on-road and standardized neuropsychological testing, respectively; "No" if Authors clearly stated that there was no association or a prediction between driving simulation and on-road and standardized neuropsychological testing, respectively; and "Mixed results" if Authors stated that a low association or prediction was found or alternatively that an association or prediction was found for a subgroup of studies included in the review but not for others, "nd" if it was not possible to detect information about the relationship between driving simulation and on-road and standardized neuropsychological testing. The answers to the third question were coded as following: "Yes" if in the text of the review, Authors clearly stated that driving simulation had the same or similar features of real driving; "No" if Authors clearly stated that driving simulation had not the same or similar features of real driving or that had different and not comparable features; and "Mixed results" if Authors stated that driving simulation had only few features comparable to those of real driving, "nd" if it was not possible to detect information about the features shared between driving simulation and real driving. The answers to the fourth question were coded as following: "1" if a critical or narrative or clinical or selective review was conducted, "2" if a systematic review was conducted following international well-established guidelines for collecting data and reporting results, such as PRISMA Statement, CONSORT Statement, QUOROM guidelines, *et cetera*, "3" if a formal meta-analysis was conducted, i.e., it was possible to obtain a pooled effect size starting from the effect sizes of primary studies and to perform publication bias as well as moderator and sensitivity analyses. In order to answer the fifth and last questions, it was decided to extract from each review the sentences highlighting critical aspects and issues specifically linked to the use of driving simulation within the covered topic (see **Supplementary Table S4**).

Results

Table 6 reports information obtained through the first four questions proposed. For each review, the following information was included in the table: cluster and keyword to which the review belongs to, the title and the reference of the review, the topic covered by the review, the discipline of the first Author, the answer to the first four questions proposed. Regarding the first question, i.e., the validity of driving simulation compared to the on-road testing, 36 out of 61 reviews reported an answer: 22 reviews reported an association between or a prediction of driving simulation with respect to on-road testing, seven did not report such an association or a prediction, and seven reported mixed results. Regarding the second question, i.e., the validity of driving simulation compared to standardized neuropsychological testing, it was possible to retrieve an answer for 24 out of 61 reviews: 21 reviews reported an association between or a prediction of driving simulation with respect to standardized neuropsychological testing, none of the reviews reported no association or prediction, while three reported mixed results. With respect to the third question, i.e., the fidelity about the experience of driving simulation with respect to the experience of real driving, 12 reviews out of 61 reported an answer: five reviews reported a comparable experience between driving simulation and real driving, three reported a non-comparable experience, and four reported mixed results. Concerning the fourth question, i.e., whether a formal meta-analysis was feasible, 47 reviews did perform a critical or narrative or clinical or selective review, 11 conducted a systematic review referring to well-established guidelines, and three were able to perform a formal meta-analysis and obtained a pooled effect size. With respect to the fifth question, i.e., what were the critical issues regarding driving simulation, several issues were reported (see **Supplementary Table S4** for the full list of critical issues regarding driving simulation for each review), and the five most frequent ones were: (a) the lack of a common set of standards in order to reduce the variability of results between different types of simulators, (b) the phenomenon of simulation sickness, (c) the need for psychometric properties and normative data for both different parameters and specific populations, (d) the lack of studies investigating predictive validity of driving simulation with respect to crash and collision rates, and (e) the lack of studies investigating

TABLE 6 | Features of the studies included in the selective review of reviews.

Cluster	Keyword	Title	Study	Topic	Discipline	Validity compared to on-road testing	Validity compared to laboratory testing	Fidelity	Systematic review
1	Attention	Driving and neurologic disorders	Drazkowski and Sirven, 2011	Neurologic Disorders	Neurology	nd	nd	nd	1
1	Attention	Parkinson disease and driving: An evidence-based review	Crizzle et al., 2012	Major Neurocognitive disorders	Health Science	nd	Yes	nd	1
1	Attention	Neural correlates of simulated driving while performing a secondary task: A review	Palmiero et al., 2019	Distraction fMRI	Health Science	nd	nd	nd	1
1	Dementia	Driving and dementia: A review of the literature	Lloyd et al., 2001	Major Neurocognitive disorders	Occupational Therapy	No	nd	nd	1
1	Dementia	Driving and dementia: A review of the literature	Brown and Ott, 2004	Major Neurocognitive disorders	Psychiatry	Yes	Yes	nd	1
1	Dementia	Systematic review of driving risk and the efficacy of compensatory strategies in persons with dementia	Man-Son-Hing et al., 2007	Major Neurocognitive disorders	Geriatry	Yes	nd	nd	1
1	Dementia	Brain morphometry and functional imaging techniques in dementia: methods, findings and relevance in forensic neurology	Klöppel, 2009	Major Neurocognitive disorders	Psychiatry	nd	Yes	nd	1
1	Dementia	Car drivers with dementia: different complications due to different etiologies?	Piersma et al., 2016	Major Neurocognitive disorders	Psychology	Yes	Yes	nd	1
1	Driving simulation	Validation of driving simulators	Davison et al., 2011	Vision	Ophthalmology	nd	nd	nd	1
1	Driving simulation	Saccadic velocity as an arousal index in naturalistic tasks	Di Stasi et al., 2013a	Workload/Fatigue	Psychology	nd	nd	nd	1
1	Driving simulation	Inside the clinical evaluation of sleepiness: Subjective and objective tools	Baiardi and Mondini, 2019	Sleepiness	Medicine	Yes	Yes	nd	1
1	Driving simulation	Driving status of patients with generalized spike–wave on EEG but no clinical seizures	Antwi et al., 2019	Epilepsy and Driving	Neurology	nd	nd	nd	1
1	Sleepiness	Neuropsychological function in obstructive sleep apnea	Engleman and Joffe, 1999	Sleepiness	Medicine	Yes	Yes	nd	1
1	Sleepiness	Daytime sleepiness and its evaluation	Cluydts et al., 2002	Sleepiness	Psychology	No	nd	nd	1
1	Sleepiness	Cognition and daytime functioning in sleep-related breathing disorders	Jackson et al., 2011	Sleepiness	Health Science	nd	nd	nd	1

(Continued)

TABLE 6 | Continued

Cluster	Keyword	Title	Study	Topic	Discipline	Validity compared to on-road testing	Validity compared to laboratory testing	Fidelity	Systematic review
1	Sleepiness	Diagnostic approach to sleep-disordered breathing	Thurnheer, 2011	Sleepiness	Pneumology	nd	nd	nd	1
1	Sleepiness	Hypersomnolence and traffic safety	Gupta et al., 2017	Sleepiness	Psychiatry	nd	nd	nd	1
1	Sleepiness	Subjective and objective assessment of hypersomnia	Murray, 2017	Sleepiness	Neurology	nd	Yes	nd	1
1	Sleepiness	Determinants of policy decisions for non-commercial drivers with OSA: An integrative review	Rizzo et al., 2018	Sleepiness	Medicine	nd	nd	nd	1
1	Sleepiness	Driving simulators in the clinical assessment of fitness to drive in sleepy individuals: A systematic review	Schreier et al., 2018	Sleepiness	Neurology	Mixed results	nd	nd	2
1	Sleepiness	Narrative review: Do spontaneous eye blink parameters provide a useful assessment of state drowsiness?	Cori et al., 2019	Sleepiness	Medicine	Yes	Yes	nd	1
2	Alcohol	Using virtual reality to study alcohol intoxication effects on the neural correlates of simulated driving	Calhoun et al., 2005	Alcohol consumption	Psychiatry	Yes	nd	nd	1
2	Alcohol	A selective review of simulated driving studies: Combining naturalistic and hybrid paradigms, analysis approaches, and future directions	Calhoun and Pearlson, 2012	Neuroimaging and Alcohol	Psychology	Yes	nd	Yes	1
2	Alcohol	The sensitivity of laboratory tests assessing driving related skills todose-related impairment of alcohol: A literature review	Jongen et al., 2016	Alcohol	Pharmacology	No	nd	nd	2
2	Alcohol	A systematic review of the evidence for acute tolerance to alcohol – the “Mellanby effect”	Holland and Ferner, 2017	Alcohol	Medicine	nd	Yes	nd	2
2	Alohol	Effects of acute alcohol consumption on measures of simulated driving: A systematic review and meta-analysis	Irwin et al., 2017	Alcohol consumption	Health Science	Yes	nd	nd	3

(Continued)

TABLE 6 | Continued

Cluster	Keyword	Title	Study	Topic	Discipline	Validity compared to on-road testing	Validity compared to laboratory testing	Fidelity	Systematic review
2	Driving	Cognitive dysfunction in sleep disorders	Fulda and Schulz, 2001	Sleepiness	Psychology	Mixed results	Mixed results	nd	1
2	Driving	Outcome measurement in sleep medicine practice and research. Part 2: assessment of neurobehavioral performance and mood	Weaver, 2001	Sleepiness	Nursing	nd	Yes	nd	1
2	Driving	Are opioid-dependent/tolerant patients impaired in driving-related skills? A structured evidence-based review	Fishbain et al., 2003	Medication assumption	Psychiatry	nd	nd	nd	1
2	Driving	Driving simulators in clinical practice	George, 2003	Sleepiness	Medicine	No	nd	Yes	1
2	Driving	Residual effects of sleep medication on driving ability	Verster et al., 2004	Medication assumption	Pharmacology	No	nd	nd	1
2	Driving	The assessment of driving abilities	Marcotte and Scott, 2004	Assessment of driving skills	Psychiatry	Yes	Mixed results	No	1
2	Driving	Conversation effects on neural mechanisms underlying reaction time to visual events while viewing a driving scene: fMRI analysis and asynchrony model	Hsieh et al., 2009	Distraction fMRI	Communication science	nd	nd	nd	1
2	Driving	Functional consequences of HIV-associated neuropsychological impairment	Gorman et al., 2009	Major Neurocognitive disorders(hiv)	Psychiatry	Yes	Yes	Mixed results	1
2	Driving	Driving ability in Parkinson's disease: Current status of research	Klimkeit et al., 2009	Major Neurocognitive Disorders	Psychology	Yes	Yes	nd	1
2	Driving	Phoning while driving II: A review of driving conditions influence	Collet et al., 2010	Distraction	Psychology	Mixed results	nd	Mixed results	1
2	Driving	A review of driving simulator parameters relevant to the operation enduring freedom/operation Iraqi freedom veteran population	Kraft et al., 2010	Iraqi Veterans/DS's parameters operation	Medicine	Yes	Yes	yes	1
2	Driving	Zopiclone as positive control in studies examining the residual effects of hypnotic drugs on driving ability	Verster et al., 2011	Medication assumption	Pharmacology	Yes	Yes	nd	3

(Continued)

TABLE 6 | Continued

Cluster	Keyword	Title	Study	Topic	Discipline	Validity compared to on-road testing	Validity compared to laboratory testing	Fidelity	Systematic review
2	Driving	Systematic review of the quality and generalizability of studies on the effects of opioids on driving and cognitive/psychomotor performance	Mailis-Gagnon et al., 2012	Medication assumption	Medicine	nd	Mixed results	No	2
2	Driving	Does personality predict driving performance in middle and older age? An evidence-based literature review	Nichols et al., 2012	Personality	Psychology	No	nd	nd	2
2	Driving	Epilepsy and driving: Potential impact of transient impaired consciousness	Chen et al., 2014	Epilepsy	Neurology	Yes	nd	nd	1
2	Driving	Saccadic peak velocity as an alternative index of operator attention: A short review	Di Stasi et al., 2013b	Saccadic velocity/attention	Psychology	nd	nd	nd	1
2	Driving	The impact of depression on driver performance	Wickens et al., 2014	Mental health	Health Science	Yes	nd	nd	1
2	Driving	Driving in Parkinson's disease	Özdilek and Uç, 2014	Major Neurodegenerative disorders	Neurology	nd	Yes	nd	1
2	Driving	The racer's brain – How domain expertise is reflected in the neural substrates of driving	Lappi, 2015	Neural substrates of driving	Psychology	nd	nd	nd	2
2	Driving	Mirtazapine as positive control drug in studies examining the effects of antidepressants on driving ability	Verster et al., 2015	Medication assumption	Pharmacology	nd	nd	nd	1
2	Driving	High risk driving in treated and untreated youth with attention deficit hyperactivity disorder: Public health implications	Jillani and Kaminer, 2016	ADHD	Medicine	nd	nd	nd	2
2	Driving	Driving with a neurodegenerative disorder: An overview of the current literature	Jacobs et al., 2017	Major Neurocognitive Disorders	Neurology	Yes	Yes	yes	1
2	Driving	Covert hepatic encephalopathy: Can my patient drive?	Shaw and Bajaj, 2017	Hepatic Encephalopathy	Gastroenterology	nd	Yes	nd	1
2	Driving	Smart in-vehicle technologies and older drivers: A scoping review	Classen et al., 2019	Aging	Occupational Therapy	No	nd	nd	2

(Continued)

TABLE 6 | Continued

Cluster	Keyword	Title	Study	Topic	Discipline	Validity compared to on-road testing	Validity compared to laboratory testing	Fidelity	Systematic review
2	Driving	Relationships between cognitive functions and driving behavior in Parkinson's disease	Ranchet et al., 2012	Cognition and driving in Parkinson	Psychology	Yes	Yes	nd	1
3	Driving simulator	The development of driving simulators: toward a multisensory solution	Pinto et al., 2008	Development of driving simulator	Psychology	Mixed results	nd	Mixed results	1
3	Driving simulator	Rehabilitation of combat-returnees with traumatic brain injury	Lew et al., 2009	Rehabilitation	Medicine	Yes	Yes	Yes	1
3	Driving simulator	Validation of driving simulators	Shechtman, 2010	Validation of driving simulators	Occupational Therapy	Mixed results	nd	Mixed results	1
3	Driving simulator	Nasal continuous positive airway pressure (nCPAP) treatment for obstructive sleep apnea, road traffic accidents and driving simulator performance: A meta-analysis	Antonopoulos et al., 2011	Effect of nasal continuous positive airway pressure	Medicine	Yes	nd	nd	3
3	Driving simulator	Effects of adaptive cruise control and highly automated driving on workload and situation awareness: A review of the empirical evidence	de Winter et al., 2014	Automated driving systems	Engineering	nd	nd	nd	2
3	Driving simulator	Establishing an evidence-base framework for driving rehabilitation in Parkinson's disease: A systematic review of on-road driving studies	Devos et al., 2015	Major Neurocognitive disorders	Health Science	Yes	Yes	nd	2
3	Driving simulator	The impact of therapeutic opioid agonists on driving-related psychomotor skills assessed by a driving simulator or an onroad driving task: A systematic review	Ferreira et al., 2018	Medication assumption	Medicine	Yes	Yes	nd	2
3	Driving simulator	Evaluation method regarding the effect of psychotropic drugs on driving performance: A literature review	Iwata et al., 2018	Drugs consumption	Psychiatry	Mixed results	nd	No	1
3	Driving simulator	Efficacy of training with driving simulators in improving safety in young novice or learner drivers: A systematic review	Martín-delosReyes et al., 2019	Training for improving driving safety	Medicine	Mixed results	nd	nd	1
3	Driving simulator	Bibliometric analysis of simulated driving research from 1997 to 2016	Guo et al., 2019	Bibliometric analysis	Business	nd	nd	nd	1

ecological validity of driving simulation in predicting real-world driving performance.

Discussion

As shown in **Table 6**, questions mainly represented in the reviews analyzed concerned the validity of driving simulation for the assessment of fitness to drive with respect to on-road and standardized neuropsychological testing.

Regarding the first question, it is noteworthy that a considerable effort has been done in order to demonstrate the validity of driving simulation techniques to those coming from ecological settings, such as on-road testing. Most of the reviews which gave a response about this question stated that it is possible to claim a significant association or prediction of driving simulation performance with respect to on-road testing performance. Nevertheless, recurring critical issues related to this question emerged. Firstly, psychometric properties of driving simulation systems are not yet firmly established (Cluydts et al., 2002). There seems to be a lack of studies in order to clearly demonstrate the validity of simulators in terms of both construct (Jongen et al., 2016) and concurrent validity with respect to on-road testing (Crizzle et al., 2012; Nichols et al., 2012). Other reviews highlighted the lack of data supporting ecological (e.g., Hsieh et al., 2009; Jacobs et al., 2017; Classen et al., 2019) as well as absolute validity (i.e., the absence of significant statistical differences between effects measured on the same scale but with different tools; Kaptein et al., 1996; Classen et al., 2019). Lew et al. (2009) pointed out the lack of evidences related to test–retest reliability and the need for establishing operating characteristics of driving simulation testing (sensibility, specificity, accuracy) for specific populations. Following Shechtman (2010), one of the reasons for these issues stands in the lack of agreement about terminology used to define the concept of validity. Indeed, such terminology primarily comes from technical discipline such as engineering and computer science, but driving simulators are widespread and employed in many others scientific fields (i.e., medicine, psychology, etc.). Other works reported the need for a consensus on (a) a common set of parameters/indicators to be included in a simulator (Kraft et al., 2010), (b) settings and assessment methods of driving skills (Schreier et al., 2018), and (c) hardware (i.e., equipment) and software (i.e., scenarios) of driving simulators (Iwata et al., 2018). The huge variability on the aforementioned features hampers the comparability between simulators and makes that every research team goes on with its own device and protocol (Iwata et al., 2018). The lack of validation studies also limits the use of simulators as a tool for rehabilitation and training of driving skills. Indeed, few studies have tried to demonstrate the efficacy of driving simulation systems as a learning tool. Results seems to be inconclusive and heterogeneous and cannot be employed in order to produce a clear statement pro or versus the use of training programs based on driving simulation (Martín-delosReyes et al., 2019). In a review on rehabilitation of driving skills, it is unclear whether a driving simulation training may restore, maintain, and ensure transferability of such skills to real-world driving, and it is also unclear whether it could produce better results with respect to classical neurocognitive rehabilitation (Devos et al., 2015).

Regarding the second question, it is possible to conclude that a clear association or prediction of driving simulation performance with respect to standardized neuropsychological testing performance is present. Driving simulators thus offer the possibility to assess the same cognitive domains involved in the evaluation of fitness to drive and usually measured throughout laboratory tests, within a more ecological sensory environment. However, also for this question, the same limitations addressed for the previous one can be put forward. The lack of both validation studies and consensus on the features, parameters, and administration settings makes it difficult to collect normative data to be used for clinical evaluation of fitness to drive. Schreier et al. (2018) proposed to use both simulators and neuropsychological tools to evaluate fitness to drive in order to minimize the biases of both methods. Also, in this case, the need to validate both neuropsychological and driving simulation tools with respect to real-world driving and to standardize them for age, gender, and specific medical conditions emerges (Engleman and Joffe, 1999; Klöppel, 2009).

The third question is the less represented in the review analyzed; indeed, only for 20% of the reviews it was possible to retrieve an answer, with a substantial equality between the three categories of answer. The fidelity of driving simulation tools refers to the extent to which they simulate real-world driving experience (Kaptein et al., 1996; de Winter et al., 2007). A low-fidelity driving simulator includes a desktop and a basic equipment for simulated vehicle control, while a high-fidelity simulator usually has a 360° visual field projected on multiple monitors, a complete cockpit of an actual vehicle and a motion-based board providing kinesthetic feedback (Kaptein et al., 1996). Following Wynne et al. (2019), also for the concept of fidelity, there are issues related to the terminology and to the classification of simulators based on fidelity level with respect to on-road driving. For example, they pointed out that some research teams used the term "physical validity" to describe the fidelity, or that the lack of a common set of standard for the evaluation of fidelity usually results in three levels of classification (i.e., high, medium, and low), but there are no clear and standardized rules in order to describe the exact features for each level. In a recent review, Murray (2017) claimed that the lack of a standard device for the assessment of driving skills in individuals coming from special population (e.g., suffering from sleep disorders; Lucidi et al., 2006, 2013) may depend on the fact that driving simulators are developed and built for other considerations of driving safety than those requested for the assessment of specific population. However, the critical issue most frequently linked to fidelity is motion sickness or simulator sickness; that is, all the physiological reactions in the form of headache, nausea, and vomiting (Pinto et al., 2008). Tolerability of simulated driving experience is a fundamental issue especially in older persons, who frequently experience simulator sickness. Following Brown and Ott (2004), there are no driving simulators tailored for older people. A low performance in such people might reflect adaptation difficulties rather than deficit in driving skills *per se*. Simulation sickness seems to be the biggest issue related to simulator fidelity since it has a significant impact on both quality of measurement and drop-out rate (Malis-Gagnon et al., 2012; Iwata et al., 2018; Schreier et al., 2018). Marcotte and Scott (2004) claimed that sickness is directly related to the degree of realism. Indeed, simulators can vary in terms of visual and auditory inputs and in complexity of simulated scenario, although a relevant issue is due to the fact that only few studies provide a detailed description of the scenario, and thus it is difficult to replicate studies and generalize the results (Irwin et al., 2017). Another recurring issue is related to the risk perception in virtual reality. Even though participants carry out the task with the utmost accuracy, they are often fully aware that a collision in simulated scenario will not result in any harm and, consequently, they could not drive with the same caution they would in the real world (Marcotte and Scott, 2004). This issue starts from fidelity of driving experience but has an impact on ecological validity of measures collected with the driving simulator.

Fourth, a relevant point which emerged from the review of reviews is the difficulty in conducting a meta-analysis in order to provide a quantitative synthesis of causal relationships, predictive ability, and/or correlation between a) cognitive variables and driving simulation performance and b) driving simulation performance and on-road test (Wynne et al., 2019). Such difficulty might be given by different sources of huge heterogeneity among studies, namely, a) the availability on the market of several types of driving simulators, the large variability in the measures taken, as well as in their fidelity and reliability, b) the variability of tools and neuropsychological batteries used to measure cognitive abilities related to driving skills, and c) the variability due to experimental designs and manipulations.

There are two other considerations that might be taken into account when results from simulated driving performance are evaluated. The first one regards the distinction between predictive validity of simulated performance with respect to on-road performance or with respect to crash and collision rates (Lucidi et al., 2014, 2019; Mallia et al., 2015; Spano et al., 2019). Neither simulated driving nor on-road testing seems to be predictive of future accidents (Man-Son-Hing et al., 2007; Drazkowski and Sirven, 2011; Gupta et al., 2017; Baiardi and Mondini, 2019), and, despite the latter is considered the gold standard for assessing fitness to drive, there are few studies which investigated which measures in simulated driving might be useful to predict the risk of collision (George, 2003; Drazkowski and Sirven, 2011; Piersma et al., 2016). Since the vast majority of research on simulated driving revolves around the topic of driving safety in a preventive perspective, it would be useful to direct research efforts to find and validate measures with high predictive validity with respect to crash and collision rates in real-world driving.

The second consideration concerns the distinction between tests of *typical* and *maximum* performance. All the methods used for assessing fitness to drive, that is, neuropsychological testing, driving simulation, and on-road testing, are tests of maximum performance, requiring the individuals to exert as much effort as possible and to obtain the best performance one can do. The real-world everyday driving activity can be instead considered a test of typical performance, requiring the individual to exert an effort enough not to incur in collisions or in major violations (Lucidi et al., 2010). Such discrepancy might be one of the reasons why all the aforementioned methods are not fully adequate to capture the variability of everyday driving. The issue here is not in the specific method used for assessing but resides in a substantial difference between the behavior elicited in these two frameworks. A possible remediation in order to get a typical evaluation of fitness to drive has been developed in multicenter longitudinal studies promoted by the AAA Foundation for Traffic Safety, namely, "The longroad study–Longitudinal research on aging drivers" (Li et al., 2017), and in another project called "The Ozcandrive Project" (Marshall et al., 2013). In these projects, in-vehicle recording devices together with a GPS system were applied within the vehicle in order to collect data from everyday driving activity (i.e., position, time of the day, speed, acceleration, safety distance, lane deviation, etc.) in real time and for a prolonged period of weeks or months.

CONCLUSION

In light of the results obtained and discussed above, some concluding remarks may be outlined.

First, driving simulation studies and reviews represent an increasingly relevant topic in the scientific literature on driving, especially in recent years and thanks to the technological innovations as well as to the increased computing power of hardware and software (e.g., Cipresso et al., 2018).

Second, it seems that driving simulation is a cross-cutting topic, present and widespread among different disciplines. It is also addressed with several approaches in virtue of a versatile methodology which allows the study of different aspects of driving simulation (e.g., from a human factor, medical, psychological, engineering-technical perspective).

Third, it is thus possible to observe a lack of shared and standardized methodologies and protocols, as well as the lack of a common language in the research field employing a driving simulation procedure (Wynne et al., 2019). All those factors act against the possibility to summarize findings from studies which investigate a similar relationship between driving-related variables, as well as to clearly compare driving simulation performance with other methods in order to assess fitness to drive in normal and special populations. Nonetheless, there are several evidences for considering driving simulation as a valid alternative to neuropsychological testing as well as to on-road testing for the assessment of fitness to drive.

Fourth, data coming from driving simulation studies are limited in providing generalizable results. Heterogeneity in simulators' types, settings, driving tasks, scenarios, specific populations, and research methodologies hampers the spread of driving simulation in clinical contexts; thus, content validity is limited for specific simulators, tasks, and populations (Kraft et al., 2010; Shechtman, 2010; Verster et al., 2011; Iwata et al., 2018). Another issue related to generalizability of results comes from the fact that several studies did not report all the data captured from the software within the simulator (Irwin et al., 2017), and this makes it difficult to establish a set of measure and consequently of normative data (Kraft et al., 2010). Moreover, simulators that warrant a complete and naturalistic assessment of driving skills are expensive, cumbersome, and hardly available (Murray, 2017; Rizzo et al., 2018). Following Schreier et al. (2018), it would be useful to conduct studies aimed at both validating the same measures with different simulators and identifying the most comparable ones.

Lastly, further research efforts could be aimed at establishing a consensus statement for protocols regarding the assessment of driving behavior and fitness to drive in order to (a) use standardized cognitive and neuropsychological tests and batteries, (b) assess and compare systematically driving simulators with regard to what they measure and to their validity and fidelity, and (c) employ shared research designs and criteria for conducting studies in a given subtopic, e.g., with special populations.

The present study has three main strengths. First, it deals with a scientometric analysis on driving simulation considering the entire population of secondary studies on that topic. Two different scientific databases were analyzed since we were aware that there could have been a reduced share of overlapping between them and we wanted to reduce the risk to exclude relevant literature. The aim was thus not to carry out a comparative analysis between databases, but an exhaustive one. Indeed, there were 228 documents classified as reviews in Scopus and 151 in WoS. The final sample was composed of 298 records. This means that there were 81 duplicate records that were present in both databases, with a consequent overlap share of about 27%. This also means that, using only one database, 70 unique records using Scopus and 147 unique records using WoS would have been excluded.

Second, as far as we know, a second-order scientometric analysis including only secondary studies has never been conducted before. The rationale of a scientometric analysis on reviews lies in the fact that the authors of primary and secondary studies do not necessarily coincide. The authors of a review may not necessarily be experts on the main topic (here, driving simulation), but they may be experts on associated topics interested in undertaking an applied study using driving simulators. Moreover, it also allowed for the comparison with the recent scientometric analysis by Guo et al. (2019) on primary studies.

Third, the present study proposes a new approach integrating scientometric analysis with a review of reviews. The latter explicitly addressed the issues of the validity of simulators with respect to the gold standard for assessing fitness to drive, which remains the on-road test. Moreover, it explicitly compares the effectiveness of simulators in replacing the neuropsychological and psychometric tests frequently used in daily practice to predict driving success in special populations. This triangulation brought out two clusters of research questions, obtaining results of interest for those who intend to undertake research or are interested in proposing to stakeholders to integrate the on-road test with driving simulator assessment. Road safety professionals can rely on data providing suggestions on how simulators preach on-road tests on the one hand and how they provide suitable experimental control over the neuropsychological tests on the other, thus giving useful indications on the neuropsychological and psychometric prerequisites for fitness to drive.

The present study has some limitations. The first one comes from an issue which is always present in review and meta-analytic studies, and it is reasonable to be also present in scientometric investigations, namely, the exclusion from the analysis of the white papers and gray literature. Such literature is usually not indexed and available in official databases and can provide a relevant source of information for disseminating studies reporting null or negative results that might not otherwise be disseminated (e.g., Paez, 2017). Currently, there are no methods to assess the impact of white papers and gray literature on the results of a scientometric analysis, unlike meta-analysis for which specific techniques have been developed. In this view, results from a scientometric analysis can be biased, especially toward positive results, and the conclusions may not be fully generalizable and need to be taken with caution.

The second limitation is due to the time coverage of the literature search. Indeed, the search did not include the second half of the 2019; this could have had an impact, albeit modest, on the last time point of **Figure 1** and on the number of the reviews included both in the scientometric analysis and in the review of reviews. For the sake of clarity, a new search was conducted on both Scopus and WoS on March 30, 2020, with the same search expression and produced the following results: Scopus yielded 238 reviews, with two more reviews in 2019 than those included in the data, and WoS yielded 164 reviews, with nine more reviews in 2019 than those included in the data. The two more reviews present on Scopus were also present on WoS, so in total, nine reviews were missed in 2019.

In conclusion, the present study represents an opportunity for broad-based methodological suggestions on a series of ideas: (a) heterogeneity of sources. It is typical for applied topics such as driving simulation. Indeed, this topic attracts the attention of scholars from very different disciplines. In addition to those largely expected such as engineers and computer scientists, with an ergonomics-oriented look, there can be found a wide range of data from medicine and allied disciplines such as neurosciences and psychology, each of these with different publication impacts and citational traditions. Such differentiation supports the need to derive the sources of analysis from multiple databases; (b) scarce bibliometric overlap between primary and secondary items and therefore the usefulness in some areas of conducting a second-order scientometric analysis. The widespread attention of several disciplines increases the variability of topics covered by the reviews, partially differentiating bibliometric characteristics (i.e., Authors, Institutes, Journals, Countries) of primary and secondary studies; and (c) usefulness to conduct a scientometric analysis together with a literature review, with the aim of providing a comprehensive picture of the topic by adopting two well-differentiated perspectives of analysis, which can be considered allied and complementary. The present study could be a good example of this broad-range approach.

AUTHOR CONTRIBUTIONS

AC, LT, ALo, and AB contributed to idea conception, data extraction, and analysis. AC, LT, ALo, AB, GS, ALi, YM, IG, and FS contributed to writing the first draft of the article. All authors contributed to article revision and approval of the final version.

FUNDING

The present article was supported by the Project "Apulia Smart Drivers," granted by the Apulia Region (Grant No. B34I19000180002).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00917/full#supplementary-material

REFERENCES

Amick, M. M., Grace, J., and Ott, B. R. (2007). Visual and cognitive predictors of driving safety in Parkinson's disease patients. *Arch. Clin. Neuropsy.* 22, 957–967. doi: 10.1016/j.acn.2007.07.004

Anstey, K. J., Wood, J., Lord, S., and Walker, J. G. (2005). Cognitive, sensory and physical factors enabling driving safety in older adults. *Clin. Psychol. Rev.* 25, 45–65.

Antonopoulos, C. N., Sergentanis, T. N., Daskalopoulou, S. S., and Petridou, E. T. (2011). Nasal continuous positive airway pressure (nCPAP) treatment for obstructive sleep apnea, road traffic accidents and driving simulator performance: a meta-analysis. *Sleep Med. Rev.* 15, 301–310. doi: 10.1016/j.smrv.2010.10.002

Antwi, P., Atac, E., Ryu, J. H., Arencibia, C. A., Tomatsu, S., Saleem, N., et al. (2019). Driving status of patients with generalized spike–wave on EEG but no clinical seizures. *Epilepsy Behav.* 92, 5–13. doi: 10.1016/j.yebeh.2018.11.031

Aria, M., and Cuccurullo, C. (2017). bibliometrix: an R-tool for comprehensive science mapping analysis. *J. Informer.* 11, 959–975.

Baiardi, S., and Mondini, S. (2019). Inside the clinical evaluation of sleepiness: subjective and objective tools. *Sleep Breath* 24, 369–377. doi: 10.1007/s11325-019-01866-8

Brooks, J. O., Goodenough, R. R., Crisler, M. C., Klein, N. D., Alley, R. L., Koon, B. L., et al. (2010). Simulator sickness during driving simulation studies. *Accid. Anal. Prev.* 42, 788–796. doi: 10.1016/j.aap.2009.04.013

Brown, L. B., and Ott, B. R. (2004). Driving and dementia: a review of the literature. *J. Geriatr. Psych. Neur.* 17, 232–240. doi: 10.1177/0891988704269825

Calhoun, V. D., Carvalho, K., Astur, R., and Pearlson, G. D. (2005). Using virtual reality to study alcohol intoxication effects on the neural correlates of simulated driving. *Appl. Psychophysiol. Biofeedback* 30, 285–306.

Calhoun, V. D., and Pearlson, G. D. (2012). A selective review of simulated driving studies: combining naturalistic and hybrid paradigms, analysis approaches, and future directions. *Neuroimage* 59, 25–35. doi: 10.1016/j.neuroimage.2011.06.037

Cantos-Mateos, G., Vargas-Quesada, B., Chinchilla-Rodríguez, Z., and Zulueta, M. A. (2012). "Stem cell research: bibliometric analysis of main research areas through KeyWords Plus," in *Aslib Proceedings*, (Bingley: Emerald Group Publishing Limited).

Chen, W. C., Chen, E. Y., Gebre, R. Z., Johnson, M. R., Li, N., Vitkovskiy, P., et al. (2014). Epilepsy and driving: potential impact of transient impaired consciousness. *Epilepsy Behav.* 30, 50–57.

Cipresso, P., Giglioli, I. A. C., Raya, M. A., and Riva G. (2018). The past, present, and future of virtual and augmented reality research: a network and cluster analysis of the literature. *Front. Psychol.* 9:2086. doi: 10.3389/fpsyg.2018.02086

Classen, S., Garvan, C., Awadzi, K., Sundaram, S., Winter, S., Lopez, E. D. S., et al. (2006). Systematic literature review and model for older driver safety. *Top. Geriatr. Rehabil.* 22, 87–98. doi: 10.1080/15389580802091199

Classen, S., Jeghers, M., Morgan-Daniel, J., Winter, S., King, L., and Struckmeyer, L. (2019). Smart in-vehicle technologies and older drivers: a scoping review. *OTJR Occup. Participat. Health* 39, 97–107. doi: 10.1177/1539449219830376

Cluydts, R., De Valck, E., Verstraeten, E., and Theys, P. (2002). Daytime sleepiness and its evaluation. *Sleep Med. Rev.* 6, 83–96.

Collet, C., Guillot, A., and Petit, C. (2010). Phoning while driving II: a review of driving conditions influence. *Ergonomics* 53, 602–616. doi: 10.1080/00140131003769092

Cori, J. M., Anderson, C., Soleimanloo, S. S., Jackson, M. L., and Howard, M. E. (2019). Narrative review: do spontaneous eye blink parameters provide a useful assessment of state drowsiness? *Sleep Med. Rev.* 45, 95–104. doi: 10.1016/j.smrv.2019.03.004

Crizzle, A. M., Classen, S., and Uc, E. Y. (2012). Parkinson disease and driving: an evidence-based review. *Neurology* 79, 2067–2074. doi: 10.1212/WNL.0b013e3182749e95

Davison, J. A., Patel, A. S., Cunha, J. P., Schwiegerling, J., and Muftuoglu, O. (2011). Recent studies provide an updated clinical perspective on blue light-filtering IOLs. *Graefes Arch. Clin. Exp. Ophthalmol.* 249, 957–968. doi: 10.1007/s00417-011-1697-6

de Winter, J. C., Happee, R., Martens, M. H., and Stanton, N. A. (2014). Effects of adaptive cruise control and highly automated driving on workload and situation awareness: a review of the empirical evidence. *Transp. Res. Part F.* 27, 196–217.

de Winter, J. C. F., De Groot, S., Mulder, M., Wieringa, P. A., Dankelman, J., and Mulder, J. A. (2009). Relationships between driving simulator performance and driving test results. *Ergonomics* 52, 137–153. doi: 10.1080/00140130802277521

de Winter, J. C. F., Wieringa, P. A., Dankelman, J., Mulder, M., Van Paassen, M. M., and De Groot, S. (2007). "Driving simulator fidelity and training effectiveness," in *In Proceedings of the 26th European Annual Conference on Human Decision Making and Manual Control*, Lyngby.

Devos, H., Ranchet, M., Emmanuel Akinwuntan, A., and Uc, E. Y. (2015). Establishing an evidence-base framework for driving rehabilitation in Parkinson's disease: a systematic review of on-road driving studies. *NeuroRehabilitation* 37, 35–52.

Devos, H., Vandenberghe, W., Nieuwboer, A., Tant, M., Baten, G., and De Weerdt, W. (2007). Predictors of fitness to drive in people with Parkinson disease. *Neurology* 69, 1434–1441.

Di Stasi, L. L., Catena, A., Cañas, J. J., Macknik, S. L., and Martinez-Conde, S. (2013a). Saccadic velocity as an arousal index in naturalistic tasks. *Neurosci. Biobehav. Rev.* 37, 968–975.

Di Stasi, L. L., Marchitto, M., Antolí, A., and Cañas, J. J. (2013b). Saccadic peak velocity as an alternative index of operator attention: a short review. *Eur. Rev. Appl. Psychol.* 63, 335–343.

Domeyer, J. E., Cassavaugh, N. D., and Backs, R. W. (2013). The use of adaptation to reduce simulator sickness in driving assessment and research. *Accid. Anal. Prev.* 53, 127–132.

Drazkowski, J. F., and Sirven, J. I. (2011). Driving and neurologic disorders. *Neurology* 76(7 Suppl. 2), S44–S49.

Elango, B., Rajendran, P., and Bornmann, L. (2015). A scientometric analysis of international collaboration and growth of literature at the macro level. *Malays. J. Libr. Inf. Sci.* 20, 41–50.

Engleman, H., and Joffe, D. (1999). Neuropsychological function in obstructive sleep apnoea. *Sleep Med. Rev.* 3, 59–78.

Ferreira, D. H., Boland, J. W., Phillips, J. L., Lam, L., and Currow, D. C. (2018). The impact of therapeutic opioid agonists on driving-related psychomotor skills assessed by a driving simulator or an on-road driving task: a systematic review. *Palliat. Med.* 32, 786–803.

Fishbain, D. A., Cutler, R. B., Rosomoff, H. L., and Rosomoff, R. S. (2003). Are opioid-dependent/tolerant patients impaired in driving-related skills? A structured evidence-based review. *J. Pain Symptom Manage.* 25, 559–577.

Fulda, S., and Schulz, H. (2001). Cognitive dysfunction in sleep disorders. *Sleep Med. Rev.* 5, 423–445.

George, C. F. (2003). Driving simulators in clinical practice. *Sleep Med. Rev.* 7, 311–320.

Gorman, A. A., Foley, J. M., Ettenhofer, M. L., Hinkin, C. H., and van Gorp, W. G. (2009). Functional consequences of HIV-associated neuropsychological impairment. *Neuropsychol. Rev.* 19, 186–203.

Grace, J., Amick, M. M., D'abreu, A., Festa, E. K., Heindel, W. C., and Ott, B. R. (2005). Neuropsychological deficits associated with driving performance in Parkinson's and Alzheimer's disease. *J. Int. Neuropsychol. Soc.* 11, 766–775.

Guo, F., Lv, W., Liu, L., Wang, T., and Duffy, V. G. (2019). Bibliometric analysis of simulated driving research from 1997 to 2016. *Traffic. Inj. Prev.* 20, 64–71.

Gupta, R., Pandi-Perumal, S. R., Almeneessier, A. S., and BaHammam, A. S. (2017). Hypersomnolence and traffic safety. *Sleep Med. Clin.* 12, 489–499.

Heikkilä, V. M., Turkka, J., Korpelainen, J., Kallanranta, T., and Summala, H. (1998). Decreased driving ability in people with Parkinson's disease. *J. Neurol. Neurosurg. Psychiatry* 64, 325–330. doi: 10.1136/jnnp.64.3.325

Holland, M. G., and Ferner, R. E. (2017). A systematic review of the evidence for acute tolerance to alcohol–the "Mellanby effect". *Clin. Toxicol.* 55, 545–556.

Hood, W., and Wilson, C. (2001). The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics* 52, 291–314.

Hsieh, L., Young, R. A., Bowyer, S. M., Moran, J. E., Genik, R. J. II, Green, C. C., et al. (2009). Conversation effects on neural mechanisms underlying reaction time to visual events while viewing a driving scene: fMRI analysis and asynchrony model. *Brain Res.* 1251, 162–175.

Irani, J., Pise, N., and Phatak, M. (2016). Clustering techniques and the similarity measures used in clustering: a survey. *Int. J. Comput. Appl.* 134, 9–14.

Irwin, C., Iudakhina, E., Desbrow, B., and McCartney, D. (2017). Effects of acute alcohol consumption on measures of simulated driving: a systematic review and meta-analysis. *Accid. Anal. Prev.* 102, 248–266.

Iwata, M., Iwamoto, K., Kawano, N., Kawaue, T., and Ozaki, N. (2018). Evaluation method regarding the effect of psychotropic drugs on driving performance: a literature review. *Psychiatry Clin. Neurosci.* 72, 747–773.

Jackson, M. L., Howard, M. E., and Barnes, M. (2011). "Cognition and daytime functioning in sleep-related breathing disorders," in *Progress in Brain Research*, Vol. 190, eds H. P. A. Van Dongen and G. A. Kerkhof (Asmterdam: Elsevier), 53–68.

Jacobs, M., Hart, E. P., and Roos, R. A. (2017). Driving with a neurodegenerative disorder: an overview of the current literature. *J. Neurol.* 264, 1678–1696.

Jillani, S. A., and Kaminer, Y. (2016). High risk driving in treated and untreated youth with attention deficit hyperactivity disorder: public health implications. *Adolesc. Psychiatry* 6, 89–99.

Jongen, S., Vuurman, E. F. P. M., Ramaekers, J. G., and Vermeeren, A. (2016). The sensitivity of laboratory tests assessing driving related skills to dose-related impairment of alcohol: a literature review. *Accid. Anal. Prev.* 89, 31–48.

Kaptein, N. A., Theeuwes, J., and Van Der Horst, R. (1996). Driving simulator validity: Some considerations. *Transport. Res. Rec.* 1550, 30–36.

Klimkeit, E. I., Bradshaw, J. L., Charlton, J., Stolwyk, R., and Georgiou-Karistianis, N. (2009). Driving ability in Parkinson's disease: current status of research. *Neurosci. Biobehav. Rev.* 33, 223–231.

Klöppel, S. (2009). Brain morphometry and functional imaging techniques in dementia: methods, findings and relevance in forensic neurology. *Curr. Opin. Neurol.* 22, 612–616.

Kraft, M., Amick, M. M., Barth, J. T., French, L. M., and Lew, H. L. (2010). A review of driving simulator parameters relevant to the operation enduring freedom/operation iraqi freedom veteran population. *Am. J. Phys. Med. Rehabil.* 89, 336–344.

Lappi, O. (2015). The racer's brain–how domain expertise is reflected in the neural substrates of driving. *Front. Hum. Neurosci.* 9:635.

Lew, H. L., Poole, J. H., Lee, E. H., Jaffe, D. L., Huang, H. C., and Brodd, E. (2005). Predictive validity of driving-simulator assessments following traumatic brain injury: a preliminary study. *Brain Inj.* 19, 177–188.

Lew, H. L., Rosen, P. N., Thomander, D., and Poole, J. H. (2009). The potential utility of driving simulators in the cognitive rehabilitation of combat-returnees with traumatic brain injury. *J. Head Trauma Rehabil.* 24, 51–56.

Li, G., Eby, D. W., Santos, R., Mielenz, T. J., Molnar, L. J., and Strogatz, D. (2017). Longitudinal research on aging drivers (LongROAD): study design and methods. *Inj. Epidemiol.* 4:22. doi: 10.1186/s40621-017-0121-z

Lloyd, S., Cormack, C. N., Blais, K., Messeri, G., McCallum, M. A., Spicer, K., et al. (2001). Driving and dementia: a review of the literature. *Can. J. Occup. Ther.* 68, 149–156.

Lucidi, F., Giannini, A. M., Sgalla, R., Mallia, L., Devoto, A., and Reichmann, S. (2010). Young novice driver subtypes: relationship to driving violations, errors and lapses. *Accid. Anal. Prev.* 42, 1689–1696.

Lucidi, F., Girelli, L., Chirico, A., Alivernini, F., Cozzolino, M., Violani, C., et al. (2019). Personality traits and attitudes toward traffic safety predict risky behavior across young, adult and older drivers. *Front. Psychol.* 10:536. doi: 10.3389/fpsyg.2019.00536

Lucidi, F., Mallia, L., Lazuras, L., and Violani, C. (2014). Personality and attitudes as predictors of risky driving among older drivers. *Accid. Anal. Prev.* 72, 318–324.

Lucidi, F., Mallia, L., Violani, C., Giustiniani, G., and Persia, L. (2013). The contributions of sleep-related risk factors to diurnal car accidents. *Accid. Anal. Prev.* 51, 135–140.

Lucidi, F., Russo, P. M., Mallia, L., Devoto, A., Lauriola, M., and Violani, C. (2006). Sleep-related car crashes: risk perception and decision-making processes in young drivers. *Accid. Anal. Prev.* 38, 302–309.

Mailis-Gagnon, A., Lakha, S. F., Furlan, A., Nicholson, K., Yegneswaran, B., and Sabatowski, R. (2012). Systematic review of the quality and generalizability of studies on the effects of opioids on driving and cognitive/psychomotor performance. *Clin. J. Pain.* 28, 542–555.

Mallia, L., Lazuras, L., Violani, C., and Lucidi, F. (2015). Crash risk and aberrant driving behaviors among bus drivers: the role of personality and attitudes towards traffic safety. *Accid. Anal. Prev.* 79, 145–151.

Man-Son-Hing, M., Marshall, S. C., Molnar, F. J., and Wilson, K. G. (2007). Systematic review of driving risk and the efficacy of compensatory strategies in persons with dementia. *J. Am. Geriatr. Soc.* 55, 878–884.

Mao, G., Liu, X., Du, H., Zuo, J., and Wang, L. (2015). Way forward for alternative energy research: a bibliometric analysis during 1994–2013. *Renew. Sust. Energ. Rev*. 48, 276–286. doi: 10.1016/j.rser.2015.03.094

Marcotte, T. D., and Scott, J. C. (2004). The assessment of driving abilities. *Adv. Transp. Stud.* 79–90.

Marshall, S. C., Man-Son-Hing, M., Bédard, M., Charlton, J., Gagnon, S., Gélinas, I., et al. (2013). Protocol for candrive II/Ozcandrive, a multicentre prospective older driver cohort study. *Accid. Anal. Prev.* 61, 245–252.

Martín-delosReyes, L. M., Jiménez-Mejías, E., Martínez-Ruiz, V., Moreno-Roldán, E., Molina-Soberanes, D., and Lardelli-Claret, P. (2019). Efficacy of training with driving simulators in improving safety in young novice or learner drivers: a systematic review. *Transp. Res. Part F.* 62, 58–65.

Mathias, J. L., and Lucas, L. K. (2009). Cognitive predictors of unsafe driving in older drivers: a meta-analysis. *Int. Psychogeriatr.* 21, 637–653. doi: 10.1017/S1041610209009119

Mayhew, D. R., Simpson, H. M., Wood, K. M., Lonero, L., Clinton, K. M., Johnson, A. G., et al. (2011). On-road and simulated driving: Concurrent and discriminant validation. *J. Safety. Res.* 42, 267–275. doi: 10.1016/j.jsr.2011.06.004

Meneghini, R., Mugnaini, R., and Packer, A. L. (2006). International versus national oriented Brazilian scientific journals. A scientometric analysis based on SciELO and JCR-ISI databases. *Scientometrics* 69, 529–538.

Muñoz-Écija, T., Vargas-Quesada, B., and Chinchilla-Rodríguez, Z. (2017). Identification and visualization of the intellectual structure and the main research lines in nanoscience and nanotechnology at the worldwide level. *J. Nanopart. Res.* 19:62. doi: 10.1007/s11051-016-3732-3

Murray, B. J. (2017). Subjective and objective assessment of hypersomnolence. *Sleep. Med. Clin.* 12, 313–322. doi: 10.1016/j.jsmc.2017.03.007

Nichols, A. L., Classen, S., McPeek, R., and Breiner, J. (2012). Does personality predict driving performance in middle and older age? An evidence-based literature review. *Traff. Inj. Prev.* 13, 133–143. doi: 10.1080/15389588.2011.644254

Noyons, E. C. M., Moed, H. F., and Luwel, M. (1999). Combining mapping and citation analysis for evaluative bibliometric purposes: a bibliometric study. *J. Am. Soc. Inf. Sci.* 50, 115–131. doi: 10.1002/(SICI)1097-4571(1999)50:2<115::AID-ASI3>3.0.CO;2-J

Özdilek, B., and Uç, E. (2014). Parkinson Hastalığında Sürücülük. *Turk. Nörol. Derg* 20, 64–71.

Paez, A. (2017). Gray literature: an important resource in systematic reviews. *J. Evid. Based. Med.* 10, 233–240. doi: 10.1111/jebm.12266

Palmiero, M., Piccardi, L., Boccia, M., Baralla, F., Cordellieri, P., Sgalla, R., et al. (2019). Neural correlates of simulated driving while performing a secondary task: a review. *Front. Psychol.* 10:1045. doi: 10.3389/fpsyg.2019.01045

Piersma, D., de Waard, D., Davidse, R., Tucha, O., and Brouwer, W. (2016). Car drivers with dementia: different complications due to different etiologies? *Traffic. Inj. Prev.* 17, 9–23.

Pinto, M., Cavallo, V., and Ohlmann, T. (2008). The development of driving simulators: toward a multisensory solution. *Trav. Humain.* 71, 62–95.

Pollack, J., and Adler, D. (2015). Emergent trends and passing fads in project management research: a scientometric analysis of changes in the field. *Int. J. Proj. Manage.* 33, 236–248.

Radford, K. A., Lincoln, N. B., and Lennox, G. (2004). The effects of cognitive abilities on driving in people with Parkinson's disease. *Disabil. Rehabil.* 26, 65–70.

Radhakrishnan, S., Erbis, S., Isaacs, J. A., and Kamarthi, S. (2017). Novel keyword co-occurrence network-based methods to foster systematic reviews of scientific literature. *PLoS One* 12:e0172778. doi: 10.1371/journal.pone.0172778

Ranchet, M., Broussolle, E., Poisson, A., and Paire-Ficout, L. (2012). Relationships between cognitive functions and driving behavior in Parkinson's disease. *Eur. Neurol.* 68, 98–107.

Ranchet, M., Paire-Ficout, L., Marin-Lamellet, C., Laurent, B., and Broussolle, E. (2011). Impaired updating ability in drivers with Parkinson's disease. *J. Neurol. Neurosurg. Psychiatry* 82, 218–223. doi: 10.1136/jnnp.2009.203166

Reger, M. A., Welsh, R. K., Watson, G., Cholerton, B., Baker, L. D., and Craft, S. (2004). The relationship between neuropsychological functioning and driving ability in dementia: a meta-analysis. *Neuropsychology* 18, 85–93.

Rizzo, D., Libman, E., Creti, L., Baltzan, M., Bailes, S., Fichten, C., et al. (2018). Determinants of policy decisions for non-commercial drivers with OSA: an integrative review. *Sleep Med. Rev.* 37, 130–137. doi: 10.1016/j.smrv.2017.02.002

Schreier, D. R., Banks, C., and Mathis, J. (2018). Driving simulators in the clinical assessment of fitness to drive in sleepy individuals: a systematic review. *Sleep Med. Rev.* 38, 86–100. doi: 10.1016/j.smrv.2017.04.004

Shaw, J., and Bajaj, J. S. (2017). Covert hepatic encephalopathy: can my patient drive? *J. Clin. Gastroenterol.* 51, 118–126. doi: 10.1097/MCG.0000000000000764

Shechtman, O., Classen, S., Awadzi, K., and Mann, W. (2009). Comparison of driving errors between on-the-road and simulated driving assessment: a validation study. *Traffic. Inj. Prev.* 10, 379–385. doi: 10.1080/15389580902894989

Shechtman, O. (2010). Validation of driving simulators. *Adv. Transp. Stud.* 53–62.

Spano, G., Caffò, A. O., Lopez, A., Mallia, L., Gormley, M., Innamorati, M., et al. (2019). Validating driver behavior and attitude measure for older Italian drivers and investigating their link to rare collision events. *Front. Psychol.* 10:368. doi: 10.3389/fpsyg.2019.00368

Stolwyk, R. J., Charlton, J. L., Triggs, T. J., Iansek, R., and Bradshaw, J. L. (2006). Neuropsychological function and driving ability in people with Parkinson's disease. *J. Clin. Exp. Neuropsyc.* 28, 898–913.

Thurnheer, R. (2011). Diagnostic approach to sleep-disordered breathing. *Expert Rev. Respir. Med.* 5, 573–589. doi: 10.1586/ers.11.46

Uc, E. Y., Rizzo, M., Anderson, S. W., Sparks, J. D., Rodnitzky, R. L., and Dawson, J. D. (2006a). Driving with distraction in Parkinson disease. *Neurology* 67, 1774–1780. doi: 10.1212/01.wnl.0000245086.32787.61

Uc, E. Y., Rizzo, M., Anderson, S. W., Sparks, J., Rodnitzky, R. L., and Dawson, J. D. (2006b). Impaired visual search in drivers with Parkinson's disease. *Ann. Neurol.* 60, 407–413. doi: 10.1002/ana.20958

Uc, E. Y., Rizzo, M., Anderson, S. W., Sparks, J. D., Rodnitzky, R. L., and Dawson, J. D. (2007). Impaired navigation in drivers with Parkinson's disease. *Brain* 130, 2433–2440. doi: 10.1093/brain/awm178

Van Eck, N. J., and Waltman, L. (2014). "Visualizing bibliometric networks," in *Measuring Scholarly Impact: Methods and Practice*, eds Y. Ding, R. Rousseau, and D. Wolfram (Berlin: Springer), 285–320.

Van Eck, N. J., Waltman, L., Dekker, R., and van den Berg, J. (2010). A comparison of two techniques for bibliometric mapping: multidimensional scaling and VOS. *J. Am. Soc. Inf. Sci. Tec.* 61, 2405–2416.

Van Raan, A. F. (2004). "Measuring science," in *Handbook of Quantitative Science and Technology Research*, eds H. F. Moed, W. Glänzel, and U. Schmoch (Dordrecht: Springer), 19–50.

Vargas-Quesada, B., Chinchilla-Rodríguez, Z., and Rodriguez, N. (2017). Identification and visualization of the intellectual structure in graphene research. *Front. Res. Metr. Anal.* 2:7.

Verster, J. C., and Roth, T. (2012). Predicting psychopharmacological drug effects on actual driving performance (SDLP) from psychometric tests measuring driving-related skills. *Psychopharmacology* 220, 293–301. doi: 10.1007/s00213-011-2484-0

Verster, J. C., Spence, D. W., Shahid, A., Pandi-Perumal, S. R., and Roth, T. (2011). Zopiclone as positive control in studies examining the residual effects of hypnotic drugs on driving ability. *Curr. Drug Saf.* 6, 209–218.

Verster, J. C., van de Loo, A. J., and Roth, T. (2015). Mirtazapine as positive control drug in studies examining the effects of antidepressants on driving ability. *Eur. J. Pharmacol.* 753, 252–256. doi: 10.1016/j.ejphar.2014.10.032

Verster, J. C., Veldhuijzen, D. S., and Volkerts, E. R. (2004). Residual effects of sleep medication on driving ability. *Sleep Med. Rev.* 8, 309–325.

Waltman, L., Van Eck, N. J., and Noyons, E. C. (2010). A unified approach to mapping and clustering of bibliometric networks. *J. Inform.* 4, 629–635.

Weaver, T. E. (2001). Outcome measurement in sleep medicine practice and research. Part 2: assessment of neurobehavioral performance and mood. *Sleep Med. Rev.* 5, 223–236.

Wickens, C. M., Smart, R. G., and Mann, R. E. (2014). The impact of depression on driver performance. *Int. J. Ment. Health Addiction.* 12, 524–537.

Worringham, C. J., Wood, J. M., Kerr, G. K., and Silburn, P. A. (2006). Predictors of driving assessment outcome in Parkinson's disease. *Mov. Disord.* 21, 230–235.

Wynne, R. A., Beanland, V., and Salmon, P. M. (2019). Systematic review of driving simulator validation studies. *Saf. Sci.* 117, 138–151. doi: 10.1016/j.ssci.2019.04.004

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 03 June 2020
doi: 10.3389/fpsyg.2020.01003

Brain Network Constancy and Participant Recognition: an Integrated Approach to Big Data and Complex Network Analysis

Lu Qiu[1,2]* *and Wenya Nan*[3]

[1] School of Finance and Business, Shanghai Normal University, Shanghai, China, [2] Department of Finance, East China University of Science and Technology, Shanghai, China, [3] Department of Psychology, College of Education, Shanghai Normal University, Shanghai, China

With the development of big data sharing and data standardization, electroencephalogram (EEG) data are increasingly used in the exploration of human cognitive behavior. Most of the existing studies focus on the changes of human brain network topology (the number of connections, degree distribution, clustering coefficient phantom) in various cognitive behaviors. However, there has been little exploration into the steady state of multi-cognitive behaviors and the recognition of multi-participant brain networks. To solve these two problems, we used EEG data of 99 healthy participants from the PhysioBank to study multi-cognitive behaviors. Specifically, we calculated the symbolic transfer entropy (STE) between 64 electrode sequences of EEG data and constructed the brain networks of various cognitive behaviors of each participant using the directed minimum spanning tree (DMST) algorithm. We then investigated the eigenvalue spectrum of the STE matrix of each individual's cognitive behavior. The results also showed that the spectrum distributions of different cognitive states of the same participant remained relatively stable, but those of the same cognitive state of different participants varied considerably, verifying the relative stability and uniqueness of the human brain network similar to a human's fingerprint. Based on these features, we used the spectral distribution set of 99 participants of various cognitive states as the original data set and developed a spectral distribution set scoring (SDSS) method to identify the brain network participants. It was found that most labels (69.35%) of the test participant with the highest score were identical to the labeled participant. This study provided further evidence for the existence of human brain fingerprints, and furnished a new approach for dynamic identification of brain fingerprints.

OPEN ACCESS

Edited by:
Pietro Cipresso,
Italian Auxological Institute (IRCCS),
Italy

Reviewed by:
Edson Filho,
University of Central Lancashire,
United Kingdom
Jamie Sleigh,
The University of Auckland, New
Zealand

****Correspondence:***
Lu Qiu
nuaaqiulu@shnu.edu.cn

Specialty section:
This article was submitted to
Quantitative Psychology and
Measurement,
a section of the journal
Frontiers in Psychology

Received: *07 April 2019*
Accepted: *22 April 2020*
Published: *03 June 2020*

Citation:
Qiu L and Nan W (2020) Brain Network Constancy and Participant Recognition: an Integrated Approach to Big Data and Complex Network Analysis. Front. Psychol. 11:1003.
doi: 10.3389/fpsyg.2020.01003

Keywords: complex network, symbolic transfer entropy (STE), directed minimum spanning tree (DMST), brain network constancy, participant recognition

1. INTRODUCTION

The human brain is a complex and dense network and as such, it has been explored with approaches ranging from 3D maps of brain circuitry (Landhuis, 2017), to communication dynamics in brain networks (Avena-Koenigsberger et al., 2018), and brain evolution (Sporns and Betzel, 2016; Thiran et al., 2016). The varied topological features of the brain network [modular structures (Hearne et al., 2017), network patterns (Vidaurre et al., 2017), nodes and edges (Kawagoe et al., 2017), and structural connectivity (Gu et al., 2018)] can be studied quantitatively (Moon et al., 2017) by techniques such as functional magnetic resonance imaging (fMRI) and electroencephalogram (EEG).

The fMRI (Kim et al., 2016; Wang et al., 2016) is an important quantitative tool to reveal regional functions of the brain. Hadley et al. used graph theory to study the change in brain network topology as a function of treatment response in schizophrenia (Hadley et al., 2016). Shi et al. applied independent component analysis to investigate the large-scale brain network connectivity underlying creativity through the task fMRI data (Shi et al., 2018). Gonzalez et al. validated the utility of the maximum entropy model in describing neurophysiological dynamics by measuring the activation rate in a separate resting state fMRI data set (Gonzalez et al., 2016). Emily et al., using the results of fMRI detection with functional connectivity as the classification standard, identified target participants from a large group of participants. Moreover, recognition was robust so that participants could be accurately identified in both the cognitive behavior and the resting state. They demonstrated that each person's brain connection profile is intrinsic and similar to a "fingerprint" that can be used for participant recognition (Huang J. et al., 2015). Takuya et al. constructed a functional connection network using fMRI detection data and defined the information transmission between the resting functional network and the cognitive behavior network as the transmission network. The information transmission characteristic was used to detect the relationship between the resting network and the cognitive network. It was concluded that the relationship between the cognitive behavior network and the static network was very close. In particular, the resting-state functional network provided a large amount of functional information for the cognitive information network (Ito et al., 2017). The above researchers, using the fMRI image processing and analysis technology, were able to detect the topological structure of participants in each cognitive state. However, the fMRI technique, with its high cost, excels mainly in spatial resolution, but is much less satisfactory with regards to time resolution, which is not conducive to studying brain network dynamics in different time periods.

By contrast, EEG is less accurate than fMRI in spatial positioning, but has a high time resolution at the scale of 1/100 s, lending itself particularly well to the time-window study of the brain network, especially to research brain network dynamics (Kluetsch et al., 2014; Yu et al., 2016; Zippo et al., 2018). Researchers often implemented filtering and independent component analysis (ICA) preprocessing on EEG data (Hatz et al., 2015), calculated the correlation between each two EEG signals, and set a threshold to create a brain network. The methods of calculating the correlation among electrode sequences include Pearson correlation coefficient, granger causality test (Farokhzadi et al., 2017), mutual information (Mikkelsen et al., 2017), and transfer entropy (Centeno and Carmichael, 2014). Among these methods, transfer entropy is the most suitable to reflect the non-linear relationship between brain electrodes. By calculating the transfer entropy between pairs of brain electrodes, one can construct the brain network of different time periods and participants by means of the threshold method or the minimum spanning tree (MST) method. Faes et al. applied entropy-based measures to quantify the predictive information in brain sub-systems and the heart system and identified a structured network of sleeping brain-brain and brain-heart interactions (Faes et al., 2014). Huang et al. calculated the transfer entropy between brain electrodes in drowsy and alert driving states. They concluded that the couplings between pairs of forehead, central lobe, and parietal areas were higher at the vigilance level than in the drowsy driving state (Huang C. S. et al., 2015). Qiao et al. constructed a brain network by fglasso and bootstrapped fglasso for both the alcoholic and the control groups. They found that links of electrodes in the frontal region were denser than those for the control group. In addition, more connected edges were detected in the left central and parietal regions of the alcoholic group (Qiao et al., 2019). Su et al. used MST to unveil the differences of brain network efficiency between young smokers and non-smokers and found that the global network efficiency decreased in young smokers (Su et al., 2017).

The above studies on EEG sequences were mainly based on the change of EEG network topology (network state, network connection number). But there is less research dedicated to quantitative grouping comparisons between EEG networks of cognitive behavior of each participant or considering individual differences among participants. In particular, to our knowledge, no studies have employed a combination of STE and SDSS in EEG analysis. In this study, we aim to investigate the EEG sequences of 99 healthy participants to verify the conclusion of Emily's study (Huang J. et al., 2015) by means of symbolic transfer entropy and spectral analysis. We also seek to explore the potential of using STE and SDSS in participant recognition based on fingerprint characteristics of EEG sequences.

2. MATERIALS AND METHODS

Ethics Approval

The datasets for this study are publicly available on https://www.physionet.org/physiobank/database/eegmmidb/ and can be used with no further permission[1] (Goldberger et al., 2000; Schalk et al., 2004). Since the data have been fully de-identified, no IRB approval is required.

EEG Data

The data set used in this study was created by the developers of the BCI2000 instrumentation system consisting of over 1,500 1- and 2-min EEG recordings, obtained from 99 healthy volunteers. For each participant, voltage values were measured from 64 electrodes as per the international 10-10 system (excluding electrodes Nz, F9, F10, FT9, FT10, A1, A2, TP9, TP10, P9, and P10), shown in **Figure 1**. All participants were required to perform 14 experimental runs listed in **Table 1**: two 1-min baseline runs (one with eyes open, one with eyes closed) and three 2-min runs of each of the four following tasks[1] (Goldberger et al., 2000; Schalk et al., 2004):

[1] https://www.physionet.org/cgi-bin/atm/ATM

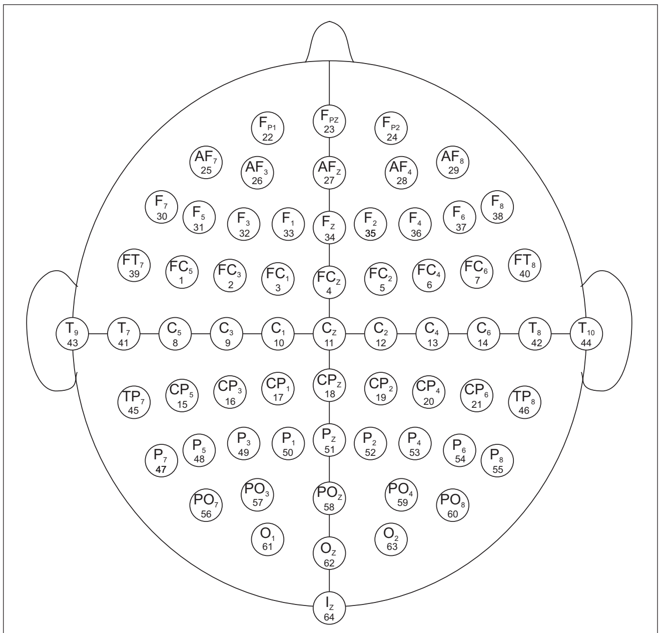


FIGURE 1 | EEG electrode diagram. The EEG are recorded from 64 electrodes which are created by international 10-10 system (excluding electrodes Nz, F9, F10, FT9, FT10, A1, A2, TP9, TP10, P9, and P10). The numbers below each electrode name demonstrate the order in which they appear in the records. The signals in the records are numbered from 1 to 64.

TABLE 1 | The 14 experimental runs constructed by different motor/imagery tasks.

NO.	Experimental runs	NO.	Experimental runs
1	Baseline, eyes open	8	Task 2
2	Baseline, eyes closed	9	Task 3
3	Task 1 (open and close left or right fist)	10	Task 4
4	Task 2 (imagine opening and closing left or right fist)	11	Task 1
5	Task 3 (open and close both fists or both feet)	12	Task 2
6	Task 4 (imagine opening and closing both fists or both feet)	13	Task 3
7	Task 1	14	Task 4

TABLE 2 | A participant of event table for task1.

Number	Event	Latency	Duration	Number	Event	Latency	Duration
1	1	1	672	**9**	1	10593	672
2	3	1313	656	**10**	3	11905	656
3	1	2609	672	**11**	1	13201	672
4	2	3921	656	**12**	2	14513	656
5	1	5217	672	**13**	1	15809	672
6	2	6529	656	**14**	3	17281	656
7	1	7825	672	**15**	1	18577	672
8	3	9297	656				

1. A target appears on either the left or the right side of the screen. The participant opens and closes the corresponding fist until the target disappears. Then the participant relaxes.
2. A target appears on either the left or the right side of the screen. The participant imagines opening and closing the corresponding fist until the target disappears. Then the participant relaxes.
3. A target appears on either the top or the bottom of the screen. The participant opens and closes either both fists (if the target is on top) or both feet (if the target is on the bottom) until the target disappears. Then the participant relaxes.
4. A target appears on either the top or the bottom of the screen. The participant imagines opening and closing either both fists (if the target is on top) or both feet (if the target is on the bottom) until the target disappears. Then the participant relaxes.

The EEG recordings were input to the EEGLAB toolbox. Each annotation includes one of three codes (e1, e2, or e3): e1 corresponds to rest, e2 corresponds to onset of motion (real or imagined) of the left fist (in runs 3, 4, 7, 8, 11, and 12) and both fists (in runs 5, 6, 9, 10, 13, and 14), and e3 refers to the onset of motion (real or imagined) of the right fist (in runs 3, 4, 7, 8, 11, and 12) and both feet (in runs 5, 6, 9, 10, 13, and 14). The 1-min-runs data of a participant in Task1 are listed in **Table 2**. Each EEG signal is sampled at 160 points per second. Events in **Table 2** include e1, e2, and e3. Latency means the start point of each event. For example, event 1 lasts until points 672, and then event 3 starts at point 1313 (with an intermission of 641 points). The duration means the time span of each event. Part of the corresponding data in Task1 is shown in **Figure 2**. The red region (event 1) indicates the opening of the eyes when the target appears. The green region (event 2) corresponds to opening the left fist when the target appears on the left. The pink one (event 3) indicates opening the right fist when the target appears on the right. The white one means rest. The horizontal and vertical axes represent the elapsed time (second) and names of electrodes, respectively.

EEG Signal Pre-Processing and Analysis

We defined the EEG data collection $\{G\}$ as follows:

$$G_p^t = \{g_{p,1}^t(c,e), g_{p,2}^t(c,e), ..., g_{p,N}^t(c,e)\} \tag{1}$$

where p is participant, t is task, c electrode, e event, and N the length of sequence. Prior to data analysis, we used eeglab (an interactive matlab toolbox) to filter the EEG sequence and ICA pretreatment. The frequency limit (Kluetsch et al., 2014) was chosen to be 1–70 HZ and 60 Hz notch filtering (Kawagoe et al., 2017). Filter order was automatically chosen (528 recommend) using the function *pop_eegfiltnew*()[2] in eeglab. We used the fully automatic algorithm based on the Independent Components analysis (ICA) algorithm (Mognon et al., 2011) to detect and remove artifacts from the filtered signals. Because the interference signals such as cardiac, eye movement artifacts, and electromyography (EMG) signals are generated by independent

[2] https://www.ccn.ucla.edu/wiki/index.php/Hoffman2:MATLAB:EEGLAB:Jobs.

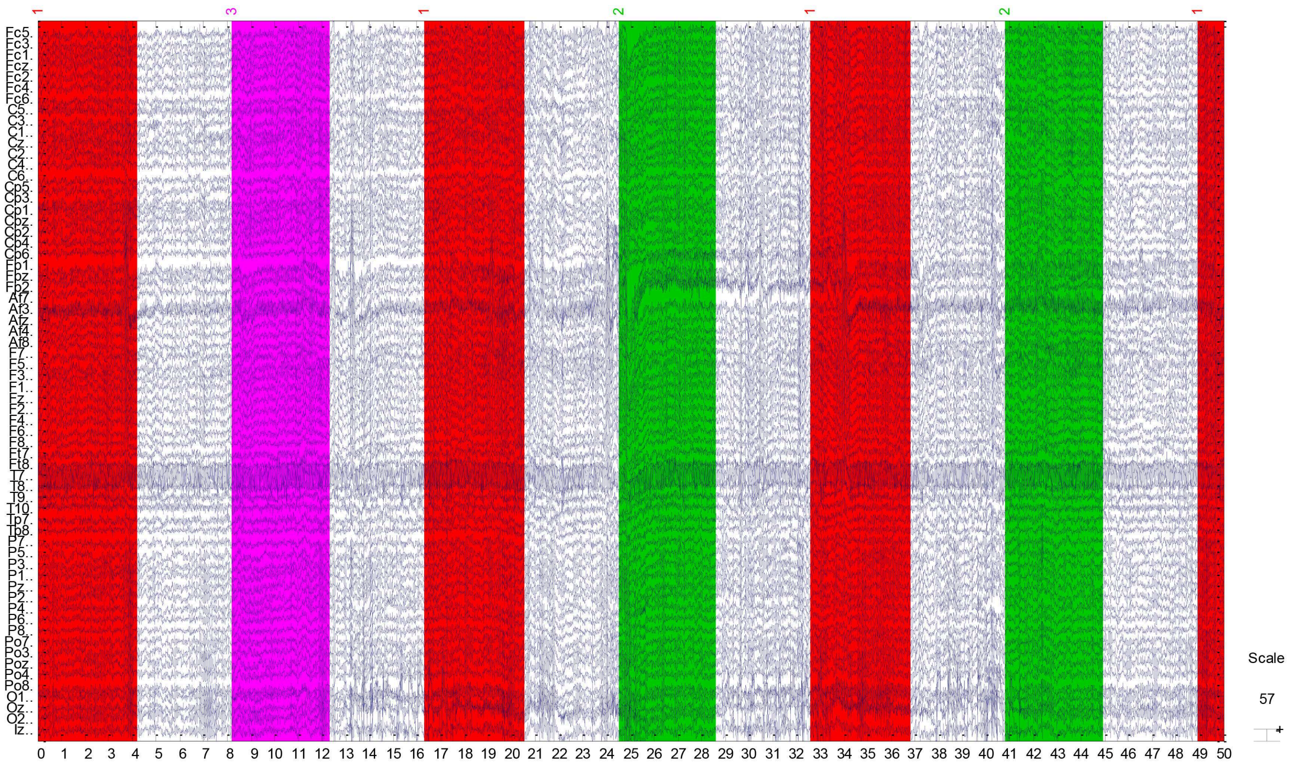


FIGURE 2 | Part signal diagram of task 1 (NO.3 in **Table 1**) of participant 1. The red, green, and violet regions indicate getting ready for the task, opening the left fist, and opening the right fist, respectively. The white regions mean rest. The horizontal axis shows the elapsed time (second) of the tasks, the vertical axis represents the names of electrodes.

sources, ICA decomposition can extract EEG signals from these interference signals. After treatment, the EEG sequence was named $\{GQ\}$,

$$GQ = \{gq^t_{p,1}(c,e), gq^t_{p,2}(c,e), ..., gq^t_{p,N}(c,e)\} \tag{2}$$

The final preprocessing was the first-order difference of the sequence $\{GQ\}$, and we obtained sequence $\{DQ\}$:

$$DQ = gq^t_{p,n+1}(c,e) - gq^t_{p,n}(c,e) = dq^t_{p,n}(c,e) \tag{3}$$

where $p = 1, 2 \ldots 99, n = 1, 2, \ldots, N-1, t = 1, 2, \ldots, 15$(14 experimental runs and 1 rest signal), $c = 1, 2, \ldots 64, e = 1, 2, 3$(events).

Symbolic Transfer Entropy (STE)

After pre-processing, we used transfer entropy to measure the dynamic non-linear relationship of sequences. Transfer entropy is used in many fields, such as the correlation between financial sequences, climate impacts, and EEG/electrocardiogram (ECG) signals. The general formula of transfer entropy is as follows:

$$TE^{(k,l)}_{y\to x} = \sum_{x_{n+1}, x^{(k)}_n, y^{(l)}_n} P(x_{n+1}, x^{(k)}_n, y^{(l)}_n) \log_2 \frac{P(x_{n+1}|x^{(k)}_n, y^{(l)}_n)}{P(x_{n+1}|x^{(k)}_n)} \tag{4}$$

where the sequence X is a Markov process of degree k, and Y is a Markov process of degree l. The element $x^{(k)}_n$ means that the sequence X is influenced by the k previous states,and $y^{(l)}_n$ indicates that the sequence Y is influenced by the l previous states. The parameters k and l are often set to 1. Then the transfer entropy from variable Y to variable X is defined as

$$\begin{aligned} TE_{y\to x} &= \sum_{x_{n+1}, x_n, y_n} P(x_{n+1}, x_n, y_n) \log_2 \frac{P(x_{n+1}|x_n, y_n)}{P(x_{n+1}|x_n)} \\ &= \sum_{x_{n+1}, x_n, y_n} P(x_{n+1}, x_n, y_n) \log_2 \frac{P(x_{n+1}, x_n, y_n)P(x_n)}{P(x_{n+1}, x_n)P(x_n, y_n)} \end{aligned} \tag{5}$$

where $P(A, B, C)$ is the joint probability of A, B, and C, and $P(A|B)$ is the conditional probability of A given by B. Before the calculation of transfer entropy, we translated the sequence $\{DQ\}$ into a symbol sequence. Specifically, we took one sequence from 64 channels for the same participant, same task, and same event as the target research object. For example, in **Figure 2**, the elapsed times from the 1st second to the 4th second (the horizontal axis) filled in red color means evet1 of task1 (shown in NO.3 of **Table 1**) of participant 1. We arranged the combined 64 signals in ascending order and divided these data points into three equal parts. The final forms were as follows:

$$B_{p,i} = \left\{ \begin{array}{c} 1: T^{t}_{p,1}(e) \leq dq^{t}_{p,n}(c,e) < T^{t}_{p,\frac{1}{3}L}(e) \\ \\ 2: T^{t}_{p,\frac{1}{3}L}(e) \leq dq^{t}_{p,n}(c,e) < T^{t}_{p,\frac{2}{3}L}(e) \\ \\ 3: T^{t}_{p,\frac{2}{3}L}(e) \leq dq^{t}_{p,n}(c,e) \leq T^{t}_{p,L}(e) \end{array} \right\} \quad (6)$$

where T is a new combined sequence of 64 signals. L means the length of sequence T. p, t, c, and e represent participant, task, electrode, and event, respectively. $p = 1, 2, 3, \ldots, 99$, $t = 1, 2, 3...15$, $c = 1, 2, 3, \ldots 64$, $e = 1, 2, 3$. We the used phase space reconstruction for symbol EEG signals and set the embedding dimension as 3 (Grassberger and Procaccia, 1983). The correlation between symbol EEG signals was expressed by the (Symbol Transfer Entropy)STE (McAuliffe, 2014).

Directed Minimum Spanning Tree (DMST)

By calculating the STE between each two EEG symbol sequences, we obtained the quantitative impact relationships between EEG signals. On this basis, the next vital step was to construct directed brain network diagrams. Using the threshold method to construct directed networks can depict certain brain network structures, but the network constructed by the threshold method is subjective and unstable. In order to ensure the consistency and objectivity of network connections, we made use of the DMST (Gabow et al., 1986; Kwon and Yang, 2008) method to construct the brain network. The minimum spanning tree (MST) algorithm (Crobe et al., 2016) is an important part of graph theory. The classical Kruskal and Prim algorithms of the undirected minimum spanning tree can solve the problem of the symmetrical adjacency matrix. Due to the asymmetry of the transfer entropy matrix, the relations between nodes can be described by DMST, also known as minimum arborescence (Hemminger, 1966). It assigns a special root node to the directed weighted graph. The DMST from the root node requires the minimum total weight of all distance weights. Steps of DMST algorithms are as follows:

1. Select a node as the root node randomly.
2. Travel all edges and find the smallest entry edges of all points except for the root node. Then sum up the weighted values of edges to form the new graph. Determine the final minimum arborescence if no cycles exist in the new graph.
3. If a ring exists in the new graph, shrink the ring into a point and change the edge weight. The way to change edge weights are as follows:

   (1). Choose a node u in the ring and set the incoming edge of this node as $in[u]$, and the outgoing edge of this one as (u, i, w). i and w refer to source node and weight, respectively.

   (2). Set the new edge weight of node u as $(u, i, w - in[u])$.

   (3). Return to Step 2 if the new weight graph contains rings.

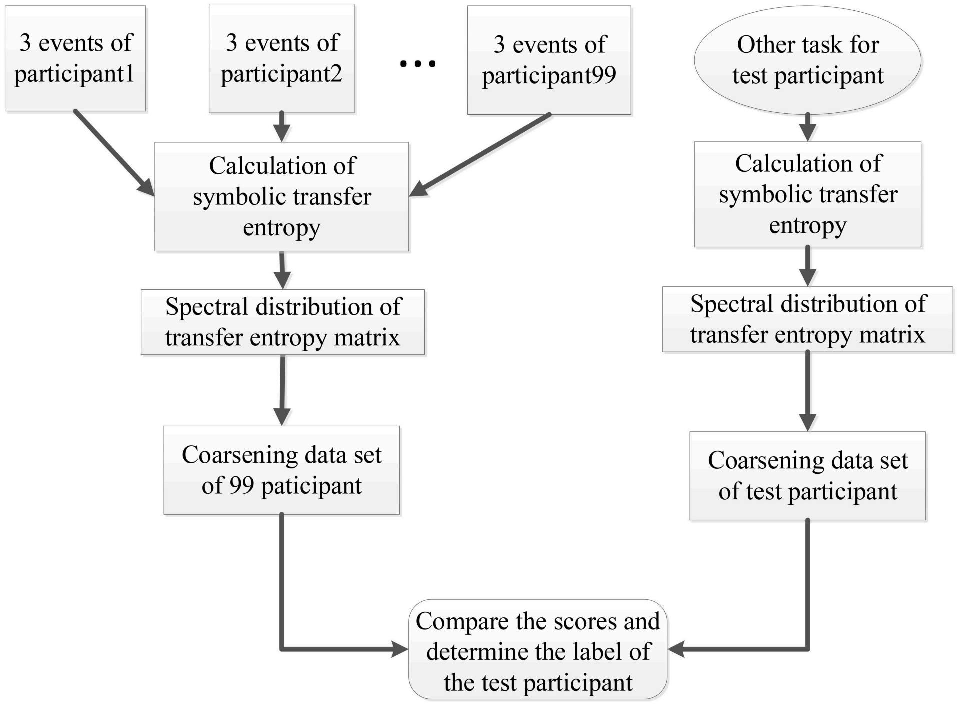

3 events of participant1
3 events of participant2
...
3 events of participant99
Other task for test participant
Calculation of symbolic transfer entropy
Calculation of symbolic transfer entropy
Spectral distribution of transfer entropy matrix
Spectral distribution of transfer entropy matrix
Coarsening data set of 99 participant
Coarsening data set of test participant
Compare the scores and determine the label of the test participant

FIGURE 3 | Participant recognition process using the SDSS method. The left part displays the process of constructing the coarsening data set, and the right one indicates the process of calculating the coarsening data set of the test participant. The last step expressed by the rounded rectangle shows the comparison of the data set of 99 participants and the data set of the test participant, to obtain the final score. This final score is used to determine the label of the test participant.

4. Expand the new graph if rings do not exist by the breaking loop method (Hemminger, 1966; Gabow et al., 1986). The steps of the breaking loop method were as follows:
   (1). Find a loop in the graph.
   (2). Remove the edge with the largest weight in the loop, but keep the graph connected.
   (3). Repeat this process until there are no loops in the graph (but they are still connected) and get the minimum spanning tree.

Average Euclidean Distance and Spectrum Distribution Set Scoring (SDSS)

The brain network constructed using the DMST method can reveal the relation between EEG channels of each participant in each action. The relative stability of events and the difference between participants can be observed in DMST graph. Because of the lack of quantitative analysis in the DMST method, we took the average Euclidean distances as the quantitative parameter indicating the distinctions between brain network patterns:

$$AD_p = \frac{\sqrt{\sum_{t=1}^{15}\sum_{e=1}^{3}\sum_{p_A=1}^{99}\sum_{p_B=1}^{99}\sum_{i=1,j=1}^{64}(TE_{i,j}^{p_A,e} - TE_{i,j}^{p_B,e})^2}}{15 \times 3 \times 99 \times 99 \times 64} \tag{7}$$

$$AD_e = \frac{\sqrt{\sum_{p=1}^{99}\sum_{t_A=1}^{15}\sum_{t_B=1}^{15}\sum_{e_A=1}^{3}\sum_{e_B=1}^{3}\sum_{i=1,j=1}^{64}[TE_{i,j}^{t_A,e_A}(p) - TE_{i,j}^{t_B,e_B}(p)]^2}}{99 \times 15 \times 15 \times 3 \times 3 \times 64} \tag{8}$$

where AD_p and AD_e indicate the average Euclidean distances of participants and the average Euclidean distances of events, respectively. e means an event in each task, e_A and e_B indicate two events in the same task or a different task ($e_A = e_B$ is allowed), p means a specific participant out of the 99 participants, p_A and p_B refer to two different participants or the same participant out of the 99 participants ($p_A = p_B$ is also allowed), and t_A and t_B correspond to two tasks from the total 15 tasks.

After quantitative analysis of differences between brain networks, we conducted a union analysis of the brain network by calculating the eigenvalue of the transfer entropy matrix for each participant and event as follows:

$$\lambda_p^t(e,c) = \alpha_p^t(e,c) + \beta_p^t(e,c) \bullet i \tag{9}$$

where α, β indicate real and imaginary parts of the eigenvalues, and p, t, e, and c represent participant, task, and event, respectively. $p = 1, 2...99$, $t = 1, 2...15$, $e = 1, 2, 3$, $c = 1, 2...64$. All the eigenvalues were normalized by the Z-Score method and the eigenvalue spectrum distribution of the transfer entropy matrix was shown by the real and imaginary eigenvalues of each action and participant on two-dimensional coordinates. On this basis, we observed and analyzed the spectral distributions of the same events of different participants and different events of the same participants.

At the same time, the eigenvalues of each action for each participant were conducted to data pre-processing through the coarse graining. First, we took the maximum ($\alpha_p^t(e)_{\max}$, $\beta_p^t(e)_{\max}$) and minimum ($\alpha_p^t(e)_{\min}$, $\beta_p^t(e,c)_{\min}$) of the real and the imaginary parts of the eigenvalues. Secondly, we defined the scale of coarse-graining θ. Then the ranges of the real part and the imaginary part were defined as $\{\alpha_p^t(e)_{\min} + \theta, \alpha_p^t(e)_{\min} + 2\theta, \ldots, \alpha_p^t(e)_{\max} - \theta, \alpha_p^t(e)_{\max}\}$ and $\{\beta_p^t(e)_{\min} + \theta, \beta_p^t(e)_{\min} + 2\theta, \ldots, \beta_p^t(e)_{\max} - \theta, \beta_p^t(e)_{\max}\}$ respectively. Finally, we counted the number of actual eigenvalues of different events and participants in this two-dimensional coarsening space. The

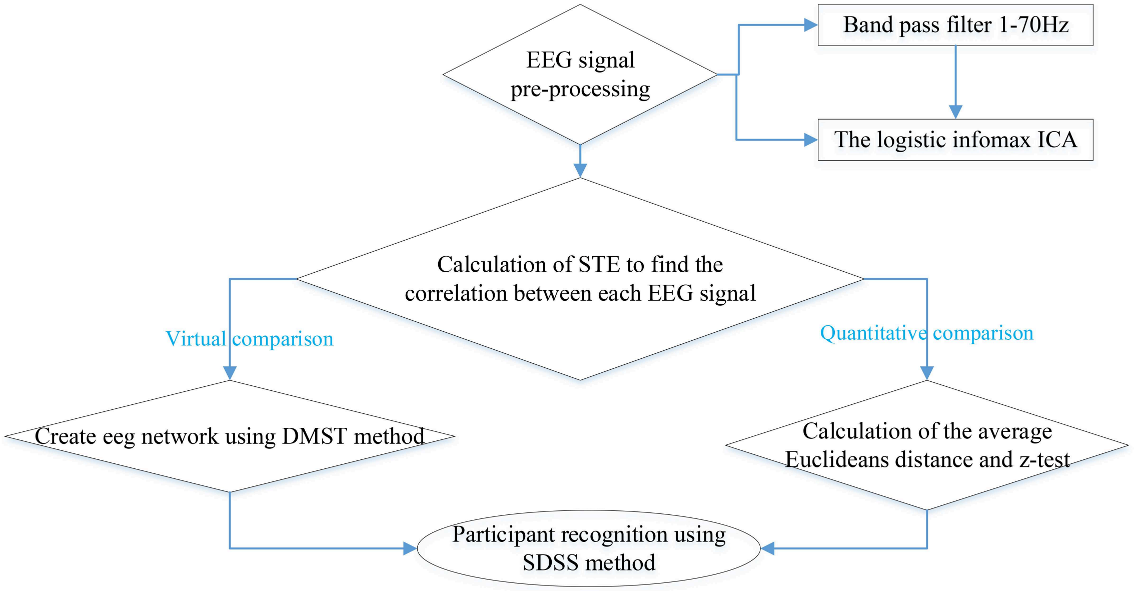

EEG signal pre-processing
Band pass filter 1-70Hz
The logistic infomax ICA
Calculation of STE to find the correlation between each EEG signal
Virtual comparison
Quantitative comparison
Create eeg network using DMST method
Calculation of the average Euclideans distance and z-test
Participant recognition using SDSS method

FIGURE 4 | Flowchart of each algorithm.

result was taken as a coarsening data set and used in participant recognition. For participant recognition, the full process of SDSS was shown in **Figure 3** with the following steps.

1. We calculated the STE between 64 electrode sequences of each event from the 99 participants and transformed the transfer entropy matrix into the spectral distribution.
2. We created a coarsening data set including the three events (task1) for each of the 99 participants.
3. We selected the data of a participant performing other tasks out of the 99 participants as the test data set and calculated the spectral distribution of the test data set.
4. Finally, we compared the scores and determined the label of the test participant.

The entire experiment process is illustrated by a flowchart (**Figure 4**).

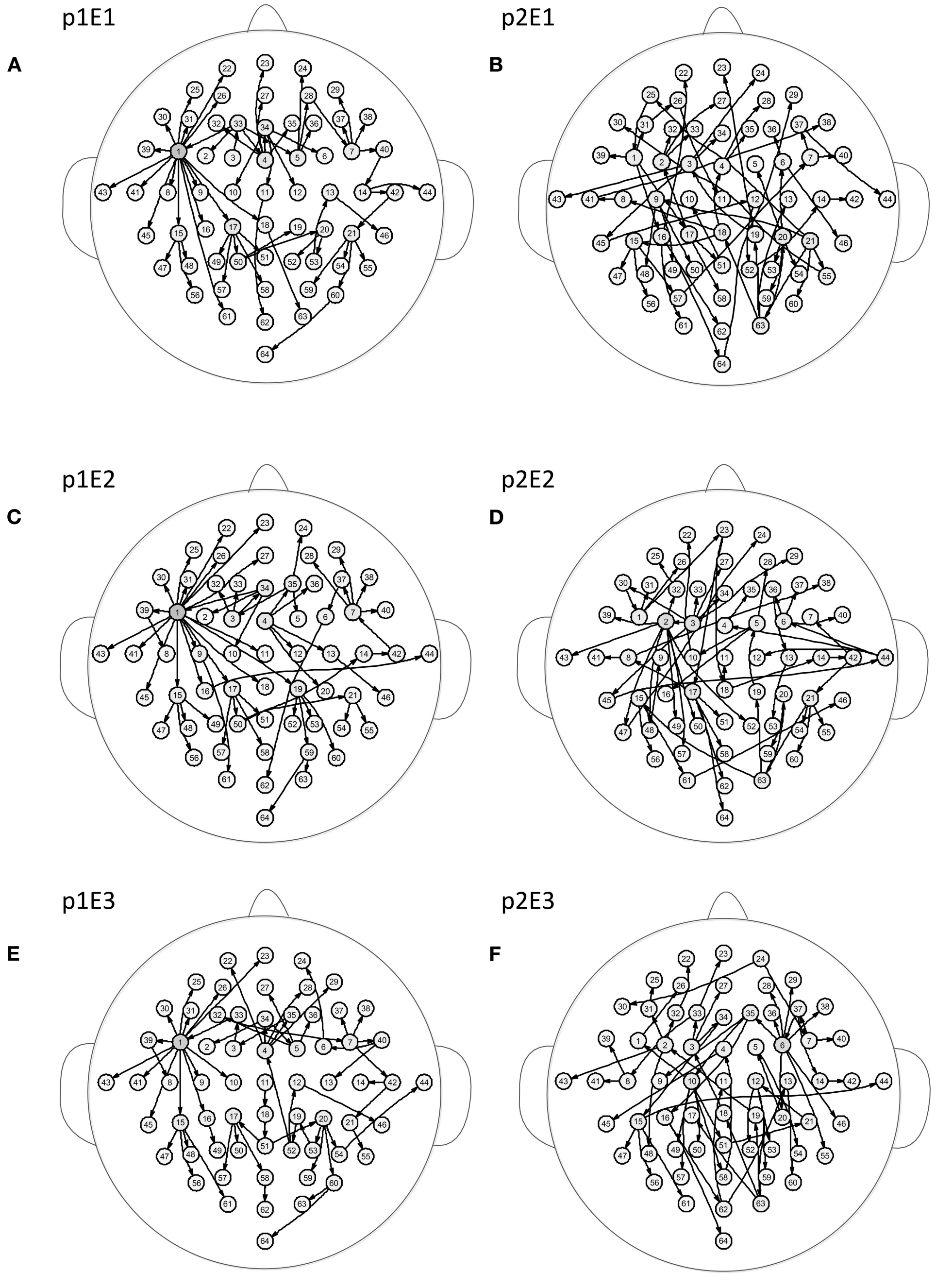


FIGURE 5 | Brain networks of three events of participant 1 and participant 2. The nodes from 1 to 64 correspond to **Figure 1**. **(A)** event 1 of participant 1 **(B)** event 1 of participant 2 **(C)** event 2 of participant 1 **(D)** event 2 of participant 2 **(E)** event 3 of participant 1 **(F)** event 3 of participant 2.

3. RESULTS

By means of the above methods, we transformed the EEG signal sequences of the 99 participants into symbolic sequences and calculated the STE of each participant and task. The transfer entropy matrix was transformed into brain networks using the DMST method.

Figure 5 shows the brain networks of the three events of task 1 for participant 1 and participant 2. For participant 1 in **Figure 5A**, the node 1(FC5) had the largest out degree which was then treated as the key node in the analysis. In this way, not only can the characteristics of the participants be studied, but the recognition of EEG fingerprints can also be facilitated. At the same time, it can be seen from **Figures 5A,C,E** that there

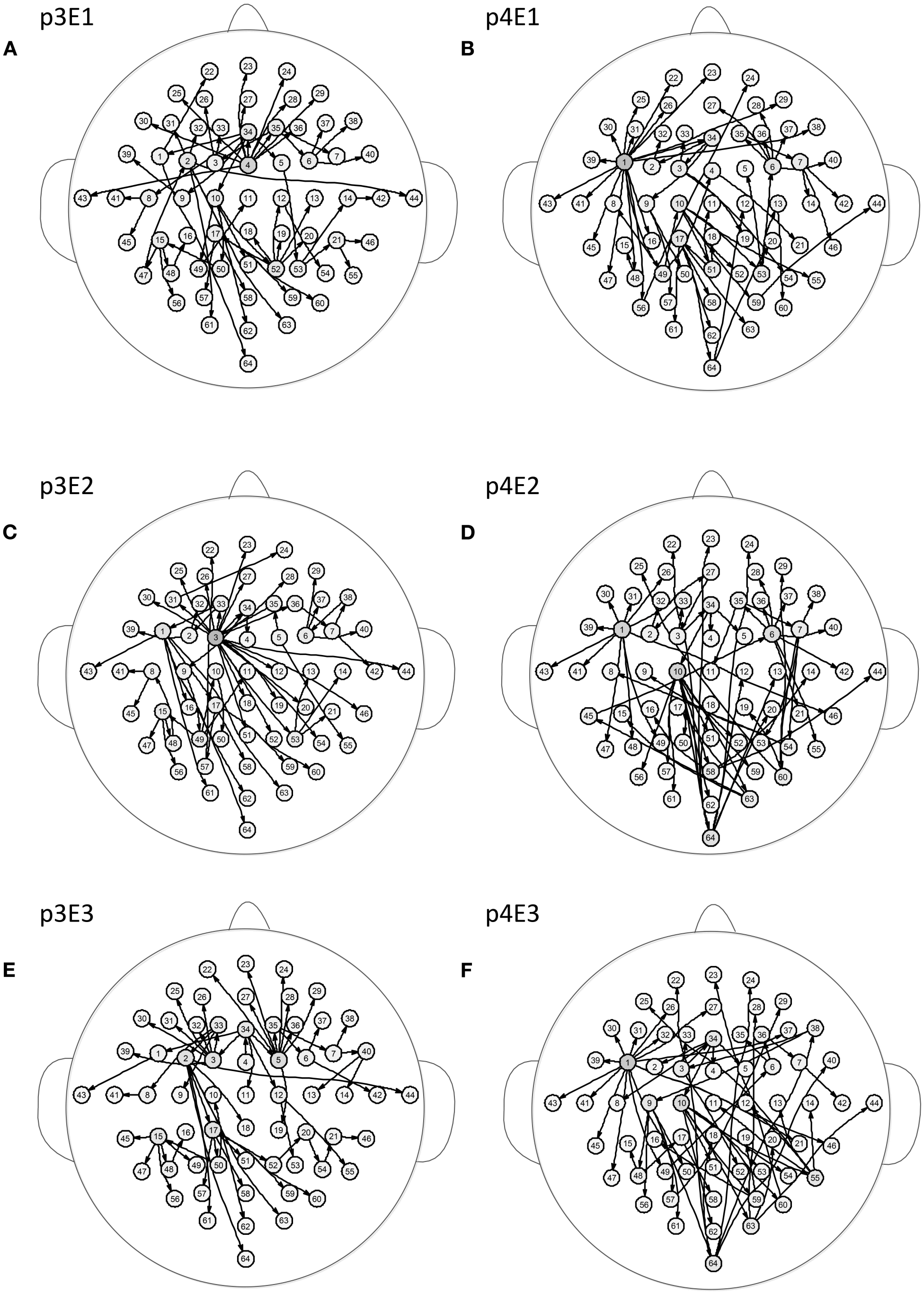


FIGURE 6 | Brain networks of three events of participant 3 and participant 4. The nodes from 1 to 64 correspond to **Figure 1**. **(A)** event 1 of participant 3 **(B)** event 1 of participant 4 **(C)** event 2 of participant 3 **(D)** event 2 of participant 4 **(E)** event 3 of participant 3 **(F)** event 3 of participant 4.

were little differences among the three brain network graphs of participant 1, which were basically in a constant state. In **Figures 5B,D,F**, the three brain network diagrams of participant 2 were also basically in a constant state, which showed that the brain network graphs of the same participant in different events had a certain degree of stability. But the same events from different participants, such as p1E1 (event1 of participant 1) and p2E1 (event1 of participant2) in **Figures 5A,B**, were widely different in structure. Similarly, in **Figures 6A,C,E**, the network diagrams of the three different events in participant 3 were similar. The three different events in participant 4 also resembled those in **Figures 6B,D,F**. But the same event of different participants, such as event 1 of participant 3 and participant 4, can be drastically different.

TABLE 3 | Euclidean distances among participants of the same event.

Event 1	**p1[a]**	**p2[a]**	**p3[a]**	**p4[a]**	**AED[b]**
p1	0	42.988	36.354	24.143	36.640
p2	42.988	0	46.512	45.141	
p3	36.354	46.512	0	24.704	
p4	24.143	45.141	24.704	0	
	p1	**p2**	**p3**	**p4**	**AED[b]**
Event 2					
p1	0	48.827	51.503	55.224	43.107
p2	48.827	0	42.520	31.579	
p3	51.503	42.520	0	28.989	
p4	55.224	31.579	28.989	0	
Event 3					
p1	0	41.039	34.286	36.366	35.767
p2	41.039	0	44.181	36.563	
p3	34.286	44.181	0	22.167	
p4	36.366	36.563	22.167	0	

[a] *p1, p2, p3, p4, indicate participant1, participant 2, participant 3, participant 4, respectively.*

[b] *AED, average of Euclidean distances.*

From the results of **Figures 5**, **6**, we can conclude that brain networks of the same participant remain constant to a certain extent regardless of task or rest. The network structures of different participants vary greatly, indicating that everyone has his or her own brain network distribution, similar to a fingerprint, thus lending support to the finding of Emily (Huang J. et al., 2015).

The superposition of brain networks can be used to verify the similarity of networks for different tasks of the same participant, but the error edges arose from the union process lead to information loss in the brain network research. In order to solve this problem, we calculated the eigenvalues of the transfer entropy matrix between EEG recordings of different tasks. The characteristic of the transfer entropy matrix was extracted and then the eigenvalue spectrum was superposed, which not only reveals the basic characteristics of the network, but also

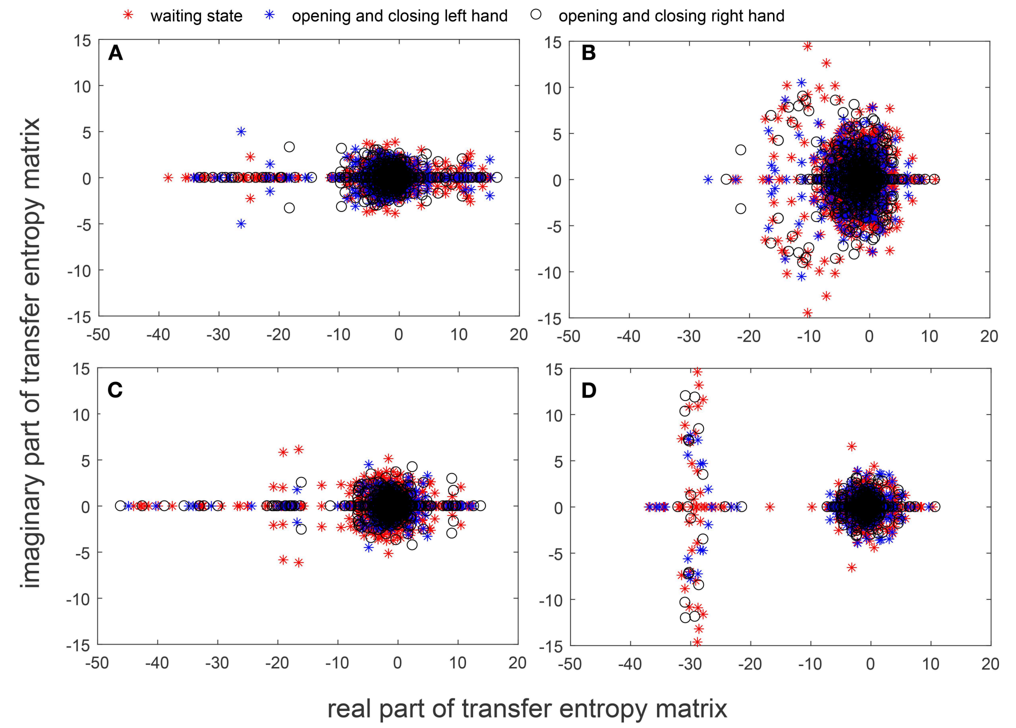


FIGURE 7 | Spectra graphs of transfer entropy matrix. The horizontal axis shows the real part of the transfer entropy matrix, while the vertical axis represents the imaginary part of the transfer entropy matrix. The red star, blue star, and black circle indicate waiting state, opening, and closing the left hand, and opening and closing right hand, respectively. **(A)** 3 events of participant 1 **(B)** 3 events of participant 2 **(C)** 3 events of participant 3 **(D)** 3 events of participant 4.

achieves the effect of superimposing the common characteristics. Because of the asymmetry of the transfer entropy matrix, the eigenvalues obtained include a real part and an imaginary part. The eigenvalues of different actions between the same participant were extracted and summarized on the coordinate axes.

Figures 7A–D show the spectral distribution of the three actions of participant 1, 2, 3, and 4, respectively. The red star means rest state, the blue star refers to moving the left hand, and the black circle indicates moving the right hand. It can be seen that the spectral structures of the network eigenvalues of the three events of the same participant were very similar, but the spectral structure of each participant obviously differed from each other. The Euclidean distances as quantitative indicators are shown in **Tables 3**, **4**. In **Table 3**, columns from 2 to 5 indicate the Euclidean distance between the first 4 participants of the same event. The results in column 6 of **Table 3** illustrates the mean value of the Euclidean distance of the first 4 participants on the same event. The results in the **Table 4** are the Euclidean distances among events of the same participant. Data in **Tables 3**, **4** are also the corresponding quantitative distances between the left and right networks in **Figures 5**, **6**. From these tables, it can be seen that the average Euclidean distances (36.640, 43.107, 35.767) of participants (from participant 1 to participant 4) in **Table 3** were all higher than those (24.792, 25.820, 9.320, 22.154) of events (event1, event2, and event3) in **Table 4**.

In order to statistically analyze the spectral distribution of all participants, we used the two-factor repeated measures ANOVA to test the differences between within-participant and between-participant spectra. Specifically, we transformed the spectrum distribution results into 5760-by-99 matrices (128*3*15 = 5760). The length of each spectrum distribution was 128 including the real part and the virtual part. The numbers of task and event were 15 and 3, respectively. Ninety-nine indicated the participant number. Then we put the matrix into the two-factor repeated measures ANOVA model and obtained the results shown in **Table 5**.

In **Table 5**, the $p - value$ of the participant factor (between-participant shown by Columns) in the second row was $1.61805 \times 10^{-10} < \alpha = 0.01$. In the third and fourth rows, the $p - value$s of the task factor (within-participant expressed with Rows) and interaction factor equaled $1 > \alpha = 0.01$. That means between-participant spectrum distributions were significantly different while the within-subject spectrum distributions had no significant difference.

We then obtained the quantitative result to confirm that inter-participant differences in the same event were more pronounced than inter-task differences of the same participant. As shown in **Figure 8**, the quantitative parameter indicating the average Euclidean distance among participants, shown by the red

TABLE 5 | The results of the two-factor repeated measures ANOVA.

Source	SS	df	MS	F	Prob>F
Columns	3137.88	98	32.0192	2.17	$1.61805E - 10$
Rows	16.24	44	0.3691	0.03	1
Interaction	757.01	4312	0.1756	0.01	1
Error	8330284.5	565785	14.7234		
Total	8334195.5	570239			

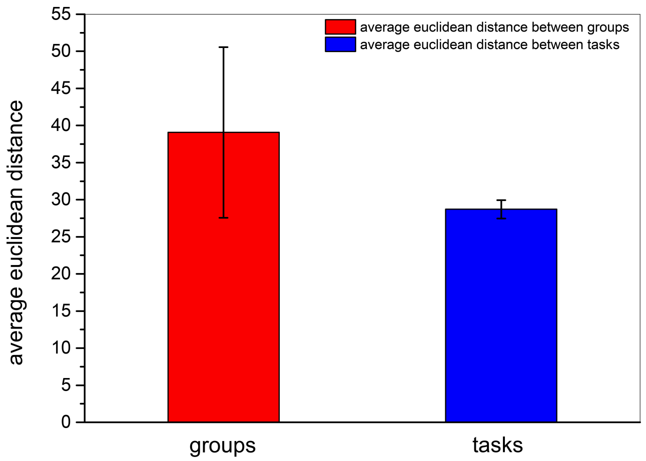


FIGURE 8 | Average Euclidean distance among different participants and among different tasks. The red column and the blue column indicate the average Euclidean distance among participants and the average Euclidean distance among tasks, respectively. The error bars indicate the standard deviations of average Euclidean distances.

TABLE 4 | Euclidean distances among events of the same participant.

Participant 1	Event 1	Event 2	Event 3	AED[a]	Participant 2	Event 1	Event 2	Event 3	AED[a]
Event1	0	20.232	27.630		event1	0	27.891	18.140	
Event2	20.232	0	26.514	24.792	event 2	27.891	0	31.430	25.820
Event3	27.630	26.514	0		event 3	18.140	31.430	0	
Participant 3	**Event 1**	**Event 2**	**Event 3**	**AED[a]**	**Participant 4**	**Event 1**	**Event 2**	**Event 3**	**AED[a]**
Event 1	0	9.954	5.977		event 1	0	24.310	17.820	
Event 2	9.954	0	12.027	9.320	event 2	24.310	0	24.333	22.154
Event 3	5.977	12.027	0		event 3	17.820	24.332	0	

[a] *AED, average of Euclidean distances.*

column, was higher than the average Euclidean distance among events represented by the blue column. The standard deviation within the participant group was also higher than that between event groups. In addition, we also compared the Euclidean distance among participants and the Euclidean distance among tasks by $z - test$. As presented in **Table 6**, the average Euclidean distance and standard deviation were the same as shown in **Figure 8**. The numbers of Euclidean distances were calculated as follows: $\frac{99*98}{2} = 4851$, $\frac{(15*3)*(15*3-1)}{2} = 990$, where 99 was the number of participants, 15 was task number, and each task contained three events. The z value was higher than the critical value of both one-tailed and two-tailed tests. The $p - value$ of the $z - test$ equaled 0. The results of the $z - test$ quantitatively demonstrated that the differences between brain networks of participants were larger than the differences between tasks.

TABLE 6 | $z - test$ analysis for two groups of Euclidean distances.

	Participants	Tasks
Average	39.471	29.362
Standard deviation	11.496	1.546
Numbers of Euclidean distances	4851	990
z	59.564	
P(Z<=z) one-tailed ($\alpha = 0.01$)	0	
z critical value of one-tailed	2.326	
P(Z<=z) two-tailed ($\alpha = 0.01$)	0	
z critical value of two-tailed	2.576	

Based on the relative stability of brain network of each participant, we used the SDSS method to create data sets using the network spectrum data of three events of 99 participants. When judging the test participants, any task of the test participants, such as moving both legs, can be used as measurement data. We compared the network spectrum structure of the measured participant with 99 participants' data set by coarsening the network spectrum. The choice of the accuracy of coarsening determines the accuracy of the final results. In this paper, we set

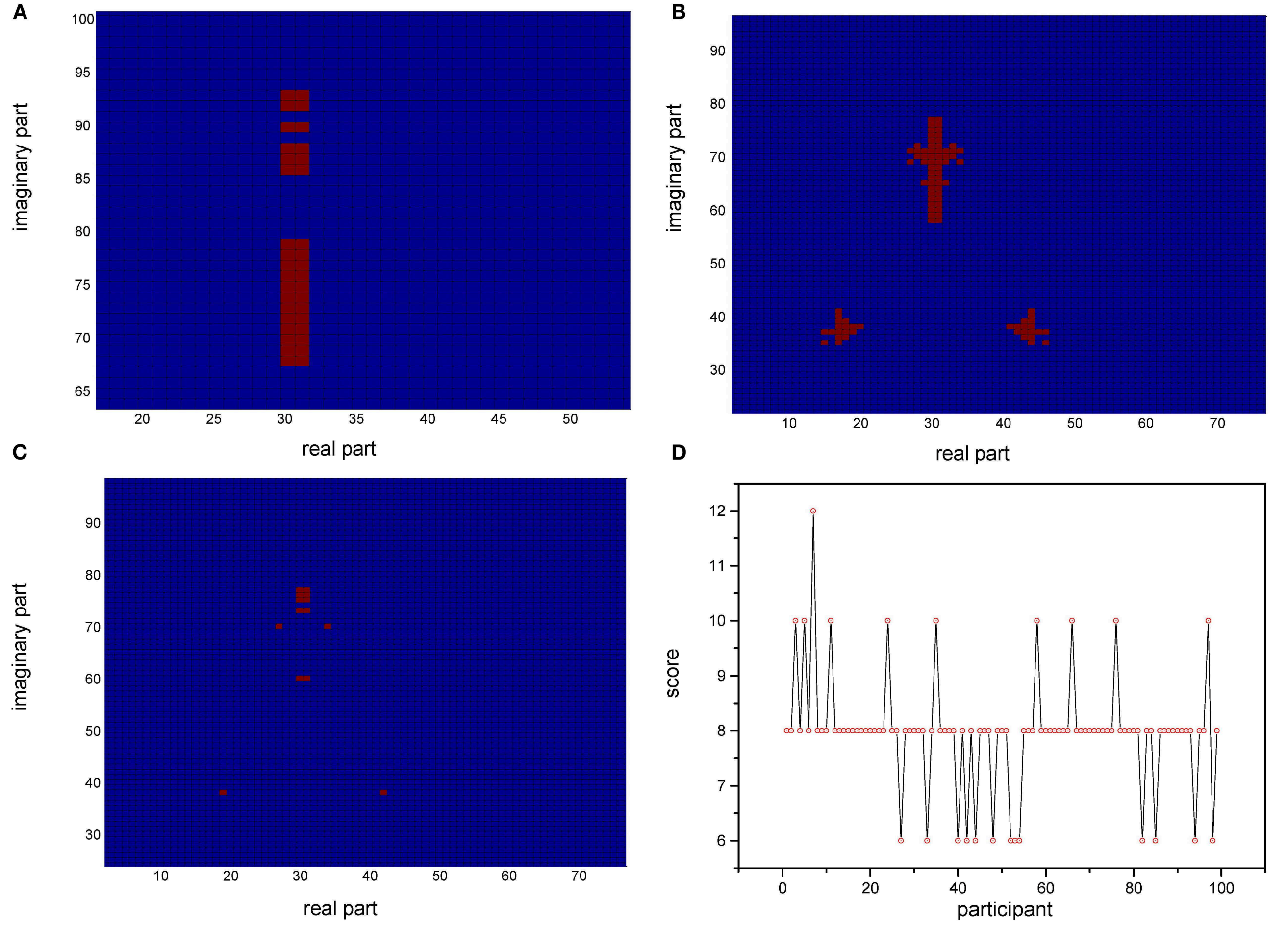


FIGURE 9 | Coarsened spectrum distributions and test scores. **(A)** Coarsened spectrum distribution of task 1 of participant 1. **(B)** Coarsened spectrum distribution of task 1 of participant 7. **(C)** Coarsened spectrum distribution of task 3 of test participant. **(D)** Score of *TXE*3 (task3 of test participant); the horizontal axis shows the participant number; the vertical axis indicates the score (overlapping part of the spectrum of *TXE*3 and data sets).

$\theta = 1$ to divide the spectrograms into various small squares and counted the number of particles in each small square. Finally, a participant test was carried out, assuming that the moving legs of participant 7 in task 3 were selected as measurement actions, labeled as *TXE*3. We calculated the transfer entropy matrix of this labeled task, whose spectrum distribution was coarsened by $\theta = 1$. By comparing the *TXE*3 coarsening data with 99 participants' coarsening data, the number of *TXE*3 was found with the highest score. **Figures 9A,B** show the spectrum distribution sets of participant 1 and participant 7, respectively. **Figure 9C** is the spectrum distribution set of *TXE*3. **Figure 9D** is the test score of *TXE*3. The horizontal axis represents the participant number, and the vertical axis score represents the overlapping part between the spectrum of *TXE*3 and data sets created by the three events of 99 participants. It can be seen that the highest score corresponds to participant 7. That is to say, the test participant was participant 7. This was consistent with the participant number selected beforehand. We also checked all participants of *T_new*1 (open and close both fists), *T_new*2 (open and close both feet), and *T_new*3 (imagine opening and closing both fists) by creating three new groups named *T_new*1_*g*, *T_new*2_*g*, and *T_new*3_*g*. Each group contained 99 participants of the new tasks (*T_new*1, *T_new*2 and *T_new*3). Thirty-three participants were selected without repetition from *T_new*1_*g*, *T_new*2_*g* and *T_new*3_*g*. A new cross test group was then created. We repeated the extraction 1,000 times and created 1,000 test groups. The 1,000 scores are shown in **Figure 10** and the average accuracy of test participants is 69.35%, which helped validate the effectiveness of the SDSS method.

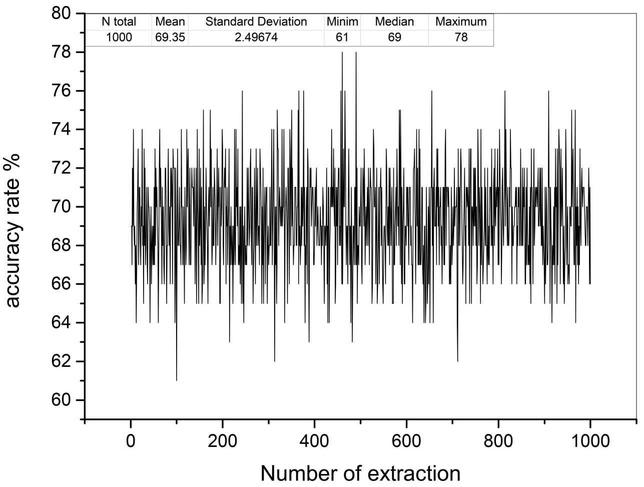


FIGURE 10 | Test accuracy rates of 1,000 test groups. The horizontal axis shows the number of extractions; the vertical axis indicates the accuracy rate (99 was divided by overlapping capacities of the spectrum of each test group and data sets).

4. DISCUSSION

EEG network research is regarded as an effective tool in identifying subject specific characteristics. As a core method for creating a network, the MST method assesses the strongest connection of individual EEG traits. Crobe et al. used MST and the k-core decomposition method to find the existence of a distinctive functional core. Their results confirmed the great impact of EEG analysis on several bioengineering applications (Crobe et al., 2016). Compared to the MST method, the DMST method can express the direction between each two nodes in the created EEG network. We can obtain the source node from the EEG network and find some features from it. Gennaro et al. found that the individual EEG-trait remains stable despite the change of sleep architecture. They proposed that EEG invariances can be related to genetic individual differences rather than sleep-dependent mechanisms (De Gennaro et al., 2005). Thomas et al. confirmed that the EEG signals are robust carriers of unique personality traits and reported that future research must focus on the uniqueness, acceptability, and robustness of EEG signals by various optimization algorithms and advanced technology (Thomas and Vinod, 2017). As mentioned in the above literature (De Gennaro et al., 2005; Huang J. et al., 2015; Thomas and Vinod, 2017): the connections in the human brain network are intrinsic and maintains a stable state, similar to the human "fingerprint." In our research, we also found these stable individual EEG traits using the graphic method (DMST) and quantitative analysis (z-test of Euclidean distance). Specifically, we used the eeglab toolbox in MATLAB to load the 20G EEG sequence data of 99 participants and preprocessed the data. The STE method was then used to calculate the transfer entropy of the three events for the 99 participants, and the DMST method was used to generate the brain networks of various cognitive behaviors for each participant. By visual inspection, brain networks of the same participant were very similar in different events, but there were great differences between different participants in the same event. For quantitative analysis, we used $z - test$ to compare Euclidean distances of participants and events. The results showed that the Euclidean distances between participants were significantly greater than those between events.

In addition, by focusing on this feature (EEG-trait remains stable), we used the SDSS method to construct the respective micro data sets (fingerprint database) based on the coarsened network spectrum of the rest, the left-hand and right-hand tasks of the 99 participants. For participant recognition, we created three groups of test data named by tasknew1 (open and close both fists), tasknew2 (open and close both feet), and tasknew3 (imagine opening and closing both fists). Each group contained 99 participants. We chose 33 different participants from group1, group2, and group3 randomly and created the new disordered group. We repeated the selection 1,000 times and obtained 1,000 new disordered groups. The average accuracy of test groups was 69.35%, which showed the effectiveness of the SDSS method.

5. LIMITATION

This present study is not without limitations: 1. In this paper, we selected the BCI2000 dataset as the research data, but BCI has a critical hurdle, in that performance varies greatly, especially in motor imagery based BCI. Researchers tried to address the problem of performance variation (Ahn and Jun, 2015) to improve reliability. In future studies, we look forward to improving the reliability and to focus the attention on task-related factors and longitudinal tracking of participants as well as integrative studies of related variables (psychological and physiological). 2. This study was limited in catching the flexible and dynamic characteristics of EEG signals when calculating the STE (McAuliffe, 2014). Further studies with the STE of short EEG sequences (about 10^2 points) (Zhang et al., 2012; Pan et al., 2014) would be required to avoid excessive reduction of brainwave features. 3. The accuracy of the coarse-grained network spectrograms of the 99 participants was likely to affect the final results, thus, in future work, we will try to select a better parameter not only to increase the accuracy of the coarse-grained network spectrogram but also to enhance the speed of identification.

6. CONCLUSION

In conclusion, the spectral analysis in complex networks can provide a very simple computational model for studying the rules of big data (multiple participants and multi-channel EEG). One can use the characteristics of the complex network spectrum to identify EEG participants. In addition, the SDSS method in this paper had important implications for the detailed comparison of network states.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The datasets for this study are publicly available on https://www.physionet.org/physiobank/database/eegmmidb/ and can be used with no further permission. Since the data have been fully de-identified, no IRB approval is required.

AUTHOR CONTRIBUTIONS

LQ designed the research and performed the calculations. LQ and WN analyzed the data and wrote the paper. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

The work is supported by The Youth Project of Humanities and Social Sciences Financed by Ministry of Education under Grant No. 19YJC190018 (WN) and 18YJC910010 (LQ), Research Projects of Humanities and Social Sciences of Shanghai Normal University under Grant No. A-7031-18-004023 (LQ) and the National Natural Science Foundation of China under Grant No. 81901830 (WN).

REFERENCES

Ahn, M., and Jun, S. C. (2015). Performance variation in motor imagery brain-computer interface: a brief review. *J. Neurosci. Methods* 243, 103–110. doi: 10.1016/j.jneumeth.2015.01.033

Avena-Koenigsberger, A., Misic, B., and Sporns, O. (2018). Communication dynamics in complex brain networks. *Nat. Rev. Neurosci.* 19, 17–33. doi: 10.1038/nrn.2017.149

Centeno, M., and Carmichael, D. W. (2014). Network connectivity in epilepsy: resting state fMRI and EEG-fMRI contributions. *Front. Neurol.* 5:93. doi: 10.3389/fneur.2014.00093

Crobe, A., Demuru, M., Didaci, L., Marcialis, G. L., and Fraschini, M. (2016). Minimum spanningtree and k-core decomposition as measure of subject-specific EEG traits. *Biomed. Phys. Eng. Express* 2:017001. doi: 10.1088/2057-1976/2/1/017001

De Gennaro, L., Ferrara, M., Vecchio, F., Curcio, G., and Bertini, M. (2005). An electroencephalographic fingerprint of human sleep. *Neuroimage* 26, 114–122. doi: 10.1016/j.neuroimage.2005.01.020

Faes, L., Nollo, G., Jurysta, F., and Marinazzo, D. (2014). Information dynamics of brain-heart physiological networks during sleep. *New J. Phys.* 16:105005. doi: 10.1088/1367-2630/16/10/105005

Farokhzadi, M., Soltanian-Zadeh, H., and Hossein-Zadeh, G. A. (2017). "Nonlinear Granger Causality using ANFIS for identification of causal couplings among EEG/MEG time series," in *2016 23rd Iranian Conference on Biomedical Engineering and 2016 1st International Iranian Conference on Biomedical Engineering, ICBME 2016* (Tehran), 69–73. doi: 10.1109/ICBME.2016.7890931

Gabow, H. N., Galil, Z., Spencer, T., and Tarjan, R. E. (1986). Efficient algorithms for finding minimum spanning trees in undirected and directed graphs. *Combinatorica* 6, 109–122. doi: 10.1007/BF02579168

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 101, 215–220. doi: 10.1161/01.cir.101.23.e215

Gonzalez, C. C., Billington, J., and Burke, M. R. (2016). The involvement of the fronto-parietal brain network in oculomotor sequence learning using fMRI. *Neuropsychologia* 87, 1–11. doi: 10.1016/j.neuropsychologia.2016.04.021

Grassberger, P., and Procaccia, I. (1983). Measuring the strangeness of strange attractors. *Phys. D Nonlinear Phenom.* 9, 189–208. doi: 10.1016/0167-2789(83)90298-1

Gu, S., Cieslak, M., Baird, B., Muldoon, S. F., Grafton, S. T., Pasqualetti, F., et al. (2018). The energy landscape of neurophysiological activity implicit in brain network structure. *Sci. Rep.* 8:2507. doi: 10.1038/s41598-018-20123-8

Hadley, J. A., Kraguljac, N. V., White, D. M., Ver Hoef, L., Tabora, J., and Lahti, A. C. (2016). Change in brain network topology as a function of treatment response in schizophrenia: a longitudinal resting-state fMRI study using graph theory. *npj Schizophr.* 2:16014. doi: 10.1038/npjschz.2016.14

Hatz, F., Hardmeier, M., Bousleiman, H., Regg, S., Schindler, C., and Fuhr, P. (2015). Reliability of fully automated versus visually controlled pre- and post-processing of resting-state EEG. *Clin. Neurophysiol.* 126, 268–274. doi: 10.1016/j.clinph.2014.05.014

Hearne, L. J., Cocchi, L., Zalesky, A., and Mattingley, J. B. (2017). Reconfiguration of brain network architectures between resting-state and complexity-dependent cognitive reasoning. *J. Neurosci.* 37, 8399–8411. doi: 10.1523/jneurosci.0485-17.2017

Hemminger, R. L. (1966). On the group of a directed graph. *Can. J. Math.* 18, 210–220. doi: 10.4153/cjm-1966-023-2

Huang, C. S., Pal, N. R., Chuang, C. H., and Lin, C. T. (2015). Identifying changes in EEG information transfer during drowsy driving by transfer entropy. *Front. Hum. Neurosci.* 9:570. doi: 10.3389/fnhum.2015.00570

Huang, J., Finn, E. S., Chun, M. M., Scheinost, D., Shen, X., Constable, R. T., et al. (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat. Neurosci.* 18, 1664–1671. doi: 10.1038/nn.4135

Ito, T., Kulkarni, K. R., Schultz, D. H., Mill, R. D., Chen, R. H., Solomyak, L. I., et al. (2017). Cognitive task information is transferred between brain regions via resting-state network topology. *Nat. Commun*. 8:1027. doi: 10.1038/s41467-017-01000-w

Kawagoe, T., Onoda, K., and Yamaguchi, S. (2017). Associations among executive function, cardiorespiratory fitness, and brain network properties in older adults. *Sci. Rep*. 7:40107. doi: 10.1038/srep40107

Kim, S. Y., Qi, T., Feng, X., Ding, G., Liu, L., and Cao, F. (2016). How does language distance between L1 and L2 affect the L2 brain network? An fMRI study of Korean-Chinese-English trilinguals. *Neuroimage* 129,25–39. doi: 10.1016/j.neuroimage.2015.11.068

Kluetsch, R. C., Ros, T., Théberge, J., Frewen, P. A., Calhoun, V. D., Schmahl, C., et al. (2014). Plastic modulation of PTSD resting-state networks and subjective wellbeing by EEG neurofeedback. *Acta Psychiatr. Scand*. 130, 123–136. doi: 10.1111/acps.12229

Kwon, O., and Yang, J. S. (2008). Information flow between stock indices. *EPL* 82:68003. doi: 10.1209/0295-5075/82/68003

Landhuis, E. (2017). Neuroscience: big brain, big data. *Nature* 541, 559–561. doi: 10.1038/541559a

McAuliffe, J. (2014). The new math of EEG: Symbolic transfer entropy, the effects of dimension. *Clin. Neurophysiol*. 125:17. doi: 10.1016/j.clinph.2013.12.017

Mikkelsen, K. B., Kidmose, P., and Hansen, L. K. (2017). On the Keyhole hypothesis: high mutual information between ear and scalp EEG. *Front. Hum. Neurosci.* 11:341. doi: 10.3389/fnhum.2017.00341

Mognon, A., Jovicich, J., Bruzzone, L., and Buiatti, M. (2011). ADJUST: an automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology* 48, 229–240. doi: 10.1111/j.1469-8986.2010.01061.x

Moon, J. Y., Kim, J., Ko, T. W., Kim, M., Iturria-Medina, Y., Choi, J. H., et al. (2017). Structure shapes dynamics and directionality in diverse brain networks: mathematical principles and empirical confirmation in three species. *Sci. Rep*. 7:46606. doi: 10.1038/srep46606

Pan, X., Hou, L., Stephen, M., Yang, H., and Zhu, C. (2014). Evaluation of scaling invariance embedded in short time series. *PLoS ONE* 9:e116128. doi: 10.1371/journal.pone.0116128

Qiao, X., Guo, S., and James, G. M. (2019). Functional graphical models. *J. Am. Stat. Assoc.* 114, 211–222. doi: 10.1080/01621459.2017.1390466

Schalk, G., McFarland, D.J., Hinterberger, T., Birbaumer, N., Wolpaw, J.R. (2004). BCI2000: a general-purpose brain-computer interface (BCI) system. *IEEE Trans. Biomed. Eng.* 51, 1034–1043.doi: 10.1109/TBME.2004.827072

Shi, L., Sun, J., Xia, Y., Ren, Z., Chen, Q., Wei, D., et al. (2018). Large-scale brain network connectivity underlying creativity in resting-state and task fMRI: cooperation between default network and frontal-parietal network. *Biol. Psychol*. 135, 102–111. doi: 10.1016/j.biopsycho.2018.03.005

Sporns, O., and Betzel, R. F. (2016). Modular brain networks. *Annu. Rev. Psychol*. 67, 613–640. doi: 10.1146/annurev-psych-122414-033634

Su, S., Yu, D., Cheng, J., Chen, Y., Zhang, X., Guan, Y., et al. (2017). Decreased global network efficiency in young male smoker: an EEG study during the resting state. *Front. Psychol*. 8:1605. doi: 10.3389/fpsyg.2017.01605

Thiran, J.-P., Fischi-Gomez, E., Eixarch, E., Batalle, D., Hüppi, P. S., Gratacós, E., et al. (2016). Structural brain network reorganization and social cognition related to adverse perinatal condition from infancy to early adolescence. *Front. Neurosci.* 10:560. doi: 10.3389/fnins.2016.00560

Thomas, K. P., and Vinod, A. P. (2017). Toward EEG-based biometric systems: the great potential of brain-wave-based biometrics. *IEEE Syst. Man Cybern. Mag*. 3, 6–15. doi: 10.1109/msmc.2017.2703651

Vidaurre, D., Smith, S. M., and Woolrich, M. W. (2017). Brain network dynamics are hierarchically organized in time. *Proc. Natl. Acad. Sci. U.S.A.* 114, 12827–12832. doi: 10.1073/pnas.1705120114

Wang, L., Wu, L., Lin, X., Zhang, Y., Zhou, H., Du, X., et al. (2016). Altered brain functional networks in people with Internet gaming disorder: evidence from resting-state fMRI. *Psychiatry Res. Neuroimaging* 254, 156–163. doi: 10.1016/j.pscychresns.2016.07.001

Yu, Q., Wu, L., Bridwell, D. A., Erhardt, E. B., Du, Y., He, H., et al. (2016). Building an EEG-fMRI multi-modal brain graph: a concurrent EEG-fMRI study. *Front. Hum. Neurosci.* 10:476. doi: 10.3389/fnhum.2016.00476

Zhang, W., Qiu, L., Xiao, Q., Yang, H., Zhang, Q., and Wang, J. (2012). Evaluation of scale invariance in physiological signals by means of balanced estimation of diffusion entropy. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys*. 86:056107. doi: 10.1103/PhysRevE.86.056107

Zippo, A. G., Della Rosa, P. A., Castiglioni, I., and Biella, G. E. M. (2018). Alternating dynamics of segregation and integration in human EEG functional networks during working-memory task. *Neuroscience* 371, 191–206. doi: 10.1016/j.neuroscience.2017.12.004

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 04 June 2020
doi: 10.3389/fpsyg.2020.01136

Identifying Qualitative Between-Subject and Within-Subject Variability: A Method for Clustering Regime-Switching Dynamics

Lu Ou[1]*, *Alejandro Andrade*[1], *Rosa A. Alberto*[2], *Arthur Bakker*[2] *and Timo Bechger*[1]

[1] ACTNext by ACT, Inc., Iowa City, IA, United States, [2] Department of Mathematics, Freudenthal Institute, Utrecht University, Utrecht, Netherlands

Technological advancement provides an unprecedented amount of high-frequency data of human dynamic processes. In this paper, we introduce an approach for characterizing qualitative between and within-subject variability from quantitative changes in the multi-subject time-series data. We present the statistical model and examine the strengths and limitations of the approach in potential applications using Monte Carlo simulations. We illustrate its usage in characterizing clusters of dynamics with phase transitions with real-time hand movement data collected on an embodied learning platform designed to foster mathematical learning.

Keywords: clustering, regime-switching model, functional data analysis, time-series data, dynamic model

Edited by:
Pietro Cipresso,
Italian Auxological Institute (IRCCS),
Italy

Reviewed by:
Rubén Maneiro,
Pontifical University of Salamanca,
Spain
Peida Zhan,
Zhejiang Normal University, China

****Correspondence:***
Lu Ou
lu.ou@act.org

Specialty section:
This article was submitted to
Quantitative Psychology and
Measurement,
a section of the journal
Frontiers in Psychology

Received: *08 December 2019*
Accepted: *04 May 2020*
Published: *04 June 2020*

Citation:
Ou L, Andrade A, Alberto RA,
Bakker A and Bechger T (2020)
Identifying Qualitative
Between-Subject and Within-Subject
Variability: A Method for Clustering
Regime-Switching Dynamics.
Front. Psychol. 11:1136.
doi: 10.3389/fpsyg.2020.01136

1. INTRODUCTION

Human dynamic processes vary within a subject over time and differ between subjects at all behavioral, physiological, emotional, attentional, and cognitive levels (Molenaar et al., 2003). Widespread examples include but not limited to change processes in belief and attitudes (van der Maas et al., 2003; Jansen et al., 2007), affective experiences (Cole et al., 2004; Kuppens et al., 2010; Hamaker et al., 2015), and executive functions (Zelazo, 2016). The within- and between-subject variabilities can be quantitative as well as qualitative in nature (Pintrich, 1988; Van Geert, 1991; van der Maas and Molenaar, 1992; van Dijk and van Geert, 2007; Stephen et al., 2009). For instance, human development is continuous and quantitative with gradual and incremental growth but simultaneously is discontinuous and qualitative as new forms and abilities emerge (Thelen and Smith, 1994). Inter-individual differences are also quantitative as no two individuals are identical within a population, and qualitative as subgroups of individuals may exist and share similar characteristics (Ram and Grimm, 2009; Bulteel et al., 2016). In order to understand the essence and drivers of human processes, researchers argue for a need to focus on studying and interpreting qualitative variability (Kelso, 2000). However, limited labor resources and subjectivity issues often put constraints on qualitative approaches (e.g., interviews and focus groups) that quest directly for qualitative findings. Alternatively, we infer qualitative changes and differences from data using quantitative methods that bring objectivity and computational accuracy and efficiency.

To this aim, we need mathematical and statistical models that represent both quantitative and qualitative within- and between-subject variability in the processes of interest and the data we collect. Mathematically, quantitative variability is often accommodated in continuous variables, while qualitative variability in categorical variables. The former refers to the within-subject numerical changes (including process noise) and between-subject random effects. In contrast, the latter refers to the within-subject regime (or phase) transitions and between-subject cluster (or group) differences. A cluster or group is a class of subjects that share similar qualities or dynamic patterns. A regime or phase is a within-subject time-varying class of dynamics that may switch from one to another as time passes. We use regime switches and phase transitions interchangeably. In our definitions, a regime or phase is different from a stage, which is a course in one-directional and non-reversible class transitions, such as age-based developmental stages.

Dynamic processes may exhibit qualitatively different cluster-wise quantitative changes interspersed with qualitative regime-switching. As an example, we consider students' learning processes that occur with dynamic sensorimotor coordination in an embodied learning environment. In a typical task, students acquire the concept of proportionality by coordinating movements of both their hands. Previous research found dynamic patterns and solution strategies from the patterns in students' action-coordination that relate to Piaget's theorized phases of reflective abstraction (Abrahamson et al., 2014; Duijzer et al., 2017; Pardos et al., 2018). The hand movements represent within-subject quantitative variability, and the strategies used form qualitatively switching regimes. High-performing and low-performing students may differ in the types and sequences of strategies used, thus displaying cluster-wise regime-switching patterns of hand movement dynamics. Hence, some interesting qualitative findings, in this case, are not only distinct regimes in students' strategy use or knowledge development but also clusters of regime-switching trajectories that are indicative of student's learning and have implications for interventions.

Many existing mathematical models only consider quantitative changes. For example, auto-regressive moving-average models and differential equations models represent quantitative within-subject changes in time-series data (Chow et al., 2005, 2011; Voelkle and Oud, 2013; Hu et al., 2014; Bulteel et al., 2016). The two types of modeling frameworks differ in whether the time in the model is discrete or continuous. They both are parametric models that exert top-down assumptions on the mechanism of change. In contrast, non-parametric methods like functional data analysis (Ramsay and Silverman, 2005) provide a bottom-up, data-driven way to approximate the dynamic changes directly using a combination of curves or smooth functions. Extensions and applications of these models allow for quantitative between-subject differences. For instance, when we apply these methods to a single subject's time-series data, we naturally allow each subject to have a unique set of parameters. When a law of change applies to the whole sample, we can include random effects that follow a specific statistical distribution to account for variability in model parameters (Oravecz et al., 2011; Lu et al., 2015; Chow et al., 2016; Ou et al., 2019).

Also, models that consider qualitative changes deal with clusters between subjects, or regimes within a subject but do not integrate the two. To capture qualitative between-subject variability, researchers use finite mixture models (McLachlan and Peel, 2004) to accommodate group differences by introducing a latent categorical variable that governs the emission of observed data. A finite mixture model assumes that a subject's data come from different latent groups with a particular set of probabilities. In each group, the emission of observed data follows different statistical distributions. In social and behavioral sciences, finite mixture models have been applied to identify latent groups with distinct means and covariance structures (Collins and Lanza, 2010), and factor structures (Lubke and Muthen, 2005; Hallquist and Wright, 2014). By incorporating assumptions on the longitudinal structure of quantitative changes, extensions of finite mixture models have been used to cluster subjects based on different growth trajectories (Colder et al., 2002; Muthen, 2004; Ram and Grimm, 2009), and dynamic emotional patterns in close relationships (Liu et al., under review).

Hidden Markov models are another standard class of models to analyze within-subject qualitative phase transitions. They have been widely applied in social and behavioral sciences to understand cognitive processes (Vermunt et al., 1999; Böckenholt, 2005; Dutilh et al., 2010; Visser, 2011; Visser and Speekenbrink, 2014; Andrade et al., 2017; Shu et al., 2017; Deonovic et al., 2018; Wang et al., 2018; Arieli-Attali et al., 2019). Hidden Markov models are extensions of the finite mixture models as the observed variables follow a mixture distribution depending on a latent categorical variable. The added feature is that the latent categorical variable can transition from one state to another in a first-order Markov chain, where the current state only depends on the previous state. Similar to finite mixture models, initial regimes and regime transitions are interpreted based on probabilities and the effects of covariates on these probabilities. As an extension of the hidden Markov model that considers longitudinal quantitative changes, regime-switching dynamic models permit modeling of manifest variables with discrete- or continuous-time equations rather than single emissions. Previous applications of the regime-switching models include the application of a regime-switching autoregressive model to facial Electromyography data to identify deactivated and activated emotional states (Yang and Chow, 2010), and the use of regime-switching differential equations to represent the regime transitions between exploration and proximity seeking of a child in mother-child interactions (Chow et al., 2018).

Despite the above developments, methods for simultaneously capturing both within- and between-subject qualitative variability (i.e., clusters and regimes) in time-series data with quantitative changes are nascent in social and behavioral sciences. In these fields, the quality of the data largely depends on the intrinsic complexity in human processes, the quality of measures (e.g., reliability and validity), and other economic, ecological, ethical, and privacy issues in data collection. As intensive longitudinal methods (Bolger and Laurenceau, 2013) become prevalent, an increasing number of data occur naturally at time points that are irregularly spaced within a subject and vary in the total number across subjects. Hence, data issues such as sample size (in terms of the number of subjects and number of measurements), noise, and missing data present challenges in applications of quantitative methods. In particular, while many clustering techniques require an equal dimension of data across subjects, data manipulation, including aggregation and imputation, is almost inevitable. Researchers that are interested in applying the methods need to understand whether the techniques are robust to the various data conditions they encounter and how the accuracy of the techniques varies with the data manipulation decisions that they have to make.

In this paper, we introduce and tailor an approach for characterizing qualitative between- and within-subject differences from quantitative changes to typical social and behavioral applications. We aim to present an elegant example for educational purposes and offer general guidance to researchers who wish to use the approach in their work. The approach is called the mixture of regressions with hidden logistic processes (mixRHLP; Chamroukhi et al., 2010, 2013; Samé et al., 2011), and was initially developed in engineering and science. It involves a complex but general modeling framework that integrates the finite mixture model for capturing group differences, a logistic regression model for explaining phase transitions, and a functional data analysis approach for non-parametrically representing the quantitative dynamics within a phase. We can estimate the mixRHLP model efficiently within the frequentist's framework. We are particularly interested in its strengths and limitations in understanding dynamic processes in social and behavioral sciences. Hence, we conduct Monte Carlo simulations to evaluate the performance of the approach and related model selection methods and test their robustness to various data limitations. We examine the fitting of the model to data with different sample sizes in terms of the number of subjects and the number of time points, proportions of missing values, and regression error variances. Then, we illustrate its usage by analyzing real-time hand movement data collected from an embodied learning platform designed to foster the learning of mathematical proportion. We offer practical guidance on data manipulation and model selection procedures based on the simulation results. Finally, we discuss the limitations, contributions, and future extensions of the current study.

2. MODELING FRAMEWORK

The mixRHLP model (Chamroukhi et al., 2010, 2013; Samé et al., 2011) is designed to analyze multi-subject time-series data. Suppose for each subject $i, i = 1, 2, 3, \cdots, N_p$, there are a total of N_t measurement occasions and N_t measurements of an interesting process (e.g., sensory data of student behavior and emotion), respectively denoted as $N_t \times 1$ vectors of $\boldsymbol{t} = (t_j)$ and $\boldsymbol{y}_i = (y_i(t_j))$. $j = 1, 2, 3, \cdots, N_t$ indexes N_t measurement occasions, and (t_j) is a set of continuous values that indicate elapsed time since each subject's onset and stay the same for all subjects. Thus, $\boldsymbol{t} = (t_j)$ represents a shared time frame for all subjects, whereas $\boldsymbol{y}_i = (y_i(t_j))$ exhibit variability across subjects and over time. We assume $\boldsymbol{y}_i$ follow a mixture distribution, whose density $p(\cdot)$ is a weighted sum of component densities $p_k(\cdot)$ as

$$p(\boldsymbol{y}_i|\boldsymbol{t}_i; \boldsymbol{\Theta}) = \sum_{k}^{K} P(Z_i = k) p_k(\boldsymbol{y}_i|\boldsymbol{t}_i, Z_i = k), \tag{1}$$

where $Z_i \in \{1, 2, \cdots, K\}$ denotes subject i's latent cluster class, with $\alpha_{ik} \overset{\Delta}{=} P(Z_i = k)$ being the probability of subject i belonging to the latent cluster class k. $\boldsymbol{\Theta_k}$ contains all parameters in the component density $p_k(\cdot)$, and $\boldsymbol{\Theta}$ contains all parameters in the density $p(\cdot)$.

At each time point t_j, we further assume $y_i(t_j)$ follows a finite Gaussian mixture regression model, whose conditional component density given cluster k and regime $r = 1, 2, \cdots, R$ is normally distributed with mean $\boldsymbol{X}_j\boldsymbol{\beta}_{kr}$ and a variance of σ^2_{kr}, denoted as $\mathcal{N}(\boldsymbol{X}_j\boldsymbol{\beta}_{kr}, \sigma^2_{kr})$. That is, in each regime, the temporal dynamics of $\boldsymbol{y}_i(\cdot)$ is captured by a linear regression model of time. While the design matrix in the regression model may take different forms, we assume that the regression model is a polynomial regression model of order d, where the design matrix $\boldsymbol{X}_j$ is $\begin{bmatrix} t_j^0 & t_j^1 & \cdots & t_j^d \end{bmatrix}$ and $\boldsymbol{\beta}_{kr}$ is a $(d+1) \times 1$ vector of regression coefficients $\begin{bmatrix} \beta_{kr0} & \beta_{kr1} & \cdots \beta_{krd} \end{bmatrix}^\top$. If we further assume $\boldsymbol{y}_i|\boldsymbol{t}_i, Z_i = k$ given subject i's latent cluster class $Z_i = k$ are serially independent, then the component density $p_k(\cdot)$ can be written as

$$p_k(\boldsymbol{y}_i|\boldsymbol{t}_i, Z_i = k) = \prod_{j=1}^{N_t} \sum_{r}^{R} P(H_{ij} = r|t_j, Z_i = k) \mathcal{N}(\boldsymbol{X}_j\boldsymbol{\beta}_{kr}, \sigma^2_{kr}), \tag{2}$$

where $H_{ij} \in \{1, 2, \cdots, R\}$ denotes subject i's latent regime at time t_j and takes categorical values of $\{1, 2, \cdots, R\}$.

The latent regime H_{ij} at each time point t_j is assumed to follow a multinomial logistic regression model such that the probability of subject i belonging to the latent regime r at time t_j under the condition that subject i belongs to the latent cluster class k is

$$P(H_{ij} = r|t_j, Z_i = k) = \frac{\exp(\omega_{kr0} + \omega_{kr1} t_j)}{\sum_{s=1}^{R} \exp(\omega_{ks0} + \omega_{ks1} t_j)} \tag{3}$$

with $\omega_{ks0} = \omega_{ks1} = 0$ in a reference class. The regression coefficients $\boldsymbol{\omega}_{kr} = [\omega_{kr0}\ \omega_{kr1}]$ control the regime switches, and thus are regime-switching parameters. For instance, if R is the reference class, $\omega_{kR0} = \omega_{kR1} = 0$. Then, $\omega_{kr0} + \omega_{kr1} t_j$ is the log-odds or relative probability of subject i belonging to regime r at time t_j compared to the reference regime R, given that the subject is in cluster k. In the log-odds, ω_{kr0} is an intercept and ω_{kr1} is a slope. If ω_{kr1} is positive, this relative probability increases over time. Hence, if the probability of being in the reference regime R stays the same across time, a positive ω_{kr1} indicates that the likelihood of being in regime r goes up with time. In this way, these parameters influence regime switches.

Assuming the observed data $\boldsymbol{Y} \overset{\Delta}{=} [\boldsymbol{y}_i]$ across subjects are independently identically distributed, we can write the log-likelihood function of $\boldsymbol{\Theta}$ given all observed data as

$$l(\boldsymbol{\Theta}) = \log \prod_{i}^{N_p} p(\boldsymbol{y}_i|\boldsymbol{t}_i; \boldsymbol{\Theta}) = \sum_{i}^{N_p} \log \sum_{k}^{K} \alpha_{ik} p_k(\boldsymbol{y}_i|\boldsymbol{t}_i, Z_i = k). \tag{4}$$

Parameter estimation can be obtained via the Expectation-Maximization algorithm (Dempster et al., 1977). To evaluate the quality of the model, we use the following information criteria:

Bayesian Information Criterion (BIC; Schwarz, 1978), sample-adjusted BIC (saBIC; Sclove, 1987), the Akaike Information Criterion (AIC; Akaike, 1973) and the corrected AIC (AICc; Hurvich and Tsai, 1989) for model selection. Each criterion is defined by the difference between the maximized log-likelihood $l_M(\mathbf{\Theta})$, and a penalty score based on the number of parameters $|\mathbf{\Theta}|$, and weights goodness of fit against model simplicity: $BIC = \log(N_t \times N_p)|\mathbf{\Theta}| - 2l_M(\mathbf{\Theta})$, $saBIC = \log(\frac{N_t \times N_p + 2}{24})|\mathbf{\Theta}| - 2l_M(\mathbf{\Theta})$, $AIC = 2|\mathbf{\Theta}| - 2l_M(\mathbf{\Theta})$, and $AICc = AIC + \frac{2|\mathbf{\Theta}|^2 + 2|\mathbf{\Theta}|}{N_t \times N_p - |\mathbf{\Theta}| - 1}$. The model yielding the lowest criterion value is perceived as the model that generalizes best (Myung and Pitt, 2018).

The estimation algorithm also computes the posterior regime and cluster probabilities at each time point as by-products. We can determine cluster and regime classifications by the highest posterior probability in posterior class probabilities at each time point.

3. SIMULATION

3.1. Simulation Design

As many naturally collected data contain a small sample size and are collected at irregular intervals, we conducted Monte Carlo simulations to evaluate the applicability of the mixRHLP model under these limitations. In particular, we were interested in (1) whether the information criteria could be useful in model selection, (2) how accurate the estimation algorithm could be in estimating parameters and making classifications, and (3) how the answers to (1) and (2) would change under different data conditions. We sought to examine the fitting of the model to data with different sample sizes in terms of the number of participants and the number of time points, proportions of missing values, and regression error variances.

We generated data from a mixRHLP model with 2 clusters ($K = 2$), 3 regimes ($R = 3$), and linear functions ($d = 1$). We wanted the K, R, d values to be as small as possible so that the model is simple enough but still exhibits minimal cluster-based regime-switching properties with time-dependent structure in each regime. We chose $R = 3$ instead of 2 to mirror the regime characteristics observed in our empirical data. The measurement occasions $\boldsymbol{t}$ were equally spaced time points within the interval of $[0, 1]$. The true parameter values are listed in **Table 1** and were selected such that in different clusters and regimes the dynamics varied but were hard to differentiate by eyes when plotted altogether. We assumed equal regression error variance σ across clusters and regimes, and that the data may be missing completely at random. We varied four factors in simulating the data: (1) the number of participants in the sample ($N_p = 20, 60, 100$), (2) the number of time points ($N_t = 20, 160, 300$), (3) the magnitude of the regression error variance ($\sigma = 0.10, 0.15, 0.20$), and (4) the proportion of missing data in each participant's data ($PMiss = 0, 0.1, 0.2$). **Figure 1** showed the simulated data in two clusters under the conditions of $N_p = 20$, $N_t = 160$, $PMiss = 0.1$, and $\sigma = 0.1$.

We carried out $M = 200$ Monte Carlo runs for each of the 81 ($= 3^4$) data conditions. Where data were missing for a participant, we replaced the missing values using linear interpolation with the *na.approx()* function from the **zoo** R package (Zeileis and Grothendieck, 2005). To each set of full data (after imputation), we fitted a total of 32 mixRHLP models with combinations of different values of $K = 1, 2, 3, 4$, $R = 1, 2, 3, 4$, and $d = 1, 2$ and heteroskedastic regression error variances using the **mixRHLP** package (Chamroukhi et al., 2010, 2013; Samé et al., 2011). When the algorithm finished successfully, we computed the four information criteria: BIC, saBIC, AIC, and AICc.

TABLE 1 | True parameter values used in the Monte Carlo simulation study.

			Regime 1		Regime 2		Regime 3	
		X	**1**	**t**	**1**	**t**	**1**	**t**
Cluster 1	$\alpha_1 = 0.5$	$\boldsymbol{\omega}_{1\cdot}$	−2.00	3.00	1.00	−2.50	0	0
		$\boldsymbol{\beta}_{1\cdot}$	0	−1.50	0.60	−0.90	1.20	−0.30
		σ_1	0.10, 0.15, 0.20					
Cluster 2	$1 - \alpha_1 = 0.5$	$\boldsymbol{\omega}_{2\cdot}$	-1.00	2.00	0.50	−2.00	0	0
		$\boldsymbol{\beta}_{2\cdot}$	0.60	0.30	1.20	0.90	1.80	1.50
		σ_2	0.10, 0.15, 0.20					

We used three sets of measures to compare the model fitting results across simulation conditions: (1) information criteria measures, (2) parameter estimate accuracy measures and (3) classification accuracy measures. Information criteria measures included a proportion measure and a rank measure. The proportion measure is the proportion of runs where a certain criterion of the true model ($K = 2, R = 3, d = 1$) indicated itself as the best-fitting model (i.e., as the smallest among those of the 32 fitted models). The rank measure is the average rank of the criterion value among ordered values of the 32 models' same criterion arranged from the smallest to the largest. To measure parameter estimate accuracy, we computed the root mean squared errors of each parameter. To simplify the presentation of the simulation results, we grouped the parameters into six sub-groups, namely, α_1, $\boldsymbol{\beta}_0 = [\beta_{kr0}]$, $\boldsymbol{\beta}_1 = [\beta_{kr1}]$, σ, $\boldsymbol{\omega}_0 = [\omega_{kr0}]$, and $\boldsymbol{\omega}_1 = [\omega_{kr1}]$ and took the average RMSE of the parameters within the same sub-group. Let $\theta_{g,\mathbb{G}}$ and $\hat{\theta}_{r,g,\mathbb{G}}$ respectively denote the true and estimated value of a parameter, where r indicates the r-th Monte Carlo run, and g indicates the g-th parameter in a parameter group $\mathbb{G}$ of size $|\mathbb{G}|$. The average RMSE was computed as $\text{rmse}_{\mathbb{G}} = \frac{1}{|\mathbb{G}|} \sum_g \sqrt{\frac{1}{M} \sum_r (\hat{\theta}_{r,g,\mathbb{G}} - \theta_{g,\mathbb{G}})^2}$. The classification accuracy measures are the proportion of correct classifications of either the clusters or the regimes of available data (before imputation).

3.2. Simulation Results

To reveal the typical characteristics of the Monte Carlo samples, we decided to remove the outliers of the simulation measures within each data condition. We used the *OutlierDetection()* function in the R package **OutlierDetection** (Tiwari and Kashikar, 2019) to identify outliers based on K-nearest neighbor graphs (K = 5% of the Monte Carlo runs, Hautamaki et al., 2004). The remaining Monte Carlo sample size ranged from 165 to 197, with a median of 191. Most outliers were found when the sample

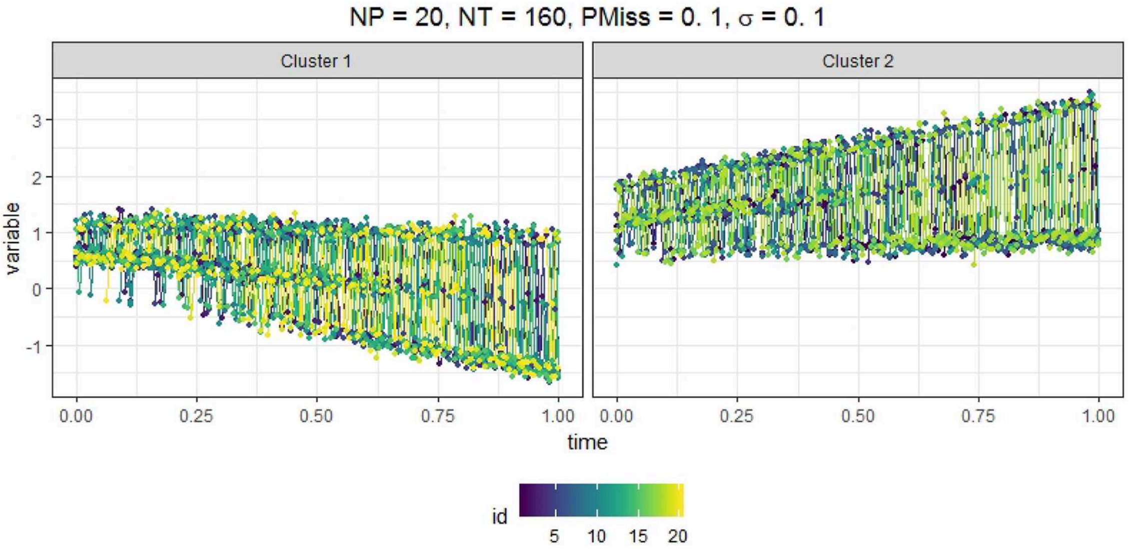


FIGURE 1 | Simulated time series in two clusters when $N_p = 20$, $N_t = 160$, $PMiss = 0.1$, and $\sigma = 0.1$.

size was the smallest ($N_p = 20$, $N_t = 20$), regression error variance high ($\sigma = 0.2$), and missing data imputation involved ($PMiss = 0.1$).

Among the rest of the Monte Carlo samples after outlier removal, BIC performed better than the other three information criteria in selecting the right model as the best-fitting model, with a success rate of 0.54, whereas the success rates of saBIC, AICc, AIC were 0.38, 0.21, and 0.20, respectively. The median rank of the true model's BIC among 32 models was 1 (i.e., the smallest), and the maximum rank was 10, both smaller than those of saBIC (median 3, maximum 12), AIC (median 5, maximum 12) and AICc (median 5, maximum 12). Although BIC could be useful for model selection under certain conditions, the smallest BIC did not always indicate the true model in simulations. When we fitted the correct model, the accuracy of the parameter estimates was high, characterized by RMSEs lower than 0.1, except for the regime-switching parameters in $\boldsymbol{\omega}_0$ and $\boldsymbol{\omega}_1$ categories. Even though some of the regime-switching parameters could not be estimated correctly, the classification accuracy was overall very high. Across all data conditions, the proportion of correct cluster classifications was invariably 1, and the proportion of correct regime classifications was 0.99, suggesting the robustness of the approach in identifying clusters and regimes in time-series data of our interest.

After examining the results from each simulation condition, we identified how the four factors considered affected the model selection and statistical inference. **Figure 2** presents the effects of the factors on the information criteria measures under typical simulation conditions. When the data were at a sufficient number of time points (e.g., $N_t \geq 160$) and without missing data, the smallest BIC could be used to select the correct model as the best-fitting model regardless of N_p and σ. As shown in **Figure 2A**, the BIC and saBIC of the true model were almost always the smallest among fitted models when N_t was higher than 160, and there was no missing data. Also, the utility of information criteria improved with an increase in N_t even though N_p was small, with higher success rates in selecting the correct model and lower rank among fitted models. When $N_t = 160$ under the same condition without missing data (e.g., in **Figure 2B**), although BIC and saBIC performed almost equally well, the utility of AIC and AICc improved as N_p increased. However, when the imputation of missing data happened, the larger the size of the missing data, either as a result of a bigger sample size or a more substantial missing proportion (e.g., partly illustrated in **Figure 2C**), the smaller the utility of all information criteria was. Nevertheless, when the regression error variance in the actual model was high, the misfit of the mixRHLP model to imputed data could be considered as regression errors, enabling the use of information criteria in model selection, as shown in **Figure 2D**.

Besides, **Figure 3** shows the effects of the four factors on the classification accuracy measures under typical simulation conditions. Generally, both the cluster and regime classifications were accurate and not affected by sample size (N_p or N_t) nor proportion of missing data (*PMiss*), unless the sample size was really small (i.e., $N_t = N_p = 20$) and the regression error variance was high, as shown in **Figures 3A,D**. However, the regime classification accuracy depended on the characteristics of the model. For example, the larger the regression error variance was, the lower the accuracy of regime classifications (see **Figure 3D**).

Moreover, **Figure 4** presents how different factors affected the accuracy of the estimates of the regime-switching parameters. As in **Figures 4A,B**, the larger the sample size was, as a result of an increase in either N_t or N_p, the more accurate the parameter estimates. When there were no missing data, a sample of size $N_p = 100$ and $N_t = 300$ was sufficient for accurate estimation of all model parameters, with the RMSEs below a threshold of 0.1. The magnitude of regression error variance did not affect the accuracy parameter estimates, as seen in

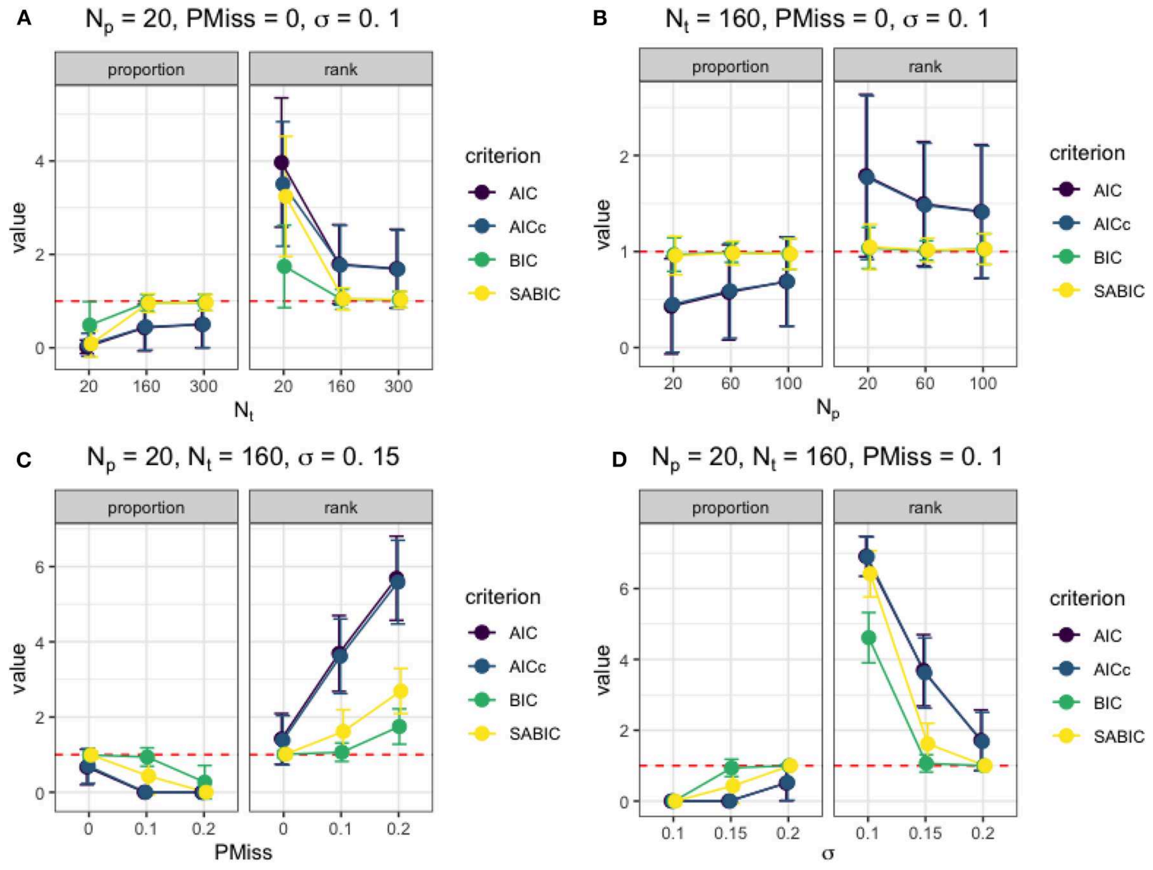


FIGURE 2 | The utility of information criteria under different simulation conditions. **(A)** The effect of N_t on the information criteria measures when $N_p = 20, PMiss = 0, \sigma = 0.1$. **(B)** The effect of N_p on the information criteria measures when $N_t = 160, PMiss = 0, \sigma = 0.1$. **(C)** The effect of *PMiss* on the information criteria measures when $N_p = 20, N_t = 160, \sigma = 0.15$. **(D)** The effect of σ on the information criteria measures when $N_p = 20, N_t = 160, PMiss = 0.1$.

Figure 4D. However, as in **Figure 4C**, the presence of missing data, although imputed, affected the parameter estimation negatively. An increase in the proportion of the missing data led to higher RMSEs of the regime-switching parameters. In **Figure 4C**, we also present the RMSEs of the parameters under the data condition that is close to the data in our empirical example.

4. EMPIRICAL EXAMPLE

To illustrate our approach with real data, we built upon the work of the Mathematics Imagery Training of Proportion (MIT-P) and analyzed secondary data collected from a previous study (Abrahamson et al., 2015; Duijzer et al., 2017) with informed consent from the legal guardians of the participants and approval of the ethical committee board of the faculty of Social Sciences at Utrecht University. In the study, 45 fifth- and sixth- graders of ages 9–11 participated in task-based semi-structured interviews at schools in the Netherlands. In the interview, the participants played with a touchscreen tablet and used their index fingers to move two parallel vertical bars up and down (see **Figure 5**). The bars changed colors between red and green based on their heights. The closer the ratio between the height of right and left bars was to a predefined value (1 : 2), the greener the bars were, which was the mysterious rule the participants did not know before the interview and needed to find out. In the beginning, the participants were given instructions to move the bars and *find as many greens as possible*. After they found the first green, the participants were encouraged to find *more*. In the end, the participants needed to *move the bars from the bottom to the top while keeping them green*. During the process, participants were probed to think aloud *why the bars turned green and what actions they were to take to solve the problem*. The same procedures applied under different task and screen conditions, where the proportional value varied from $\frac{1}{2}$ to $\frac{3}{4}$ or grids with and without numbers appeared on the screen. Screen recordings of participants' hand movements, together with tracking of their eye movements and concurrent verbalization, were captured during the whole interview. The data of 38 participants were of sufficiently high quality to include them in the analysis. The mean age of the participants was 11.3 years old ($SD = 0.70$), and there were 17 females in the sample. For retaining time series data

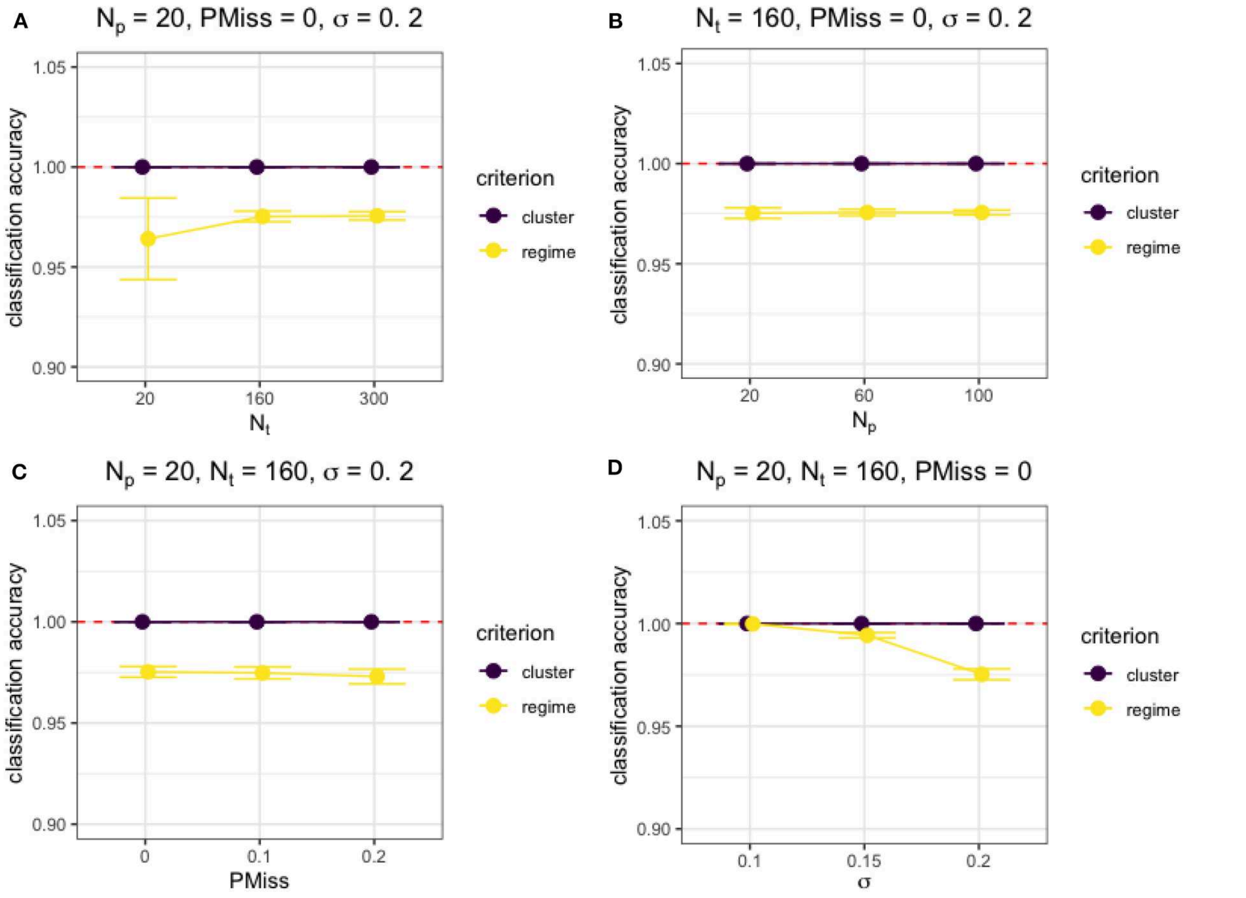


FIGURE 3 | The accuracy of classification under different simulation conditions. **(A)** The effect of N_t on the cluster and regime classification accuracy measures when $N_p = 20, PMiss = 0, \sigma = 0.2$. **(B)** The effect of N_p on the cluster and regime classification accuracy measures when $N_t = 160, PMiss = 0, \sigma = 0.2$. **(C)** The effect of *PMiss* on the cluster and regime classification accuracy measures when $N_p = 20, N_t = 160, \sigma = 0.2$. **(D)** The effect of σ on the cluster and regime classification accuracy measures when $N_p = 20, N_t = 160, PMiss = 0$.

under the same task and screen condition from all participants, we only focused on hand movement data collected in the task with the proportion of 1 : 2 and on blank screen background without grids.

The original real-time capture of hand movement data happened as the participants moved their fingers on the tablet, and hence the data contained missing values and were irregularly spaced. To prepare the data for our analysis, we first removed data with a partial recording of only one hand's movement, which took up <5% of the available data and was missing largely because of off-task behavior and technical errors. The remaining data for each participant varied in the number of measurement occasions (6,132–32,543) and the total period they covered (3.28–13.71 min, with a mean of 6.74). To construct a common time frame for all participants, we re-scaled individuals' measurement occasions to a range of [0, 1] by subtracting the initial time point and dividing the times by the total period of each individual. We then aggregated data at the individual level in 200 equally spaced intervals in [0, 1) using their mean to create a data set of 38 participants on the same 201 occasions equally spaced in [0, 1]. In cases where there was no recording in a certain time interval for an individual, missing data would occur in the aggregation. In the new data, the proportion of missing data ranged from 0 to 0.17, with a median of 0.05, across individuals. We replaced the missing values with linear interpolation via the *na.approx()* function in the R package **zoo**. We took the ratios between right and left-hand positions as our variable of interest and winsorized the data by substituting the extreme ratios that are above the 95 percentile of the ratios with the 95 percentile. **Figure 6** shows the aggregated time series of two individuals in points, and the imputed and winsorized data in lines. We marked the imputed data with squares.

We fitted the mixRHLP models to the time series data with different values of K (1–4), R (1–4), and d (1–2). Among all 32 models, we chose the more parsimonious model with the top three smallest BIC values, which consisted of two clusters, three regimes, and linear regressions. The parameter estimates from fitting the chosen model to the data are summarized in **Table 2**. The probability of an individual being in Cluster 1 was estimated to be 0.42, a little smaller than that of being in Cluster

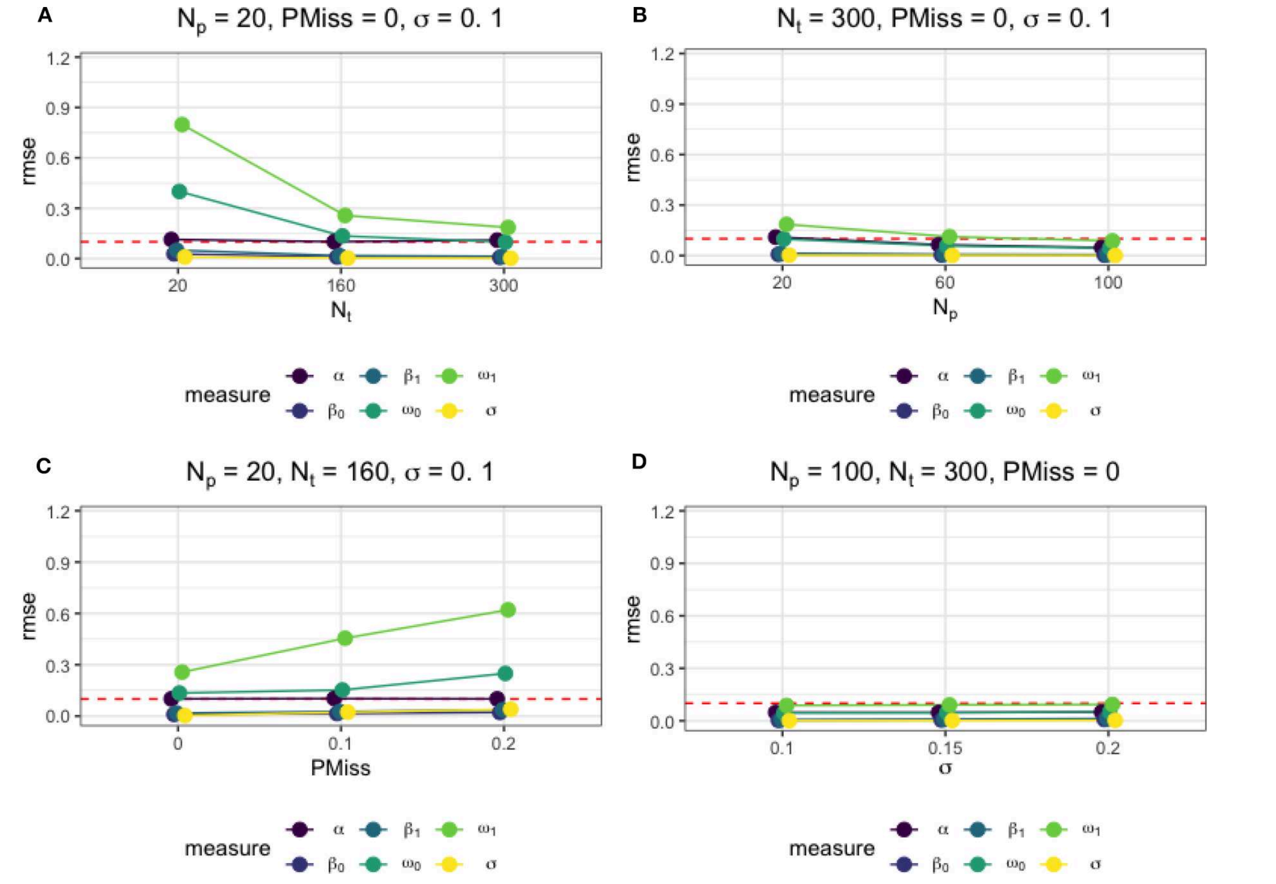


FIGURE 4 | The average root mean square error (RMSE) of parameter estimates under different simulation conditions. **(A)** The effect of N_t on the parameter estimate accuracy measures when $N_p = 20, PMiss = 0, \sigma = 0.1$. **(B)** The effect of N_p on the parameter estimate accuracy measures when $N_t = 300, PMiss = 0, \sigma = 0.1$. **(C)** The effect of *PMiss* on the parameter estimate accuracy measures when $N_p = 20, N_t = 160, \sigma = 0.1$. **(D)** The effect of σ on the parameter estimate accuracy measures when $N_p = 100, N_t = 300, PMiss = 0$.

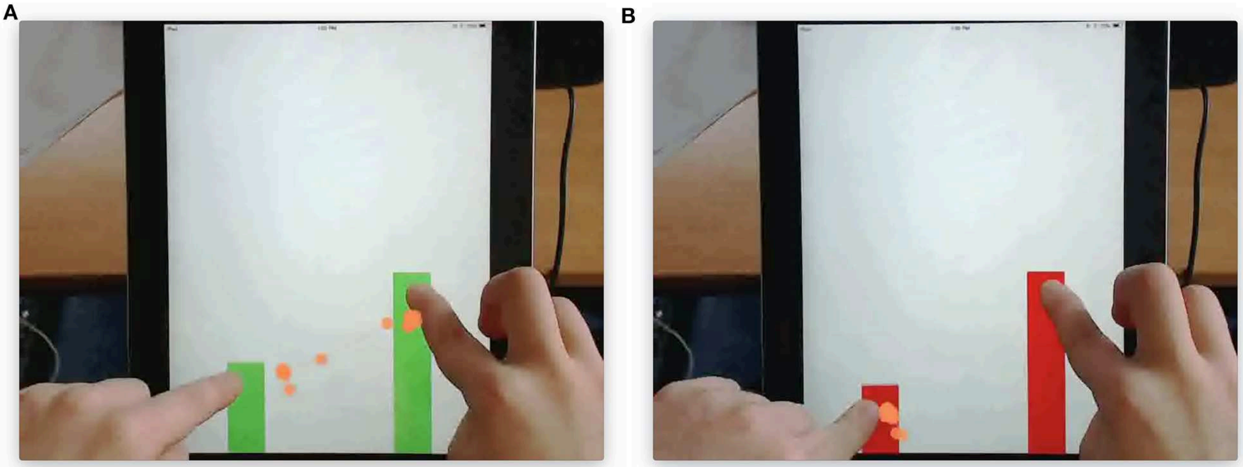


FIGURE 5 | The touchscreen tablet version of the Mathematical Imagery Trainer for Proportion (MIT-P). **(A)** Fingers maintain a 1:2 ratio to make the bars green. **(B)** Fingers do not maintain such ratio and therefore the bars are red.

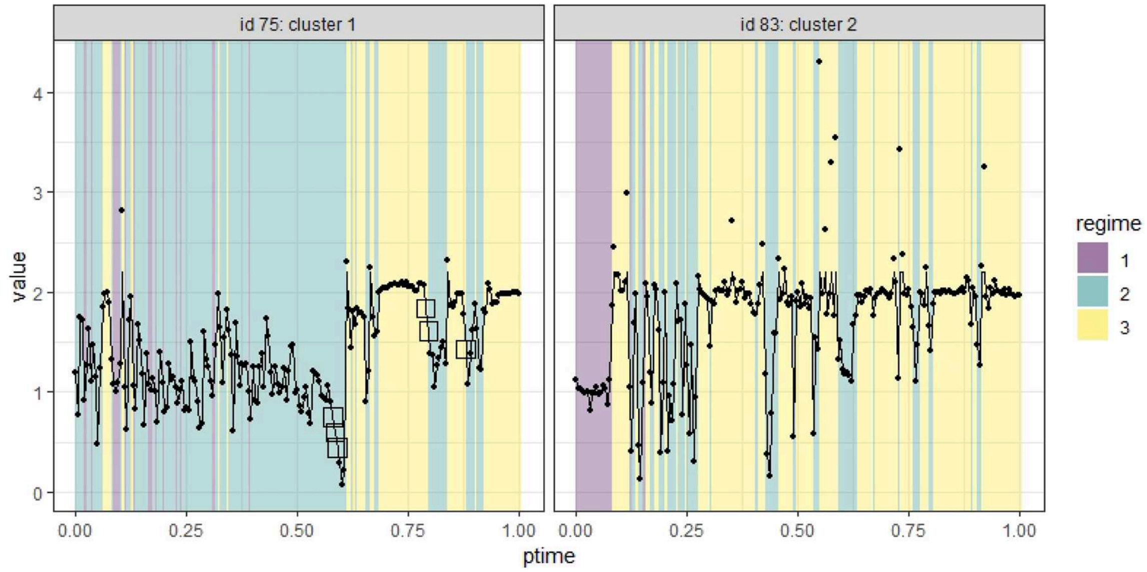


FIGURE 6 | Time series data of two individuals identified in two clusters. The imputed data are marked as squares.

TABLE 2 | Parameter estimates from the empirical example.

		Regime 1		Regime 2		Regime 3	
	X	**1**	**t**	**1**	**t**	**1**	**t**
	$\boldsymbol{\omega}_{1.}$	1.899	−7.871	1.864	−2.597	0	0
cluster 1 $\alpha_1 = 0.42$	$\boldsymbol{\beta}_{1.}$	1.014	−0.019	1.149	0.016	2.017	−0.061
	σ_1^2	0.003		0.180		0.015	
	$\boldsymbol{\omega}_{2.}$	1.423	−14.651	0	0	−0.386	1.794
cluster 2 $1 - \alpha_1 = 0.58$	$\boldsymbol{\beta}_{2.}$	1.018	−0.039	0.896	0.594	2.027	−0.043
	σ_2^2	0.009		0.213		0.009	

2. Although the regimes' regression parameters differed across clusters, the respective three regimes were comparable in the two clusters. In particular, Regime 1 in both clusters had regression intercepts near one with small error variances, indicating hands were moving at the same height. Regime 2 in both clusters had significant error variances with intercepts around one, indicating hands moving with a noticeable variability. Regime 3 in both clusters had intercepts about two with small error variances, suggesting hands moving at the desired heights of 1:2 ratio to keep the bars green.

After completing our statistical analysis, we also wanted a qualitative interpretation of the results in light of the possible solution strategies subjects were following during each regime. Accordingly, Regime 1 corresponded to an initial phase of the embodied interaction. During this regime, the hands were at the same height; perhaps the student awaited to see what happened next. Regime 2 corresponded to an intermediate phase of the interaction. During this regime, it seems as though the participant was actively exploring different hand ratios, perhaps attempting to find how to make the bars green. From prior qualitative observations, we know that this regime contains a mixture of strategies in that changes in the hands' ratio not only happens when the two hands move independently but also when they move at fixed distances. As our analysis missed this distinction, this seems to be one of the limitations of our current approach. Regime 3 corresponded to a later phase of the interaction and was the desired outcome of the interview. During this phase, the hands maintained a 1:2 ratio. However, as the task asked students to find green in as many ways as they could, from time to time, this particular ratio was lost, and the student fell back into Regime 2. Note that to keep the same ratio as the hands move up, the one hand has to move twice as fast as the other hand, which proves to be a challenging bodily coordination exercise for participants even though they have figured out the proportion rule. Further analysis of the participants' verbalization during the interview using natural language processing techniques confirmed our interpretation of the different regimes to some extent [see Ou et al. (2020) for more details].

Additionally, **Figure 7** illustrates the estimated expected logistic curves of the probabilities of an individual being in a regime during the interview. In Cluster 1, the probability of being in Regime 1 was the highest at the start of the session but close to the probability of being in Regime 2, which grew slowly but soon became the highest until Regime 3 became the most probable regime at around 70% into the interview session. In Cluster 2, the probability of being in Regime 1 was the highest until approximately 10% into the session, when the probability of being in Regime 2 took the lead but was only slightly higher than that of being in Regime 3; then, Regime 3 became the most probable state at about 20% into the session, much sooner compared to Cluster 1. Indeed, what the logistic curves tell us

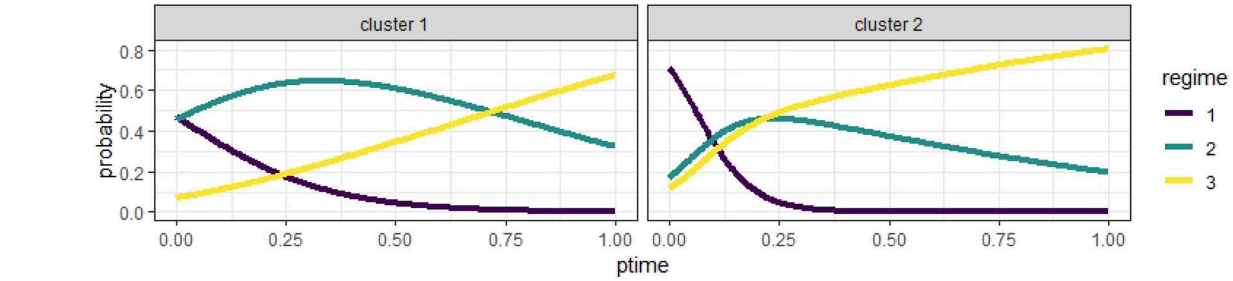


FIGURE 7 | The estimated logistic curves of different regimes in two clusters.

is that a student in Cluster 1 has about the same likelihood to find the rule than not to find it, as indicated by the high logistic curve of Regime 2 for most of the task segment. Instead, students in Cluster 2 have a much higher probability of finding the proportional rule, especially after the first half. It is apparent that, in Cluster 2, the probability of being in Regime 1 goes down a lot more quickly than that in Cluster 1, and almost disappears after the first quarter. On the other hand, the probability of Regime 2 goes down but still lingers on, albeit low, until the end of the task segment.

To exemplify these results, **Figure 6** shows different hand movement dynamics of participants with IDs 75 and 83. Participant 75, classified in Cluster 1, spent a substantial proportion (> 50%) of the session exploring various ratios (Regime 2) or merely moving her hands at the same speed (Regime 1). Participant 83, classified in Cluster 2, spent only 10% of the time moving hands at the same heights (Regime 1) and quickly switched to a 1:2 ratio phase (Regime 3), interspersed with chunks of short periods of Regime 2.

5. LIMITATIONS

The modeling framework and our empirical illustration have some limitations. First, a shared time frame of measurement occasions needs to apply to all subjects, which is often unrealistic in data collection. As participants differed in their time spent on the task, we were only able to construct a proportional time relative to their respective elapsed time such that the time frame is within $[0, 1]$. Besides, we had to involve data aggregation and missing data imputation based on the shared time frame, which could affect the accuracy of parameter estimates.

Second, the modeling framework and estimation algorithm only apply to univariate time-series data at this moment. Our example only took into account the hands' ratio, so we were not able to identify from the ratio data some of the strategies discussed in prior studies, such as the fixed-distance strategy with which participants kept their hands at the same speed. We could utilize eye gaze data and other hand-movement variables such as speed and distance between hands to study how hand and eye movements coordinate in such activities.

Third, the logistic transition process in Equation (3) assumes that the log odds of being in a regime relative to the reference regime change monotonically with time. It ignores the local context of a regime switch such as the current regime from which a switch is happening. Further, it lacks some flexibility in modeling bidirectional regime switches that are more common in hidden Markov type models and may apply under different circumstances.

Despite the limitations, the mixRHLP model is useful in extracting qualitative clusters and regimes from quantitative time-series data, and the illustrative example furthers our knowledge of qualitative differences in how students approach the mathematical concept of proportion physically.

6. DISCUSSION

Advancements in real-time data capture technology revolutionized the type and amount of data we collect about human dynamic processes. In this paper, we have introduced the mixRHLP model for clustering multi-subject time-series data with regime-switching properties. In a Monte Carlo simulation study, we examined the accuracy of the approach in parameter estimation and cluster and regime classification under various data conditions. We tested the feasibility of using information criteria for model selection. We showed how different factors such as the number of time points, the number of participants, the proportion of missing data, and the error variance in the model could affect the performance and applicability of the approach and had a deeper understanding of the strengths and limitations of the approach.

To illustrate the use of this approach in real scenarios, we applied it to studying students' behavior in an action-based learning environment for mathematical learning. We based our data aggregation and model selection decisions on the Monte Carlo simulation results. We discovered qualitative differences in students' hand movements on a tablet during the task and across students, as they explored the concept of proportion using physical actions. This type of analysis helped reveal between and within-subject differences in dynamic processes not seen with prior qualitative analyses (Duijzer et al., 2017). That is, although qualitative analysis may help reveal phase transitions in strategy use, efficiently comparing students' experiences and performing grouping exceeded human capacity. Using the approach, we can not only extract strategies directly and efficiently from data but also identify clusters of students with homogeneous dynamics and potentially similar needs for intervention.

In the future, we should extend the estimation algorithm to fit multivariate time series data to account for systematic changes in dynamic systems. For instance, Kuppens et al. (2007) found that the extent to which individuals experience qualitatively different feelings in the core affect space is a consistent measure to their trait measures of self-esteem and depression. We need cluster-based multivariate dynamic models potentially with regime-switching features to help reveal systematic emotion dynamics that may have implications for psychological well-being and adjustment. Besides, we need to compare the mixRHLP modeling approach to other model-based and data-driven approaches for clustering regime-switching dynamics in simulations and applications. Candidate approaches include but not limited to the mixture of hidden Markov models (Chamroukhi and Nguyen, 2019) and potential extensions of existing data-driven methods that identify clusters or regimes (e.g., Cabrieto et al., 2018). Moreover, it is worthwhile to examine different imputation methods for missing data, for example, the newly developed ones that depend on machine learning approaches (Yoon et al., 2018).

Finally, the model contributes to the tools to extract qualitative cluster and regime patterns from quantitative time-series of human dynamics. We anticipate its broader usage in analyzing the increasingly prevalent multi-modal time-series data in social and behavioral sciences beyond mathematics learning. In applications, developers and practitioners may use the qualitative findings from time-series data to inform intervention and training programs. For instance, in collaborative learning environments such as classrooms, we might be able to monitor students' real-time behavior with various sensors and utilize the technique to generate learners' qualitative profiles and tailor personalized or group-based feedback to facilitate learning and shift students from one cluster to another.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the ethical committee board of the Faculty of Social Sciences at Utrecht University. Written informed consent to participate in this study was provided by the participants' legal guardian or next of kin.

AUTHOR CONTRIBUTIONS

LO worked on simulations and data analysis and prepared the first draft. RA and AB provided the raw data. All authors contributed to writing and editing the paper.

ACKNOWLEDGMENTS

This work was funded by ACT and NWO grant number 652.001.004. We appreciate the support of the research from ACTNext by ACT and suggestions on the manuscript from Gunter Maris and Benjamin Deonovic.

REFERENCES

Abrahamson, D., Lee, R. G., Negrete, A. G., and Gutiérrez, J. F. (2014). Coordinating visualizations of polysemous action: values added for grounding proportion. *ZDM Int. J. Math. Educ.* 46, 79–93. doi: 10.1007/s11858-013-0521-7

Abrahamson, D., Shayan, S., Bakker, A., and van der Schaaf, M. (2015). Eye-tracking piaget: capturing the emergence of attentional anchors in the coordination of proportional motor action. *Hum. Dev.* 58, 218–244. doi: 10.1159/000443153

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle," in *Second International Symposium on Information Theory*, eds B. N. Petrov and F. Csaki (Budapest: Akademiai Kiado), 267–281.

Andrade, A., Danish, J., and Maltese, A. (2017). A measurement model of gestures in an embodied learning environment: accounting for temporal dependencies. *J. Learn. Anal.* 4, 18–45. doi: 10.18608/jla.2017.43.3

Arieli-Attali, M., Ou, L., and Simmering, V. R. (2019). Understanding test takers' choices in a self-adapted test: a hidden Markov Modeling of process data. *Front. Psychol.* 10:8. doi: 10.3389/fpsyg.2019.00083

Böckenholt, U. (2005). A latent Markov model for the analysis of longitudinal data collected in continuous time: States, durations, and transitions. *Psychol. Methods* 10, 65–83. doi: 10.1037/1082-989X.10.1.65

Bolger, N., and Laurenceau, J.-P. (2013). *Intensive Longitudinal Methods: An Introduction to Diary and Experience Sampling Research*. New York, NY: Guilford Press.

Bulteel, K., Tuerlinckx, F., Brose, A., and Ceulemans, E. (2016). Clustering vector autoregressive models: capturing qualitative differences in within-person dynamics. *Front. Psychol.* 7:1540. doi: 10.3389/fpsyg.2016.01540

Cabrieto, J., Adolf, J., Tuerlinckx, F., Kuppens, P., and Ceulemans, E. (2018). Detecting long-lived autodependency changes in a multivariate system via change point detection and regime switching models. *Sci. Rep.* 8, 1–15. doi: 10.1038/s41598-018-33819-8

Chamroukhi, F., and Nguyen, H. D. (2019). Model-based clustering and classification of functional data. *Wiley Interdiscipl. Rev. Data Mining Knowledge Discov.* 9:e1298. doi: 10.1002/widm.1298

Chamroukhi, F., Samé, A., Govaert, G., and Aknin, P. (2010). A hidden process regression model for functional data description. Application to curve discrimination. *Neurocomputing* 73, 1210–1221. doi: 10.1016/j.neucom.2009.12.023

Chamroukhi, F., Trabelsi, D., Mohammed, S., Oukhellou, L., and Amirat, Y. (2013). Joint segmentation of multivariate time series with hidden process regression for human activity recognition. *Neurocomputing* 120, 633–644. doi: 10.1016/j.neucom.2013.04.003

Chow, S.-M., Lu, Z., Sherwood, A., and Zhu, H. (2016). Fitting nonlinear ordinary differential equation models with random effects and unknown initial conditions using the stochastic approximation expectation-maximization (SAEM) algorithm. *Psychometrika* 81, 102–134. doi: 10.1007/s11336-014-9431-z

Chow, S.-M., Ou, L., Ciptadi, A., Prince, E., You, D., Hunter, M. D., et al. (2018). Representing sudden shifts in intensive dyadic interaction data using differential equation models with regime switching. *Psychometrika* 83, 476–510. doi: 10.1007/s11336-018-9605-1

Chow, S.-M., Ram, N., Boker, S. M., Fujita, F., and Clore, G. (2005). Emotion as a thermostat: representing emotion regulation using a damped oscillator model. *Emotion* 5, 208–225. doi: 10.1037/1528-3542.5.2.208

Chow, S.-M., Zu, J., Shifren, K., and Zhang, G. (2011). Dynamic factor analysis models with time-varying parameters. *Multivar. Behav. Res.* 46, 303–339. doi: 10.1080/00273171.2011.563697

Colder, C. R., Campbell, R. T., Ruel, E., Richardson, J. L., and Flay, B. R. (2002). A finite mixture model of growth trajectories of adolescent alcohol use: Predictors and consequences. *J. Consult. Clin. Psychol.* 70, 976–985. doi: 10.1037/0022-006X.70.4.976

Cole, P. M., Martin, S. E., and Dennis, T. A. (2004). Emotion regulation as a scientific construct: methodological challenges and directions for child development research. *Child Dev.* 75, 317–333. doi: 10.1111/j.1467-8624.2004.00673.x

Collins, L. M., and Lanza, S. T. (2010). *Latent class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. Hoboken, NJ: Wiley.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. Ser. B* 39, 1–38. doi: 10.1111/j.2517-6161.1977.tb01600.x

Deonovic, B., Yudelson, M., Bolsinova, M., Attali, M., and Maris, G. (2018). Learning meets assessment. *Behaviormetrika* 45, 457–474. doi: 10.1007/s41237-018-0070-z

Duijzer, C. A., Shayan, S., Bakker, A., Van der Schaaf, M. F., and Abrahamson, D. (2017). Touchscreen tablets: Coordinating action and perception for mathematical cognition. *Front. Psychol.* 8:144. doi: 10.3389/fpsyg.2017.00144

Dutilh, G., Wagenmakers, E., Visser, I., and van der Maas, H. L. J. (2010). A phase transition model for the speed-accuracy trade-off in response time experiments. *Cogn. Sci.* 35, 211–250. doi: 10.1111/j.1551-6709.2010.01147.x

Hallquist, M. N., and Wright, A. G. C. (2014). Mixture modeling methods for the assessment of normal and abnormal personality, part I: cross-sectional models. *J. Pers. Assess.* 96, 256–268. doi: 10.1080/00223891.2013.845201

Hamaker, E. L., Ceulemans, E., Grasman, R. P., and Tuerlinckx, F. (2015). Modeling affect dynamics: state of the art and future challenges. *Emot. Rev.* 7, 316–322. doi: 10.1177/1754073915590619

Hautamaki, V., Karkkainen, I., and Franti, P. (2004). "Outlier detection using k-nearest neighbour graph," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004* (Cambridge, UK), 430–433. doi: 10.1109/ICPR.2004.1334558

Hu, Y., Boker, S. M., Neale, M. C., and Klump, K. L. (2014). Coupled latent differential equation with moderators: simulation and application. *Psychol. Methods* 19, 56–71. doi: 10.1037/a0032476

Hurvich, C. M., and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika* 76, 297–307. doi: 10.1093/biomet/76.2.297

Jansen, B. R. J., Raijmakers, M. E. J., and Visser, I. (2007). Rule transition on the balance scale task: a case study in belief change. *Synthese* 155, 211–236. doi: 10.1007/s11229-006-9142-9

Kelso, J. A. S. (2000). "Principles of dynamic pattern formation and change for a science of human behavior," in *Developmental Science and the Holistic Approach: Proceedings of a conference at Wiks Castle and the Nobel Institute*, eds L. Bergman, R. B. Cairns, L. G. Nilsson, and L. Nystedt (Stockholm; Mahwah, NJ: Erlbaum), 63–83.

Kuppens, P., Oravecz, Z., and Tuerlinckx, F. (2010). Feelings change: accounting for individual differences in the temporal dynamics of affect. *J. Pers. Soc. Psychol.* 99, 1042–1060. doi: 10.1037/a0020962

Kuppens, P., van Mechelen, I., Nezlek, J. B., Dossche, D., and Timmermans, T. (2007). Individual differences in core affect variability and their relationship to personality and psychological adjustment. *Emotion* 7, 262–274. doi: 10.1037/1528-3542.7.2.262

Lu, Z.-H., Chow, S.-M., Sherwood, A., and Zhu, H. (2015). Bayesian analysis of ambulatory blood pressure dynamics with application to irregularly spaced sparse data. *Ann. Appl. Stat.* 9, 1601–1620. doi: 10.1214/15-AOAS846

Lubke, G. H., and Muthen, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychol. Methods* 10, 21–39. doi: 10.1037/1082-989X.10.1.21

McLachlan, G. J., and Peel, D. (2004). *Finite Mixture Models*. New York, NY: John Wiley & Sons.

Molenaar, P. C. M., Huizenga, H. M., and Nesselroade, J. R. (2003). "The relationship between the structure of interindividual and intraindividual variability: a theoretical and empirical vindication of developmental systems theory," in *Understanding Human Development: Dialogues with Lifespan Psychology*, eds U. M. Staudinge and U. E. R. Lindenberger (Dordrecht), 339–360. doi: 10.1007/978-1-4615-0357-6_15

Muthen, B. (2004). *Latent Variable Analysis: Growth Mixture Modeling and Related Techniques for Longitudinal Data*. Newbury Park, CA: Sage.

Myung, J. I., and Pitt, M. A. (2018). "Model comparison in psychology," in *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, 1–34. doi: 10.1002/9781119170174.epcn503

Oravecz, Z., Tuerlinckx, F., and Vandekerckhove, J. (2011). A hierarchical latent stochastic differential equation model for affective dynamics. *Psychol. Methods* 16, 468–490. doi: 10.1037/a0024375

Ou, L., Andrade, A., Alberto, R., van Helden, G., and Bakker, A., (2020). "Using a cluster-based regime-switching dynamic model to understand embodied mathematical learning," in *Proceedings of the 10th International Conference on Learning Analytics and Knowledge (LAK '20), March 23–27, 2020, Frankfurt, Germany* (New York, NY: ACM), 1–6. doi: 10.1145/3375462.3375513

Ou, L., Hofman, A. D., Simmering, V. R., Bechger, T., Maris, G., and van der Maas, H. L. J. (2019). "Modeling person-specific development of math skills in continuous time: new evidence for mutualism," in *The 12th International Conference on Educational Data Mining*, eds M. Desmarais, C. F. Lynch, A. Merceron, and R. Nkambou (Montreal, QC).

Pardos, Z. A., Hu, C., Meng, P., Neff, M., and Abrahamson, D. (2018). "Classifying learner behavior from high frequency touchscreen data using recurrent neural networks," in *UMAP'18 Adjunct: 26th Conference on User Modeling, Adaptation and Personalization Adjunct* (Singapore; New York, NY: ACM), 317–322. doi: 10.1145/3213586.3225244

Pintrich, P. R. (1988). "A process-oriented view of student motivation and cognition," in *Improving Teaching and Learning Through Research: New Directions for Institutional Research*, Vol. 57, eds L. A. Mets and J. S. Stark (San Francisco, CA: Jossey-Bass, Spring), 65–79. doi: 10.1002/ir.37019885707

Ram, N., and Grimm, K. J. (2009). Growth mixture modeling: a method for identifying differences in longitudinal change among unobserved groups. *Int. J. Behav. Dev.* 33, 565–576. doi: 10.1177/0165025409343765

Ramsay, J. O., and Silverman, B. W. (2005). *Functional Data Analysis, 2nd edn*. New York, NY: Springer-Verlag. doi: 10.1007/b98888

Samé, A., Chamroukhi, F., Govaert, G., and Aknin, P. (2011). Model-based clustering and segmentation of time series with changes in regime. *Adv. Data Anal. Classificat.* 5, 301–321. doi: 10.1007/s11634-011-0096-5

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika* 52, 333–343. doi: 10.1007/BF02294360

Shu, Z., Bergner, Y., Zhu, M., Hao, J., and von Davier, A. A. (2017). An item response theory analysis of problem-solving processes in scenario-based tasks. *Psychol. Test Assess. Model.* 59, 109–131.

Stephen, D. G., Dixon, J. A., and Isenhower, R. W. (2009). Dynamics of representational change: entropy, action, and cognition. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 1811–1832. doi: 10.1037/a0014510

Thelen, E., and Smith, L. B. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge, MA: MIT Press.

Tiwari, V., and Kashikar, A. (2019). *OutlierDetection: Outlier Detection*. R package version 0.1.1.

van der Maas, H. L. J., Kolstein, R., and van der Pligt, J. (2003). Sudden transitions in attitudes. *Sociol. Methods Res.* 32, 125–152. doi: 10.1177/0049124103253773

van der Maas, H. L. J., and Molenaar, P. C. M. (1992). Stagewise cognitive development: an application of catastrophe theory. *Psychol. Rev.* 99, 395–417. doi: 10.1037/0033-295X.99.3.395

van Dijk, M., and van Geert, P. (2007). Wobbles, humps and sudden jumps: a case study of continuity, discontinuity and variability in early language development. *Infant Child Dev.* 16, 7–33. doi: 10.1002/icd.506

Van Geert, P. (1991). A dynamic systems model of cognitive and language growth. *Psychol. Rev.* 98, 3–53. doi: 10.1037/0033-295X.98.1.3

Vermunt, J. K., Langeheine, R., and Böckenholt, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *J. Educ. Behav. Stat.* 24, 179–207. doi: 10.3102/10769986024002179

Visser, I. (2011). Seven things to remember about hidden markov models: a tutorial on Markovian models for time series. *J. Math. Psychol.* 55, 403–415. doi: 10.1016/j.jmp.2011.08.002

Visser, I., and Speekenbrink, M. (2014). "It's a Catastrophe! Testing dynamics between competing cognitive states using mixture and hidden Markov models," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 36, 1688–1693. Available online at: https://escholarship.org/uc/item/0h89f38v

Voelkle, M. C., and Oud, J. H. L. (2013). Continuous time modelling with individually varying time intervals for oscillating and non-oscillating processes. *Brit. J. Math. Stat. Psychol.* 66, 103–126. doi: 10.1111/j.2044-8317.2012.02043.x

Wang, S., Yang, Y., Culpepper, S. A., and Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: a higher-order, hidden Markov model with covariates. *J. Educ. Behav. Stat.* 43, 57–87. doi: 10.3102/1076998617719727

Yang, M., and Chow, S.-M. (2010). Using state-space model with regime switching to represent the dynamics of facial electromyography (EMG) data. *Psychometrika Appl. Case Stud.* 74, 744–771. doi: 10.1007/s11336-010-9176-2

Yoon, J., Jordon, J., and Schaar, M. V. D. (2018). GAIN: missing data imputation using generative adversarial nets. *arXiv* arXiv:1806.02920.

Zeileis, A., and Grothendieck, G. (2005). zoo: S3 infrastructure for regular and irregular time series. *J. Stat. Softw.* 14, 1–27. doi: 10.18637/jss.v014.i06

Zelazo, P. D., Blair, C. B., and Willoughby, M. T. (2016). *Executive Function: Implications for Education. NCER 2017-2000*. National Center for Education Research.

Conflict of Interest: LO, AA, and TB were employed by the non-profit company ACT, Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 10 June 2020
doi: 10.3389/fpsyg.2020.01085

Using Two-Step Cluster Analysis and Latent Class Cluster Analysis to Classify the Cognitive Heterogeneity of Cross-Diagnostic Psychiatric Inpatients

Mariagrazia Benassi[1], Sara Garofalo[1], Federica Ambrosini[1], Rosa Patrizia Sant'Angelo[2], Roberta Raggini[2], Giovanni De Paoli[2], Claudio Ravani[2], Sara Giovagnoli[1], Matteo Orsoni[1] and Giovanni Piraccini[2]*

[1] Department of Psychology, University of Bologna, Bologna, Italy, [2] AUSL della Romagna, SPDC Psychiatric Emergency Unit, Cesena, Italy

OPEN ACCESS

Edited by:
Jason C. Immekus,
University of Louisville, United States

Reviewed by:
Andrea Rose Zammit,
Albert Einstein College of Medicine, United States
Ana Maria Hernandez Baeza,
University of Valencia, Spain

****Correspondence:***
Sara Garofalo
sara.garofalo@unibo.it

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *30 November 2019*
Accepted: *28 April 2020*
Published: *10 June 2020*

Citation:
Benassi M, Garofalo S, Ambrosini F, Sant'Angelo RP, Raggini R, De Paoli G, Ravani C, Giovagnoli S, Orsoni M and Piraccini G (2020) Using Two-Step Cluster Analysis and Latent Class Cluster Analysis to Classify the Cognitive Heterogeneity of Cross-Diagnostic Psychiatric Inpatients. Front. Psychol. 11:1085. doi: 10.3389/fpsyg.2020.01085

The heterogeneity of cognitive profiles among psychiatric patients has been reported to carry significant clinical information. However, how to best characterize such cognitive heterogeneity is still a matter of debate. Despite being well suited for clinical data, cluster analysis techniques, like the Two-Step and the Latent Class, received little to no attention in the literature. The present study aimed to test the validity of the cluster solutions obtained with Two-Step and Latent Class cluster analysis on the cognitive profile of a cross-diagnostic sample of 387 psychiatric inpatients. Two-Step and Latent Class cluster analysis produced similar and reliable solutions. The overall results reported that it is possible to group all psychiatric inpatients into Low and High Cognitive Profiles, with a higher degree of cognitive heterogeneity in schizophrenia and bipolar disorder patients than in depressive disorders and personality disorder patients.

Keywords: two-step cluster analysis, latent class cluster analysis, cognitive functioning, psychiatric inpatients, cluster analyses

INTRODUCTION

The traditional categorical nosology which mostly characterizes both research and clinical activity in psychology and psychiatry has been largely criticized in favor of a dimensional approach, which may better reflect the overlapping features of different disorders (Ivleva et al., 2012; Owoeye et al., 2013; van Os and Reininghaus, 2016). Cognitive impairment reflects one of the aspects shared by many psychiatric disorders, and it presents important overlaps with epidemiological, symptomatologic, and biological measures, as well as other risk factors (Smith and Weissman, 1992; Berrettini, 2000; Cosgrove and Suppes, 2013; Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013; Owoeye et al., 2013; Tamminga et al., 2014; Pearlson, 2015). The heterogeneity of cognitive profiles found among psychiatric patients has been reported to carry significant information about biomarkers, etiologies, and clinical factors (Mesholam-Gately et al., 2009; Bora, 2016), and about prognosis and treatment planning (Burdick et al., 2014; Lewandowski et al., 2014), which might have important implications for their treatment and prognosis (Cochrane et al., 2012). Interestingly, these findings are in line with the so-called genetic overlap among schizophrenia, bipolar disorder, depression, and personality disorder diagnosis that has been documented so far in different studies (Witt et al., 2017; Gandal et al., 2019). However, how to best characterize such cognitive heterogeneity across or within specific diagnostic categories in an informative way is still a matter of debate, and the use of well-suited statistical techniques to achieve stable and robust conclusions on this issue appears critical.

Clustering techniques can serve this purpose by identifying homogeneous subgroups presenting similar characteristics within a large cross-diagnostic sample (Allen and Goldstein, 2013). Amongst the several approaches available, the Two-Step cluster analysis (Chiu et al., 2001; Bacher et al., 2004) and the Latent Class cluster analysis appear to be well suited for clinical data, as they can handle ordinal as well as nominal variables, which can be more informative for clinical practice (Kent et al., 2014). Indeed, data obtained from classical neuropsychological tests are not purely quantitative and are better represented as nominal measures, i.e., classifying subjective performance according to normative values that specify whether the score is "above," "within," or "below" the normative range. Nevertheless, the most commonly used clustering methods adopted by previous studies investigating cognitive profiles of psychiatric inpatients are either hierarchical (Goldstein and Shelly, 1987; Hermens et al., 2011; Cotrena et al., 2017; Van Rheenen et al., 2017; Crouse et al., 2018; Lewandowski et al., 2018) or k-means (Lee et al., 2017). However, such methods present several limitations, like applicability to continuous variables only, assumption of normality of distribution, and an arbitrary choice of the number of clusters (Bacher et al., 2004; Matthiesen, 2010; Everitt, 2011; Mooi and Sarstedt, 2011).

From a detailed examination of the cluster solutions proposed from previous literature (**Supplementary Table S1**) on major psychiatric diagnoses, most studies reported either three (Hermens et al., 2011; Lee et al., 2015; Cotrena et al., 2017; Van Rheenen et al., 2017; Crouse et al., 2018) or four clusters (Goldstein and Shelly, 1987; Lewandowski et al., 2014, 2018; Reser et al., 2015), while only a few found two clusters (Lee et al., 2017). In all these studies, executive functions seemed to be the most important measures to explain the heterogeneity of psychiatric patients' cognitive profiles. Most studies focused on only one or two diagnostic categories, like schizophrenia and bipolar disorder (Goldstein and Shelly, 1987; Heinrichs and Zakzanis, 1998; Dawes et al., 2011; Hermens et al., 2011; Allen and Goldstein, 2013; Burdick et al., 2014; Cotrena et al., 2017; Lee et al., 2017; Roux et al., 2017; Van Rheenen et al., 2017; Crouse et al., 2018; Kollmann et al., 2019), with a few exceptions (Hermens et al., 2011; Lewandowski et al., 2014, 2018; Lee et al., 2015; Reser et al., 2015), thus limiting potential information about the differences and similarities between different diagnoses. Indeed, despite personality disorder being characterized by cognitive impairments similar to those presented by other psychiatric dysfunctions, like memory, attention, language, and executive functions (Dinn and Harris, 2000; Morgan and Lilienfeld, 2000; Dell'Osso et al., 2010; Cochrane et al., 2012; Rosell et al., 2014; Fineberg et al., 2015; Koch and Exner, 2015; McClure et al., 2016), these patients have been inexplicably neglected in this line of research.

Based on these considerations, the general goal of the present study was to identify subgroups of psychiatric inpatients based on cognitive nominal measures assessed in a large cross-diagnostic cohort (*N* = 387) including Schizophrenia Spectrum and Other Psychotic Disorders (SZ), personality disorders (PD), bipolar and related disorders (BD), and depressive disorders (DD). More specifically, we aimed to verify the best solution among those previously reported in the literature (ranging from two to four clusters; see **Supplementary Table S1**). The presence of a single cluster for all the diagnoses would suggest that all patients share a unique cognitive profile. The presence of two or more clusters would suggest the presence of different cognitive endophenotypes (e.g., preserved/impaired performances in specific cognitive domains or within specific diagnoses). To achieve a stable and robust solution, we provided several methodological and statistical improvements that allowed overcoming the limitations of previous similar studies (Hermens et al., 2011; Reser et al., 2015; Van Rheenen et al., 2017; Crouse et al., 2018). In particular: the stability of the clustering solution (Kraus et al., 2011) was checked by directly comparing two different techniques—Two-step and Latent Class cluster analysis—on several indexes of fit [Akaike information criterion (AIC), Bayes information criterion (BIC), and entropy]; the external validity of the solution was tested by comparing the obtained clustering solution on a different set of cognitive tests; the internal validity of the clustering solution was evaluated by running the same cluster analysis within each diagnostic subsample.

MATERIALS AND METHODS

Participants

Three hundred and eighty-seven participants were recruited from the Psychiatric Emergency Unit of the Health Clinical Service Azienda USL della Romagna (Cesena, Italy). Following the DSM-5 and ICD-10 criteria, patients with SZ, PD, BD, and DD were included in the study. The Mini-International Neuropsychiatric Interview (Sheehan et al., 1998) and the Structured Clinical Interview (First Michael et al., 1996) were used to confirm the psychiatric diagnosis. Exclusion criteria were insufficient Italian language skills, presence of neurological disorders, and severe visual or verbal impairments.

The participants were 189 males and 198 females with a mean age of 45.7 years. All the four diagnoses included were sufficiently represented numerically: 28% (*n* = 110) of the subjects had a diagnosis of SZ, 35% (*n* = 134) had a diagnosis of BD, 24% (*n* = 93) had a diagnosis of DD, and 13% (*n* = 50) had a diagnosis of PD. The demographic and clinical characteristics of the whole sample are reported in **Table 1**. Differences in cognitive performance among diagnoses are reported in the **Supplementary Information** and **Supplementary Figure S1**.

All procedures complied with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. The study was approved by the Research Ethical Committee of the AUSL Romagna (Regional Health Clinical Service). Written informed consent was acquired from each participant or, whenever necessary, from a parent or legal guardian.

TABLE 1 | Demographic and clinical characteristics of the whole sample.

		Participants *n = 387*
Age *mean (S.D.; range)*		45.7 (14.1; 17–80)
Gender *n* (%) M/F (% M)		189/198 (48.8)
Nationality *n* (%) Italian/others (% Italian)		292/95 (76.2)
Education *n* (%)	Primary school	17 (4.3)
	Secondary school	114 (29.4)
	High school	116 (30.0)
	Degree	23 (6.0)
	Missing	117 (30.3)
Diagnosis *n* (%)	Schizophrenia Spectrum and Other Psychotic Disorders	110 (28)
Bipolar and Related Disorders		134 (35)
Depressive Disorders		93 (24)
Personality Disorders		50 (13)
BPRSa *mean (S.D.)*		48.2 (10.3)
BPRSd *mean (S.D.)*		35.2 (7.5)
HoNOS *mean (S.D.)*		30.4 (6.5)

HoNOS, Health of the Nation Outcome Scales; BPRSa, Brief Psychiatric Rating Scale administered at admission; BPRSd, Brief Psychiatric Rating Scale administered at discharge.

Information about medication at the time of assessment was obtained from the medication list. All the patients were taking various combinations of mood stabilizers, antipsychotics, and antidepressants.

Cognitive and Clinical Assessment

The inpatients, admitted during the acute phase of illness, were recruited during the hospitalization. A team of psychologists and psychiatrists performed cognitive and clinical assessments. The complete assessment lasted approximately 3 h (see **Supplementary Information** for a comprehensive description of the tests used in the study).

The severity of symptomatology was measured at admission and at discharge with the Brief Psychiatric Rating Scale Expanded Version 4.0 (BPRS) (Ventura et al., 1993), while health and social functioning were measured with the Health of the Nation Outcome Scales—Roma (HoNOS) (Morosini et al., 2003).

Each patient completed two self-report questionnaires concerning the quality of life and the level of disability experienced during their daily life, respectively, the *World Health Organization Quality of Life—BREF* (WhoQoL) (Skevington et al., 2004) and the *World Health Organization Disability Assessment Schedule 2.0—36 items* (WhoDAS) (Üstün, 2010). The *UKU Side Effect rating scale* (Lingjaerde et al., 1987) was administered to evaluate the severity of pharmacological treatment side effects.

The *Tower of London—Drexel University* (ToL) (Culbertson and Zillmer, 2001) was used to assess planning abilities and problem-solving. The *Modified Wisconsin Card Sorting Test* (MCST) (Caffarra et al., 2004) was used to analyze the tendency toward perseveration and shifting. The *Attentional Matrices* (AM) (Spinnler and Tognoni, 1987) test was applied to evaluate selective visual attention. The *Stroop Word Interference Test* (STROOP) (Caffarra et al., 2002) was used as an index of selective attention, inhibitory control, and processing speed. The Italian standardized version of *Raven's Colored Progressive Matrices* (CPM-47) (Pruneti et al., 1996) was used to evaluate fluid intelligence.

A set of other cognitive measures was collected to explore the external validity of the clusters. Global cognitive functioning was assessed using the Mini Mental State Examination (MMSE) (Folstein et al., 1975) and the Clock Drawing Test (CDT) (Watson et al., 1993). Mental flexibility and verbal intelligence were assessed using Test dei Giudizi verbali e dei Compiti Astratti (Verbal abilities and abstract thinking test, GCA) (Spinnler and Tognoni, 1987). The Digit Span (Orsini et al., 1987) was used to assess short-term memory (SPAN Forward) and working memory (SPAN Backward).

For each test included in the cognitive assessment, detailed information about the purpose of the instrument, number of items and subscales, response recording method, administration time, scores, and psychometric properties is reported in the **Supplementary Information**.

Statistical Analysis

The variables used in the present study were standardized according to the normative scores available for each test (see **Supplementary Information**) by applying the following formula: $z = (x - \mu)/\sigma$, where x is the subject's raw score, μ represents the average obtained in the normative population, and σ is the normative population standard deviation. Then, following the indication of common clinical practice and the general guidelines for neuropsychological assessment (Mitrushina et al., 2005), the standardized scores were transformed into three categories: scores below the 10th percentile (corresponding to z score < -1.3) indicated cognitive deficit; scores equal or above the 10th and below the 90th percentile (corresponding to z score $> = -1.3$ and < 1.3) indicated normal cognitive functioning; and scores equal to or above the 90th percentile (corresponding to z score $> = 1.3$) indicated superior cognitive ability.

The variables included in both cluster analyses were: ToL (Total Number of Moves, Number of Correct Moves, Rule Violations, and Time Violations subscales), MCST (number of categories and Perseverative Errors subscales), CPM-47 total score, AM total score, and STROOP (Time and Errors subscales). The *Two-Step cluster analysis* is a hybrid approach which first uses a distance measure to separate groups and then a probabilistic approach (similar to latent class analysis) to choose the optimal subgroup model (Gelbard et al., 2007; Kent et al., 2014). Such a technique presents several advantages compared to more traditional techniques, like determining the number of clusters based on a statistical measure of fit (AIC or BIC) rather than on an arbitrary choice, using categorical and continuous variables simultaneously, analyzing atypical values (i.e., outliers), and being able to handle large datasets (Chiu et al., 2001; Bacher et al., 2004; Gelbard et al., 2007; Mooi and Sarstedt, 2011; Kent et al., 2014). Comparative studies regarded Two-Step cluster analysis as one of the most reliable in terms of the number of subgroups detected, classification probability of individuals to subgroups, and reproducibility of findings on clinical and other types of data (Bacher et al., 2004; Gelbard et al., 2007; Kent et al., 2014). The Two-Step cluster analysis was implemented in IBM SPSS Statistics (version 23.0) (Chiu et al., 2001; Bacher et al., 2004). In the first step (pre-clustering), a sequential approach is used to pre-cluster the cases based on the definition of dense regions in the analyzed attribute-space. In the second step (clustering), the pre-clusters are statistically merged in a stepwise way until all clusters are in one cluster.

The *Latent Class cluster analysis* consists of finding latent factors or class referred to a specific model that, from manifest variables, determines the differences among groups of subjects (Vermunt and Magidson, 2002, 2009; Allen and Goldstein, 2013; Kent et al., 2014). This approach is a model-based clustering technique in which, starting from the distribution of the data, each case or observation is probabilistically clustered into a latent class (McLachlan and Peel, 2000; Vermunt and Magidson, 2009). The model parameters are estimated as the proportion of observations in each latent class, and they are determined by the conditional probability of observing each response for each manifest variable in a given class. The cases presenting similar responses to the manifest variables are more likely included within the same latent class. Importantly, this approach is suitable for fitting ordinal manifest variables as well as nominal. The Latent Class cluster analysis was implemented using the R package "poLCA" (Haughton et al., 2009; Linzer and Lewis, 2011; Flynt and Dean, 2016). This procedure aims to fit a model in which any confounding between the manifest variables can be explained by a single unobserved "latent" categorical variable. Local independence is assumed to estimate a mixture model of latent multi-way tables.

Following a parsimony criterion, the best clustering solution was considered the one with the best balance between the number of clusters considered and the corresponding fit. Based on previous literature (see **Supplementary Table S1**), solutions ranging from two to four clusters were considered. BIC, AIC, and entropy were first calculated for each cluster solution and then used to find the greatest change in distance between two cluster solutions. BIC, AIC, and entropy change were calculated as the difference between two cluster solutions starting from the most parsimonious (one cluster) to the less parsimonious (four clusters), thus obtaining three values (2vs1, 3vs2, and 4vs3). The best cluster solution was considered the one with the strongest change and the lower number of clusters. This allowed evaluating the most parsimonious cluster solution presenting the best fit. Such a procedure was performed automatically for the Two-Step cluster analysis and implemented via a custom-made script implemented in R for the Latent Class cluster analysis.

Aiming for a detailed description of the selected clustering solution, the clusters were compared based on clinical and psychosocial functioning using a general linear model on the following continuous variables: severity of psychiatric symptoms (HoNOS and BPRS), side effects of pharmacological treatment (UKU), duration of hospitalization, number of hospitalizations, and quality of life (WhoQoL and WhoDAS). A chi-squared test was used to compare the frequency of diagnosis between the two clusters.

The external validity of the clustering solutions was verified by comparing the clusters (independent variable) on a different set of cognitive tests (dependent variables), including global cognitive functioning (MMSE and CDT), mental flexibility and verbal intelligence (GCA), short-term memory (Digit Span Forward), and working memory (Digit Span Backward). General linear models were used for normally distributed variables (MMSE and CDT). Mann–Whitney tests were used for non-normally distributed variables (GCA and Digit Span Forward and Backward).

The internal validity of the clustering solution was evaluated by dividing the sample according to the diagnosis and running both the Two-Step and Latent Class cluster analysis on each subsample. Cohen's Kappa statistic was calculated to test the degree of agreement between the cluster assignment for each subject when considered in the cross-diagnostic sample and within the single diagnostic subsample.

RESULTS

The results that emerged from both the Two-Step and the Latent Class cluster analysis reported a two-cluster classification as the optimal solution for the data considered in the present study. That is, following a parsimony criterion (see the *Statistical Analysis* section), the two-cluster solution presented the greatest BIC, AIC, and entropy change between the two closest clusters at each stage (**Figure 1** and **Supplementary Table S2**). Following the principle of parsimony, the best cluster solution is the one with the highest value of the difference between two indexes of n cluster and n plus one cluster. This way to select the best cluster solution allows evaluating the improvement of homogeneity within each cluster and the heterogeneity between the clusters from one cluster to n cluster by adding one cluster at each step.

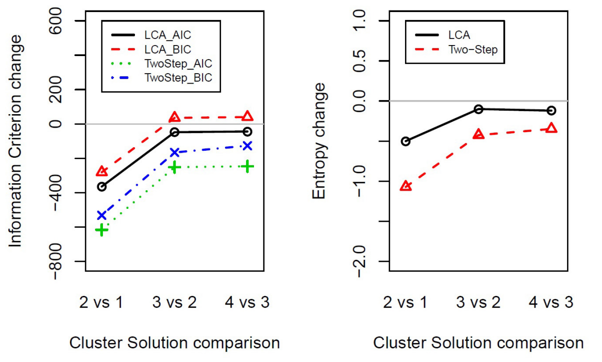


FIGURE 1 | Indexes of fit changes obtained from Latent Class cluster analysis and Two-Step cluster analysis for solutions ranging from one to four clusters. The panels show the change in information criterion (left) or entropy (right) between two close clusters' solutions (e.g., 2vs1 shows two-cluster solution minus one-cluster solution). LCA, Latent Class cluster analysis; TwoStep, Two-Step cluster analysis; BIC, Bayesian information criterion; AIC, Akaike information criterion.

TABLE 2 | Description of the two clusters according to the number and percentage of cases scoring below, within, and above the normative scores for each cognitive test.

		Cluster 1 *Low Cognitive Profile*			Cluster 2 *High Cognitive Profile*			
	Tests N (%)	**Below**	**Within**	**Above**	**Below**	**Within**	**Above**	χ^2
Two-Step	MCST categories	58 (29)	69 (34)	76 (37)	17 (9)	26 (14)	141 (77)	60.56; $p < 0.001$
	MCST errors	55 (27)	82 (40)	66 (33)	14 (8)	61 (33)	109 (59)	37.17; $p < 0.001$
	CPM-47	58 (29)	91 (45)	54 (27)	12 (7)	60 (33)	112 (61)	56.06; $p < 0.001$
	AM	81 (40)	64 (32)	58 (28)	28 (15)	64 (35)	92 (50)	32.62; $p < 0.001$
	ToL Rule Violations	173 (85)	27 (13)	3 (1)	60 (33)	112 (61)	12 (7)	111.52; $p < 0.001$
	ToL N of correct moves	52 (26)	147 (72)	4 (2)	10 (5)	124 (67)	50 (27)	68.82; $p < 0.001$
	ToL Time Violations	167 (82)	35 (17)	1 (1)	43 (23)	133 (72)	8 (4)	135.22; $p < 0.001$
	ToL total N of moves	150 (74)	52 (26)	1 (0)	13 (7)	138 (75)	33 (18)	183.7; $p < 0.001$
	STROOP Time	112 (55)	55 (27)	36 (18)	39 (21)	75 (41)	70 (38)	48.46; $p < 0.001$
	STROOP Errors	64 (32)	75 (37)	64 (32)	20 (11)	67 (36)	97 (53)	29.4; $p < 0.001$
Latent Class	MCST categories	69 (38)	67 (36)	48 (26)	6 (3)	28 (14)	169 (83)	135.8; $p < 0.001$
	MCST errors	60 (33)	82 (45)	42 (22)	9 (4)	61 (30)	133 (66)	87.38; $p < 0.001$
	CPM-47	63 (34)	92 (50)	29 (16)	7 (3)	59 (29)	137 (68)	121.64; $p < 0.001$
	AM	91 (49)	56 (31)	37 (20)	18 (9)	72 (35)	113 (56)	88.68; $p < 0.001$
	ToL Rule Violations	157 (85)	25 (14)	2 (1)	76 (37)	114 (56)	13 (7)	92.50; $p < 0.001$
	ToL N of correct moves	47 (26)	131 (71)	6 (4)	15 (7)	140 (69)	48 (24)	48.67; $p < 0.001$
	ToL Time Violations	138 (75)	45 (24)	1 (1)	72 (35)	123 (61)	8 (4)	61.62; p < 0.001
	ToL total N of moves	118 (64)	62 (34)	4 (2)	45 (22)	128 (63)	30 (15)	74.75; $p < 0.001$
	STROOP Time	109 (59)	40 (22)	35 (19)	42 (21)	90 (44)	71 (35)	60.40; $p < 0.001$
	STROOP Errors	66 (36)	58 (32)	60 (33)	18 (9)	84 (41)	101 (50)	41.80; $p < 0.001$

Performance class of score below (z score < −1.3) average (z score between −1.3 and +1.3) and above (z score > 1.3). MCST, Modified Wisconsin Card Sorting Test; CPM-47, Raven's Colored Progressive Matrix; AM, Attentional Matrices; ToL, Tower of London—Drexel University test; STROOP, Stroop Word Interference Test.

The frequency distribution of performances scoring below, within, and above the normative sample for each cognitive test was examined to define the composition of the two clusters (**Table 2**). The results showed a significantly higher presence of performances classified as "below" in one cluster and "within" or "above" in the other cluster, for both the Two-Step and the Latent Class clustering solutions (**Table 2**). Consequently, one group was defined as the Low Cognitive Profile cluster (including 48% of subjects for the Two-Step clustering solution and 52% of subjects for the Latent Class clustering solution), and the other group was defined as the High Cognitive Profile cluster. The contribution of each cognitive test to such a clustering solution is represented in **Figure 2**. For the Latent Class cluster analysis, the major cognitive differences between clusters concerned perseveration and shifting abilities (MCST), fluid intelligence (CPM-47), and selective visual attention (AM), while for the Two-Step cluster analysis, the major cognitive differences between clusters concerned planning abilities and problem-solving (ToL). Since the two clusters reported differences in age ($F_{2,304} = 0.63$; $p = 0.533$; partial $\eta^2 = 0.004$) and education ($F_{2,304} = 2.64$; $p = 0.073$; partial $\eta^2 = 0.017$), these two variables were introduced as covariates in all analyses. A general linear model was applied to verify whether the clusters differed in clinical and psychosocial functioning. Although with some discrepancies between the Two-Step and the Latent Class clustering solutions, the Low Cognitive Profile cluster generally reported higher severity of symptoms (HoNOS and BPRS at admission and discharge), higher side effects of pharmacological treatment (UKU), lower improvement in BPRS symptom severity between admission and discharge, and longer duration of hospitalization than the High Cognitive Profile cluster (**Table 3**). No differences were found on measures of quality of life (WhoQoL and WhoDAS) and the number of hospitalizations (**Table 3**). The diagnoses were differently represented in the two clusters. Most of the schizophrenia and bipolar disorder patients were similarly distributed between the High and Low Cognitive Profile clusters, while most depressive disorder and personality disorder patients were more represented in the High Cognitive Profile cluster (**Table 3**).

The analysis for the external validity confirmed the presence of poorer global functioning, short-term memory, working memory, and mental flexibility and verbal intelligence in the Low Cognitive Profile cluster as compared to the High Cognitive Profile cluster. Such differences were present in both the clustering solutions identified by means of Two-Step cluster analysis (MMSE, $F_{1,297} = 60.72$, $p < 0.001$, partial $\eta^2 = 0.170$; CDT, $F_{1,123} = 19.21$, $p < 0.001$, partial $\eta^2 = 0.135$; GCA, $U = 6{,}314.00$, $p < 0.001$; SPAN Forward, $U = 8{,}130.50$, $p = 0.018$; SPAN Backward, $U = 7{,}181.50$, $p < 0.001$) and Latent Class cluster analysis (MMSE, $F_{1,296} = 65.83$, $p < 0.001$, partial $\eta^2 = 0.18$; CDT, $F_{1,122} = 24.67$, $p < 0.0001$, partial $\eta^2 = 0.17$; GCA, $U = 6{,}314.00$, $p < 0.001$; SPAN Forward, $U = 8{,}000$, $p < 0.001$; SPAN Backward, $U = 7{,}000$, $p < 0.001$). The Low Cognitive

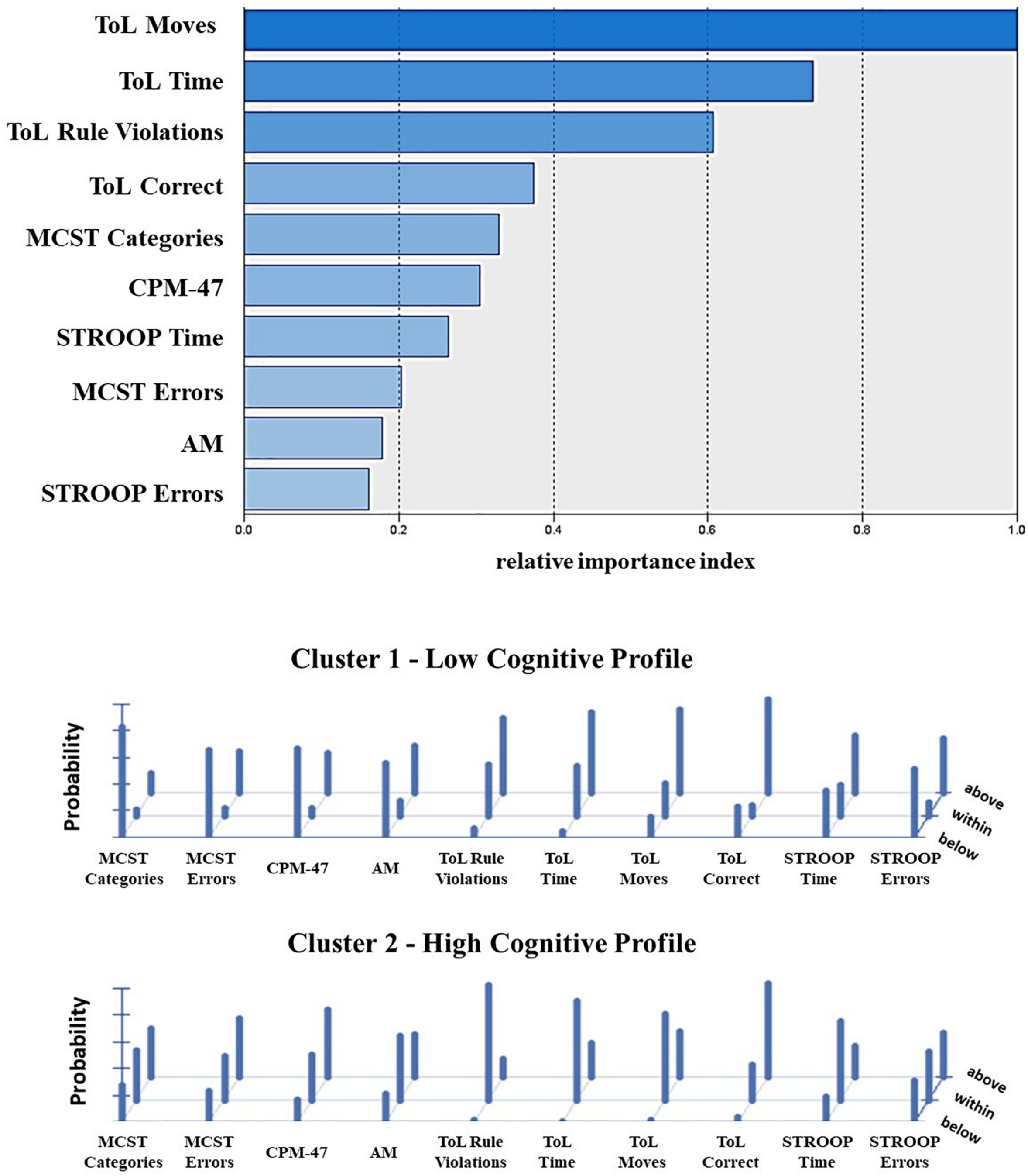


FIGURE 2 | Contribution of the single cognitive tests to the clustering solution as reported from the Two-Step **(top)** and Latent Class cluster analysis **(bottom)**. The top panel shows the index of relative importance of each cognitive test as identified by the Two-Step cluster analysis. The panel on the bottom shows the conditional item response probabilities for the two clusters identified by the Latent Class cluster analysis. Performance class of score below (z score < –1.3) average (z score between –1.3 and +1.3) and above (z score > 1.3) the normative sample. MCST, Modified Wisconsin Card Sorting Test; CPM-47, Raven's Colored Progressive Matrix; AM, attentional matrices; ToL, tower of London—Drexel University test; STROOP, Stroop Word Interference Test.

Profile performed worse than the High Cognitive Profile in all the tests: MMSE, Low Cognitive Profile = 26.16 (S.E. = 0.21) vs High Cognitive Profile 28.36 (SE.19); CDT, Low Cognitive Profile = 10.27 (S.E. = 0.39) vs High Cognitive Profile 12.70 (SE.39); GCA, Low Cognitive Profile mean rank = 114.47 vs High Cognitive Profile mean rank = 189.12; SPAN-Forward, Low Cognitive Profile mean rank = 128.04 vs High Cognitive mean rank = 160.50; SPAN Forward mean rank = 120.74 vs high cognitive mean rank = 166.39. These results showed that in all the tests, the Low Cognitive Profile obtained with

TABLE 3 | Clinical characteristics and distribution of diagnoses in the two clusters.

			Cluster 1	**Cluster 2**		
			Low cognitive profile	***High cognitive profile***	**Statistic**	
Two-Step	Test *mean (s.e.)*	HoNOS	31.87 (0.51)	29.28 (0.56)	$F_{1,321} = 11.86$	$p = 0.001$
		BPRSa	49.67 (0.74)	46.55 (0.78)	$F_{1,321} = 8.26$	$p = 0.004$
		BPRSd	36.44 (0.57)	34.31 (0.63)	$F_{1,321} = 6.23$	$p = 0.013$
		BPRSa-d	12.32 (0.52)	12.78 (0.57)	$F_{1,313} = 8.81$	$p = 0.003$
		UKU	3.30 (0.19)	3.01 (0.19)	$F_{1,175} = 1.2$	$p = 0.276$
		WhoQoL	81.68 (1.44)	78.5 (1.54)	$F_{1,338} = 2.22$	$p = 0.137$
		WhoDAS	80.72 (2.27)	83.45 (2.68)	$F_{1,261} = 0.60$	$p = 0.438$
	Hosp.	Duration	13.98	12.27	$F_{1,385} = 3.99$	$p = 0.05$
		Number	1.68	1.82	$F_{1,385} = 0.83$	$p = 0.36$
	Diagnosis *N (%)*	*BD*	77 (57)	57 (43)	$\chi^2_3 = 16.58$	$p = 0.001$
		DD	41 (44)	52 (56)		
		PD	18 (36)	32 (64)		
		SZ	67 (61)	43 (39)		
Latent Class	Test *mean (s.e.)*	HoNOS	32.72 (0.52)	28.81 (0.042)	$F_{1,321} = 28.56$	$p < 0.001$
		BPRSa	50.20 (0.74)	46.45 (0.79)	$F_{1,321} = 11.53$	$p = 0.001$
		BPRSd	37.13 (0.06)	33.91 (0.49)	$F_{1,321} = 14.75$	$p < 0.001$
		BPRSa-d	13.50 (0.53)	11.80 (0.56)	$F_{1,321} = 1.81$	$p = 0.17$
		UKU	3.51 (0.14)	2.88 (0.12)	$F_{1,175} = 5.74$	$p = 0.018$
		WhoQoL	80.85 (1.34)	79.64 (1.44)	$F_{1,338} = 0.33$	$p = 0.568$
		WhoDAS	79.64 (0.10)	84.03 (2)	$F_{1,261} = 1.61$	$p = 0.21$
	Hosp.	Duration	14.4	12	$F_{1,385} = 7.56$	$p = 0.006$
		Number	1.8	1.7	$F_{1,385} = 0.48$	$p = 0.49$
	Diagnosis *N (%)*	*BD*	73 (54)	61 (46)	$\chi^2_3 = 30$	$p < 0.001$
		DD	26 (28)	67 (72)		
		PD	19 (38)	31 (62)		
		SZ	66 (60)	44 (40)		

Hosp., hospitalization; BPRSa-d, Brief Psychiatric Rating Scale difference between BPRSa and BPRSd; UKU, UKU side effect rating scale; WhoQoL, world health organization quality of life—BREF scale; Who DAS, world health organization disability assessment schedule; SZ, schizophrenia spectrum and other psychotic disorders; BD, bipolar and related disorders; DD, depressive disorders; PD, personality disorders.

Two-Step cluster analysis performed worse than the High Cognitive Profile.

The internal validity of the clustering solution was verified by applying the same cluster procedures on each of the four diagnostic groups separately. The results reported the two-cluster classification as the optimal solution within each diagnosis (**Supplementary Figure S2** and **Supplementary Table S2**), thus confirming the result obtained on the cross-diagnostic sample as stable and consistent. Cohen's Kappa statistics showed a significant agreement between the results of the whole cross-diagnostic sample and those emerging from the single diagnostic subsamples for both the Two-Step (Kappa = 0.66; $p < 0.001$) and the Latent Class (Kappa = 0.72; $p < 0.001$) cluster analysis. Patients were re-classified according to the cross-diagnostic solution in 83% of cases for the Two-Step clustering solution and in the 87% of cases for the Latent Class clustering solution. Overall, the two clusters obtained within each diagnosis were confirmed as being characterized by a lower and a higher cognitive profile (**Supplementary Tables S3, S4**). However, important differences were observed between the diagnoses. Indeed, for both clustering techniques, while schizophrenia and bipolar disorder patients showed a clear-cut separation and a fairly even distribution of subjects between the two clusters, depressive disorder and personality disorder patients were more represented in the High Cognitive Profile cluster (**Figure 3**; see also **Table 3**), thus showing lower cognitive heterogeneity.

To support of the validation of the two cluster solutions obtained with categorical variables, we applied the Two-Step cluster analysis to quantitative data (i.e., standardized scores). Results showed that the two cluster solutions remained the best option according to AIC and BIC changes (see **Supplementary Table S5**).

DISCUSSION

The main findings here reported responded to our general aim to find reliable and robust cognitive clusters of psychiatric inpatients by comparing Two-Step and Latent Class cluster analysis. To our knowledge, despite the wide use of different cluster analyses in former literature, no study compared different clustering approaches that can handle nominal data on a cross-diagnostic sample of psychiatric inpatients. The two cluster analyses converged on finding the presence of two separate clusters (Low and High) as the most efficient and robust description of the whole sample's cognitive profile. Importantly,

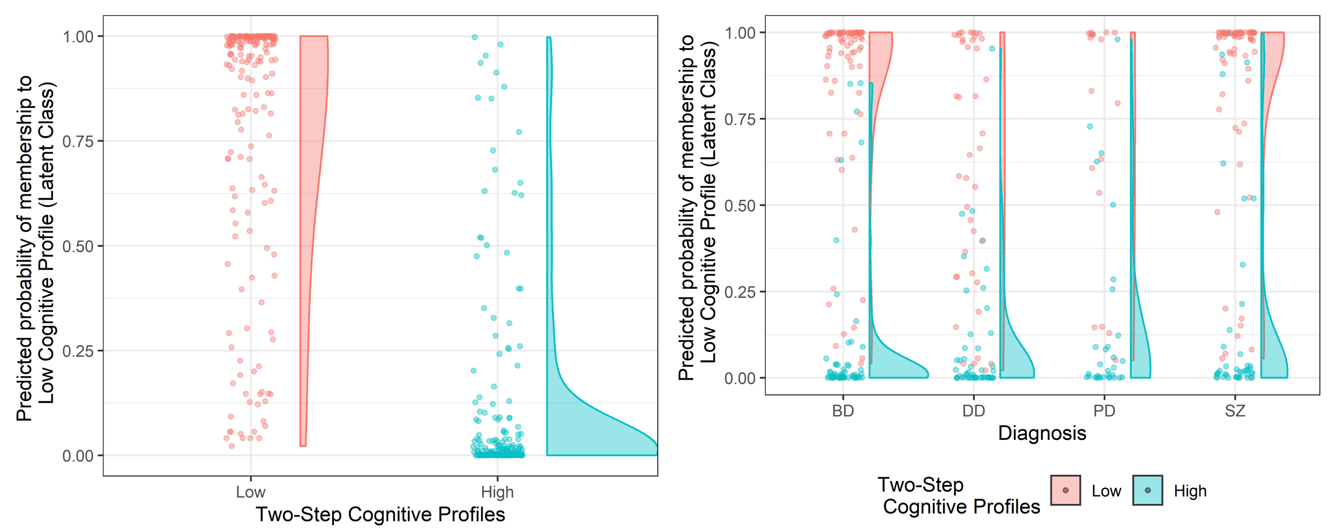


FIGURE 3 | Cluster assignment according to the Two-Step clustering solution as a function of the predicted probability of cluster membership of the Latent Class clustering solution, on the cross-diagnostic sample and within each diagnosis. The left panel represents the clustering solutions obtained on the cross-diagnostic sample. The panel on the right represents the clustering solutions obtained within each diagnosis. SZ, Schizophrenia Spectrum and Other Psychotic Disorders; BD, Bipolar and Related Disorders; DD, Depressive Disorders; PD, Personality Disorders.

clustering was not dependent on pharmacological treatment side effects, as the two clusters reported comparable levels of iatrogenic effects. Measures of internal and external validity also confirmed the two-cluster classification as the best solution.

The analysis performed within each diagnostic sample showed that while schizophrenia and bipolar disorder were similarly represented in the two clusters, depressive disorder and personality disorder patients were overrepresented in the High Cognitive Profile cluster (**Figure 3** and **Table 3**), thus indicating a higher cognitive heterogeneity in the first two diagnostic categories than in the last two. Crucially, given the known link with biomarkers, etiologies, and clinical factors reported in the literature about cognitive heterogeneity (Burdick et al., 2014; Lewandowski et al., 2014), such differentiation can be informative for clinical practice in terms of both prognosis and treatment planning (Cochrane et al., 2012; Burdick et al., 2014; Lewandowski et al., 2014). Indeed, the two clusters resulted as different in terms of severity and improvement of the symptomatology, side effects of pharmacological treatment, and duration of hospitalization.

The number of clusters here obtained is dissimilar to most of the previous studies using cross-diagnostic samples. A direct comparison between different cluster analytic studies is always problematic, as the clustering solutions are highly sensitive to the input data and the algorithm chosen (Marquand et al., 2016). For example, due to the marked variability of neuropsychological measures used by the previous studies above mentioned, any consideration would be limited by the absence of cluster analytic studies based on the same input data but extended to different cohorts. Nevertheless, we will try to examine the main differences and similarities with previous studies, in the attempt to obtain a more general overview of the currently available evidence (**Supplementary Table S1**). A recent study from Lee et al. (2017) in schizophrenia and bipolar disorder patients reported two clusters (for a complete overview, see **Supplementary Table S1**). Conversely, most studies reported either three (Hermens et al., 2011; Lee et al., 2015; Cotrena et al., 2017; Van Rheenen et al., 2017; Crouse et al., 2018) or four clusters (Goldstein and Shelly, 1987; Lewandowski et al., 2014, 2018; Reser et al., 2015). The main reason for obtaining more than two clusters could be attributed to the inclusion of healthy subjects within the cluster analysis and the presence of verbal reasoning tests, which we excluded in favor of a deeper evaluation of executive functions, as classically reported as the most important measures to explain the heterogeneity of cognitive profiles (Goldstein and Shelly, 1987; Hermens et al., 2011; Lewandowski et al., 2014, 2018; Lee et al., 2015, 2017; Reser et al., 2015; Cotrena et al., 2017; Van Rheenen et al., 2017; Crouse et al., 2018). Relatedly, some authors indicated that intermediary clusters could reflect a degree of normal variability across measures of cognitive functioning (Binder et al., 2009) that may underpin different brain abnormalities as far as nature and severity are concerned (Demjaha et al., 2012; Woodward, 2016). However, whether the clusters characterized by selective cognitive impairment represent distinct profiles or only reflect artificial divisions along a continuum of severity is a matter of debate (Wykes and Reeder, 2005). Indeed, the results reported may, at least in part, be confounded by the statistical and methodological limitations of these studies. Indeed, in contrast with previous literature, the robustness of the selected cluster solution was here tested by comparing two clustering techniques, namely Two-Step and Latent Class cluster analysis, that can both handle nominal data and continuous data and are based on optimal BIC and AIC indexes of fit (Chiu et al., 2001; Haughton et al., 2009). These two critical points are the main strengths of the two approaches. Moreover, some specific features of each technique should be mentioned. While the Two-Step cluster analysis is based on a fixed model procedure, in the Latent Class, a probability-based classification is computed for each subject according to the specific model selected by the researcher. Therefore, in the Latent Class cluster analysis, it is possible to obtain the subjective probability membership to each cluster (**Figure 3**). These aspects already have been discussed in previous literature (Chiu et al., 2001), but no previous study attempted to use them as a validation method for determining the stability of the selected cluster solution. Furthermore, given the known limitations of the cluster analysis, internal and external validation of a clustering solution, as reported in the present study, is always crucial (Marquand et al., 2016). A review by Marquand et al. (2016) has well explained that applying a cluster analysis necessarily entails some heuristics, concerning the choice of algorithm, distance function, and model order, which influence the clustering solution and complicate potential quantitative comparisons between different studies and cohorts. Unfortunately, only a few cross-diagnostic studies provided a validation of the clustering solution obtained (Hermens et al., 2011; Lee et al., 2015; Reser et al., 2015; Van Rheenen et al., 2017; Crouse et al., 2018). The two clusters identified in the present study can be considered as robust since both the external and internal validity of the clustering solution were verified. That is, the Low and High Cognitive Profiles were distinguishable also when compared based on a set of cognitive measures not considered during the cluster analysis and when applying the same cluster procedure on each of the four diagnostic groups separately.

Some limitations of the present study should also be mentioned. Personality Disorder patients are slightly underrepresented in the whole sample. This limitation may have biased the results; therefore, additional studies are needed to better understand if it is possible to find specific cognitive profiles in Personality Disorder patients. Although we attempted to analyze the contribution of pharmacological treatment in the clustering solution, we could only evaluate the iatrogenic effect. Further studies are required to investigate the effect of pharmacological treatment in grouping the cognitive performance of psychiatric patients.

CONCLUSION

Despite the large variety of solutions proposed by previous literature, the application and comparison of Two-Step and Latent Class cluster analysis on four possible clustering solutions (one to four clusters) allowed confirmation of the robustness of two clusters as the best representation of the cognitive heterogeneity characterizing large cross-diagnostic psychiatric inpatients. The presence of similar solutions obtained with two separate procedures suggests a combined use for future applications to maximize the criteria selection efficiency. These results have also important clinical implications. By clarifying that two subgroups of patients with low or high cognitive abilities can be identified in all the diagnostic groups, we envision the possibility to find specific phenotypes connected to executive functions. These two groups, irrespectively from the diagnosis, present different symptom severity and prognosis (better outcome and lower duration of hospitalization for those patients who are not cognitively impaired as compared to the ones with cognitive deficits). This result informs clinical practice about the fact that specific cognitive training could be proposed to psychiatric patients with low cognitive profile, and suggests that a specific cognitive evaluation could enhance the clinical effectiveness for personalized intervention.

DATA AVAILABILITY STATEMENT

The data supporting the findings of the present study can be found in the **Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ausl della Romagna, Ethical Committee. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MB, SGi, CR, and GP developed the main hypothesis, and all authors contributed to the study design. Patient screening, testing, and the data collection were carried on by FA, RS, RR, GD, and GP. Data analysis was performed by SGa, FA, and MO under the supervision of MB, GP, and SGi. The main text was written by MB, SGa, and FA. All authors approved the final version.

ACKNOWLEDGMENTS

We are grateful to the patients for their participation. Our thanks go to the Psychiatric Emergency Unit team (U.O. SPDC, AUSL della Romagna, Cesena). Special thanks go to Prof. Roberto Bolzani, for his support in the statistical analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01085/full#supplementary-material

REFERENCES

Allen, D. N., and Goldstein, G. (2013). *Cluster Analysis in Neuropsychological Research: Recent Applications*. New York, NY: Springer Science & Business Media.

Bacher, J., Wenzig, K., and Vogler, M. (2004). SPSS twostep cluster - a first evaluation. *Univ. Erlangennürnb.* 1, 1–20.

Berrettini, W. H. (2000). Are schizophrenic and bipolar disorders related? A review of family and molecular studies. *Biol. Psychiatry* 48, 531–538. doi: 10.1016/S0006-3223(00)00883-0

Binder, L. M., Iverson, G. L., and Brooks, B. L. (2009). To err is human: "abnormal" neuropsychological scores and variability are common in healthy adults. *Arch. Clin. Neuropsychol.* 24, 31–46. doi: 10.1093/arclin/acn001

Bora, E. (2016). Differences in cognitive impairment between schizophrenia and bipolar disorder: considering the role of heterogeneity. *Psychiatry Clin. Neurosci.* 70, 424–433. doi: 10.1111/pcn.12410

Burdick, K. E., Russo, M., Frangou, S., Mahon, K., Braga, R. J., Shanahan, M., et al. (2014). Empirical evidence for discrete neurocognitive subgroups in bipolar disorder: clinical implications. *Psychol. Med.* 44, 3083–3096. doi: 10.1017/S0033291714000439

Caffarra, P., Vezzadini, G., Dieci, F., Zonato, F., and Venneri, A. (2002). A short version of the Stroop test: normative data in an Italian population sample. *Nuova Riv. Neurol.* 12, 111–115.

Caffarra, P., Vezzadini, G., Dieci, F., Zonato, F., and Venneri, A. (2004). Modified card sorting test: normative data. *J. Clin. Exp. Neuropsychol.* 26, 246–250. doi: 10.1076/jcen.26.2.246.28087

Chiu, T., Fang, D., Chen, J., Wang, Y., and Jeris, C. (2001). "A robust and scalable clustering algorithm for mixed type attributes in large database environment," in *Proceedings of The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '01*, (New York, NY: ACM Press), 263–268.

Cochrane, M., Petch, I., and Pickering, A. D. (2012). Aspects of cognitive functioning in schizotypy and schizophrenia: evidence for a continuum model. *Psychiatry Res.* 196, 230–234. doi: 10.1016/j.psychres.2012.02.010

Cosgrove, V. E., and Suppes, T. (2013). Informing DSM-5: biological boundaries between bipolar I disorder, schizoaffective disorder, and schizophrenia. *BMC Med.* 11:127. doi: 10.1186/1741-7015-11-127

Cotrena, C., Damiani Branco, L., Ponsoni, A., Milman Shansis, F., and Paz Fonseca, R. (2017). Neuropsychological clustering in bipolar and major depressive disorder. *J. Int. Neuropsychol. Soc.* 23, 584–593. doi: 10.1017/S1355617717000418

Cross-Disorder Group of the Psychiatric Genomics Consortium (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* 381, 1371–1379. doi: 10.1016/S0140-6736(12)62129-1

Crouse, J. J., Moustafa, A. A., Bogaty, S. E. R., Hickie, I. B., and Hermens, D. F. (2018). Parcellating cognitive heterogeneity in early psychosis-spectrum illnesses: a cluster analysis. *Schizophr. Res.* 202, 91–98. doi: 10.1016/j.schres.2018.06.060

Culbertson, W. C., and Zillmer, E. A. (2001). *Tower of London-Drexel University (TOLDX)*. North Tonawada: Multi-Health Systems.

Dawes, S. E., Jeste, D. V., and Palmer, B. W. (2011). Cognitive profiles in persons with chronic schizophrenia. *J. Clin. Exp. Neuropsychol.* 33, 929–936. doi: 10.1080/13803395.2011.578569

Dell'Osso, B., Berlin, H. A., Serati, M., and Altamura, A. C. (2010). Neuropsychobiological aspects, comorbidity patterns and dimensional models in borderline personality disorder. *Neuropsychobiology* 61, 169–179. doi: 10.1159/000297734

Demjaha, A., MacCabe, J. H., and Murray, R. M. (2012). How genes and environmental factors determine the different neurodevelopmental trajectories of schizophrenia and bipolar disorder. *Schizophr. Bull.* 38, 209–214. doi: 10.1093/schbul/sbr100

Dinn, W. M., and Harris, C. L. (2000). Neurocognitive function in antisocial personality disorder. *Psychiatry Res.* 97, 173–190.

Everitt, B. (2011). *Cluster Analysis*, 5th Edn. Thousand Oaks, CA: SAGE Publications.

Fineberg, N. A., Day, G. A., de Koenigswarter, N., Reghunandanan, S., Kolli, S., Jefferies-Sewell, K., et al. (2015). The neuropsychology of obsessive-compulsive personality disorder: a new analysis. *CNS Spectr.* 20, 490–499. doi: 10.1017/S1092852914000662

First Michael, B., Spitzer Robert, L., Gibbon, M., and Janet, B. W. (1996). *Structured Clinical Interview for DSM-IV Axis I Disorders, Clinician Version (SCID-CV)*. Washington, DC: American Psychiatric Press Inc.

Flynt, A., and Dean, N. (2016). A survey of popular R packages for cluster analysis. *J. Educ. Behav. Stat.* 41, 205–225. doi: 10.3102/1076998616631743

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). Mini-mental state. A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12, 189–198.

Gandal, M. J., Haney, J. R., Parikshak, N. N., Leppa, V., Ramaswami, G., Hartl, C., et al. (2019). Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Focus* 17, 66–72. doi: 10.1176/appi.focus.17103

Gelbard, R., Goldman, O., and Spiegler, I. (2007). Investigating diversity of clustering methods: an empirical comparison. *Data Knowl. Eng.* 63, 155–166. doi: 10.1016/j.datak.2007.01.002

Goldstein, G., and Shelly, C. (1987). The classification of neuropsychological deficit. *J. Psychopathol. Behav. Assess.* 9, 183–202. doi: 10.1007/BF00960574

Haughton, D., Legrand, P., and Woolford, S. (2009). Review of three latent class cluster analysis packages: latent gold, poLCA, and MCLUST. *Am. Stat.* 63, 81–91. doi: 10.1198/tast.2009.0016

Heinrichs, R. W., and Zakzanis, K. K. (1998). Neurocognitive deficit in schizophrenia: a quantitative review of the evidence. *Neuropsychology* 12, 426–445. doi: 10.1037/0894-4105.12.3.426

Hermens, D. F., Redoblado Hodge, M. A., Naismith, S. L., Kaur, M., Scott, E., and Hickie, I. B. (2011). Neuropsychological clustering highlights cognitive differences in young people presenting with depressive symptoms. *J. Int. Neuropsychol. Soc.* 17, 267–276. doi: 10.1017/S1355617710001566

Ivleva, E. I., Morris, D. W., Osuji, J., Moates, A. F., Carmody, T. J., Thaker, G. K., et al. (2012). Cognitive endophenotypes of psychosis within dimension and diagnosis. *Psychiatry Res.* 196, 38–44. doi: 10.1016/j.psychres.2011.08.021

Kent, P., Jensen, R. K., and Kongsted, A. (2014). A comparison of three clustering methods for finding subgroups in MRI, SMS or clinical data: SPSS twostep cluster analysis, latent Gold and SNOB. *BMC Med. Res. Methodol.* 14:113. doi: 10.1186/1471-2288-14-113

Koch, J., and Exner, C. (2015). Selective attention deficits in obsessive-compulsive disorder: the role of metacognitive processes. *Psychiatry Res.* 225:550. doi: 10.1016/j.psychres.2014.11.049

Kollmann, B., Yuen, K., Scholz, V., and Wessa, M. (2019). Cognitive variability in bipolar I disorder: a cluster-analytic approach informed by resting-state data. *Neuropharmacology* 156:107585. doi: 10.1016/j.neuropharm.2019.03.028

Kraus, J. M., Müssel, C., Palm, G., and Kestler, H. A. (2011). Multi-objective selection for collecting cluster alternatives. *Comput. Stat.* 26, 341–353. doi: 10.1007/s00180-011-0244-6

Lee, J., Rizzo, S., Altshuler, L., Glahn, D. C., Miklowitz, D. J., Sugar, C. A., et al. (2017). Deconstructing bipolar disorder and schizophrenia: a cross-diagnostic cluster analysis of cognitive phenotypes. *J. Affect. Disord.* 209, 71–79. doi: 10.1016/j.jad.2016.11.030

Lee, R. S. C., Hermens, D. F., Naismith, S. L., Lagopoulos, J., Jones, A., Scott, J., et al. (2015). Neuropsychological and functional outcomes in recent-onset major depression, bipolar disorder and schizophrenia-spectrum disorders: a longitudinal cohort study. *Transl. Psychiatry* 28:e555. doi: 10.1038/tp.2015.50

Lewandowski, K. E., Baker, J. T., McCarthy, J. M., Norris, L. A., and Öngür, D. (2018). Reproducibility of cognitive profiles in psychosis using cluster analysis. *J. Int. Neuropsychol. Soc.* 24, 382–390. doi: 10.1017/S1355617717001047

Lewandowski, K. E., Sperry, S. H., Cohen, B. M., and Öngür, D. (2014). Cognitive variability in psychotic disorders: a cross-diagnostic cluster analysis. *Psychol. Med.* 44, 3239–3248. doi: 10.1017/S0033291714000774

Lingjaerde, O., Ahlfors, U. G., Bech, P., Dencker, S. J., and Elgen, K. (1987). The UKU side effect rating scale. A new comprehensive rating scale for psychotropic drugs and a cross-sectional study of side effects in neuroleptic-treated patients. *Acta Psychiatr Scand Suppl* 334, 1–100.

Linzer, D. A., and Lewis, J. B. (2011). poLCA: an R package for polytomous variable latent class analysis. *J. Stat. Softw.* 42, 1–29. doi: 10.18637/jss.v042.i10

Marquand, A. F., Wolfers, T., Mennes, M., Buitelaar, J., and Beckmann, C. F. (2016). Beyond lumping and splitting: a review of computational approaches for stratifying psychiatric disorders. *Biol. Psychiatry* 1:433. doi: 10.1016/j.bpsc.2016.04.002

Matthiesen, R. (2010). *Bioinformatics Methods in Clinical Research*. Cham: Springer.

McClure, G., Hawes, D. J., and Dadds, M. R. (2016). Borderline personality disorder and neuropsychological measures of executive function: a systematic review. *Pers. Ment. Health* 10, 43–57. doi: 10.1002/pmh.1320

McLachlan, G., and Peel, D. (2000). "Mixtures of factor analyzers," in *Proc. Seventeenth Int. Conf. Mach. Learn*, San Francisco, CA.

Mesholam-Gately, R. I., Giuliano, A. J., Goff, K. P., Faraone, S. V., and Seidman, L. J. (2009). Neurocognition in first-episode schizophrenia: a meta-analytic review. *Neuropsychology* 23, 315–336. doi: 10.1037/a0014708

Mitrushina, M., Boone, K. B., Razani, J., and D'Elia, L. F. (2005). *Handbook of Normative Data for Neuropsychological Assessment*. Oxford: Oxford University Press.

Mooi, E., and Sarstedt, M. (2011). *A Concise Guide to Market Research*. Berlin: Springer.

Morgan, A. B., and Lilienfeld, S. O. (2000). A meta-analytic review of the relation between antisocial behavior and neuropsychological measures of executive function. *Clin. Psychol. Rev.* 20, 113–136.

Morosini, P., Gigantesco, A., Mazzarda, A., and Gibaldi, L. (2003). HoNOS-Rome: an expanded, customized, and longitudinally oriented version of the HoNOS. *Epidemiol. Psychiatry Sci.* 12, 53–62.

Orsini, A., Grossi, D., Capitani, E., Laiacona, M., Papagno, C., and Vallar, G. (1987). Verbal and spatial immediate memory span: normative data from 1355 adults and 1112 children. *Ital. J. Neurol. Sci.* 8, 539–548.

Owoeye, O., Kingston, T., Scully, P. J., Baldwin, P., Browne, D., Kinsella, A., et al. (2013). Epidemiological and clinical characterization following a first psychotic episode in major depressive disorder: comparisons with schizophrenia and bipolar disorder in the cavan-monaghan first episode psychosis study (CAMFEPS). *Schizophr. Bull.* 39, 756–765. doi: 10.1093/schbul/sbt075

Pearlson, G. D. (2015). Etiologic, phenomenologic, and endophenotypic overlap of schizophrenia and bipolar disorder. *Annu. Rev. Clin. Psychol.* 11, 251–281. doi: 10.1146/annurev-clinpsy-032814-112915

Pruneti, C. A., Fenu, A., Freschi, G., Rota, S., Cocci, D., Marchionni, M., et al. (1996). Aggiornamento della standardizzazione italiana del test delle Matrici Progressive Colorate di Raven. *Boll. di Psicol. Appl.* 217:7.

Reser, M. P., Allott, K. A., Killackey, E., Farhall, J., and Cotton, S. M. (2015). Exploring cognitive heterogeneity in first-episode psychosis: what cluster analysis can reveal. *Psychiatry Res.* 229, 819–827. doi: 10.1016/j.psychres.2015.07.084

Rosell, D. R., Futterman, S. E., McMaster, A., and Siever, L. J. (2014). Schizotypal personality disorder: a current review. *Curr. Psychiatry Rep.* 16:452. doi: 10.1007/s11920-014-0452-1

Roux, P., Raust, A., Cannavo, A. S., Aubin, V., Aouizerate, B., Azorin, J. M., et al. (2017). Cognitive profiles in euthymic patients with bipolar disorders: results from the FACE-BD cohort. *Bipolar Disord.* 19, 146–153. doi: 10.1111/bdi.12485

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., et al. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J. Clin. Psychiatry* 59, 22–57.

Skevington, S. M., Lotfy, M., and O'Connell, K. A. (2004). The World Health Organization's WHOQOL-BREF quality of life assessment: psychometric properties and results of the international field trial. A Report from the WHOQOL Group. *Qual. Life Res.* 13, 299–310. doi: 10.1023/B:QURE.0000018486.91360.00

Smith, A. L., and Weissman, M. M. (1992). "Epidemiology," in *Handbook of Affective Disorders*, ed. E. S. Paykel (Edinburgh: Churchill-Livingstone), 111–129.

Spinnler, H., and Tognoni, G. (1987). Italian standardization and classification of Neuropsychological tests. *Ital. J. Neurol. Sci.* 8, 1–120.

Tamminga, C. A., Pearlson, G., Keshavan, M., Sweeney, J., Clementz, B., and Thaker, G. (2014). Bipolar and schizophrenia network for intermediate phenotypes: outcomes across the psychosis continuum. *Schizophr. Bull.* 40, S131–S137. doi: 10.1093/schbul/sbt179

Üstün, T. B. (2010). *World Health Organization. Measuring health and disability: Manual for WHO Disability Assessment Schedule WHODAS 2.0*. Genève: World Health Organization.

van Os, J., and Reininghaus, U. (2016). Psychosis as a transdiagnostic and extended phenotype in the general population. *World Psychiatry* 15, 118–124. doi: 10.1002/wps.20310

Van Rheenen, T. E., Lewandowski, K. E., Tan, E. J., Ospina, L. H., Ongur, D., Neill, E., et al. (2017). Characterizing cognitive heterogeneity on the schizophrenia–bipolar disorder spectrum. *Psychol. Med.* 47, 1848–1864. doi: 10.1017/S0033291717000307

Ventura, J., Lukoff, D., Nuechterlein, K. H., Liberman, R. P., Green, M., and Shaner, A. (1993). Manual for the expanded brief psychiatric rating scale. *Int. J. Methods Psychiatry* 3:221.

Vermunt, J. K., and Magidson, J. (2002). Latent class models for clustering: a comparison with K-means. *Can. J. Mark. Res.* 20, 36–43.

Vermunt, J. K., and Magidson, J. (2009). Latent class cluster analysis. *Appl. Latent Cl. Anal.* 11, 89–106. doi: 10.1017/cbo9780511499531.004

Watson, Y. I., Arfken, C. L., and Birge, S. J. (1993). Clock completion: an objective screening test for dementia. *J. Am. Geriatr. Soc.* 41, 1235–1240.

Witt, S. H., Streit, F., Jungkunz, M., Frank, J., Awasthi, S., Reinbold, C. S., et al. (2017). Genome-wide association study of borderline personality disorder reveals genetic overlap with bipolar disorder, major depression and schizophrenia. *Transl. Psychiatry* 7:e1155. doi: 10.1038/tp.2017.115

Woodward, N. D. (2016). The course of neuropsychological impairment and brain structure abnormalities in psychotic disorders. *Neurosci. Res.* 102, 39–46. doi: 10.1016/j.neures.2014.08.006

Wykes, T., and Reeder, C. (2005). *Cognitive Remediation Therapy for Schizophrenia: Theory and Practice*. Milton Park: Taylor & Francis.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

METHODS
published: 18 June 2020
doi: 10.3389/fpsyg.2020.01077

Development of a Computerized Adaptive Test for Separation Anxiety Disorder Among Adolescents

Yiyuan Hu, Yan Cai, Dongbo Tu, Yingying Guo and Siyang Liu*

School of Psychology, Jiangxi Normal University, Nanchang, China

Background: Separation anxiety disorder (SAD) is one of the most common mental disorders among children and adolescents, and it may seriously affect their growth, daily life, and learning. Self-report scales have been used for diagnosis, which require lengthy testing and personnel.

Methods: A total of 1,241 adolescents were recruited from 16 junior- and senior-high schools in China. The initial item bank was selected from classical SAD scales according to the *DSM-5*. First, the optimal model was selected using item response theory (IRT) according to data fit. Then, per the IRT analysis, items that did not meet the psychometric requirements were deleted (e.g., discriminating values < 0.2). Consequently, a computerized adaptive test (CAT) for SAD was formed (CAT-SAD).

Results: An average of 17 items per participant was required to achieve and maintain a 0.3 standard error of measurement in the SAD severity estimate. The estimated correlation of the CAT-SAD with the total 68-item test score was 0.955. CAT-SAD scores were strongly related to the probability of a SAD diagnosis with the Separation Anxiety Assessment Scale—Child and Adolescent Version. Therefore, SAD could be accurately predicted by the CAT-SAD.

Conclusions: Exploratory factor analyses revealed that SAD was unidimensional. The CAT-SAD, which has good reliability and validity and high sensitivity and specificity, provides an efficient test for adolescents with SAD as compared to standard paper-and-pencil tests. It can be used to diagnose varying degrees of SAD quickly and reliably and ease the burden on adolescents. Potential applications for inexpensive, efficient, and accurate screening of SAD are discussed.

Keywords: separation anxiety disorder, adolescent, computerized adaptive testing, item response theory, *DSM-5*

OPEN ACCESS

Edited by:
Pietro Cipresso,
Italian Auxological Institute (IRCCS), Italy

Reviewed by:
Caterina Primi,
University of Florence, Italy
Andrea Giordano,
Carlo Besta Neurological Institute (IRCCS), Italy

****Correspondence:***
Yan Cai
cy1979123@aliyun.com

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *13 October 2019*
Accepted: *28 April 2020*
Published: *18 June 2020*

Citation:
Hu Y, Cai Y, Tu D, Guo Y and Liu S (2020) Development of a Computerized Adaptive Test for Separation Anxiety Disorder Among Adolescents. Front. Psychol. 11:1077. doi: 10.3389/fpsyg.2020.01077

INTRODUCTION

Separation anxiety disorder (SAD) is one of the most common mental disorders among children and adolescents—and its frequently reported symptoms are separation-related distress, avoidance of being alone/without an adult, and distress when sleeping away from caregivers/home (Allen et al., 2010)—as well as among some parents and patients undergoing psychotherapy. Currently, SAD begins (on average) at age 8 years, and it may persist into mid-childhood or adolescence (Last et al., 1992; Costello et al., 2003). SAD brings difficulties for both children and caregivers including undue worry, sleep problems, stress in social and academic environments, and a variety of physical symptoms that lower quality of life (Brand et al., 2011). Symptoms typically persist for more than 4 weeks, significantly interfering with children's daily learning, which hinders their growth and development such as in interpersonal communication and learning efficiency (Eisen and Schaefer, 2007; Chessa et al., 2012).

Recently, some studies (e.g., Kossowsky et al., 2012) tracked the anxiety disorders of children and adolescents and showed that SAD was persistent and patients deteriorated steadily. Moreover, Lipsitz et al. (1994) suggested that early separation anxiety may constitute a non-specific vulnerability to a wide range of anxiety disorders in adulthood, including panic disorder. Some separation anxiety is a normal part of development in children aged 1–3 years. The lifetime prevalence is between 4 and 7.6% (Kessler et al., 2005; Shear et al., 2006; Merikangas et al., 2010; Milrod et al., 2014), and Manicavasaga et al. (1997) suggest that it may be possible to identify adults whose SAD mirrors the constellation of symptoms observed in childhood, even though some of the specific features are modified by maturation. Therefore, the early detection and intervention treatment of separation anxiety among children and adolescents are vital.

The definition of SAD has undergone significant changes in the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (*DSM-5*)—the most consequential being the lifting of the age restriction (i.e., 18 years old) to assign a diagnosis. Why do clinicians traditionally redefine applicable ages? Because the construct of SAD has long been central to developmental theories that exert a strong influence in guiding clinical practice. In psychoanalytic and attachment theories, SAD is regarded as representative of neurophysiological, psychological, and behavioral responses designed to protect children from danger by ensuring close relationships with adult caregivers, typically mothers (Battaglia et al., 2009). Within the development framework of attachment theory, heightened expressions of SAD are regarded as indicating disturbances in children's working models or internal representations of attachment figures, which are shaped by past and ongoing bonding experiences with primary caretakers (Bowlby, 1960).

According to the *DSM-5* definition, separation anxiety refers to individuals' separation anxiety concerning their family and the related developmental problems. Significant symptoms such as physical symptoms (vomiting, stomachaches, etc.), emotional symptoms (anxiety and fear), and social functioning problems (declined learning efficiency) present themselves when adolescents are separated from their caregivers. A description of SAD symptoms in the *DSM-5* is shown in **Table 1**. If individuals meet any three symptoms of SAD, and they persist for at least 4 weeks, they are considered to have SAD. SAD comprises a repertoire of neurophysiological, intrapsychic, and behavioral responses. Therefore, experts hold different ideas about the dimensions of SAD; for example, one study suggested that separation anxiety was a multidimensional trait and that it should be divided into six dimensions (Hahn et al., 2003). However, In-Albon et al. (2013) suggested that a two-factor structure provided an adequate fit for the Separation Anxiety Avoidance Inventory—Child Version (SAAI-C). While an exploratory factor analysis (EFA) of the structure of children's separation anxiety revealed a two-factor structure, a confirmatory factor analysis showed that the correlation between the two factors was 0.62 in a school-aged sample [standard error (*SE*) = 0.05, $p = 0.01$; In-Albon et al., 2013]. In other words, these dimensions measure different domains of SAD, and there is a significant correlation between them; i.e., they measure different domains of the same trait.

This study considers the arguments in favor of and against this definition change in the hope of stimulating debate and research aimed at achieving a consensus. We aimed to show that separation anxiety is unidimensional and provide a new perspective to the cross-cultural study of SAD measurements by using a Chinese sample. In fact, the scales that measure separation anxiety in previous studies have been developed according to Classical Test Theory. The purpose of the norm-referenced test is to distinguish the degree of separation of anxiety by maximizing the total score of the scale. At this point, how much more appropriate is the difficulty of each item on the test, and is the difficulty distribution of the item wider or narrower? A computer adaptive test (CAT) based on item response theory (IRT) may solve this problem. Furthermore, this study developed an assessment instrument of SAD based on CAT (SAD-CAT) in hopes of providing an effective instrument to measure SAD. CAT is more than just effective, due to cost and time effectiveness, less need for personnel to administer the test, and accurate and efficient diagnosis.

Computer adaptive test is an effective and fast measurement to evaluate participants' individual latent traits (θ). CAT starts with randomly selecting one item from the test database and then selects the next item with lower or higher difficulty/threshold according to the previous responses. The process will continue when the uncertainty of the estimation capability does not reach the set value, or it will stop when the number of items reaches the predefined threshold. The paradigm shift is to manage items of different lengths to provide limited information to participants, depending on their specific level of the latent trait. Concurrently, CAT allows researchers to adaptively select a small set of items from a multi-item test based on participants' prior latent trait estimation. Although only a small number of items are administered during this process, the information is comparable to several items. Therefore, compared with traditional tests of fixed length and topic, CAT has many remarkable advantages: (a) the length and test items differ among individuals; (b) it can effectively solve problems including long testing times and ineffective information for participants; and (c) it can present scores immediately after the test and has several practical implications, including the American Graduate School Humanities Test, the American Graduate School Admission Test, the American Nurses' License Test, the American Military Occupational Direction Test, and so on.

Adolescents usually complete self-reports with the help of computer technology; therefore, a computerized adaptive application is advantageous for use with teenagers. In medical diagnoses, mental disorders usually rely on patients' self-reports (or report to the diagnostician) to assess disorder presence and severity. Therefore, it is important to help patients complete self-reports effectively and accurately.

TABLE 1 | The diagnostic criteria of separation anxiety disorder (SAD) in *DSM-5* and the initial item bank structure.

	Number of items
(1) Developmentally inappropriate and excessive fear or anxiety concerning separation from those to whom the individuals is attached, as evidenced by at least three of the following:	
(a) Recurrent excessive distress when anticipating or experiencing separation from home or from major attachment figures.	11
(b) Persistent and excessive worry about losing major attachment figures or about possible harm to them, such as illness, injury, disasters, or death.	12
(c) Persistent and excessive worry about experiencing an untoward event (e.g., getting lost, being kidnapped, having an accident, becoming ill) that causes separation from a major attachment figure.	10
(d) Persistent reluctance or refusal to go out, away from home, to school, to work, or elsewhere because of fear of separation.	11
(e) Persistent and excessive fear of or reluctance about being alone or without major attachment figures at home or in other settings.	14
(f) Persistent reluctance or refusal to sleep away from home or to go to sleep without being near a major attachment figure.	12
(g) Repeated nightmares involving the theme of separation.	12
(h) Repeated complaints of physical symptoms (e.g., headaches, stomachaches, nausea, vomiting) when separation from major attachment figures occurs or is anticipated.	11
(2) The fear, anxiety, or avoidance is persistent, lasting at least 4 weeks in children and adolescents and typically 6 months or more in adults.	
(3) The disturbance causes clinically significant distress or impairment in social, academic, occupational, or other important areas of functioning.	
(4) The disturbance is not better explained by another mental disorder, such as refusing to leave home because of excessive resistance to change in autism spectrum disorder; delusions or hallucinations concerning separation in psychotic disorders; refusal to go outside without a trusted companion in agoraphobia; worries about ill health or other harm befalling significant others in generalized anxiety disorder; or concerns about having an illness in illness anxiety disorder.	

Further, CAT and IRT have been widely used in education measurements and competency assessment; however, their use in the field of personality and mental health needs to be expanded. To the best of our knowledge, using CAT and IRT to effectively assess SAD has not yet been formally discussed in the literature. We wanted to use CAT to achieve the goal of developing shorter and more effective tools to measure SAD and analyze the characteristics of teenagers. Specifically, we aimed to develop a new tool, an alternative to traditional paper-and-pencil (P&P) testing, that measures SAD with CAT and to examine its accuracy, reliability, and effectiveness.

METHOD

Sample

A total of 1,241 Chinese adolescents were recruited from 16 junior- and senior-high schools across nine cities in China. All adolescents and their guardians provided informed consent to participate, and their privacy was protected. Any participant with language issues was assisted, and participants completed the tests anonymously. The survey consisted of basic demographic questions, SAD measurement items, and exclusion criteria (see **Table 1**). To screen out individuals who randomly responded, four lie-detection items were embedded in the survey. For example, for an original item of the child version of the Revised Child Anxiety and Depression Scale (RCADS-C) such as "I am afraid of being alone at home," its corresponding lie-detection item was "I am not afraid of being alone at home." Participants who responded to any one of the four paired items using the same answer were eliminated from analyses.

Next, 1,161 respondents completed the P&P tests. Of those, 56 (5.60%) participants were eliminated owing to lie-detection items, and 15 (1.4%) participants were excluded owing to meeting any of the pre-established exclusion criteria: (1) adolescents had inappropriate fear or anxiety that persisted for at least 4 weeks; (2) clinically significant distress or impaired social, learning, work, or other important functions caused this inappropriate fear or anxiety; and (3) the inappropriate fear or anxiety was better explained by other mental disorders, like infantile autism, psychotic disorder, agoraphobia disorder, or generalized anxiety disorder. In addition, there were 76 (5.6%) partial completers—most of the missing values concerned gender, age, and region. The MissMech R package (Jamshidian et al., 2014) was employed to test the assumption that data were missing completely at random (Rubin, 1976).

After eliminating missing values using the listwise deletion method, the final sample comprised 1,014 (effective response rate = 81.71%) participants. Participants' ages ranged from 12 to 18 years (mean age = 15.42 ± 1.57 years). All participants were of Chinese ethnicity, and 55.82% (*n* = 566) were male. Moreover, 21.40% (*n* = 217) of the sample were from urban areas. Participants' demographics are shown in **Table 2**.

Measures

Initially, we reviewed the contents of six questionnaires that are commonly used to measure SAD to develop an item bank: the SAAI-C, the Multidimensional Anxiety Scale for Children, the Separation Anxiety Assessment Scale—Child and Adolescent Version (SAAS-C), the Separation Anxiety Symptom Inventory (SASI), the Screen for Child Anxiety Related Disorders (SCARED), and the Spence Children's Anxiety Scale—Child and Adolescent Version (SCAS-C).

TABLE 2 | Demographic characteristics (*N* = 1,014).

Variables	Category	Frequency	Percent (%)
Gender	Male	556	54.84
	Female	458	45.16
Age	Under 16 years	610	60.16
	16 and above	405	39.94
Region	Rural	797	78.56
	Urban	217	21.4

Separation Anxiety Avoidance Inventory—Child Version

The SAAI-C (Schneider and In-Albon, 2005) is a 12-item self-report scale that is rated on a five-point scale ranging from 0 (*never*) to 4 (*always*). According to In-Albon et al. (2013), the internal consistency coefficients ranged from 0.81 to 0.84, and the test–retest reliability was 0.80 ($p < 0.01$) in a school-aged sample. Among a sample of 49 participants with SAD, the SAAI-C total score correlated significantly with the separation anxiety subscale of the SCAS ($r = 0.49$). In this study, Cronbach's α was 0.86.

Separation Anxiety Assessment Scale—Child and Adolescent Version

The SAAS-C (Hahn et al., 2003), which is suitable for children aged 6–18 years, is a 34-item self-report scale. All items have a four-point rating scale ranging from 1 (*never*) to 4 (*all the time*). The SAAS-C has six subscales including fear of being alone (five items), fear of abandonment (five items), fear of physical illness (five items), being worried about calamitous events (five items), frequency of calamitous events (five items), and a safety signals index (nine items). The SAAS-C possesses good internal consistency: αs = 0.91 and 0.85 in Hahn et al. (2003), and in this study.

Separation Anxiety Symptom Inventory

The SASI (Silove et al., 1993) is a 22-item self-report scale, and all items are rated on a four-point scale: *always*, *often*, *occasionally*, and *never*. In Silove et al. (1993), the SASI construct validity with symptoms of SAD was 0.79 ($p < 0.00l$). In this study, the Cronbach's α was 0.81.

Screen for Child Anxiety Related Disorders

The SCARED (Birmaher et al., 1999) is a 37-item self-report scale that measures anxiety disorders among children and adolescents aged 9–18 years. Each item is rated on a three-point scale ranging from 0 (*not true*) to 2 (*certainly true*). In Birmaher et al. (1999), the Cronbach's α for the SCARED total score was 0.89, and its subscale αs ranged from 0.43 to 0.77. In this study, the SAD subscale Cronbach's α was 0.73.

Spence Children's Anxiety Scale—Child and Adolescent Version

The SCAS-C (Spence, 1998) is a 44-item self-report scale that was designed to assess children's anxiety symptoms. Items are rated on a four-point scale ranging from *never* to *always*. There are six subscales reflecting six symptoms: social phobia (six items), panic disorder and agoraphobia (nine items), generalized anxiety disorder (six items), obsessive–compulsive disorder (six items), SAD (six items), and fear of physical injury (five items). The total score was summed to reflect overall anxiety symptoms. The SCAS possessed good internal consistency (total scale > 0.90; subscales = 0.60–0.90; Spence et al., 2003; Essau et al., 2011; Zhao et al., 2012). In this study, Cronbach's α was 0.75.

RCADS

The RCADS (Chorpita et al., 2000) is a 47-item child self-report scale to assess anxiety and depression disorder symptoms. It is rated on a four-point scale (0 = *never* to 3 = *always*). In addition to a depression scale (10 items), the RCADS has five anxiety scales: separation anxiety (7 items), generalized anxiety (6 items), panic disorder (9 items), social phobia (9 items), and obsessive–compulsive (6 items). Cronbach's α for the total RCADS-C total was 0.92, and Cronbach's α for its subscales are as follows: 0.81 for separation anxiety, 0.82 for generalized anxiety, 0.89 for social phobia, 0.76 for panic disorder, 0.68 for obsessive–compulsive, 0.71 for depression, and 0.91 for total anxiety (Chorpita et al., 2000). In this study, Cronbach's α was 0.86.

Procedure

First, according to the symptom criterion of SAD as defined in the *DSM-5*, experts from Wuhan Mental Health Center judged which symptoms were measured by each item of the SAD scales, and items fitting at least one symptom criterion were considered for selection. Moreover, to ensure there were enough items measuring each symptom of SAD, according to content balance guidelines, experts selected items from these scales to form the initial item bank of the CAT-SAD. Second, participants completed the initial item bank via P&P testing, and their response data were used for later IRT analyses, construction of the final item bank, and CAT simulation research.

Item Bank

We intended to keep the original scoring of all items to verify the effectiveness of each scale in a cross-culture setting. Ninety-three items of the above six measures met the criteria and comprised our initial CAT-SAD item bank. As shown in **Table 1**, each symptom was measured by at least 10 items, which indicated that there were sufficient items to cover all symptoms of SAD as defined in the *DSM-5*. Moreover, a series of analyses under the framework of IRT were performed to choose the acceptable items from the initial item bank, which embraced the unidimensionality test, item fit test, and differential item function detection.

Unidimensionality

Unidimensionality of the 93-item P&P version of the SAD from the above six measures was first demonstrated using an EFA. The ratio of the first eigenvalue to the second eigenvalue was greater than 3 in EFA indicating unidimensionality (Lord, 1980; Hattie, 1984), and the percentage of variance explained by the first factors exceeded 20% (Reckase, 1979). According to Nunnally (1978), who observed that factor loadings smaller than 0.30 should not be taken seriously and that ones smaller than 0.30 could easily be over-interpreted, we first eliminated items whose factor loadings on the first factor were below 0.30 to confirm acceptable unidimensionality of the dataset; then, the EFA was conducted again to test unidimensionality.

IRT Model Selection

We considered IRT models with polytomous items including the graded response model (GRM; Samejima, 1969), the nominal response model (NRM; Bock, 1972), and the generalized partial credit model (GPCM; Muraki, 1992). Akaike's information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwarz, 1978) of the three models were employed to compare model fit. The smaller the value of the AIC or BIC, the better the model fit; thus, the IRT model with the smallest AIC and BIC value was chosen for the IRT analysis in this study.

Item Calibration

Item fit

Evaluating model fit generally requires an evaluation of both test and item fit. Test fit was evaluated for whether the selected model was consistent with the actual data at the test level; item fit was evaluated as whether the model was consistent with the actual data at the item level, which can be used to screen items in the test. Item fit was evaluated as an absolute fit test, and this kind of method calculates some statistics between the model to be selected and the actual data. The $S\text{-}X^2$ index (Orlando and Thissen, 2000, 2003) tested item-level fit. Items with a *p*-value of $S\text{-}X^2$ less than 0.05 were considered to have poor item fit and were deleted. The R package MIRT (version 1.29; Chalmers, 2012) was utilized to evaluate item fit.

Discrimination parameter

According to the IRT, the item discrimination parameter defined the degree to which an item distinguishes between individuals with similar scores. An item with a high discrimination parameter *t* is high quality and could more accurately estimate the potential characteristics of the participants in the test. In addition, item discrimination had an important impact on item information, which was used to decide which item was selected in the CAT environment; therefore, items with low discrimination (i.e., less than 0.8) were excluded from the initial item bank (Tan et al., 2018).

Differential item functioning

Measurement bias is an important indicator of the validity of a questionnaire survey, and qualified items had no measurement bias for different groups (region, gender, age, health condition). This study used a differential item functioning (DIF) analysis to evaluate the systematic error caused by group bias (Zumbo, 1999). We used ordinal logit regression analysis (Crane et al., 2006) under the optimal model through *R* package Lordif (version 0.3-3; Choi, 2015) based on test-level model fitting checks. Items with changes in McFadden's pseudo $R^2 < 0.2$ were deemed as DIF (Flens et al., 2017) and were deleted from the initial item bank. DIF was independently evaluated by region (rural, urban), gender (male, female), age ($<$16 years, $\geq$16 years), and health condition (SAD, normal) groups.

CAT-SAD Simulation Study

We performed a simulation study with the 1,014 adolescents to investigate the properties of the developed item bank. We examined four properties: reliability, validity, sensitivity, and specificity.

We simulated a CAT in the item bank from the real responses obtained from adolescents' P&P data. At the beginning of the CAT, we did not know prior information about the adolescents (Kreitzberg and Jones, 1980). The first item that the CAT simulation started on was randomly selected from the item bank (Magis and Barrada, 2017). Then, base item parameters and adolescents' item responses estimated their SAD latent trait (θ) and measurement precision. Here, the expected *a posteriori* method (Bock and Mislevy, 1982) was used to update adolescents' SAD latent trait (θ) based on their real P&P responses. The maximum Fisher information criterion (Baker, 1992) selection strategy was adopted to select the next question for adolescents in the simulation of CAT-SAD, and three different stopping rules were set: 0.3, 0.4, and 0.5, respectively. When measurement accuracy or the pre-set test length (i.e., 20 items) was reached, the program would terminate (Magis and Raiche, 2012).

CAT-SAD Properties

To evaluate CAT-SAD properties, three statistic criteria were investigated to evaluate test estimation accuracy: the number of items used, SE, and marginal reliability (Smits et al., 2011). The number of items used was the number of items each adolescent answered when completing the test. The SE for trait level can be defined as the reciprocal of the square root of the value of the test information function at that trait level (Magis and Raiche, 2012); the formula is defined as follows:

$$\mathrm{SE}(\theta) = \frac{1}{\sqrt{I(\theta_i)}}, \text{ in which } I(\theta_i) \text{ is the test information at } \theta_i$$

The corresponding reliability $r_{xx}(\theta_i)$ of each individual can be derived via the following formula (Samajima, 1994) when the mean and standard deviation (SD) of the score are fixed to 0 and 1, respectively:

$$r_{xx}(\theta_i) = 1 - \frac{1}{I(\theta_i)}$$

Validity

Criterion-related validity refers to the degree to which the measure is consistent with its measurement objectives. Taking the total SAAS-C score as the criterion, the correlation between separation anxiety level (θ), as estimated by the CAT-SAD, and the criterion data calculated was regarded as the criterion-related validity of the CAT-SAD. The high correlation indicated that the CAT-SAD had good criterion-related validity. We also investigated the content validity of the CAT-SAD by analyzing whether the items in the final item bank adequately measured all symptoms of SAD as defined in the *DSM-5*.

Sensitivity and Specificity

In medical diagnosis, sensitivity and specificity are usually used as an important reference index for the accuracy of delimitation scores. Sensitivity refers to the probability of a patient being diagnosed with a disease, and specificity refers to the probability that ordinary people will be diagnosed without the disease (Smits et al., 2011). Here, sensitivity and specificity were used to investigate the predictive utility of the CAT-SAD. In addition, the Youden index (*YI* = sensitivity + specificity – 1) was also used to assess the effect of the diagnosis by CAT-SAD, which reflected the difference between the rate of true positives and false positives. The larger the value of *YI*, the better the diagnostic capacity (Schisterman et al., 2005).

To calculate sensitivity and specificity, participants were classified as SAD samples and non-SAD samples by the SAAS-C.

Specifically, 40 participants with total SAAS-C scores $\geq$ 75 were classified as the SAD sample, while the other 974 participants with SAAS-C scores $<$ 75 were classified as the non-SAD sample (Eisen and Schaefer, 2007).

RESULTS

Item Bank

Unidimensionality

Results of unidimensionality showed that the factor loadings on the first factor were less than 0.3. After excluding the 15 items, the EFA was conducted to analyze unidimensionality with the remaining items. The results indicated that the ratio between the first eigenvalue of 25.08 and the second eigenvalue of 5.59 was 4.49, and the first factor accounted for 25.08% (more than 20%). The above results indicated that the remaining 78 items met unidimensionality.

IRT Model Selection

The IRT model with the smallest value of AIC and BIC was finally chosen and applied (see **Table 3**). The AIC and BIC values in the GRM were the smallest compared with the GPCM and NRM, which showed that the GRM fit the data better than the others. Accordingly, the GRM was selected as the IRT analysis for the CAT-SAD.

Item Fit and DIF

Results of the $S\text{-}X^2$ suggested that two items (ps $<$ 0.05) were deleted from the item bank. Regarding DIF, there were no items in the regional, sex, age, and health condition groups (all items' McFadden's pseudo R^2 were less than 0.2). In addition, the discrimination values of 15 items were less than 0.8; thus, they were deleted from the item bank (Tan et al., 2018).

The remaining 68 items in the item bank met unidimensionality, fit the GRM well, possessed high discrimination, and had no DIF. **Table 4** shows the estimated item parameter values of GRM in the item bank. The discrimination parameters showed considerable variation and similar patterns for all scales, ranging from 0.83 (Item 2, "I feared that one of my parents might come to harm when I was away from home") to 2.14 (Item 51, "I am afraid to be alone in the house"). The threshold parameters showed considerable variation for all scales; for example, all four Likert items ranged from −1.12 (Item 2, "I imagined that monsters or animals might attack me when I was alone at night") to 6.82 (Item 13, "I am afraid my

TABLE 3 | Fitting models.

Model	AIC	BIC
GRM	140,506	142,026.8
GPCM	141,084.2	142,605
NRM	140,832.7	143,106.5

GRM, graded response model; GPCM, generalized partial credit model; NRM, nominal response model; AIC, Akaike's information criterion; BIC, Bayesian information criterion.

TABLE 4 | Location and discrimination parameter values and the descriptive statistics of the responses of each item for the item bank.

Item number	Item parameters					Descriptive statistics of the responses			
	a	b1	b2	b3	b4	Mean	*SD*	Skewness	Kurtosis
1	0.98	−0.32	2.04	3.04		0.79	0.87	1.06	0.58
2	0.83	−0.3	1.68	2.67		0.35	0.75	2.37	5.05
3	0.9	−0.82	1.01	2.12		1.11	1.06	0.57	−0.89
4	1.02	−0.16	1.79	2.54		0.8	0.95	1.09	0.25
5	1.23	0.42	1.64	2.33		0.66	0.96	1.34	0.64
6	0.88	−1.12	0.73	1.8		1.27	1.1	0.37	−1.17
7	0.92	1.75	3.52	4.38		0.31	0.61	2.32	5.9
8	1.36	0.93	2.48	3.61		0.36	0.65	1.96	3.74
9	1.27	0.83	2.48	3.59		0.4	0.69	1.84	3.21
10	1.08	0.6	2.58	3.5		0.49	0.75	1.65	2.45
11	1.17	−0.41	1.47	2.76		0.86	0.88	0.83	−0.03
12	1.05	−0.42	1.59	2.92		0.86	0.89	0.86	−0.01
13	1.85	0.34	1.46	2.1		0.61	0.88	1.42	1.15
14	1.4	−0.33	1.2	1.95		0.92	0.98	0.87	−0.26
15	0.88	0.18	2.13	3.18		0.71	0.92	1.21	0.55
16	1.32	−0.51	1.31	2.36		0.91	0.9	0.81	−0.05
17	1.04	0.93	2.75	3.96		0.41	0.71	1.85	3.14
18	1.61	0.51	1.53	2.15		0.58	0.91	1.52	1.26
19	1.64	−0.31	1.38	2.34		0.82	0.84	0.92	0.32
20	1.52	−0.4	1.42	2.53		0.82	0.83	0.87	0.3
21	1.02	0.6	2.55	3.6		0.51	0.77	1.59	2.11
22	1.61	−0.12	1.16	1.93		0.84	0.97	0.95	−0.13
23	1.56	0.67	2.13	2.99		0.42	0.71	1.79	3.03
24	1.82	0.14	1.75	2.62		0.59	0.75	1.26	1.37
25	1.63	0.25	1.8	2.86		0.56	0.75	1.31	1.34
26	1.5	−0.55	1.3	2.35		0.9	0.86	0.8	0.11
27	1.18	1.29	2.76	3.95		0.3	0.63	2.31	5.26
28	1.29	0.97	2.29	3.09		0.39	0.74	2.05	3.75
29	1.64	0.15	1.59	2.29		0.65	0.86	1.3	1.01
30	1.45	0.17	1.82	2.65		0.62	0.81	1.33	1.27
31	1.48	0.89	2.45	3.12		0.35	0.66	2.15	4.85
32	1.55	1	2.42	3.21		0.32	0.63	2.24	5.21
33	1.5	1.55	2.75	3.49		0.2	0.54	3.13	10.55
34	1.53	1.08	2.59	3.58		0.29	0.58	2.27	5.56
35	1.52	1.27	2.81	3.81		0.24	0.53	2.51	6.97
36	1.75	1.29	2.57	3.39		0.22	0.54	2.76	8.35
37	1.38	0.24	1.84	2.66		0.62	0.84	1.35	1.21
38	1.31	1.66	3.1	3.84		0.2	0.53	3.17	11.07
39	1.1	−0.01	3.01			0.56	0.6	0.56	−0.61
40	1.46	0.73	3.02			0.34	0.53	1.22	0.48
41	1.48	0.28	2.15			0.51	0.64	0.89	−0.28
42	1.7	−0.2	1.43			0.72	0.72	0.48	−0.95
43	1.1	−1	1.74			0.88	0.67	0.14	−0.77
44	1.53	0.14	1.84			0.58	0.68	0.76	−0.58
45	0.87	−0.49	2.32			0.74	0.7	0.4	−0.9
46	1.07	1.14	3.43			0.31	0.54	1.56	1.5
47	1.01	0.79	3.38			0.39	0.58	1.19	0.41
48	1	1.28	3.5			0.3	0.55	1.64	1.74
49	0.95	1.3	3.38			0.32	0.58	1.64	1.65

(Continued)

TABLE 4 | Continued

Item number	Item parameters					Descriptive statistics of the responses			
	a	b1	b2	b3	b4	Mean	*SD*	Skewness	Kurtosis
50	1.39	0.51	2.45			0.43	0.61	1.09	0.15
51	1.95	0.49	2			0.43	0.61	1.14	0.23
52	1.92	0.36	2.04			0.46	0.61	0.98	−0.07
53	1.27	0.41	3.01			0.44	0.57	0.86	−0.26
54	1.24	0.93	3.19			0.32	0.54	1.42	1.07
55	1.58	0.55	1.95			0.45	0.66	1.17	0.15
56	1.62	0.29	2.03			0.5	0.64	0.91	−0.26
57	1.82	0.05	1.65			0.6	0.68	0.7	−0.65
58	1.37	−0.6	1.8			0.78	0.66	0.26	−0.75
59	1.7	−0.11	1.52			0.68	0.71	0.55	−0.86
60	1.26	−0.6	0.71	2.12	3.02	1.13	1.09	0.79	−0.01
61	1.32	−0.67	0.36	1.32	2.24	1.38	1.29	0.58	−0.77
62	1.11	−0.11	1.06	2.27	3.34	0.96	1.13	1.03	0.2
63	0.89	0.03	1.6	3.22	4.32	0.83	1.03	1.24	0.98
64	1.07	0.18	1.3	2.38	3.04	0.87	1.17	1.3	0.78
65	0.97	−0.45	0.58	1.41	2.25	1.38	1.44	0.64	−0.98
66	1.55	0.14	0.92	1.74	2.29	0.92	1.22	1.19	0.35
67	1.38	−0.31	0.7	1.74	2.59	1.12	1.2	0.85	−0.23
68	1.38	−0.27	0.66	1.65	2.47	1.12	1.22	0.86	−0.29

a is the discrimination parameter; the bs are location parameters. Mean is the mean of all the participants' response in each item, SD is the standard deviation of all the participants' response in each items, skewness is the skewness of all the participants' response in each item, and kurtosis is the kurtosis of all the participants' response in each item.

family might abandon me"). Therefore, the final item bank of the CAT-SAD included 68 items after 25 items were excluded for the abovementioned psychometric reasons.

CAT-SAD Simulation Study

Properties of the CAT-SAD

A description of the termination rules and the results are provided in **Table 5**. A CAT algorithm was run with no termination rules ("none" in **Table 5**) to generate scores based on administration of the full item bank for comparison. **Table 5** reveals that the stop rule with the SE was less than 0.3 [i.e., SE (θ) < 0.3], an average of 17.04 items per participant was required with a marginal reliability of 0.89, and the correlation between the 17-item average CAT severity score and the total 68-item score was 0.953. In this study, seeking for a reliable and shorter measure, we specified that when the SE < 0.3, the CAT simulation terminated the latent trait estimate of an adolescent, and the marginal reliability was 0.89 (Green et al., 1984). **Table 5** also indicated that, as the SE increased (i.e., less precise), the average amount of items decreased. For example, when SE increased from 0.3 to 0.4, the number of items, on average, decreased from 17.04 to 10.89, and the marginal reliability also decreased.

The descriptive statistics of the responses to each item in the final item bank are presented in **Table 4**. The mean score for four Likert items ranged from 0.22 to 1.27 (SD ranged from 0.53 to 1.10), the mean score for three Likert items ranged from 0.30 to 0.88 (SD ranged from 0.32 to 1.88), and the mean score for five Likert items ranged from 0.83 to 1.38 (SD ranged from 1.03 to 1.44). The skewness values were all greater than 0 (range 0.14 to 3.17; *SD* = 0.077), and the kurtosis values ranged from −1.17 to 11.07 (*SD* = 0.153); for example, Item 38 had the highest skewness (3.17) and kurtosis (11.07).

TABLE 5 | Characteristics of the computerized adaptive test (CAT) under several stopping rules.

Stopping rule	Number of items used		Mean *SE* (θ)	Marginal reliability	Correlation[b]
	Mean	*SD*			
None	68	0	0.19	0.96[a]	1.000
SE (θ) < 0.3	17.04	2.43	0.31	0.89	0.953
SE (θ) < 0.4	10.89	3.96	0.40	0.84	0.924
SE (θ) < 0.5	7.456	3.61	0.48	0.77	0.892

None = all of the item bank was used. [a]Coefficient alpha for the full test was 0.960. [b]Correlation between CAT_θ and complete test θ.

Figure 1 displays the reliability and test information of the final CAT-SAD item bank for the final estimate under stopping rule SE (θ) < 0.3. Furthermore, the precision of test information function was expounded, which measured adolescents' latent traits whose location given was estimated as well. **Figure 1** shows that the CAT-SAD provided ideal test information quantity on the latent trait ranging from −2 to 4.

Validity

The Pearson correlations between the full-scale SAAS-C score and the estimated score under different stopping rules (*SE* < 0.3, *SE* < 0.4, and *SE* < 0.5) for the CAT-SAD were 0.705, 0.685, and 0.650, respectively. These high or moderate significant correlations indicated that the CAT-SAD had acceptable criterion-related validity with the SAAS-C. In addition, the final item bank with 68 items covered all symptoms of SAD, as defined in the *DSM-5*, and each symptom was assessed by at least seven items. Therefore, the CAT-SAD also had acceptable content validity.

Sensitivity and Specificity

To make the scores more intuitive, the CAT-SAD scores, which used an average of 17 adaptively administered items [*SE* (θ) < 0.3], were strongly related to total SAAS-C scores (r = 0.706, p < 0.001). This relationship is shown in **Figure 2**. **Figure 2** also displays the CAT-SAD score percentile ranking for adolescents who were classified as having SAD by the SAAS-C. For example, an adolescent with a CAT-SAD score of 1.78 had a 0.50 probability of meeting the SAD criteria—specifically, at the upper 94th percentile of the CAT-SAD distribution. In contrast, if an adolescent had a CAT-SAD score of −0.32, the probability of meeting criteria for SAD was Close to 0, and would be at the 50th percentile for the sample of adolescent.

The results of the sensitivity and specificity for CAT-SAD under different stopping rules are displayed in **Table 5**. The CAT-SAD area under the curve (AUC) value, based on the SAAS-C scale, was 0.958 under the "none" stopping rule (sensitivity = 0.900, specificity = 0.925, YI = 0.825), 0.925 under the stopping rule SE (θ) < 0.3 (sensitivity = 0.850,

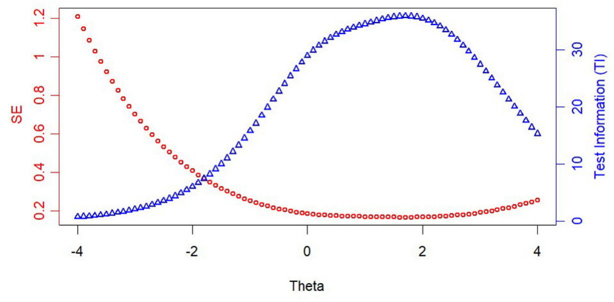


FIGURE 1 | Test information and standard error (SE) curve of the CAT-SAD.

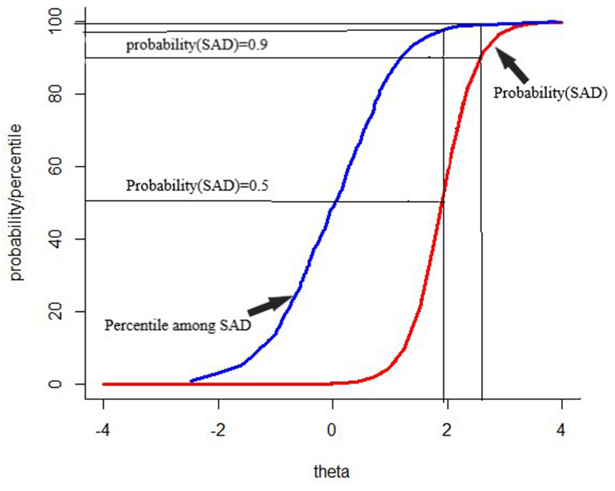


FIGURE 2 | Percentile rank among patients with separation anxiety disorder and probability of separation anxiety disorder diagnosis for the range of scores on the computerized adaptive testing–separation anxiety disorder.

specificity = 0.900, YI = 0.749), 0.921 under the stopping rule SE (θ) < 0.4 (sensitivity = 0.850, specificity = 0.865, YI = 0.714), and 0.912 under the stopping rule SE (θ) < 0.5 (sensitivity = 0.900, specificity = 0.815, YI = 0.715). Overall, the sensitivity and specificity under different stopping rules were acceptable. Taking CAT-SAD under the stopping rule of SE (θ) < 0.3 as an example, the SAAS-C scale was regarded as the classification criteria of SAD in which sensitivity was 0.850 and specificity was 0.900.

DISCUSSION

In this study, the steps to establish an item bank in a Chinese sample were unidimensionality, IRT model selection, item fit, DIF, and discrimination; the development of the CAT-SAD used a GRM to conduct simulation research. To obtain high-quality CAT-SAD development, the item bank consisted of six subscales to measure SAD, which comprehensively covered all criteria for adolescents with SAD per the *DSM-5*. Then, the most appropriate model could be selected from four common IRT models based on real data when strict unidimensionality was met. Results revealed that the final item bank included 68 items, the ratio between the first eigenvalue and the second eigenvalue displayed strict unidimensionality, and each symptom (which had eight criteria of separation anxiety per the *DSM-5*) was assessed by at least seven items. Further, the $S\text{-}X^2$ of the 68 items fit the GRM well, and the IRT discrimination of the item bank exhibited that the final item bank of the CAT-SAD was high quality.

Although the item bank contains eight symptoms of SAD, which all measure the same latent factor (i.e., SAD), the EFA demonstrated that the item bank formed six scales, and thus, SAD was unidimensional. Consistently, the first and second eigenvalues and first factor variance that was accounted for conformed to the standards of unidimensionality (Reckase, 1979).

The length of measurement can vary during the CAT process; therefore, the number of items and items answered by each participant differed. Further investigations presented that (1) the CAT-SAD had an acceptable marginal reliability, (2) the CAT-SAD had reasonable and acceptable criterion-related validity with the SAAS-C, (3) the sensitivity and specificity of the CAT-SAD were both acceptable under stopping rule $SE < 0.3$, and (4) the ROC curves showed that the AUC had an appropriate range under different stopping rules. Further, the number of items managed under the CAT format has been reduced by an average of 75% compared with P&P tests, and the correlation between scores obtained from the CAT-SAD and P&P tests was high and significant, which indicates that there is no significant loss of information. Consequently, the CAT-SAD is an effective and efficient measure to screen for varying degrees of SAD among adolescents, even without clinician assistance.

The scientific contribution of this study lies in the fact that we discovered an efficient method to assess SAD among adolescents that reduces the time and number of items to complete as compared to earlier measures. The test results have certain reference values for patients when they visit doctors; e.g., patients with mood disorders, who are difficult to assess over the long-term, can benefit from the efficiency of the CAT-SAD. Additionally, studies have shown that the suspension rule of $SE < 0.3$ is feasible for using CAT with adolescents, which has high validity, sensitivity, and specificity.

Of course, some limitations of this study are worth mentioning. First, concerning participant distribution, the number of abnormal participants obtained was very small, and the sample coverage was not diverse enough. In future studies, the sample distribution should be expanded to improve the representation of adolescents in cross-cultural studies of separation anxiety. Second, the title of the test bank targeted all participants, which may generate systematic bias when using CAT. Third, this research only notes CAT simulations; in the future, researchers should thoroughly validate the efficiency of the CAT-SAD in large-scale clinical trials; the simulated and actual CAT administration may have different results because there are many factors, such as answer time, individual emotion, test environment, and so on, that can affect individual responses in actual situations (Smits et al., 2011). Fortunately, as Kocalevent et al. (2009) showed that the simulated CAT and the actual CAT results were consistent, this paper still has some practical significance. However, the item bank can be used to construct short forms in situations in which researchers lack the equipment to complete a CAT, that is, to select a fixed set of items for optimal measurement in future studies. Indeed, CAT is supported for use in a special group (SAD) to investigate its practicality. Lastly, although the results showed that a test database established with the one-dimensional CAT can effectively diagnose SAD among adolescents, we focused only on diagnostic classification, which is of great help in clinical treatments, but the cognitive process mechanism underlying SAD remains unclear. In the future, the researchers, through cognitive diagnosis, can explore the cognitive process mechanism of SAD. SAD's attributes are multidimensional, and it is difficult to determine which attributes have caused the patient to suffer from SAD. The CAT-SAD provides certain item bank information for the cognitive diagnosis of SAD, which can diagnose attributes for each patient quickly and also improve the efficiency and help the treatment.

SAD is one of the most common mental disorders among children and adolescents, and it may seriously affect their growth, daily life, and learning. There are two ways to diagnose SAD: clinical diagnosis based on doctors' experience-based assessment and measurement. Nevertheless, the feasibility of clinical diagnosis has been questionable in some psychiatric and mental health clinics. Thus, it was necessary to relieve the pressure through measurement based on experience assessment. Psychometric tools are effective ways to screen for mental disorders in the field of clinical and mental health. This article reported on the development of a CAT version of SAD that involves shorter and more effective tools to measure SAD and analyze teenagers' characteristics. Self-report scales, which require considerable time and personnel, have previously been used for diagnosis. The CAT-SAD could be used as a routine clinical assessment, to save clinicians' time and ease patients' burden. At the same time, it can serve as a tool for follow-up treatment and effective review. Moreover, the CAT-SAD can measure SAD for all Chinese adolescents, regardless of region, gender, age, or health condition. The current research provides an efficient and accurate psychometric tool for researchers and clinicians to measure SAD among adolescents. At present, there is no research, other than this paper, on the CAT version of SAD with Chinese adolescents. Of course, this study used well-known international SAD; therefore, the CAT-SAD may have some applicability to other countries' adolescents.

DATA AVAILABILITY STATEMENT

The datasets generated in this article are not publicly available to maintain respondents' anonymity. Requests to access the datasets should be directed to 651804834@qq.com.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Research Center of Mental Health, Jiangxi Normal University. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

YH: thesis writing. YC and DT: guided the thesis writing and data processing. YG and SL: data processing.

FUNDING

This work was supported by the National Natural Science Foundation of China (31760288, 31660278, and 31960186).

REFERENCES

Akaike, H. T. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Allen, J. L., Lavallee, K. L., Herren, C., Ruhe, K., and Schneider, S. (2010). DSM IV criteria for childhood separation anxiety disorder: informant, age, and sex differences. *J. Anxiety Disord.* 24, 946–952. doi: 10.1016/j.janxdis.2010.06.022

Baker, F. B. (1992). *Item Response Theory: Parameter Estimation Techniques*. New York, NY: Marcel Dekker.

Battaglia, M., Pesenti-Gritti, P., Medland, S. E., Ogliari, A., Tambs, K., and Spatola, C. A. M. (2009). A genetically informed study of the association between childhood separation anxiety, sensitivity to CO_2, panic disorder, and the effect of childhood parental loss. *Arch. Gen. Psychiatry* 66, 64–71. doi: 10.1001/archgenpsychiatry.2008.513

Birmaher, B., Brent, D. A., Chiappetta, L., Bridge, J., Monga, S., and Baugher, M. (1999). Psychometric properties of the screen for child anxiety related emotional disorders (scared): a replication study. *J. Am. Acad. Child Adolesc. Psychiatry* 38, 1230–1236. doi: 10.1097/00004583-199910000-00011

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* 37, 29–51. doi: 10.1007/BF02291411

Bock, R. D., and Mislevy, R. J. (1982). Adaptive eap estimation of ability in a microcomputer environment. *Appl. Psychol. Meas.* 6, 431–444. doi: 10.1177/014662168200600405

Bowlby, J. (1960). Separation anxiety. *Int. J. Psychoanal.* 41, 89–113. doi: 10.1080/00140136008930499

Brand, S., Wilhelm, F. H., Kossowsky, J., Holsboer-Trachsler, E., and Schneider, S. (2011). Children suffering from separation anxiety disorder (SAD) show increased HPA axis activity compared to healthy controls. *J. Psychiatr. Res.* 45, 452–459. doi: 10.1016/j.jpsychires.2010.08.014

Chalmers, R. P. (2012). Mirt: a multidimensional item response theory package for the r environment. *J. Stat. Softw.* 48, 1–29. doi: 10.18637/jss.v048.i06

Chessa, D., Di, R. D., Delvecchio, E., and Lis, A. (2012). Assessing separation anxiety in italian youth: preliminary psychometric properties of the separation anxiety assessment scale. *Percept. Mot. Skills* 115, 811–832. doi: 10.2466/03.10.15.PMS.115.6.811-832

Choi, S. W. (2015). *Lordif: Logistic Ordinal Regression Differential Item Functioning Using IRT. Version 0.3-3.*

Chorpita, B. F., Yim, L., Moffitt, C., Umemoto, L. A., and Francis, S. E. (2000). Assessment of symptoms of DSM-IV anxiety and depression in children: A revised child anxiety and depression scale. *Behav. Res. Ther.* 38, 835–855. doi: 10.1016/S0005-7967(99)00130-8

Costello, E. J., Mustillo, S., Erkanli, A., Keeler, G., and Angold, A. (2003). Prevalence and development of psychiatric disorders in childhood and adolescence. *Arch. Gen. Psychiatry* 60:837. doi: 10.1001/archpsyc.60.8.837

Crane, P. K., Gibbons, L. E., Jolley, L., and Van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. *Med. Care* 44, S115–S123. doi: 10.1097/01.mlr.0000245183.28384.ed

Eisen, A. R., and Schaefer, C. E. (2007). Separation anxiety in children and adolescents: an individualized approach to assessment and treatment. *J. Can. Acad. Child Adolesc. Psychiatry* 16, 39–41. doi: 10.1089/cap.1999.9.277

Essau, C. A., Sasagawa, S., Anastassiou-Hadjicharalambous, X., Guzmán, B. O., and Ollendick, T. H. (2011). Psychometric properties of the spence child anxiety scale with adolescents from five european countries. *J. Anxiety Disord.* 25, 19–27. doi: 10.1016/j.janxdis.2010.07.001

Flens, G., Smits, N., Terwee, C. B., Dekker, J., Huijbrechts, I., and de Beurs, E. (2017). Development of a computer adaptive test for depression based on the dutch-flemish version of the PROMIS Item Bank. *Eval. Health Prof.* 40, 79–105. doi: 10.1177/0163278716684168

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., and Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *J. Educ. Meas.* 21, 347–360. doi: 10.1111/j.1745-3984.1984.tb01039.x

Hahn, I., Hajinlian, J., Eisen, A. R., Winder, B., and Pincus, D. B. (2003). "Measuring the dimensions of separation anxiety and early panic in children and adolescents: the separation anxiety assessment Scale," in *Paper Presented at the 37th Annual Convention of the Association for the Advancement of Behavior Therapy: Recent Advances in the Treatment of Separation Anxiety and Panic in Children and Adolescents*, ed. A. R. Eisen, Boston, MA.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behav. Res.* 19, 49–78. doi: 10.1207/s15327906mbr1901_3

In-Albon, T., Meyer, A. H., and Schneider, S. (2013). Separation anxiety avoidance inventory-child and parent version: psychometric properties and clinical utility in a clinical and school sample. *Child Psychiatry Hum. Dev.* 44, 689–697. doi: 10.1007/s10578-013-0364-z

Jamshidian, M., Jalal, S., and Jansen, C. (2014). Missmech: an R package for testing homoscedasticity, multivariate normality, and missing completely at random (MCAR). *J. Stat. Softw.* 56, 1–31. doi: 10.18637/jss.v056.i06

Kessler, R. C., Chiu, W. T., Demler, O., Merikangas, K. R., and Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication (NCS-R). *Arch. Gen. Psychiatry* 62, 617–627. doi: 10.1001/archpsyc.62.6.617

Kocalevent, R.-D., Rose, M., Becker, J., Walter, O. B., Fliege, H., Bjorner, J. B., et al. (2009). An evaluation of patient-reported outcomes found computerized adaptive testing was efficient in assessing stress perception. *J. Clin. Epidemiol.* 62, 278–287. doi: 10.1016/j.jclinepi.2008.03.003

Kossowsky, J., Wilhelm, F. H., Roth, W. T., and Schneider, S. (2012). Separation anxiety disorder in children: disorder-specific responses to experimental separation from the mother. *J. Child Psychol. Psychiatry* 53, 178–187. doi: 10.1111/j.1469-7610.2011.02465.x

Kreitzberg, C. B., and Jones, D. H. (1980). An empirical study of the broad range tailored test of verbal ability. *ETS Res. Rep. Ser.* 1980, 1–232. doi: 10.1002/j.2333-8504.1980.tb01195.x

Last, C. G., Perrin, S., Hersen, M., and Kazdin, A. E. (1992). Dsm-iii-r anxiety disorders in children: sociodemographic and clinical characteristics. *J. Am. Acad. Child Adolesc. Psychiatry* 31, 1070–1076. doi: 10.1097/00004583-199211000-00012

Lipsitz, J. D. Martin, L. Y., Mannuzza, S., Chapman, T. F., Liebowitz, M. R., Klein, D. F., et al. (1994). Childhood separation anxiety disorder in patients with adult anxiety. *Am. J. Psychiatry* 151, 927–929. doi: 10.1176/ajp.151.6.927

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. New Jersey: Lawrence Erlbaum.

Magis, D., and Barrada, J. R. (2017). Computerized adaptive testing with R: recent updates of the package catR. *J. Stat. Softw.* 76, 1–19. doi: 10.18637/jss.v076.c01

Magis, D., and Raiche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *J. Stat. Softw.* 48, 1–31. doi: 10.18637/jss.v048.i08

Manicavasaga, V., Silove, D., and Curtis, J. (1997). Separation anxiety in adulthood: a phenomenological investigation. *Compr. Psychiatry* 38, 274-282. doi: 10.1016/s0010-440x(97)90060-2

Merikangas, K. R., He, J.-P., Brody, D., Fisher, P. W., Bourdon, K., and Koretz, D. S. (2010). Prevalence and treatment of mental disorders among US children in the 2001–2004 NHANES. *Pediatrics* 125, 75–81. doi: 10.1542/peds.2008-2598

Milrod, B., Markowitz, J. C., Gerber, A. J., Cyranowski, J., Altemus, M., Shapiro, T., et al. (2014). Childhood separation anxiety and the pathogenesis and treatment of adult anxiety. *Am. J. Psychiatr*y 171, 34–43. doi: 10.1176/appi.ajp.2013.13060781

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Appl. Psychol. Meas.* 16, 159–176. doi: 10.1002/j.2333-8504.1992.tb01436.x

Nunnally, J. C. (1978). Psychometric theory. *Am. Educ. Res. J.* 5:83. doi: 10.2307/1161962

Orlando, M., and Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Appl. Psychol. Meas.* 24, 50–64. doi: 10.1177/01466216000241003

Orlando, M., and Thissen, D. (2003). Further investigation of the performance of S-X2: an item fit index for use with dichotomous item response theory models. *Appl. Psychol. Meas.* 27, 289–298. doi: 10.1177/0146621603027004004

Purperouakil, D., and Franc, N. (2010). [Separation anxiety in children]. *La Revue Du Praticien* 60, 783–787.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *J. Educ. Stat.* 4, 207–230. doi: 10.3102/10769986004003207

Rubin, D. B. (1976) Inference and missing data. *Biometrika* 63, 581–592. doi: 10.1093/biomet/63.3.581

Samajima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Appl. Psychol. Meas.* 18, 229–244. doi: 10.1177/014662169401800304

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika* 34, 1–97. doi: 10.1007/BF03372160

Schisterman, E. F., Perkins, N. J., Liu, A., and Bondell, H. (2005). Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology* 16, 73–81. doi: 10.1097/01.ede.0000147512.81966.ba

Schneider, S., and In-Albon, T. (2005). *Separation Anxiety Avoidance Inventory, Child and Parent Version*. Basel: University of Basel.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176345415

Shear, K., Jin, R., Ruscio, A. M., Walters, E. E., and Kessler, R. C. (2006). Prevalence and correlates of estimated DSM-IV child and adult separation anxiety disorder in the national comorbidity survey replication. *Am. J. Psychiatry* 163, 1074–1083. doi: 10.1176/ajp.2006.163.6.1074

Silove, D., Manicavasagar, V., O'Connell, D., Blaszczynski, A., Wagner, R., and Henry, J. (1993). The development of the separation anxiety symptom inventory (sasi). *Aust. N. Z. J. Psychiatry* 27, 477–488. doi: 10.3109/00048679309075806

Smits, N., Cuijpers, P., and Straten, A. V. (2011). Applying computerized adaptive testing to the CES-D scale: a simulation study. *Psychiatry Res.* 188, 147–155. doi: 10.1016/j.psychres.2010.12.001

Spence, S. H. (1998). A measure of anxiety symptoms among children. *Behav. Res. Ther.* 36, 545–566. doi: 10.1016/S0005-7967(98)00034-5

Spence, S. H., Barrett, P. M., and Turner, C. M. (2003). Psychometric properties of the spence children's anxiety scale with young adolescents. *J. Anxiety Disord.* 17, 605–625. doi: 10.1016/S0887-6185(02)00236-0

Tan, Q., Cai, Y., Li, Q., Zhang, Y., and Tu, D. (2018). Development and validation of an item bank for depression screening in the Chinese population using computer adaptive testing: a simulation study. *Front. Psychol.* 9:1225. doi: 10.3389/fpsyg.2018.01225

Zhao, J., Xing, X., and Wang, M. (2012). Psychometric properties of the spence children's anxiety scale (scas) in mainland chinese children and adolescents. *J. Anxiety Disord.* 26, 728–736. doi: 10.1016/j.janxdis.2012.05.006

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Ottawa, ON: National Defense Headquarters.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

BRIEF RESEARCH REPORT
published: 21 July 2020
doi: 10.3389/fpsyg.2020.01575

Positive Affect Over Time and Emotion Regulation Strategies: Exploring Trajectories With Latent Growth Mixture Model Analysis

Margherita Brondino, Daniela Raccanello, Roberto Burro and Margherita Pasini*

Department of Human Science, University of Verona, Verona, Italy

The influence of Positive Affect (PA) on people's well-being and happiness and the related positive consequences on everyday life have been extensively described by positive psychology in the past decades. This study shows an application of Latent Growth Mixture Modeling (LGMM) to explore the existence of different trajectories of variation of PA over time, corresponding to different groups of people, and to observe the effect of emotion regulation strategies on these trajectories. We involved 108 undergraduates in a 1-week daily on-line survey, assessing their PA. We also measured their emotion regulation strategies before the survey. We identified three trajectories of PA over time: a constantly high PA profile, an increasing PA profile, and a decreasing PA profile. Considering emotion regulation strategies as covariates, reappraisal showed an effect on trajectories and class membership, whereas suppression regulation strategy did not.

Keywords: latent growth mixture modeling, trajectories, positive affect, emotion regulation strategies, longitudinal data

OPEN ACCESS

Edited by:
Pietro Cipresso,
Italian Auxological Institute (IRCCS),
Italy

Reviewed by:
Yuki Nozaki,
Konan University, Japan
Wolfgang Rauch,
Ludwigsburg University, Germany

****Correspondence:***
Margherita Brondino
margherita.brondino@univr.it

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

***Received:** 30 September 2019*
***Accepted:** 12 June 2020*
***Published:** 21 July 2020*

Citation:
Brondino M, Raccanello D,
Burro R and Pasini M (2020) Positive
Affect Over Time and Emotion
Regulation Strategies: Exploring
Trajectories With Latent Growth
Mixture Model Analysis.
Front. Psychol. 11:1575.
doi: 10.3389/fpsyg.2020.01575

INTRODUCTION

Nowadays, the relevance of Positive Affect (PA) for many aspects of people's life is well recognized, mainly on the basis of the positive psychology approach. Positive affect seems to influence people's cognition and behaviors, to improve physical and mental health, and to promote good social relationships, with many consequences also on the quality of life and life satisfaction (see Lyubomirsky et al., 2005, for a review).

In this work, we focus on positive affect, defined as "the extent to which a person feels enthusiastic, active, and alert. High positive affect is a state of high energy, full concentration, and pleasurable engagement, whereas low positive affect is characterized by sadness and lethargy" (Watson et al., 1988, p. 1065). We refer to the theoretical framework distinguishing positive affect and negative affect (or activating and deactivating affect, according to more recent literature), being them both the structural dimensions (Burro, 2016) of affect more frequently characterizing English mood terms and the emotional dimensions underlying subjective well-being (Diener et al., 1985; Watson et al., 1988, 1999).

Positive affect is connected with many positive outcomes, such as psychological growth (e.g., Sheldon and Houser-Marko, 2001), mental health (e.g., Taylor and Brown, 1988; Tugade and Fredrickson, 2004), and physical health (e.g., Rasmussen et al., 2009). Positive affective states also contribute to an individual's long term well-being, and they broaden individuals' perspective making them more disposed to appreciate positive aspects in their lives, also influencing life satisfaction (Bryant, 2003; Quoidbach et al., 2010; Lyubomirsky and Layous, 2013; Farquharson and MacLeod, 2014; Douglass and Duffy, 2015).

In this paper, we focused on the study of changes of positive affect over time through an application of Latent Growth Mixture Modeling (LGMM), as a way to identify unobserved groupings in a longitudinal dataset permitting to capture temporal trends.

TRAJECTORIES OF AFFECT OVER TIME

Positive affect has been largely studied; however, only recently, attention has been paid to the description of its trajectories over time; this perspective should be more considered, given the fact that, as a state, positive affect fluctuates largely over time and across situations. Fluctuations in daily mood in adolescents, for instance, have been studied to identify distinct developmental trajectories, finding that adolescents with an increasing mood variability trajectory showed stable depressive and delinquency symptoms in early to middle adolescence compared with adolescents with a decreasing mood variability trajectory (Maciejewski et al., 2019). Patterns of change and stability in positive emotions, connected with physical education, assessed in secondary school students were found, and these patterns of variations were related with satisfaction of basic psychological needs and quality of motivation (Løvoll et al., 2019). Cece et al. (2019), using a three-wave design, found different emotional trajectories in athletes.

Some researches looking at changes in emotions along time are focused on weekly changes. Studies of variation of daily mood found an increasing of mood on the weekend relative to Monday through Thursday (Rossi and Rossi, 1977; Larsen and Kasimatis, 1990; Egloff et al., 1995; Reid et al., 2000; Reis et al., 2000; Helliwell and Wang, 2014; Young and Lim, 2014).

These findings suggest to deeply explore weekly changes in positive emotions, searching for different trajectories. We use a longitudinal design, assessing positive affect at seven time points, that is, seven days along one week, from Monday to Sunday. Longitudinal research studies with panel data are often applied to analyze processes of stability and change in individuals or groups. Working on this kind of data allows to explore individual differences and changes of patterns in variables over time. On the basis of the structural equation modeling methodology, it is possible to analyze longitudinal data using the latent class methods (Muthén, 2004; Green, 2014). This statistical approach models heterogeneity by classifying individuals into groups with similar patterns, called latent classes. In Growth Mixture Modeling (GMM), repeated measurements of observed variables are used as indicators of latent variables that describe specific characteristics of individuals' changes. A special type of GMM is Latent Class Growth Analysis (LCGA) whereby all individual growth trajectories within a class are assumed to be homogeneous.

With this methodology, intercept and slope are considered two latent variables (also called random coefficients), which, respectively, represent the level of the studied variable when time is equal to zero, and the rate of change in the same variable over time. Given that few studies examined the trajectories of positive emotions over a week, no specific hypotheses were advanced regarding the number of trajectories, their characteristics (e.g., intercepts), or their evolution through time (e.g., linear and/or quadratic slopes).

These models also allow the inclusion of covariates (conditional model) as part of the same model of estimation of the trajectories (Nagin, 1999; Roeder et al., 1999; Muthén, 2004), evaluating the covariates' impact on the longitudinal trajectory. In the present study, the conditional model evaluated the impact of emotion regulation strategies, assessed one week before the one-week daily positive affect assessment, on the trajectories.

EMOTION REGULATION

Little is known about how emotion regulation strategies are associated with changes in positive affect in daily life, even if some emotion regulation strategies are shown to be related with changes in positive and negative affect (Brans et al., 2013; Gunaydin et al., 2016).

Emotion regulation strategies refer to the process through which people modify how they feel or express emotions they are experiencing (Gross, 1998, 2014, 2015; Gross and Thompson, 2007). This process can consist in the downregulation of negative emotions (that is, decreasing them) or in the upregulation of positive emotions (that is, increasing them) or in maintaining stable one's own emotions. Upregulation of positive emotions has been shown to have a moderation effect on the relation between daily positive events and momentary happy mood (Jose et al., 2012). Furthermore, frequent use of positive upregulation strategies also seems to be associated with higher levels of happiness, life satisfaction, and positive emotions (Bryant, 2003; Quoidbach et al., 2010).

In the present study, we examined the relations between positive affect and emotion regulation strategies in terms of reappraisal and suppression emotion regulation strategies. Reappraisal is "a form of cognitive change that involves construing a potentially emotion-eliciting situation in a way that changes its emotional impact," while suppression is "a form of response modulation that involves inhibiting ongoing emotion-expression behavior" (Gross and John, 2003, p. 349). Reappraisal and suppression strategies play a key role within the process of emotion regulation and are among the two emotion regulation strategies that are more investigated in the literature (Gross, 1998, 2014, 2015; Gross and Thompson, 2007). Taking as a framework Gross' theoretical model, we know that people use them, respectively, when focusing on the antecedents of an emotion, for reappraisal, and when they focus on ways to modulate their responses, for suppression. These two strategies are particularly relevant in relation to positive affect, given empirical evidence documenting that people who frequently use reappraisal emotion regulation strategy experience more positive emotions, better relationships, a higher quality of life, and higher levels of well-being, compared to those who tend to prefer suppression (e.g., Kelley et al., 2019). For this reason, we hypothesize that reappraisal could affect positive emotion trajectories, whereas we expect that suppression does not.

AIMS

The main aim of this study is to show an application of LGMM, as a way to identify unobserved groupings in a longitudinal dataset. This technique was applied to the exploration of different trajectories of positive affect over a week, which corresponded to different profiles. Furthermore, we aimed at verifying whether the identified trajectories were affected by emotion regulation strategies, such as reappraisal and suppression.

METHOD

Participants

The participants were 108 undergraduate students (mean age: 22.2 years, *SD* = 6.2, range: 18–52 years; 84% female, 16% male) at the University of Verona, in Northern Italy, coming from a wide range of socio-economic status. They all took part to a larger micro-longitudinal study for which daily measures of students' affect had been planned (e.g., Pasini et al., 2016; Raccanello et al., 2017, 2018; Burro et al., 2018). Students' participation was voluntary, and all of them signed an informed consent form. The research was approved by the Ethics Committee of the Department of Human Sciences at University of Verona.

Procedure

The study included two questionnaires. The first questionnaire was administered in group sessions in a pre-assessment phase, which took place 1 week before the beginning of the daily affect assessment. It included measures on emotion regulation strategies, as well as some demographic characteristics. The second questionnaire was administered through an on-line survey; it was presented daily for 1 week from Monday to Sunday. The participants received an e-mail message daily, at 10 a.m., in which they were reminded to answer to the on-line questionnaire between 6 p.m. and midnight. The on-line procedure permitted participants to answer with different kinds of devices. The use of different devices in psychological on-line research surveys has been proved to be connected with a good measurement quality and also with high participants' compliance in a longitudinal design, with a low level of sample attrition (Pasini et al., 2016). This could be particularly true in affect assessment, because of the possibility, in contrast to what happens in laboratory studies, to record the construct of interest within the individual's environment, increasing ecological validity (Shiffman et al., 2008). Furthermore, assessing affective states as they naturally occur, permits to avoid some biases connected with retrospective self-report methods, for instance, "peak," and "recency" effects (Kahneman, 1999).

Instruments

Daily Positive Affect

We used the 20-item Positive and Negative Affect Schedule (PANAS, Watson et al., 1988; Italian adaptation by Terracciano et al., 2003), considering only the subscale for the assessment of positive affective states. It consists of 10 items (e.g., active, enthusiastic, and excited), and participants rated on a 7-point scale the extent to which they had experienced each affect term (1 = *not at all* and 7 = *very much*), referring to the current day. This instrument was administered daily for 1 week.

Emotion Regulation Strategies

To assess emotion regulation strategies, we used the 10-item Emotion Regulation Questionnaire (ERQ, Gross and John, 2003; Italian adaptation by Balzarotti et al., 2010). Items had to be evaluated on a 7-point Likert scale (1 = *I completely disagree* and 7 = *I completely agree*). This scale assesses two different strategies: reappraisal (with six items, e.g., "I control my emotions by changing the way I think about the situation I'm in") and expressive suppression (with four items, e.g., "I control my emotions by not expressing them"). The ERQ was administered in the pre-assessment group session.

Socio-Demographic Variables

We collected data on participants' gender, age, and socio-economic status during the pre-assessment group session.

Data Analyses

We carried out some preliminary analyses to assess the stability of the psychometric properties of the Positive Affect scale. This is an important preliminary step to properly conduct LCGA. First, we used a Confirmatory Factor Analysis (CFA) to evaluate the measurement model. The following combination of fit indices was used to evaluate the models (Brown and Moore, 2012; Kline, 2015): Chi-square degree of freedom ratio (χ^2/df), the comparative fit index (CFI), the Tucker-Lewis index (TLI), the root-mean-square error of approximation (RMSEA), and the standardized root mean residual (SRMR), with $\chi^2/df \leq 2.0$, CFI and TLI ≥ 0.90, RMSEA ≤ 0.08, and SRMR ≤ 0.06 as threshold values.

The second step concerned Measurement Invariance (MI), investigated in order to check the stability of the measurement model over the 7 days. Measurement Invariance (MI) analyses examined hypotheses on the similarity of the covariance structure over the 7 days, considering: (1) configural invariance, allowing all the parameters to be freely estimated; (2) metric invariance, requiring invariant factor loadings; (3) scalar invariance, also requiring invariant intercepts; and (4) uniqueness invariance, requiring invariant item uniqueness. Comparisons among models were based on differences in CFI and RMSEA, sample size independent; support for no changes in goodness of fit indexes requires a change in CFI and RMSEA less or equal than 0.010 and 0.015, respectively (Chen, 2007).

Then, we conducted LGMMs as a means of identifying growth trajectories of PA over seven time points, that is the 7 days of the on-line affect assessment, and to test predictors of the trajectories and of membership in these classes, using maximum-likelihood estimation to estimate class parameters (Muthén and Muthén, 2006). The analysis was performed in Mplus using the guidelines of Jung and Wickrama (2007).

First, as a preliminary analysis, we ran a single class latent growth curve model to define the best baseline model. We compared two growth curves: a first curve with the PA measures repeated on the 7 days as indicators and intercept and linear slope as higher-order latent factors, and a second one adding a quadratic parameter. Second, we specified a latent class model without covariates (unconditional). We evaluated the best-fitting model on the basis of the number of latent classes and the best-fitting parameters (linear vs. linear and quadratic). In order to compare the models, we used the information criteria and the fit indices. In addition, we followed the recommendations from the literature (e.g., Nylund et al., 2007), considering parsimony and interpretability as relevant criteria. We evaluated the Bayesian information criterion (BIC), and the Bootstrap Likelihood Ratio Test (B-LRT). We also assessed entropy values, to compare the degree of separation among the classes in the models, where scores closer to 1 highlight better fit of the data; the proportions for the latent classes (not less than 1% of total count in a class); and the posterior latent class probabilities (near to 1.00). After identifying the best unconditional model (free from covariates), we added into the model the two emotion regulation strategies as covariates (conditional model).

RESULTS

The Stability of the Psychometric Properties of the Positive Affect Scale

Our results supported the goodness of fit of the hypothesized model analyzed running a CFA for each of the 7 days (see **Supplementary Table S1**). In the seven models, χ^2/df ranged from 1.77 to 2.36, CFI from 0.92 to 0.95, RMSEA ranged from 0.087 to 0.114, and SRMR was always below 0.06, as recommended. The standardized loadings were all statistically significant at the 0.001 level. Therefore, our findings confirmed that the psychometric properties of the positive affect measure were acceptable across the 7 days.

MI analyses examined hypotheses on the similarity of the covariance structure across the different days. When we tested simultaneously the model over days, not imposing equality constraints between them (configural invariance), the goodness of fit of the models was confirmed. When all factor loadings were constrained to be equal (metric invariance), the models resulted invariant. When also the intercepts of the observed variables were constrained to be equal over days (scalar invariance), the models were invariant, as well as when factor loadings, intercepts, and residuals were constrained to be equal (uniqueness invariance). To sum up, the results of the sequence of gradually more restrictive tests of MI supported all the steps of invariance, confirming the stability of the measure of positive affect over the 7 days (fit indices of measurement invariance tests are reported in **Supplementary Table S2**).

The Single Class Latent Growth Curve Model

At first, we conducted the analysis only with the estimate of the intercept. Then, we ran the model with the intercept and the slope parameters, and finally the model with the addition of the quadratic factor. The model with the estimation of the quadratic parameter did not converge, and the linear model, which considered intercept and slope, reported better fit indexes than the one with intercept only (see **Table 1**). Furthermore, examining the trajectories of the observed data, the linear growth curve seemed the more appropriate, and so it was retained. Moreover, estimates of variance related to the intercept and slope were significant, which justified an examination of interindividual differences in PA over time.

Determining the Number of Classes

In the next step, we conducted the analyses to determine the number of latent classes. We compared progressive unconditional models from one to four classes, examining them on the basis of different elements. Model testing indicated that some parameters' variance needed to be fixed to zero for the models to converge. We rejected the four-class model because the B-LRT highlighted that the three-class model was favored ($p > 0.05$). Even if for the two-class solution entropy was higher and BIC was slightly lower, the three-class solution seemed the more adequate on the basis of B-LRT (**Table 2**). Furthermore, the exploration of the trajectories confirmed the goodness of the three-class model, showing two trajectories with a constant level of PA along time, one high and one medium, and a decreasing PA trajectory.

Adding the Covariates

In the next step, the influence of covariates on trajectories and class membership was analyzed. **Figure 1** shows the three PA trajectories along the week, from Monday (day 0) to Sunday (day 6) for the conditional model.

Table 3 shows the parameters' estimates for the unconditional model and for the conditional one with reappraisal and suppression as covariates. Reappraisal regulation strategy showed an effect on trajectories for Constant High PA and Increasing PA trajectories, whereas suppression regulation strategy only showed an effect on intercept for Decreasing PA trajectory. About the effect on class membership, because the Constant High PA group comprises the largest number of participant (49), we decided to designate it as the reference class, and used logistic regressions to assess the degree to which the probability of being in the Constant High PA class was associated with each of the two covariates. Compared to the Constant High PA group, the coefficient of −0.74 ($p = 0.037$) for the Increasing PA class indicated that subjects were 0.74 times less likely to be assigned to the Increasing PA class. Relative to the Constant High PA class, the probabilities of latent class membership were significantly different by reappraisal regulation strategy. This means that

TABLE 1 | Fit indexes, means, and variances of the parameters for the linear growth models (LGM).

Model	χ^2 (*df*)	CFI	RMSEA	SRMR	BIC	*M* intercept	Var intercept	*M* slope	Var slope
LGM (only intercept)	46.56 (26)	0.889	0.086	0.137	2074.72	3.75***	0.47***		
LGM (intercept, slope)	31.07 (23)	0.956	0.057	0.090	2073.28	3.87***	0.47***	−0.04	0.02*

*N = 108. df, degree of freedom; CFI, comparative fix index; RMSEA, root mean square error of approximation; SRMR, standardized root mean square residual; BIC, Bayesian Information Criterion; M, mean; Var, variance. *p < 0.05, ***p < 0.001.*

TABLE 2 | Information criteria and fit indexes for the unconditional GMM models.

Linear unconditional model (no covariates)						
Number of profiles	**Parameters***	**BIC**	**B-LRT *p*-value**	**Entropy**	**Number of subjects (%) in each class**	**Posterior probability estimate of class membership**
2	16	2069.88	<0.05	0.77	70 (65%), 38 (35%)	0.95, 0.91
3	15	2072.80	<0.001	0.71	39 (36%), 37 (34%), 32 (30%)	0.81, 0.90, 0.88
4	20	2085.39	1.00	0.65	24 (22%), 19 (18%), 31 (29%), 34 (31%)	0.81, 0.73, 0.92, 0.73

*BIC, Bayesian Information Criterion; B-LRT, Bootstrap Likelihood Ratio Test. *Intercept and slope variance fixed to zero in some cases.*

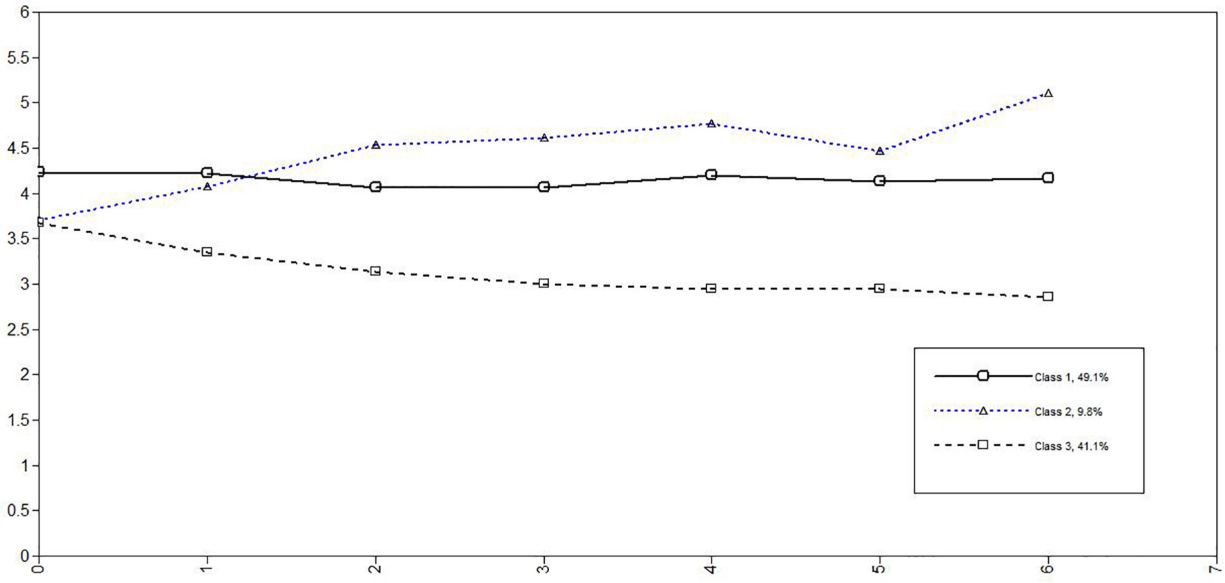


FIGURE 1 | The three trajectories of Positive Affect along the week, from Monday (0) to Sunday (6), identified by the conditional model (observed means).

Increasing PA class probability is lowered by high reappraisal values, relative to Constant High PA class. On the contrary, this regulation strategy did not show any effect on Decreasing PA class membership. No effect was found for suppression.

DISCUSSION

In the past decades, a large corpus of literature has amply documented how positive affect influences a variety of aspects within people's everyday life, in terms of cognitive, behavioral, and also biological domains, in some cases identifying the nature of underlying mechanisms (Lyubomirsky et al., 2005). Positive psychology has shown its adaptive role for people's health, describing the links between positive affect and both physical and psychological well-being in a variety of contexts (e.g., Taylor and Brown, 1988; Sheldon and Houser-Marko, 2001; Bryant, 2003; Tugade and Fredrickson, 2004; Rasmussen et al., 2009; Quoidbach et al., 2010; Lyubomirsky and Layous, 2013; Farquharson and MacLeod, 2014; Douglass and Duffy, 2015). However, only recently attention has been paid to the study of changes of affect over time, a highly relevant issue in light of the transient nature of affect, and of positive affect in particular, as a state.

TABLE 3 | Parameters' estimates, information criteria, and fit indexes for the unconditional and conditional models.

	Trajectories of positive affect over time					
	Unconditional LGMM			**Conditional LGMM[1]**		
	Constant high PA	**Constant medium PA**	**Decreasing PA**	**Constant high PA**	**Increasing PA**	**Decreasing PA**
Mean intercept	4.69***	3.49***	3.41***	0.81	8.59***	4.41***
Mean slope	−0.02	0.04	−0.14*	0.28*	−1.15*	−0.45**
Intercept on reappraisal				0.79***	−0.47**	−0.07
Slope on reappraisal				−0.07*	0.37***	0.06
Intercept on suppression				−0.04	0.33	−0.21*
Slope on suppression				0.004	−0.03	0.03
Number of subjects (%) in each class	37 (34)	39 (36)	32 (30)	49 (49)	10 (10)	41 (41)
Posterior probability of class membership	0.91	0.81	0.88	0.90	0.92	0.90
Estimated parameters	15			32		
BIC	2072.80			1950.09		
B-LRT *p*-value	<0.001			0.08		
Entropy	0.71			0.78		

*[1] In the conditional model, the variance of slope was constrained to 0. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.*

Therefore, in order to extend current literature, we focused on the identification of trajectories of affect—specifically, of positive affect—examining micro-longitudinal data gathered within a larger project for which daily affect assessments had been planned (Pasini et al., 2016; Raccanello et al., 2017, 2018; Burro et al., 2018). We applied the LGMM analysis, a methodology that permitted to better understand the phenomenon of positive affect changes along a week, from Monday to Sunday. We identified three different trajectories which characterized three profiles of students: a profile with constant high levels of positive affect (Constant High PA), a profile showing an increasing trend of positive affect over the 7 days of assessment (Increasing PA), and a profile showing a decreasing trend (Decreasing PA).

According to Ryan et al. (2010), activities along the week, such as working or not working, as well as the weekday can have an effect on mood. People generally experience a higher level of positive emotions during weekends and non-working times. Furthermore, Fritz et al. (2010) found that relaxing activities carried on during the weekend, as well as the possibility to spend more time with family and friends, lead to more positive emotions. Our results seem to suggest that a deeper understanding of this effect is needed. In fact, in addition to the increasing PA profile, a constant high PA profile and a decreasing PA profile emerged. Mood variations across days during the week can result from many different factors, such as lifestyle, working condition, and social relationships. This allows us to suppose the stability of positive affect over time during a week for some people, whereas for other people, with the approaching of the weekend, the mood can change, improving in some cases, and getting worse in other cases.

We also examined how these profiles were affected by two among the most investigated emotion regulation strategies, reappraisal and suppression (Gross and John, 2003). From a theoretical perspective, this can shed some light on whether emotion regulation strategies play a role as protective or risk factors for people's well-being (Diener, 2000). To do this, we estimated a conditional model, adding these two emotion regulation strategies as covariates. Results from the conditional model showed that the addition of reappraisal strategies affected the trajectories and, partially, the class membership, in particular decreasing the probability to be assigned to Increasing PA class, relative to the Constant High PA class. No effect was found for with the addition of suppression. In other terms, emotion regulation strategies played a role in characterizing the changes of positive affect over time in a differentiated way for reappraisal and suppression strategies. This result could be interpreted speculating that, on the whole, reappraisal could be responsible not only for feeling positively in a more intense way, but also for a higher stability of positive affect over time. However, further data should be considered to confirm this interpretation, considering for example the relations between emotion regulation strategies and profiles of negative affect over time.

Our study suffers from limitations related, for example, to the nature of self-report data. Furthermore, given the problems connected to the computational complexity and the reduced sample size, our results must be carefully considered. More research should be done to verify, for example, whether the three identified profiles could be generalized to other samples. Future researches should be also focused to check the stability of these results, looking at more than 1 week, and comparing the trajectories week by week. Despite these limitations, we think that, on the whole, we exemplified how the LGMM methodological approach could be used to identify and describe different trajectories of affect changes over time, extending what is currently known in the literature on positive affect. At an applied level, knowledge on such changes can be helpful when devising interventions aiming at favoring people's well-being based on the awareness of their real inner states.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of the Department of Human Sciences, University of Verona. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MB and MP: project administration, conceptualization, data curation, formal analysis, investigation, data analysis, supervision, and writing original draft. DR: conceptualization, data curation, formal analysis, investigation, project administration, supervision, and writing original draft. RB: conceptualization, data curation, formal analysis, investigation, and supervision. All authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01575/full#supplementary-material

REFERENCES

Balzarotti, S., John, O. P., and Gross, J. (2010). An Italian adaptation of the emotion regulation questionnaire. *Eur. J. Psychol. Assess.* 26, 61–67. doi: 10.1027/1015-5759/a000009

Brans, K., Koval, P., Verduyn, P., Lim, Y. L., and Kuppens, P. (2013). The regulation of negative and positive affect in daily life. *Emotion* 13, 926–939. doi: 10.1037/a0032400

Brown, T. A., and Moore, M. T. (2012). "Confirmatory factor analysis," in *Handbook of Structural Equation Modeling*, ed. R. H. Hoyle (New York, NY: Guilford), 361–379.

Bryant, F. (2003). Savoring beliefs inventory (SBI): a scale for measuring beliefs about savoring. *J. Ment. Health* 12, 175–196. doi: 10.1080/0963823031000103489

Burro, R. (2016). To be objective in experimental phenomenology: a psychophysics application. *SpringerPlus* 5, 1720. doi: 10.1186/s40064-016-3418-4

Burro, R., Raccanello, D., Pasini, M., and Brondino, M. (2018). An estimation of a nonlinear dynamic process using Latent Class extended Mixed Models. Affect profiles after terrorist attacks. *Nonlinear Dyn. Psychol. Life Sci.* 22, 35–52.

Cece, V., Guillet-Descas, E., Nicaise, V., Lienhart, N., and Martinent, G. (2019). Longitudinal trajectories of emotions among young athletes involving in intense training centres: do emotional intelligence and emotional regulation matter? *Psychol. Sport Exerc.* 43, 128–136. doi: 10.1016/j.psychsport.2019.01.011

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equ. Model. Multidiscip. J.* 14, 464–504. doi: 10.1080/10705510701301834

Diener, E. (2000). Subjective well-being: the science of happiness, and a proposal for national index. *Am. Psychol.* 55, 34–43. doi: 10.1037/0003-066X.55.1.34

Diener, E., Emmons, R. A., Larsen, R. J., and Griffin, S. (1985). The satisfaction with life scale. *J. Pers. Assess.* 49, 71–75. doi: 10.1207/s15327752jpa4901_13

Douglass, R. P., and Duffy, R. D. (2015). Strengths use and life satisfaction: a moderated mediation approach. *J. Happin. Stud.* 16, 619–632. doi: 10.1007/s10902-014-9525-4

Egloff, B., Tausch, A., Kohlmann, C. W., and Krohne, H. W. (1995). Relationships between time of day, day of the week, and positive mood: exploring the role of the mood measure. *Motiv. Emot.* 19, 99–110. doi: 10.1007/bf02250565

Farquharson, L., and MacLeod, A. K. (2014). A brief goal-setting and planning intervention to improve well-being for people with psychiatric disorders. *Psychother. Psychos.* 83, 122–124. doi: 10.1159/000356332

Fritz, C., Sonnentag, S., Spector, P. E., and McInroe, J. A. (2010). The weekend matters: relationships between stress recovery and affective experiences. *J. Organ. Behav.* 31, 1137–1162. doi: 10.1002/job.672

Green, M. J. (2014). Latent class analysis was accurate but sensitive in data simulations. *J. Clin. Epidemiol.* 67, 1157–1162. doi: 10.1016/j.jclinepi.2014.05.005

Gross, J. J. (1998). Antecedent- and response-focused emotion regulation: divergent consequences for experience, expression, and physiology. *J. Pers. Soc. Psychol.* 74, 224–237. doi: 10.1037/0022-3514.74.1.224

Gross, J. J. (2014). "Emotion regulation: conceptual and empirical foundations," in *Handbook of Emotion Regulation*, 2nd Edn, ed. J. J. Gross (New York, NY: Guilford), 3–20.

Gross, J. J. (2015). Emotion regulation: current status and future prospects. *Psychol. Inq.* 26, 1–26. doi: 10.1080/1047840X.2014.940781

Gross, J. J., and John, O. P. (2003). Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *J. Pers. Soc. Psychol.* 85, 348–362. doi: 10.1037/0022-3514.85.2.348

Gross, J. J., and Thompson, R. A. (2007). "Emotion regulation: conceptual foundations," in *Handbook of Emotion Regulation*, ed. J. J. Gross (New York: Guilford Press), 3–24.

Gunaydin, G., Selcuk, E., and Ong, D. A. (2016). Trait reappraisal predicts affective reactivity to daily positive and negative events. *Front. Psychol.* 7:1000. doi: 10.3389/fpsyg.2016.01000

Helliwell, J. F., and Wang, S. (2014). Weekends and subjective well-being. *Soc. Indic. Res.* 116, 389–407. doi: 10.1007/s11205-013-0306-y

Jose, P. E., Lim, B. T., and Bryant, F. B. (2012). Does savoring increase happiness? A daily diary study. *J. Posit. Psychol.* 7, 176–187. doi: 10.1080/17439760.2012.671345

Jung, T., and Wickrama, K. A. S. (2007). An introduction to latent class growth analysis and growth mixture modeling. *Soc. Pers. Psychol. Compass* 2, 302–317. doi: 10.1111/j.1751-004.2007.00054.x

Kahneman, D. (1999). "Objective happiness," in *Well-Being: The Foundations of Hedonic Psychology*, eds D. Kahneman, E. Diener, and N. Schwarz (New York: Russell Sage Foundation), 3–25.

Kelley, N. J., Glazer, J. E., Pornpattananangkul, N., and Nusslock, R. (2019). Reappraisal and suppression emotion-regulation tendencies differentially predict reward-responsivity and psychological well-being. *Biol. Psychol.* 140, 35–47. doi: 10.1016/j.biopsycho.2018.11.005

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling*, 4th Edn. New York, NY: Guilford publications.

Larsen, R. J., and Kasimatis, M. (1990). Individual differences in entrainment of mood to the weekly calendar. *J. Pers. Soc. Psychol.* 58, 164–171. doi: 10.1037/0022-3514.58.1.164

Løvoll, H. S., Bentzen, M., and Säfvenbom, R. (2019). Development of positive emotions in physical education: person-centred approach for understanding motivational stability and change. *Scand. J. Educ. Res.* 1–16. doi: 10.1080/00313831.2019.1639818

Lyubomirsky, S., King, L., and Diener, E. (2005). The benefits of frequent positive affect: does happiness lead to success? *Psychol. Bull.* 131, 803–855. doi: 10.1037/0033-2909.131.6.803

Lyubomirsky, S., and Layous, K. (2013). How do simple positive activities increase well-being? *Curr. Direct. Psychol. Sci.* 22, 57–62. doi: 10.1177/0963721412469809

Maciejewski, D. F., Keijsers, L., van Lier, P. A., Branje, S. J., Meeus, W. H., and Koot, H. M. (2019). Most fare well—But some do not: distinct profiles of mood variability development and their association with adjustment during adolescence. *Dev. Psychol.* 55, 434–448.

Muthén, B. (2004). "Latent variable analysis: growth mixture modeling and related techniques for longitudinal data," in *Handbook of Quantitative Methodology for the Social Sciences*, ed. D. Kaplan (Newbury Park, CA: Sage Publications), 345–368.

Muthén, L. K., and Muthén, B. (2006). *Mplus user's guide*, 4th Edn. Los Angeles, CA: Muthén & Muthén.

Nagin, D. S. (1999). Analyzing developmental trajectories: a semi-parametric, group-based approach. *Psychol. Methods* 4, 139–157. doi: 10.1037//1082-989X.4.2.139

Nylund, K. L., Asparouhov, T., and Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Struct. Equ. Model.* 14, 535–569. doi: 10.1080/10705510701575396

Pasini, M., Brondino, M., Burro, R., Raccanello, D., and Gallo, S. (2016). The use of different multiple devices for an ecological assessment in psychological research: an experience with a daily affect assessment. *Adv. Intellig. Soft Comput.* 478, 121–129.

Quoidbach, J., Berry, E. V., Hansenne, M., and Mikolajczak, M. (2010). Positive emotion regulation and well-being: comparing the impact of eight savoring and dampening strategies. *Pers. Individ. Diff.* 49, 368–373. doi: 10.1016/j.paid.2010.03.048

Raccanello, D., Burro, R., Brondino, M., and Pasini, M. (2017). Use of internet and wellbeing: a mixed-device survey. *Adv. Intellig. Soft Comput.* 617, 65–73. doi: 10.1007/978-3-319-60819-8_8

Raccanello, D., Burro, R., Brondino, M., and Pasini, M. (2018). Relevance of terrorism for Italian students not directly exposed to it: the affective impact of the 2015 Paris and the 2016 Brussels attacks. *Stress Health* 34, 338–343. doi: 10.1002/smi.2793

Rasmussen, H. N., Scheier, M. F., and Greenhouse, J. B. (2009). Optimism and physical health: a meta-analytic review. *Ann. Behav. Med.* 37, 239–256. doi: 10.1007/s12160-009-9111-x

Reid, S., Towell, A. D., and Golding, J. F. (2000). Seasonality, social zeitgebers and mood variability in entrainment of mood: implications for seasonal affective disorder. *J. Affect. Disord.* 59, 47–54.

Reis, H. T., Sheldon, K. M., Gable, S. L., Roscoe, J., and Ryan, R. M. (2000). Daily well-being: the role of autonomy, competence, and relatedness. *Pers. Soc. Psychol. Bull.* 26, 419–435.

Roeder, K., Lynch, K., and Nagin, D. (1999). Modeling uncertainty in latent class membership: a case study in criminology. *J. Am. Stat. Assoc.* 94, 766–776.

Rossi, A. S., and Rossi, P. E. (1977). Body time and social time: mood patterns by menstrual cycle phase and day of the week. *Soc. Sci. Res.* 6, 273–308.

Ryan, R. M., Bernstein, J. H., and Brown, K. W. (2010). Weekends, work, and well-being: psychological need satisfactions and day of the week effects on mood, vitality, and physical symptoms. *J. Soc. Clin. Psychol.* 29, 95–122.

Sheldon, K. M., and Houser-Marko, L. (2001). Self-concordance, goal attainment, and the pursuit of happiness: can there be an upward spiral? *J. Pers. Soc. Psychol.* 80, 152–165. doi: 10.1037/0022-3514.80.1.152

Shiffman, S., Stone, A. A., and Hufford, M. R. (2008). Ecological momentary assessment. *Annu. Rev. Clin. Psychol.* 4, 1–32. doi: 10.1037/a0017074

Taylor, S. E., and Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychol. Bull.* 103, 193–210. doi: 10.1037/0033-2909.103.2.193

Terracciano, A., McCrae, R. R., and Costa, P. T. (2003). Factorial and construct validity of the Italian positive and negative affect schedule (PANAS). *Eur. J. Psychol. Assess.* 19, 131–141. doi: 10.1027//1015-5759.19.2.131

Tugade, M. M., and Fredrickson, B. L. (2004). Resilient individuals use positive emotions to bounce back from negative emotional experiences. *J. Pers. Soc. Psychol.* 86, 320–333. doi: 10.1037/0022-3514.86.2.320

Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *J. Pers. Soc. Psychol.* 54, 1063–1070. doi: 10.1037/0022-3514.54.6.1063

Watson, D., Wiese, D., Vaidya, J., and Tellegen, A. (1999). The two general activation systems of affect: structural findings, evolutionary considerations, and psychobiological evidence. *J. Pers. Soc. Psychol.* 76, 820–838. doi: 10.1037/0022-3514.76.5.820

Young, C., and Lim, C. (2014). Time as a network good: evidence from unemployment and the standard workweek. *Sociol. Sci.* 1, 10–27.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 05 August 2020
doi: 10.3389/fpsyg.2020.01791

A General Three-Parameter Logistic Model With Time Effect

Zhaoyuan Zhang[†], *Jiwei Zhang*[*†], *Jian Tao and Ningzhong Shi*[*]

Key Laboratory of Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University, Changchun, China

OPEN ACCESS

Edited by:
Jason C. Immekus,
University of Louisville, United States

Reviewed by:
Rubén Maneiro,
Pontifical University of Salamanca, Spain
Dandan Liao,
American Institutes for Research, United States

Correspondence:
Jiwei Zhang
zhangjw713@nenu.edu.cn
Ningzhong Shi
shinz@nenu.edu.cn

[†]*These authors have contributed equally to this work and share first authorship*

Specialty section:
This article was submitted to Quantitative Psychology and Measurement,
a section of the journal Frontiers in Psychology

Received: *11 October 2019*
Accepted: *29 June 2020*
Published: *05 August 2020*

Citation:
Zhang Z, Zhang J, Tao J and Shi N (2020) A General Three-Parameter Logistic Model With Time Effect. Front. Psychol. 11:1791. doi: 10.3389/fpsyg.2020.01791

Within the framework of item response theory, a new and flexible general three-parameter logistic model with response time (G3PLT) is proposed. The advantage of this model is that it can combine time effect, ability, and item difficulty to influence the correct-response probability. In contrast to the traditional response time models used in educational psychology, the new model incorporates the influence of the time effect on the correct-response probability directly, rather than linking them through a hierarchical method via latent and speed parameters as in van der Linden's model. In addition, the Metropolis–Hastings within Gibbs sampling algorithm is employed to estimate the model parameters. Based on Markov chain Monte Carlo output, two Bayesian model assessment methods are used to assess the goodness of fit between models. Finally, two simulation studies and a real data analysis are performed to further illustrate the advantages of the new model over the traditional three-parameter logistic model.

Keywords: Bayesian inference, deviance information criterion (DIC), item response theory (IRT), logarithm of the pseudomarginal likelihood (LPML), Markov chain Monte Carlo (MCMC), three-parameter logistic model

1. INTRODUCTION AND MOTIVATION

Computerized assessment has become a widely accepted method of testing owing to the fact that the results produced by examinees can be quickly and accurately evaluated by virtue of the computational power that is now available. In addition, with the help of computer technology, the response times of examinees are easier to collect than in the case of traditional paper-and-pencil tests. The collected response times provide a valuable source of information on examinees and test items. For example, response times can be used to improve the accuracy of ability estimates (van der Linden, 2007; Klein Entink et al., 2009a; van der Linden and Glas, 2010; Wang et al., 2013, 2018a; Wang and Xu, 2015; Fox and Marianti, 2016; Bolsinova and Tijmstra, 2018; De Boeck and Jeon, 2019), to detect rapid guessing and cheating behavior (van der Linden and Guo, 2008; van der Linden, 2009; Wang and Xu, 2015; Pokropek, 2016; Qian et al., 2016; Skorupski and Wainer, 2017; Wang et al., 2018a,b; Lu et al., 2019; Sinharay and Johnson, 2019; Zopluoglu, 2019), to evaluate the speededness of tests (Schnipke and Scrams, 1997; van der Linden et al., 2007), and to design more efficient tests (Bridgeman and Cline, 2004; Chang, 2004; Choe et al., 2018).

1.1. Advantages of Our Model Over Traditional Response Time Models in Educational Psychology Research

Although response times in both educational and psychological research have been studied widely and in depth, there are still some deficiencies in the existing literature. Here, we compare existing response time models with our new model and analyze the advantages of our model from multiple aspects.

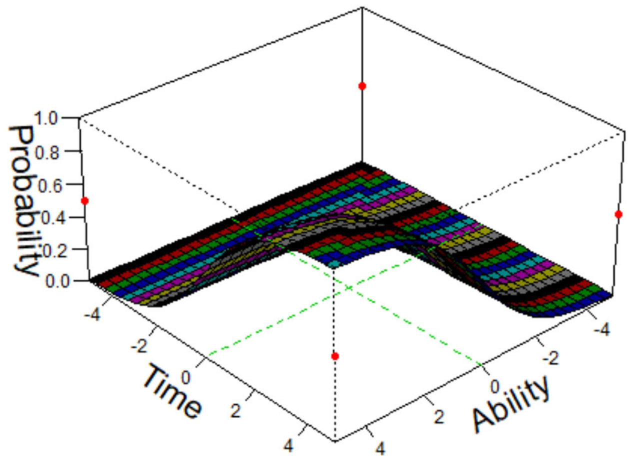


FIGURE 1 | The 3-D diagram of ability, time, and correct response probability.

Thissen (1983) proposed a joint model of response time and accuracy to describe the speed-accuracy relationship. In his model, the speed-accuracy trade-off is reflected by letting response accuracy depend on the time devoted to an item: spending more time on an item increases the probability of a correct response. Thissen's joint model can be expressed as follows:

$$\log T_{ij} = u + \eta_i + \varsigma_j - \rho(a_j\theta_i - b_j) + \varepsilon_{ij},$$

where T_{ij} is the response time of the ith examinee answering the jth item, u is a general intercept parameter, η_i and ς_j can be interpreted, respectively as the speed of examinee i and the amount of time required by item j, ρ is a regression parameter, a_j and b_j are, respectively the item discrimination and difficulty parameters, θ_i is the ability parameter for the ith examinee, and $\varepsilon_{ij} \sim N(0, \sigma^2)$. The speed–accuracy trade–off is represented by the term $a_j\theta_i - b_j$ when $\rho < 0$. When $\rho > 0$, the speed-accuracy relation is reversed. However, the way in which this model incorporates personal-level and item-level parameters means that it is unable to fully reflect the direct impact of the response time on the correct-response probability. Our new model solves this problem. The response time and the ability and item difficulty parameters are combined in an item response model that reflects the way in which the interactions among the three factors influences the correct-response probability. To provide an intuitive explanation, we use a three-dimensional diagram (**Figure 1**) to illustrate the effect of the ability and response time on the correct-response probability. A similar modeling method was proposed by Verhelst et al. (1997).

Roskam (1987, 1997) proposed a Rasch response time model integrating response time and correctness. According to this model, the probability of a correct response for the ith examinee answering the jth item can be written as

$$p(Y_{ij} = 1 \mid T_{ij}, i, j) = \frac{\theta_i T_{ij}}{\theta_i T_{ij} + \delta_j} = \frac{\exp(\xi_i + \tau_{ij} - \kappa_j)}{1 + \exp(\xi_i + \tau_{ij} - \kappa_j)},$$

where Y_{ij} denotes the response of the ith examinee answering the jth item, θ_i is the ability parameter for the ith examinee. δ_j is the item difficulty parameter for the jth item, and ξ_i, τ_{ij}, and κ_j are the logarithms of θ_i, T_{ij}, and δ_j, respectively. We can see that when T_{ij} goes to infinity, the correct-response probability $p(Y_{ij} = 1 \mid T_{ij}, i, j)$ approaches 1, no matter how difficult the item is. In fact, this type of model can only be applied to speeded tests, because a basic characteristic of such tests is that test items are quite easy, so, with unlimited time available, the answers are almost always correct. However, our new model is designed for a power test. This means that even if the examinees are given enough time, they cannot be sure to answer an item correctly, but rather they answer the item correctly with the probability of a three-parameter logistic (3PL) model.

Although there is some similarity between our model and the item response model proposed by Wang and Hanson (2005) with regard to the incorporation of response time into the traditional 3PL model, there are some major differences in concept and construction. Wang and Hanson give the probability of a correct response to item j by examinee i as

$$p(Y_{ij} = 1 \mid a_j, b_j, c_j, d_j, \theta_i, \eta_i, T_{ij}) = c_j + \frac{1 - c_j}{1 + \exp[-1.7a_j(\theta_i - b_j - \eta_i d_j/T_{ij})]},$$

where a_j, b_j, and c_j are, respectively the item discrimination, difficulty, and guessing parameters for the jth item, as in the regular 3PL model. θ_i and η_i are, respectively the ability and slowness parameters for the ith examinee, and d_j is the slowness parameter for the jth item. The item and personal slowness parameters determine the rate of increase in the probability of a correct answer as a function of response time. We will now analyze the differences between the two models.

From the perspective of model construction, the response time and the item and personal parameters are all incorporated into the same exponential function in Wang and Hanson's model, namely, $\exp[-1.7a_j(\theta_i - b_j - \eta_i d_j/T_{ij})]$, whereas in our model, the parameters and time effect appear in two different exponential functions (see the following section for a detailed description of the model): $\exp[-1.7a_j(\theta_i - b_j)] + \exp(-t^*_{ij})$. Our model considers not only the influence of the personal and item factors on the correct-response probability, but also that of the time effect. In Wang and Hanson's model, two slowness parameters associated with persons and items are introduced on the basis of the traditional 3PL model, which increases the complexity of the model. The model can be identified only by imposing stronger constraints on the model parameters. The accuracy of parameter estimation may be reduced owing to the increase in the number of model parameters. However, in our model, no such additional parameters related to items and persons are introduced, and therefore the model is more concise and easy to understand. In terms of model identifiability, our model is similar to the traditional 3PL model in that no additional restrictions need to be imposed. More importantly, parameter estimation becomes more accurate because of the addition of time information. Besides the personal ability parameter, a personal slowness parameter is included Wang and Hanson's model. In fact, their model is a multidimensional item response theory model incorporating response time. In their model, it is assumed that these two personal parameters are independent, but this assumption may not necessarily be true in practice. For example, the lower a person's ability, the slower is their response. That is to say, there is a negative correlation between the ability parameter and the slowness parameter. More research is needed to verify this. Like other models based on the traditional 3PL model (see the next subsection), Wang and Hanson's model cannot distinguish between different abilities under different time intensities when examinees have the same response framework. However, our new model can deal with this problem very well.

In addition, our model introduces the concept of a time weight. Depending on the importance of a test (e.g., whether it is a high-stakes or a low-stakes test), the effect of the time constraint on the whole test is characterized by a time weight. This is something that cannot be dealt with by Wang and Hanson's model.

van der Linden (2007) proposed a hierarchical framework in which responses and response times are modeled separately at the measurement model level, while at a higher level, the ability and speed parameters are included in a population model to account for the correlation between them. In his approach, the latent speed parameter directly affects the response time, while the speed parameters and ability parameters are linked by the hierarchical model. It is known that in item response theory models, ability has a direct impact on the correct-response probability. Thus, we can see that the correct-response probability is related to the response time via the personal parameters (speed and ability). Van der Linden's hierarchical modeling method is unrealistic in that it includes the response time and the ability parameters in the item response model, whereas our model represents the relationships among response time, ability, and correct-response probability more simply and directly. Several other models have a similar structure to van der Linden's hierarchical model, including those of Fox et al. (2007), Klein Entink et al. (2009a,b), van der Linden and Glas (2010), Marianti et al. (2014), Wang and Xu (2015), Wang et al. (2018a), Fox and Marianti (2016), and Lu et al. (2019).

1.2. Advantages of Our Model Compared With the Traditional 3PL Model

Item response theory (IRT) models have been extensively used in educational testing and psychological measurement (Lord and Novick, 1968; van der Linden and Hambleton, 1997; Embretson and Reise, 2000; Baker and Kim, 2004). The most popular IRT model that includes guessing is the 3PL model (Birnbaum, 1968), which has been discussed in many papers and books (see e.g., Hambleton et al., 1991; van der Linden and Hambleton, 1997; Baker and Kim, 2004; von Davier, 2009; Han, 2012). However, several studies have revealed that the 3PL model has technical and theoretical limitations (Swaminathan and Gifford, 1979; Zhu et al., 2018). In this paper, we focus on another defect of the traditional 3PL model, namely, that it cannot distinguish between different abilities under different time intensities when the examinees have the same response framework. Here, we give a simulation example to illustrate the shortcomings of the traditional 3PL model and the advantages of our model (which is a general three-parameter logistic model with response time: G3PLT). We assume that 24 examinees answer three items and that the examinees can be divided into three groups of eight, with the examinees in each group having response frameworks $(1, 0, 0)$, $(0, 1, 0)$, and $(1, 1, 0)$, respectively. Here, 0 indicates that the item is answered correctly and 1 indicating that it is answered incorrectly. The item parameters of the three items are calibrated in advance and known. The discrimination, difficulty, and guessing parameters are set as in **Table 1**.

TABLE 1 | The setting of the true values of discrimination, difficulty, and guessing parameters.

Item	Discrimination	Difficulty	Guessing
1	0.8	−1	0
2	1	0	0.05
3	1.2	1	0.1

To consider the influence of different time effects on the ability of the examinees, eight time transformation values are considered: -0.2, 0.2, 0.5, 1, 2, 3, and 8. The specific settings for the time transformation values can be found in section 2. **Table 2** shows the estimated ability values from the 3PL model and from our model under different response frameworks, with the maximum likelihood method being used to estimate the ability parameter.

The following conclusions can be drawn from **Table 2**.

1. The estimated ability under the G3PLT model with the same response framework will gradually increase as the transformed time decreases from 8 to -0.2. This indicates that the examinees may have different proficiencies in responding to items. Less time is taken if the examinee has greater ability. The time effect captures exactly the information that the traditional 3PL model cannot provide. Specifically, the 3PL model cannot distinguish between abilities when there are different response times under the same response framework.
2. As an illustration, we consider the case where the transformed time is -0.2. The ability estimates under the three response frameworks $(1, 0, 0)$, $(0, 1, 0)$, and $(1, 1, 0)$ are -0.8863, 0.1408, and 1.3109, respectively. We find that the more difficult the item and the greater the number of items answered correctly, the higher are the ability estimates. Without considering the time effect, the ability estimates based on the 3PL model under the three response frameworks are -0.9339, -0.7207, and 0.6659, respectively.
3. Under the three response frameworks, the ability estimates obtained from the G3PLT model and the 3PL model are almost the same when the transformed time reaches 8. This indicates that even if the examinees are allowed enough time, they cannot be certain of answering an item correctly, but can do so only with the correct-response probability given by the 3PL model.

TABLE 2 | The comparisons of ability estimates under the frameworks of 3PL model and G3PLT model.

Examinees	Fitting model	Response framework	Transformated time t^*	Estimation of ability
1	G3PLT	(1, 0, 0)	−0.2	−0.8863
2		(1, 0, 0)	0	−0.8970
3		(1, 0, 0)	0.2	−0.9052
4		(1, 0, 0)	0.5	−0.9142
5		(1, 0, 0)	1	−0.9232
6		(1, 0, 0)	2	−0.9305
7		(1, 0, 0)	3	−0.9327
8		(1, 0, 0)	8	−0.9339
–	3PL	(1, 0, 0)	–	−0.9339
9	G3PLT	(0, 1, 0)	−0.2	0.1408
10		(0, 1, 0)	0	0.0614
11		(0, 1, 0)	0.2	−0.0139
12		(0, 1, 0)	0.5	−0.1233
13		(0, 1, 0)	1	−0.2945
14		(0, 1, 0)	2	−0.5397
15		(0, 1, 0)	3	−0.6515
16		(0, 1, 0)	8	−0.7202
–	3PL	(0, 1, 0)	–	−0.7207
17	G3PLT	(1, 1, 0)	−0.2	1.3109
18		(1, 1, 0)	0	1.0990
19		(1, 1, 0)	0.2	0.9791
20		(1, 1, 0)	0.5	0.8706
21		(1, 1, 0)	1	0.7752
22		(1, 1, 0)	2	0.7016
23		(1, 1, 0)	3	0.6785
24		(1, 1, 0)	8	0.6660
–	3PL	(1, 1, 0)	–	0.6659

We now give another example to further explain the advantages of the G3PLT model. Under the condition that the correct-response probability is the same, we consider the response times of examinees i and j when they answer the same item, and we find that these are 1 and 2 min, respectively. In general, we think that the examinee with shorter response times has a higher ability. Thus, here the ability of examinee i should be higher than that of examinee j. However, since the 3PL model does not consider response time, the difference in ability cannot be distinguished. This problem can be solved by using the G3PLT model. Because this model takes into account the information provided by response time, it can estimate the ability of examinees more objectively and accurately. As shown in **Figure 2**, for the same item, L1 represents the item characteristic curve corresponding to the case where examinees need a long response time ($t_1^* = 4.41$), and L2 represents the item characteristic curve corresponding to the case where examinees need a short response time ($t_1^* = 1.94$). When $p = 0.86$ is given as the correct-response probability, the estimated ability under L1 is 0, while the estimated ability under L2 is 0.88. Therefore, according to the evaluation results from the G3PLT model, the examinees with shorter times should have higher abilities, whereas the 3PL model is unable to distinguish between the two cases. In addition, it can be seen from the figure that when the ability is fixed at 0, the probabilities of a correct response under the two characteristic curves L1 and L2 are 0.86 and 0.52, respectively. This indicates that under the same ability condition, the correct-response probability of the examinees with short response times is lower than that of the examinees with long response times.

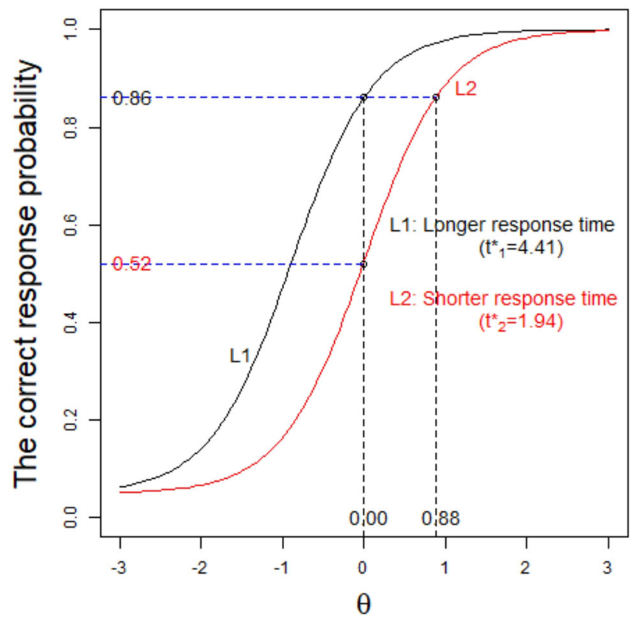


FIGURE 2 | The item characteristic curve based on different time intensities.

The remainder of this paper is organized as follows. Section 2 presents a detailed introduction to the proposed G3PLT model. Section 3 provides a computational strategy based on a Metropolis–Hastings within Gibbs sampling algorithm to meet computational challenges for the proposed model. Two Bayesian model comparison criteria are also discussed in section 3. In section 4, simulation studies are conducted to examine the performance of parameter recovery using the Bayesian algorithm and to assess model fit using the deviance information criterion (DIC) and the logarithm of the pseudomarginal likelihood (LPML). A real data analysis based on the Program for International Student Assessment (PISA) is presented in section 5. We conclude with a brief discussion and suggestions for further research in section 6.

2. THE MODEL AND ITS IDENTIFICATION

2.1. The General Three-Parameter Logistic Model With Response Time (G3PLT)

Let the examinees be indexed by $i = 1, 2, \ldots, N$ and the items by $j = 1, 2, \ldots, J$. Let θ denote the parameters representing the effects of the abilities of the examinees, and let a_j, b_j, and c_j denote the item effects, which are generally interpreted, respectively as discrimination power, difficulty, and success probability in the case of random guessing. If Y_{ij} denotes the response of the ith examinee answering the jth item, then the corresponding correct-response probability can be expressed as

$$\begin{aligned} p_{ij} &= p(Y_{ij} = 1 \mid a_j, b_j, c_j, \theta_i, t^*_{ij}) \\ &= c_j + \frac{1 - c_j}{1 + \exp[-Da_j(\theta_i - b_j)] + \exp(-t^*_{ij})}, \end{aligned} \tag{2.1}$$

where D is a constant equal to 1.7. The influence of the time effect on the probability is described by the term $\exp(-t^*_{ij})$.

2.2. Time Transformation Function

It is obvious that when the response time of each item is very short, the correct-response probability of an item is reduced. In addition, we know that it is impossible for an examinee to answer an item 100% correctly even if they are given enough time to think about the item, and this can be attributed to limitations of the examinee's ability. When examinees are given enough time to answer each item, our model will reduce to the traditional 3PL model, and each item is answered correctly with the corresponding 3PL model correct-response probability. To make the model fully represent the requirement that the correct-response probability varies with time and to eliminate the effects of different average response times for each item in different tests, we consider the following time transformation:

$$t^*_{ij} = f(t_{ij}) = \frac{\log t_{ij} - \mu_t}{\sigma_t} + W, \tag{2.2}$$

where μ_t is the logarithm of the average time spent by all examinees in answering all items, and σ_t is the corresponding standard deviation. W denotes the time weight, which is equal to zero or a positive integer. From the simulation study and real data analysis, we find that the G3PLT model reduces to the traditional 3PL model when the time weight increases to 8, and therefore we restrict the weight to values in the range 0–8. An increase in the time weight indicates that the time factor of the test has a small influence on the correct-response probability of the examinee.

Proposition 1. *Suppose that the correct-response probability* $p(Y_{ij} = 1 \mid a_j, b_j, c_j, \theta_i, t^*_{ij})$*is given by Equation (2.1). Then, we have the following results:*

1. *As the transformed time* $t^*_{ij} \to +\infty$*, the G3PLT model reduces to the 3PL model. That is,*

$$p_{ij} \to c_j + \frac{1 - c_j}{1 + \exp[-Da_j(\theta_i - b_j)]}. \tag{2.3}$$

In other words, it is impossible for the examinee to answer the item 100% correctly even if they are given enough time to think about the item, which can be attributed to the limitations of the examinee's ability.

2. *As the transformed time* $t^*_{ij} \to -\infty$ *(the original time* $t_{ij} \to 0$*), the correct-response probability of the G3PLT model tends to zero. That is,*

$$p_{ij} = c_j + \frac{1 - c_j}{1 + \exp[-Da_j(\theta_i - b_j) + \exp(-t^*_{ij})]} \downarrow 0. \tag{2.4}$$

When there is not enough time to answer items (e.g., at the end of the examination), any item answered by the examinee must be one that requires only a very short time to finish. As the response time continues to shorten, the correct-response probability is reduced.

3. *The G3PLT model can be reduced to a G2PLT model by constraining the lower asymptote parameter* c_j *to be zero, and a G1PLT model can be obtained by further constraining* a_j *to be the same across all items.*

2.3. Asymptotic Properties of the Model

Let p_j be the correct-response rate for the jth item. When the transformed time $t^*_{ij} \to +\infty$, the model in Equation (2.1) can be written as

$$\begin{aligned} &\lim_{t^*_{ij} \to +\infty} \left\{ c_j + \frac{1 - c_j}{1 + \exp[-Da_j(\theta_i - b_j)] + \exp(-t^*_{ij})} \right\} \\ &= c_j + \frac{1 - c_j}{1 + \exp[-Da_j(\theta_i - b_j)]} = p_j. \end{aligned} \tag{2.5}$$

The ability can be obtained as

$$\theta_i = b_j - \frac{1}{Da_j} \log\left(\frac{1 - p_j}{p_j - c_j} \right). \tag{2.6}$$

Next, we will use a specific example to explain the meaning of Equations (2.5) and 2.6. Assuming that $p_j = 0.5$, $a_j = 1.5$, $b_j = 1$, and $c_j = 0.1$, we obtain $\theta_i = 0.8$ from Equation (2.6). This result indicates that even if examinee i has sufficient response time to finish item j, the examinee's ability should be at least 0.8 (the intersection of the vertical asymptote and the x-axis in **Figure 3**) if the correct response probability reaches 0.5; otherwise, no matter how long a response time is allowed, the examinee's correct-response probability cannot reach 0.5. This is like a primary school pupil attempting to solve a college math problem, because the pupil's ability is so low that no matter how much time he is given, he cannot get a correct answer to item j other than by guessing. Moreover, when the ability $\theta_i \to +\infty$, the model in Equation (2.1) can be written as

$$\begin{aligned} &\lim_{\theta_i \to +\infty} \left\{ c_j + \frac{1 - c_j}{1 + \exp[-Da_j(\theta_i - b_j)] + \exp(-t^*_{ij})} \right\} \\ &= c_j + \frac{1 - c_j}{1 + \exp(-t^*_{ij})} = p_j. \end{aligned} \tag{2.7}$$

The transformed time t^*_{ij} can be obtained as

$$t^*_{ij} = -\log\left(\frac{1 - p_j}{p_j - c_j} \right). \tag{2.8}$$

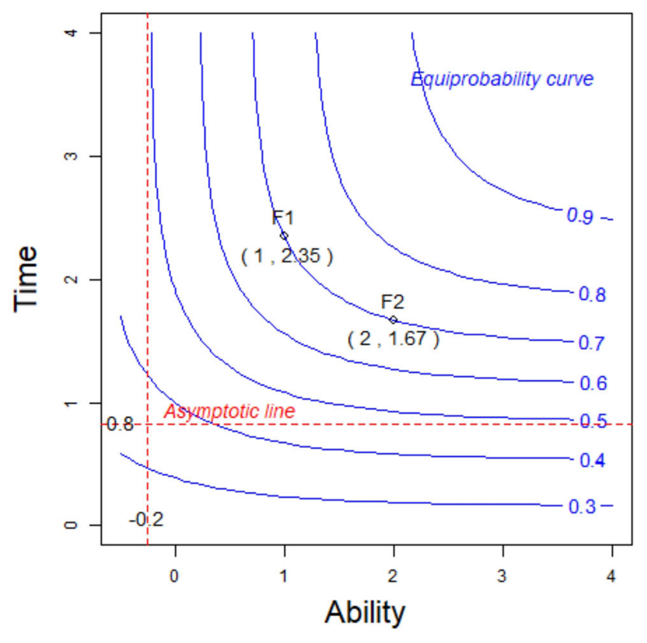


FIGURE 3 | The equiprobability curve based on the ability and time.

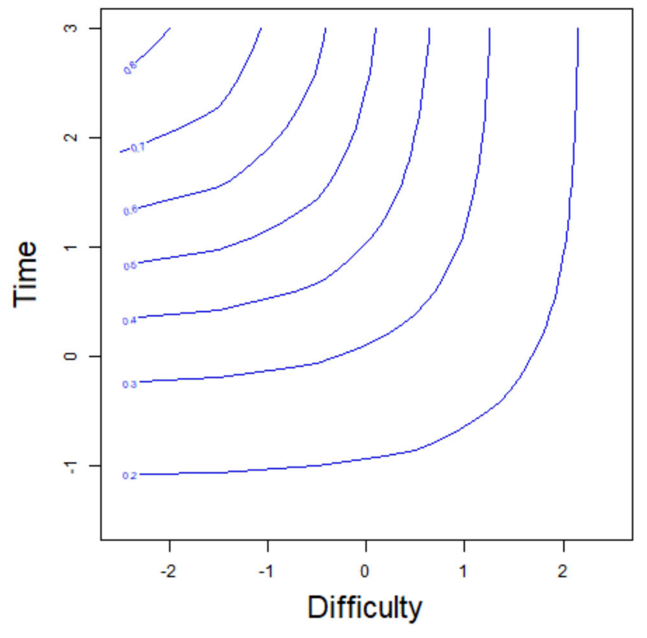


FIGURE 4 | The equiprobability curve based on the item difficulty and time.

We again assume that $p_j = 0.5$, $a_j = 1.5$, $b_j = 1$, and $c_j = 0.1$. From (2.8), the transformed time t^*_{ij} is about -0.2. This result indicates that even if the examinee i has a strong ability, the transformed time required to answer item j should not be less than -0.2 (the intersection of the horizontal asymptote and the y-axis in **Figure 3**) if the correct-response probability reaches 0.5; otherwise, no matter how strong the ability of the examinee, it is impossible to reach a correct-response probability of 0.5. This is like a college student solving a primary school math problem. Although the college student's ability is very strong, she cannot finish the item in a very short time. In addition, the correct-response probability of the examinees is the same for two points on the equiprobability curve. For example, for the two examinees F1 and F2 with the same correct-response probability 0.7 in **Figure 3**, the examinee F1 with low ability (1) takes a long time (2.35), while the response time (1.67) of the examinee F2 with high ability (2) is short to obtain the same correct-response probability. Similarly, the equiprobability curve based on item difficulty and time is shown in **Figure 4**. The correct-response probability is the same for two points on the equiprobability curve. The item with high difficulty takes a long time, while the response time of the item with low difficulty is short, giving the same correct-response probability.

2.4. Model Identification

To ensure identification of the G3PLT model, either the scale of latent traits or the scale of item parameters has to be restricted (Birnbaum, 1968; Lord, 1980; van der Linden and Hambleton, 1997). In this paper, we set the mean and variance of the latent traits to zero and one, respectively (Bock and Aitkin, 1981). The mean of the latent trait is fixed to remove the trade-off between θ_i and b_j in location, and the variance of the latent trait is fixed to remove the trade-off among θ_i, b_j, and a_j in scale.

3. BAYESIAN INFERENCE

3.1. Prior and Posterior Distributions

In a Bayesian framework, the posterior distribution of the model parameters is obtained based on the observed data likelihood (sample information) and prior distributions (prior information). In general, these two kinds of information have an important influence on the posterior distribution. However, in large-scale educational assessment, the number of examinees is often very large. Therefore, the likelihood information plays a dominant role, and the selection of different priors (informative or non-informative) has no significant influence on the posterior inference (van der Linden, 2007; Wang et al., 2018a). Based on previous results (Wang et al., 2018a), we adopt the informative prior distribution to analyze the following simulation studies and real data. The specific settings are as follows. For the latent ability, we assume a standardized normal prior, i.e., $\theta_i \sim N(0, 1)$ for $i = 1, \ldots, N$. The prior distribution for the discrimination parameter a_j is a lognormal distribution, i.e., $a_j \sim \log N(0, 1)$ for $j = 1, \ldots, J$. The prior distribution for the difficulty parameter b_j is a standardized normal distribution, i.e., $b_j \sim N(0, 1)$ for $j = 1, \ldots, J$. For the guessing parameter, we assume a Beta distribution, i.e., $c_j \sim \text{Beta}(2, 10)$ for $j = 1, \ldots, J$. Then, the joint posterior distribution of the parameters given the data is as follows:

$$p(\boldsymbol{\theta},\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}\mid \boldsymbol{Y},\boldsymbol{T}) \propto \left[\prod_{i=1}^{N}\prod_{j=1}^{J} p(Y_{ij}\mid \theta_i,a_j,b_j,c_j,T_{ij})\right]\prod_{i=1}^{N}p(\theta_i) \times \prod_{j=1}^{J}p(a_j)p(b)p(c_j). \tag{3.1}$$

3.2. Bayesian Estimation

Bayesian methods have been widely applied to estimate parameters in complex IRT models (see e.g., Albert, 1992; Patz and Junker, 1999a,b; Béguin and Glas, 2001; Rupp et al., 2004). In this study, the Metropolis-Hastings within Gibbs algorithm (Metropolis et al., 1953; Hastings, 1970; Tierney, 1994; Chib and Greenberg, 1995; Chen et al., 2000) is used to draw samples from the full conditional posterior distributions because the parameters of interest do not have conjugate priors within the framework of the IRT model.

Detailed MCMC Sampling Process

Step 1: Sample the ability parameter θ_i for the ith examinee. We independently draw θ_i^* from the normal proposal distribution, i.e., $\theta_i^* \sim N(\theta_i^{(r-1)}, v_\theta^2)$. The prior of θ_i is assumed to follow a normal distribution with mean μ_θ and variance σ_θ^2, i.e., $\theta_i \sim N(\mu_\theta, \sigma_\theta^2)$. Therefore, the acceptance probability is given by

$$\alpha(\theta_i^{(r-1)},\theta_i^*) = \min\left\{1, \frac{p(\boldsymbol{Y}_i\mid \theta_i^*, \boldsymbol{a}^{(r-1)}, \boldsymbol{b}^{(r-1)}, \boldsymbol{c}^{(r-1)}, \boldsymbol{T}_i)p_{\text{prior}}(\theta_i^*\mid \mu_\theta,\sigma_\theta^2)}{p(\boldsymbol{Y}_i\mid \theta_i^{(r-1)}, \boldsymbol{a}^{(r-1)}, \boldsymbol{b}^{(r-1)}, \boldsymbol{c}^{(r-1)}, \boldsymbol{T}_i)p_{\text{prior}}(\theta_i^{(r-1)}\mid \mu_\theta,\sigma_\theta^2)}\right\}. \tag{3.2}$$

Otherwise, the value of the preceding iteration is retained, i.e., $\theta_i = \theta_i^{(r-1)}$. Here, $\boldsymbol{Y}_i = (Y_{i1}, Y_{ti2}, \ldots, Y_{iJ})$, $\boldsymbol{T}_i = (Y_{i1}, Y_{ti2}, \ldots, Y_{iJ})$, $\boldsymbol{a} = (a_1, a_2, \ldots, a_J)$, $\boldsymbol{b} = (b_1, b_2, \ldots, b_J)$, and $\boldsymbol{c} = (c_1, c_2, \ldots, c_J)$. In Equation (3.3), $p(\boldsymbol{Y}_i \mid \theta_i, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{T}_i) = \prod_{j=1}^{J}(p_{ij})^{y_{ij}}(1-p_{ij})^{1-y_{ij}}$, where p_{ij} is given in Equation (2.1).

Step 2: Sample the difficulty parameter b_j for the jth item. We independently draw b_j^* from the normal proposal distribution, i.e., $b_j^* \sim N(b_j^{(r-1)}, v_j^2)$. The prior of b_j is assumed to follow a normal distribution with mean μ_b and variance σ_b^2, i.e., $b_j \sim N(\mu_b, \sigma_b^2)$. The acceptance probability is given by

$$\alpha(b_j^{(r-1)},b_j^*) = \min\left\{1, \frac{p(\boldsymbol{Y}_j\mid \boldsymbol{\theta}^{(r)}, a_j^{(r-1)}, b_j^*, c_j^{(r-1)}, \boldsymbol{T}_j)p_{\text{prior}}(b_j^*\mid \mu_b,\sigma_b^2)}{p(\boldsymbol{Y}_j\mid \boldsymbol{\theta}^{(r)}, a_j^{(r-1)}, b_j^{(r-1)}, c_j^{(r-1)}, \boldsymbol{T}_j)p_{\text{prior}}(b_j^{(r-1)}\mid \mu_b,\sigma_b^2)}\right\}. \tag{3.3}$$

Otherwise, the value of the preceding iteration is retained, i.e., $b_j = b_j^{(r-1)}$. Here, $\boldsymbol{Y}_j = (Y_{1j}, Y_{2j}, \ldots, Y_{Nj})$, $\boldsymbol{T}_j = (T_{1j}, T_{2j}, \ldots, T_{Nj})$, and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_N)$. In Equation (3.3), $p(\boldsymbol{Y}_j \mid \boldsymbol{\theta}, a_j, b_j, c_j, \boldsymbol{T}_j) = \prod_{i=1}^{n}(p_{ij})^{y_{ij}}(1-p_{ij})^{1-y_{ij}}$.

Step 3: Sample the discrimination parameter a_j for the jth item. We independently draw a_j^* from the log-normal proposal distribution, i.e., $a_j^* \sim \log N(\log a_j^{(r-1)}, v_a^2)$. In addition, $p_{\text{prior}}(a_j)$ is a lognormal prior distribution, i.e., $a_j \sim \log N(\mu_a, \sigma_a^2)$. The acceptance probability is given by

$$\alpha(a_j^{(r-1)},a_j^*) = \min\left\{1, \frac{p(\boldsymbol{Y}_j\mid \boldsymbol{\theta}^{(r)}, a_j^*, b_j^{(r)}, c_j^{(r-1)}, \boldsymbol{T}_j)p_{\text{prior}}(a_j^*\mid \mu_a,\sigma_a^2)a_j^*}{p(\boldsymbol{Y}_j\mid \boldsymbol{\theta}^{(r)}, a_j^{(r-1)}, b_j^{(r)}, c_j^{(r-1)}, \boldsymbol{T}_j)p_{\text{prior}}(a_j^{(r-1)}\mid \mu_a,\sigma_a^2)a_j^{(r-1)}}\right\}. \tag{3.4}$$

Otherwise, the value of the preceding iteration is retained, i.e., $a_j = a_j^{(r)}$. In Equation (3.4), $(\boldsymbol{Y}_j \mid \boldsymbol{\theta}, a_j, b_j, c_j, \boldsymbol{T}_j) = \prod_{i=1}^{n}(p_{ij})^{y_{ij}}(1-p_{ij})^{1-y_{ij}}$.

Step 4: Sample the guessing parameter c_j for the jth item. We independently draw c_j^* from the uniform proposal distribution, i.e., $c_j^* \sim U(c_j^{(r-1)} - 0.01, c_j^{(r-1)} + 0.01)$. The prior of c_j is assumed to follow a Beta distribution, i.e., $c_j \sim \text{Beta}(\upsilon_1, \upsilon_2)$. Therefore, the acceptance probability is given by

$$\alpha(c_j^{(r-1)},c_j^*) = \min\left\{1, \frac{p(\boldsymbol{Y}_j\mid \boldsymbol{\theta}^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^*, \boldsymbol{T}_j)p_{\text{prior}}(c_j^*\mid \upsilon_1,\upsilon_2)}{p(\boldsymbol{Y}_j\mid \boldsymbol{\theta}^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r-1)}, \boldsymbol{T}_j)p_{\text{prior}}(c_j^{(r-1)}\mid \upsilon_1,\upsilon_2)}\right\}. \tag{3.5}$$

Otherwise, the value of the preceding iteration is retained, i.e., $c_j = c_j^{(r)}$. In Equation (3.5), $p(\boldsymbol{Y}_j \mid \boldsymbol{\theta}, a_j, b_j, c_j, \boldsymbol{T}_j) = \prod_{i=1}^{n}(p_{ij})^{y_{ij}}(1-p_{ij})^{1-y_{ij}}$.

3.3. Bayesian Model Assessment

Spiegelhalter et al. (2002) proposed the deviance information criterion (DIC) for model comparison when the number of parameters is not clearly defined. The DIC is an integrated measure of model fit and complexity. It is defined as the sum of a deviance measure and a penalty term for the effective number of parameters based on a measure of model complexity. We write $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_{ij},\ i = 1, \ldots, N,\ j = 1, \ldots, J)$, where $\boldsymbol{\Omega}_{ij} = (\theta_i, a_j, b_j, c_j)'$. Let $\{\boldsymbol{\Omega}^{(1)}, \ldots, \boldsymbol{\Omega}^{(R)}\}$, where $\boldsymbol{\Omega}^{(r)} = (\boldsymbol{\Omega}_{ij}^{(r)},\ i = 1, \ldots, N,\ j = 1, \ldots, J)$, $\boldsymbol{\Omega}_{ij}^{(r)} = (\theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)})'$ for $i = 1, \ldots, N, j = 1, \ldots, J$, and $r = 1, \ldots, R$, denote an Markov chain Monte Carlo (MCMC) sample from the posterior distribution in Equation (3.1). The joint likelihood function of the responses can be written as

$$L(\boldsymbol{Y}\mid \boldsymbol{\Omega},\boldsymbol{T}) = \prod_{i=1}^{N}\prod_{j=1}^{J} f(y_{ij}\mid \theta_i, a_j, b_j, c_j, t_{ij}), \tag{3.6}$$

where $f(y_{ij} \mid \theta_i, a_j, b_j, c_j, t_{ij})$ is the response probability of the G3PLT model. The logarithm of the joint likelihood function in Equation (3.6) evaluated at $\boldsymbol{\Omega}^{(r)}$ is given by

$$\log L(\boldsymbol{Y}\mid \boldsymbol{\Omega}^{(r)},\boldsymbol{T}) = \sum_{i=1}^{N}\sum_{j=1}^{J} \log f(y_{ij}\mid \theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)}, t_{ij}). \tag{3.7}$$

The joint log-likelihoods for the responses, $\log f(y_{ij} \mid \theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)}, t_{ij})$, $i = 1, \ldots, N$ and $j = 1, \ldots, J$, are readily available from MCMC sampling outputs, and therefore $\log f(y_{ij} \mid \theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)}, t_{ij})$ in Equation (3.7) is easy to compute. The effective number of parameters in the models is defined by

$$p_D = \overline{\text{Dev}(\boldsymbol{\Omega})} - \text{Dev}(\widehat{\boldsymbol{\Omega}}), \tag{3.8}$$

where $\overline{\text{Dev}(\boldsymbol{\Omega})}$ is a Monte Carlo estimate of the posterior expectation of the deviance function $\text{Dev}(\boldsymbol{\Omega}) = -2\log L(\boldsymbol{Y} \mid \boldsymbol{\Omega}, \boldsymbol{T})$, and the term $\text{Dev}(\widehat{\boldsymbol{\Omega}})$ is computed by plugging the mean of the simulated values of $\boldsymbol{\Omega}$ into $\text{Dev}(\boldsymbol{\cdot})$, where $\widehat{\boldsymbol{\Omega}} = \sum_{r=1}^{R} \boldsymbol{\Omega}^{(r)}/R$. More specifically,

$$\begin{aligned} \overline{\text{Dev}(\boldsymbol{\Omega})} &= -\frac{2}{R}\sum_{r=1}^{R} \log L(\boldsymbol{Y} \mid \boldsymbol{\Omega}^{(r)}), \\ \text{Dev}(\widehat{\boldsymbol{\Omega}}) &= -2\log L(\boldsymbol{Y} \mid \widehat{\boldsymbol{\Omega}}). \end{aligned} \tag{3.9}$$

The DIC can now be formulated as follows:

$$\text{DIC} = \widehat{\text{Dev}(\boldsymbol{\Omega})} + 2p_D = \widehat{\text{Dev}(\boldsymbol{\Omega})} + 2\left[\overline{\text{Dev}(\boldsymbol{\Omega})} - \widehat{\text{Dev}(\boldsymbol{\Omega})}\right], \tag{3.10}$$

A model with a smaller DIC value fits the data better.

Another method is to use the logarithm of the pseudomarginal likelihood (LPML) (Geisser and Eddy, 1979; Ibrahim et al., 2001) to compare different models. This is also based on the log-likelihood functions evaluated at the posterior samples of model parameters. The detailed calculation process is as follows.

We let $U_{ij,\max} = \max_{1 \le r \le R}[-\log f(y_{ij} \mid \theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)}, t_{ij})]$, and a Monte Carlo estimate of the conditional predictive ordinate (CPO) (Gelfand et al., 1992; Chen et al., 2000) is then given by

$$\begin{aligned} \log(\widehat{\text{CPO}_{ij}}) &= -U_{ij,\max} \\ &- \log\left\{\frac{1}{R}\sum_{r=1}^{R} \exp[-\log f(y_{ij} \mid \theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)}, t_{ij}) - U_{ij,\max}]\right\}. \end{aligned} \tag{3.11}$$

Note that the maximum value adjustment used in $\log(\widehat{\text{CPO}_{ij}})$ plays an important role in numerical stabilization in the computation of $\exp[-\log f(y_{ij} \mid \theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, c_j^{(r)}, t_{ij}) - U_{ij,\max}]$ in Equation (3.11). A summary statistic of the $\widehat{\text{CPO}_{ij}}$ is the sum of their logarithms, which is called the LPML and is given by

$$\text{LPML} = \sum_{i=1}^{N}\sum_{j=1}^{J} \log(\widehat{\text{CPO}_{ij}}). \tag{3.12}$$

A model with a larger LPML has a better fit to the data.

3.4. Accuracy Evaluation of Parameter Estimation

To implement the MCMC sampling algorithm, chains of length 10,000 with an initial burn-in period 5,000 are chosen. In the following simulation study, 200 replications are used. Five indices are used to assess the accuracy of the parameter estimates. Let ϑ be the parameter of interest. Assume that $M = 200$ data sets are generated. Also, let $\widehat{\vartheta}^{(m)}$ and $\text{SD}^{(m)}(\vartheta)$ denote the posterior mean and the posterior standard deviation of ϑ obtained from the mth simulated data set for $m = 1, \ldots, M$.

The bias for the parameter ϑ is defined as

$$\text{Bias}(\vartheta) = \frac{1}{M}\sum_{m=1}^{M}(\widehat{\vartheta}^{(m)} - \vartheta), \tag{3.13}$$

and the mean squared error (MSE) for ϑ is defined as

$$\text{MSE}(\vartheta) = \frac{1}{M}\sum_{m=1}^{M}(\widehat{\vartheta}^{(m)} - \vartheta)^2. \tag{3.14}$$

The simulation SE is the square root of the sample variance of the posterior estimates over different simulated data sets. It is defined as

$$\text{Simulation SE}(\vartheta) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\vartheta}^{(m)} - \frac{1}{M}\sum_{\ell=1}^{M}\widehat{\vartheta}^{(\ell)}\right)^2}, \tag{3.15}$$

and the average of posterior standard deviation is defined as

$$\text{SD}(\vartheta) = \frac{1}{M}\sum_{m=1}^{M}\text{SD}^{(m)}(\vartheta). \tag{3.16}$$

The coverage probability based on the 95% highest probability density (HPD) intervals is defined as

$$\begin{aligned} &\text{CP}(\vartheta) \\ &= \frac{\#\text{ of 95\% (HPD) intervals containing } \vartheta \text{ in } M \text{ simulated data sets}}{M}. \end{aligned} \tag{3.17}$$

4. SIMULATION STUDY

4.1. Simulation 1

We conduct a simulation study to evaluate the recovery performance of the combined MCMC sampling algorithm based on different simulation conditions.

Simulation Design

The following manipulated conditions are considered: (a) test length $J = 20$ or 60 and (b) number of examinees $N = 500$, 1,000, or 2,000. Fully crossing different levels of these two factors yields six conditions (two test lengths × three sample sizes). Next, the true values of the parameters are given. True item discrimination parameters a_j are generated from a truncated normal distribution, i.e., $a_j \sim N(1, 0.2)\text{I}(a_j > 0), j = 1, 2, \ldots, N$, where the indicator function $\text{I}(A)$ takes a value of 1 if A is true and a value of 0 if A is false. The item difficulty parameters b_j are generated from a standardized normal distribution. The item guessing parameters c_j are generated from a Beta distribution, i.e., $c_j \sim \text{Beta}(2, 10)$. In addition, the ability parameters of the examinees, θ_i, are also generated from a standardized normal distribution. In each simulation condition, 200 replications (replicas) are considered. Next, we generate the response time data for each examinee based on the following facts:

1. The difficulty of each item has a direct impact on the response time. That is to say, the time spent on simple items is shorter, and the time spent on difficult items is longer.
2. In addition, the ability of each examinee also has a direct impact on the response time. That is to say, examinees with higher ability spend less time on an item.
3. Depending on the importance of the test (high-stakes test or low-stakes test), the effect of the time constraint on the whole test should be characterized by the time weighting.

In Wang and Xu (2015, p. 459), the average logarithms of the response times for each item based on the solution behavior follow a normal distribution. That is, $\log t_j \sim N(0.5, 0.25)$, $j = 1, 2, \ldots, J$, where the average time t_j spent on item j is about 1.64872 $(= e^{0.5})$ min. We take the standardized transformation $t_j^* = f(t_j) = (\log t_j - 0.5)/0.5$, so that $t_j^* \sim N(0, 1)$, where $-\infty < t_j^* < +\infty$.

Next, we consider the premise that the easier an item, the shorter is the response time. The true values of the difficulty parameter and the transformed time t_j^* for each item are arranged in order from small to large, i.e., $b_1 < b_2 < \cdots < b_{J-1} < b_J$ and $t_1^* < t_2^* < \cdots < t_{J-1}^* < t_J^*$. The corresponding item–time pairs can be written as $(b_1, t_1^*) < (b_2, t_2^*) < \cdots < (b_{J-1}, t_{J-1}^*) < (b_J, t_J^*)$. The response time of each examinee is generated from a normal distribution, i.e., $t_{ij}^* \sim N(t_j^*, 0.5)$, where $j = 1, \ldots, J$. Moreover, for a given item j, the premise that examinees with higher ability spend less time on the item needs to be satisfied. Therefore, we arrange $\theta_{1j} > \theta_{2j} > \cdots > \theta_{N-1,j} > \theta_{N,j}$, and $t_{1j}^* < t_{2j}^* < \cdots < t_{N-1,j}^* < t_{N,j}^*$. The corresponding ability–time pairs can be obtained by arranging the true values of the ability parameter and the transformed time t_{ij}^*, i.e., (θ_{ij}, t_{ij}^*). The time weights range from 0 to 8. The higher the value of the time weight, the weaker is the influence of the time factor of the test on the correct-response probability of the examinee. In this simulation study, we assume that the time factor of the test has an important influence on the correct-response probability of the examinee. Therefore, we set the time weight to 1 in this simulation. Based on the true values of the parameters and the response time data, the response data can be simulated using the G3PLT model given by Equation (2.1).

Convergence Diagnostics

To evaluate the convergence of the parameter estimations, we only consider convergence in the case of minimum sample sizes. That is, the test length is fixed at 20, and the number of examinees is 500. Two methods are used to check the convergence of our algorithm. One is the "eyeball" method to monitor convergence by visually inspecting the history plots of the generated sequences (Zhang et al., 2007; Hung and Wang, 2012), and the other is the Gelman–Rubin method (Gelman and Rubin, 1992; Brooks and Gelman, 1998) for checking the convergence of the parameters.

The convergence of the Bayesian algorithm is checked by monitoring the trace plots of the parameters for consecutive sequences of 10,000 iterations. The trace plots show that all parameter estimates stabilize after 5,000 iterations and then converge quickly. Thus, we set the first 5,000 iterations as the burn-in period. As an illustration, four chains started at overdispersed starting values are run for each replication. The trace plots of three randomly selected items are shown in **Figure 5**. In addition, we find that the potential scale reduction factor (PSRF) (Brooks and Gelman, 1998) values for all parameters are less than 1.2, which ensures that all chains converge as expected.

Recovery of Item Parameters

The average bias, MSE, SD, SE, and CP for discrimination, difficulty, and guessing parameters based on six different simulation conditions are shown in **Table 3**. The following conclusions can be drawn.

1. Given the total test length, when the number of individuals increases from 500 to 2,000, the average MSE, SD, and SE for the discrimination, difficulty, and guessing parameters show a decreasing trend. For example, for a total test length of 20 items, when the number of examinees increases from 500 to 2,000, the average MSE of all discrimination parameters decreases from 0.0088 to 0.0072, the average SE of all discrimination parameters decreases from 0.0022 to 0.0014, and the average SD of all discrimination parameters decreases from 0.0085 to 0.0066. The average MSE of all difficulty parameters decreases from 0.0436 to 0.0213, the average SE of all difficulty parameters decreases from 0.0272 to 0.0122, and the average SD of all difficulty parameters decreases from 0.0362 to 0.0143. The average MSE of all guessing parameters decreases from 0.0019 to 0.0013, the average SE of all guessing parameters decreases from 0.0007 to 0.0006, and the average SD of all guessing parameters decreases from 0.0013 to 0.0008.
2. The average SDs of the item parameters are larger than their average SEs. This indicates that the fluctuations of the posterior means of item parameters between different replications are small compared with their fluctuations within each replication.
3. Under the six simulated conditions, the average CPs of the discrimination, difficulty, and guessing parameters are about 0.950.
4. When the number of examinees is held fixed but the number of items increases from 20 to 40, the average MSE, SD, and SE show that the recovery results for the discrimination, difficulty and guessing parameters do not change much, which indicates that the Bayesian algorithm is stable and there is no reduction in accuracy due to an increase in the number of items.

In summary, the Bayesian algorithm provides accurate estimates of the item parameters for various numbers of examinees and items. Therefore, it can be used as a guide to practice.

Recovery of Ability Parameters

Next, we evaluate the recovery of latent ability from the plots of the true values and the estimates in **Figure 6**. For a fixed number

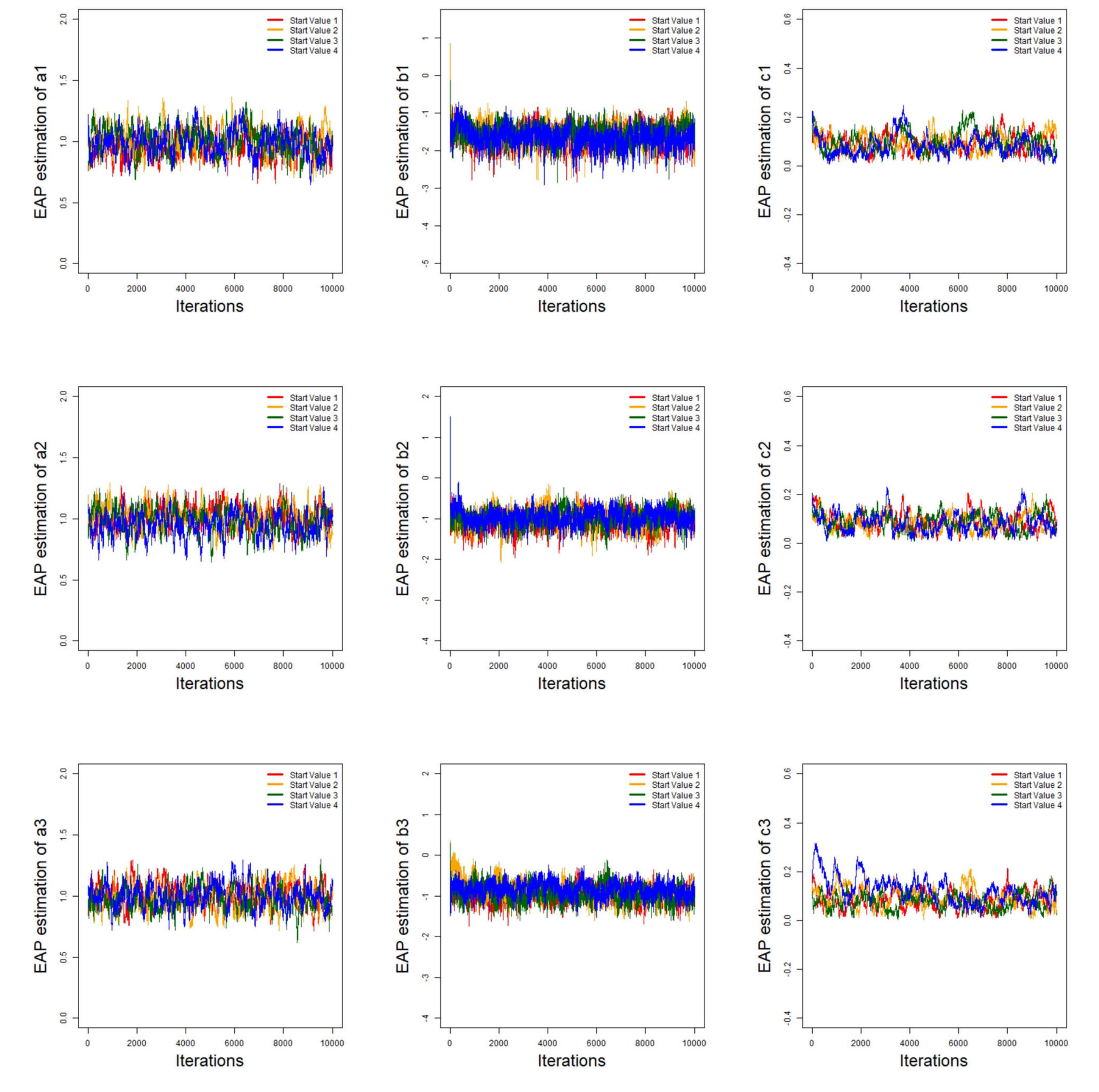


FIGURE 5 | The trace plots of three randomly selected items for the simulation study 1.

of examinees (500 or 1,000), when the number of items increases from 20 to 60, the ability estimates become more accurate, with the true values and the estimates basically lying on the diagonal line. Note that the estimated abilities are the average of 200 replication estimates. Because of the increase in the number of items, the probability of the situation in which all items are answered correctly by the high-ability examinees and incorrectly by the low-ability examinees, leading to a large deviation of the ability estimators, is reduced. Therefore, the estimated values and the true values of the ability at the end of the curve are closer to the diagonal line when the number of items is 60. In summary, the Bayesian sampling algorithm also provides accurate estimates of the ability parameters in term of the plots of the true values and the estimates.

4.2. Simulation 2

In this simulation study, we use the DIC and LPML model assessment criteria to evaluate model fitting. Two issues need further study. The first is whether the two criteria can accurately identify the true model that generates data from numerous fitting

TABLE 3 | Evaluating the accuracy of parameters based on six different simulated conditions in simulation study 1.

	No. of items=20														
	No. of examinees 500					No. of examinees 1,000					No. of examinees 2,000				
Item parameter	**Bias**	**MSE**	**SE**	**SD**	**CP**	**Bias**	**MSE**	**SE**	**SD**	**CP**	**Bias**	**MSE**	**SE**	**SD**	**CP**
Discrimination ***a***	−0.0162	0.0088	0.0022	0.0085	0.9513	−0.0081	0.0083	0.0019	0.0076	0.9503	−0.0038	0.0072	0.0014	0.0066	0.9480
Difficulty ***b***	−0.0134	0.0436	0.0272	0.0362	0.9385	−0.0103	0.0290	0.0166	0.0213	0.9410	−0.0068	0.0213	0.0122	0.0143	0.9287
Guessing ***c***	−0.0031	0.0019	0.0007	0.0013	0.9315	−0.0026	0.0016	0.0006	0.0010	0.9378	−0.0014	0.0013	0.0006	0.0008	0.9283
	No. of items=60														
	No. of examinees 500					No. of examinees 1,000					No. of examinees 2,000				
Item parameter	**Bias**	**MSE**	**SE**	**SD**	**CP**	**Bias**	**MSE**	**SE**	**SD**	**CP**	**Bias**	**MSE**	**SE**	**SD**	**CP**
Discrimination ***a***	0.0159	0.0082	0.0023	0.0082	0.9543	0.0132	0.0081	0.0019	0.0071	0.9393	0.0005	0.0074	0.0014	0.0059	0.9343
Difficulty ***b***	−0.0345	0.0447	0.0245	0.0339	0.9574	−0.0112	0.0233	0.0153	0.0205	0.9497	−0.0086	0.0163	0.0098	0.0121	0.9296
Guessing ***c***	0.0071	0.0016	0.0005	0.0011	0.9389	0.0061	0.0013	0.0005	0.0008	0.9328	0.0025	0.0011	0.0005	0.0006	0.9484

The Bias, MSE, SD, SE, and CP denote the average; Bias, MSE, SE, SD, and CP for the parameters. ***a*** *represents all discrimination parameters,* ***b*** *represents all difficulty parameters and* ***c*** *represents all guessing parameters.*

models. The second concerns the influence of different time weights in the G3PLT model on model fitting.

Simulation Design

In this simulation, the number of examinees is $N = 1,000$ and the test length is fixed at 20. Six item response models will be considered: the traditional 3PL model and the G3PLT model with time weights $W = 0$, 2, 4, 6, and 8. Thus, we evaluate the model fitting in the following five cases:

- Case 1. True model: G3PLT model with time weight 0 vs. Fitted model: 3PL model, G3PLT model with time weight 0.
- Case 2. True model: G3PLT model with time weight 2 vs. Fitted model: 3PL model, G3PLT model with time weight 2.
- Case 3. True model: G3PLT model with time weight 4 vs. Fitted model: 3PL model, G3PLT model with time weight 4.
- Case 4. True model: G3PLT model with time weight 6 vs. Fitted model: 3PL model, G3PLT model with time weight 6.
- Case 5. True model: G3PLT model with time weight 8 vs. Fitted model: 3PL model, G3PLT model with time weight 8.

The true values and prior distributions for the parameters are the same as in Simulation 1. To implement the MCMC sampling algorithm, chains of length 10,000 with an initial burn-in period 5,000 are chosen. The results of Bayesian model assessment based on the 200 replications are shown in **Table 4**. Note that the following results for DIC and LPML are based on the average of 200 replications.

From **Table 4**, we find that when the G3PLT model with time weight 0 (G3PLT0) is the true model, the G3PLT0 model is chosen as the better-fitting model according to the results for DIC and LPML, which is what we expect to see. The medians of DIC and LPML are respectively 25 324.43 and −13231.77. The differences between the G3PLT0 model and 3PL model in the medians of DIC and LPML are −33.72 and 199.23, respectively. Similarly, when the G3PLT model with time weight 2 (G3PLT2) is the true model, the G3PLT2 model is also chosen as the better-fitting model according to the results for DIC and LPML. The medians of DIC and LPML are respectively 22 777.38 and −12221.93. The differences between the G3PLT2 model and 3PL model in the medians of DIC and LPML are −74.07 and 21.75, respectively. However, when the time weight increases from 4 to 8, the medians of DIC for the 3PL model and G3PLT model are basically the same. This shows that the 3PL model is basically the same as the G3PLT model with time weights 4, 6, and 8, which is attributed to the fact that the G3PLT model reduces to the traditional 3PL model when the time weight increases from 4 to 8. Based on the results for LPML, we find that the power of LPML to distinguish between the true G3PLT4 (6, 8) model and the 3PL model is stronger than that of DIC, because the LPMLs of the two models differ greatly. For example, the difference between the G3PLT8 model and 3PL model in the median of LPML is 46.45.

In summary, the two Bayesian model assessment criteria can accurately identify the true model that generates data. In addition, the process of transformation of the G3PLT model into the traditional 3PL model is also reflected by the differences in DIC and LPML. Therefore, the two Bayesian model assessment criteria are effective and robust and can guide practice.

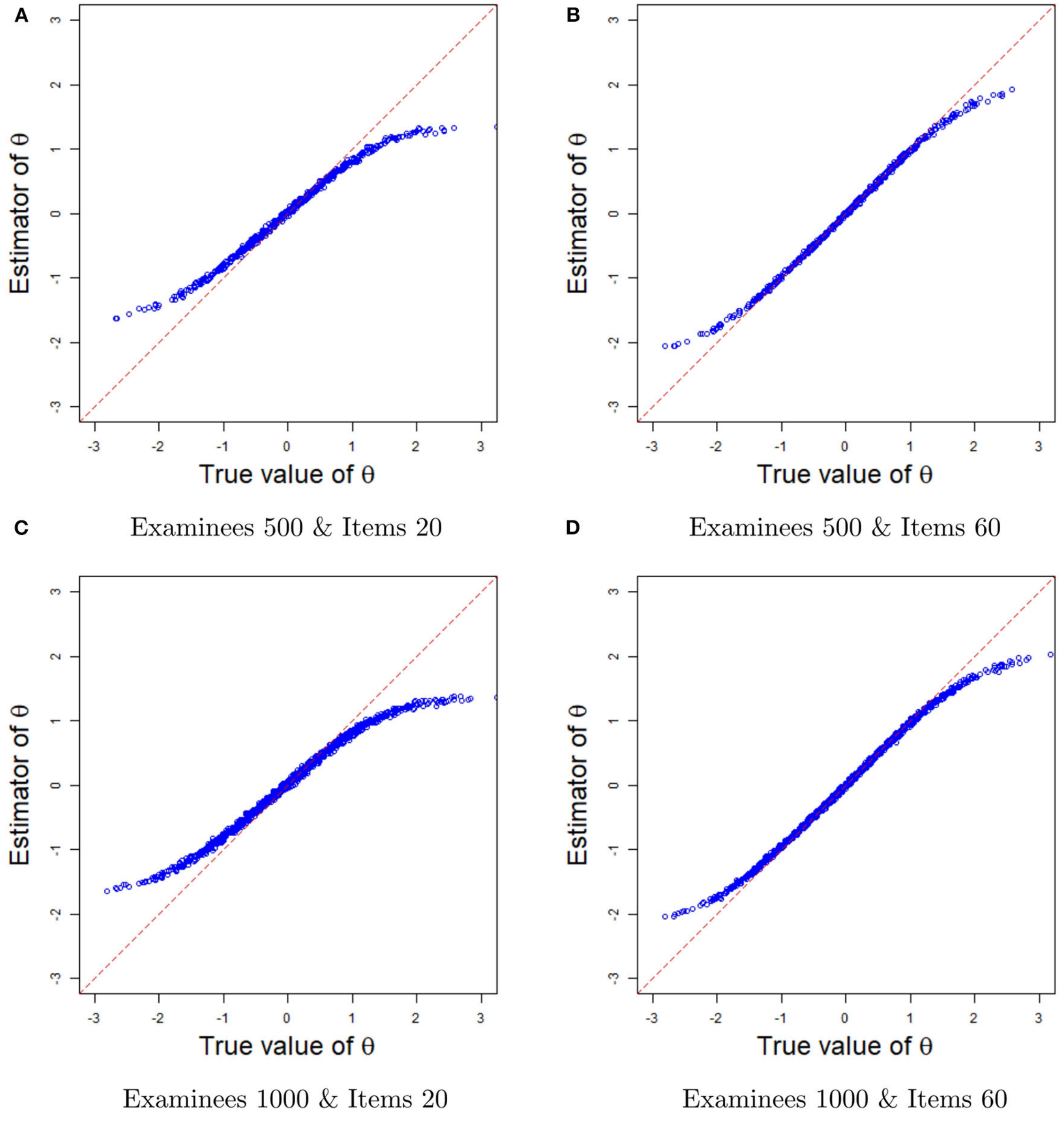


FIGURE 6 | The comparisons between ability estimates and true values in different sample sizes. **(A)** The comparisons between ability estimates and true values based on 500 examinees and 20 items. **(B)** The comparisons between ability estimates and true values based on 500 examinees and 60 items. **(C)** The comparisons between ability estimates and true values based on 1,000 examinees and 20 items. **(D)** The comparisons between ability estimates and true values based on 1,000 examinees and 60 items.

5. REAL DATA

5.1. Data Description

In this example, the 2015 computer-based Program for International Student Assessment (PISA) science data are used. From among the many countries that have participated in the computer-based assessment of the sciences, we choose the students from the USA as the object of analysis. Students with Not Reached (original code 6) or Not Response (original code 9) are removed in this study, where Not Reached and Not Response (omitted) are treated as missing data. The final 548 students are used to answer 16 items, and the corresponding response times are recorded. All 16 items are scored using a dichotomous scale. The 16 items are respectively CR083Q01S, CR083Q02S, CR083Q03S, CR083Q04S, DR442Q02C, DR442Q03C, DR442Q05C, DR442Q06C, CR442Q07S, CR245Q01S, CR245Q02S, CR101Q01S, CR101Q02S, CR101Q03S, CR101Q04S, and CR101Q05S. The frequency histogram of logarithmic response times and the correct rate for each item are shown in **Figure 7**.

5.2. Bayesian Model Assessment

To evaluate the impact of different time weights on the PISA data and to analyze the differences between the G3PLT model and

TABLE 4 | The results of Bayesian model assessment in Simulation 2.

	Fitted model			3PL	G3PLT0	G3PLT2	G3PLT4	G3PLT6	G3PLT8
True Model	G3PLT0	DIC	Q_1	25297.63	**25270.07**	–	–	–	–
			Median	25358.15	**25324.43**	–	–	–	–
			Q_3	25412.52	**25379.70**	–	–	–	–
			IQR	114.88	**109.63**	–	–	–	–
		LPML	Q_1	−13456.37	**−13251.64**	–	–	–	–
			Median	−13431.01	**−13231.77**	–	–	–	–
			Q_3	−13406.19	**−13218.86**	–	–	–	–
			IQR	50.17	**32.77**	–	–	–	–
	G3PLT2	DIC	Q_1	22742.44	–	**22677.65**	–	–	–
			Median	22851.46	–	**22777.38**	–	–	–
			Q_3	22953.34	–	**22890.79**	–	–	–
			IQR	210.89	–	**213.14**	–	–	–
		LPML	Q_1	−12274.46	–	**−12246.10**	–	–	–
			Median	−12243.68	–	**−12221.93**	–	–	–
			Q_3	−12221.43	–	**−12200.33**	–	–	–
			IQR	53.02	–	**45.76**	–	–	–
	G3PLT4	DIC	Q_1	20529.71	–	–	**20522.24**	–	–
			Median	20614.41	–	–	**20613.60**	–	–
			Q_3	20711.15	–	–	**20708.31**	–	–
			IQR	181.44	–	–	**186.06**	–	–
		LPML	Q_1	−11322.69	–	–	**−11263.87**	–	–
			Median	−11300.75	–	–	**−11239.84**	–	–
			Q_3	−11273.01	–	–	**−11219.60**	–	–
			IQR	49.67	–	–	**44.26**	–	–
	G3PLT6	DIC	Q_1	20210.35	–	–	–	**20206.43**	–
			Median	20295.34	–	–	–	**20294.27**	–
			Q_3	20386.09	–	–	–	**20384.67**	–
			IQR	175.73	–	–	–	**178.23**	–
		LPML	Q_1	−11102.84	–	–	–	**−11144.73**	–
			Median	−11079.08	–	–	–	**−11121.81**	–
			Q_3	−11052.10	–	–	–	**−11098.77**	–
			IQR	50.74	–	–	–	**45.96**	–
	G3PLT8	DIC	Q_1	20014.40	–	–	–	–	**20013.64**
			Median	20111.34	–	–	–	–	**20112.86**
			Q_3	20191.08	–	–	–	–	**20189.52**
			IQR	176.68	–	–	–	–	**175.87**
		LPML	Q_1	−11083.24	–	–	–	–	**−11032.39**
			Median	−11053.93	–	–	–	–	**−11007.48**
			Q_3	−11026.44	–	–	–	–	**−10981.35**
			IQR	56.79	–	–	–	–	**51.03**

Note that the 3PL denotes three parameter logistic model, and the G3PLTw denotes the general three parameter logistic model with time weight w, where w = 0, 2, 4, 6, 8.

the traditional 3PL model in fitting the data, both models are used to fit the data. G3PLT models with different time weights $W = 0, 1, 2, 3, 4, 5, 6, 7$, and 8 are considered. In the estimation procedure, the setting of the prior distributions is the same as in Simulation 1. In all of the Bayesian computations, we use 10,000 MCMC samples after a burn-in of 5,000 iterations for each model to compute all posterior estimates.

Table 5 shows the results for DIC and LPML under the 3PL model and the G3PLT model with different time weights. According to DIC and LPML, we find that the G3PLT model with time weight 6 is the best-fitting model, with DIC and LPML values of 8389.316 and -4196.672, respectively. The G3PLT model with time weight 0 is the worst-fitting model, with DIC and LPML values of 9708.940 and -4792.301, respectively. That the G3PLT model with time weight 0 is the worst fitting model can be attributed to the fact that the influence of the time effect on the correct-response probability is relatively weak for the PISA data. This is consistent with the the evaluation purpose of the PISA test, which is a nonselective and low-stakes test. Examinees lack motivation to answer each item carefully, and therefore the

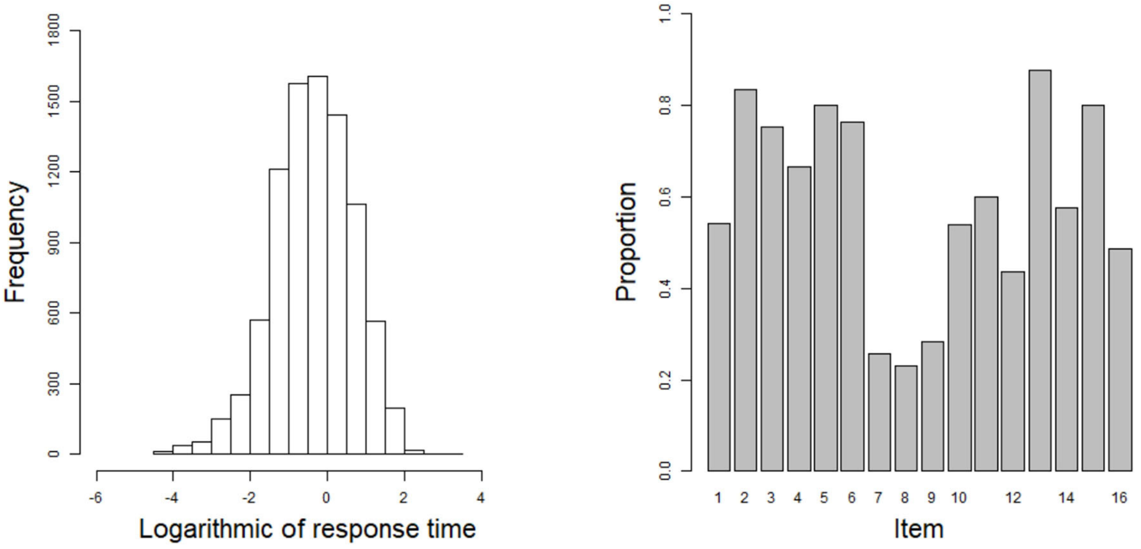


FIGURE 7 | The frequency histogram of logarithmic response times and the correct rate for each item in the real data.

time effect cannot be reflected. However, when the time weight of the G3PLT model increases from 5 to 8, the DIC and LPML values are basically the same as those in the case of the 3PL model. The model fitting results once again verify that our G3PLT model reduces to the traditional 3PL model when the time weight increases to a certain value. Next, we will analyze the PISA data based on the G3PLT model with time weight 6.

5.3. Analysis of Item Parameters

The estimated results for the item parameters are shown in **Table 6**. We can see that the expected a posteriori (EAP) estimates of the nine item discrimination parameters are greater than one. This indicates that these items can distinguish well between different abilities. In addition, the EAP estimates of the 11 difficulty parameters are less than zero, which indicates that 10 items are slightly easier than the other six. The three most difficult items are items 8 (DR442Q06C), 7 (DR442Q05C), and 9 (CR442Q07S). The EAP estimates of the difficulty parameters for these three items are, respectively 1.085, 0.900, and 0.839. The corresponding correct rates for the three items in **Figure 7** are 0.231, 0.257, and 0.285. The most difficult three items have the lowest correct rates, which is consistent with our intuition. The six EAP estimates of the guessing parameters are larger than 0.1. The three items that the examinees are most likely to answer correctly by guessing are items 11 (CR245Q02S), 12 (CR101Q01S), and 10 (CR245Q01S). The EAP estimates of the guessing parameters for these three items are respectively 0.132, 0.128, and 0.117. Among the 16 items, item 7 is the best design item owing to the fact that it has high discrimination and difficulty estimates, and the guessing parameter has the lowest estimate in all of the items. Next, we use the posterior standard deviation (SD) to evaluate the degree of deviation from the EAP estimate. The average SD of all discrimination parameters is about 0.005, the average SD of all difficulty parameters is about 0.010, and the average SD of all guessing parameters is about 0.001. We can see that the average SD values of the three parameters are very small, indicating that the estimated values fluctuate near the posterior mean.

5.4. Analysis of Personal Parameters

Next, we analyze the differences between the estimated abilities of examinees in the 3PL model and in the G3PLT model under the same response framework, together with the reasons for these differences. We consider four examinees with same response framework for the 16 items, (1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1). They are examinee 60, examinee 313, examinee 498, and examinee 210, and the corresponding response times for these examinees to answer the 16 items are 25.80, 29.36, 35.48, and 41.44 min. Under the framework of the 3PL model, the estimated abilities of the four examinees are the same, 1.45. However, taking into account the time factors for the four examinees, the estimated abilities are different according to the G3PLT model with time weight 6. The estimated abilities are 1.46, 1.42, 1.41, and 1.38, respectively. We find that under the same response framework, as the response times of the examinees increase from 25.80 to 41.44 min, the estimated abilities of the examinees show a decreasing trend. This indicates that examinees with short response times are more proficient in answering these items than examinees with long response times. Therefore, the ability of examinees with short response times to answer 15 items correctly should be higher than that of examinees with long times. This once again shows that our G3PLT model is reasonable. By incorporating the time effect into the IRT model, the interpretation of the latent construct essentially shifts: before we were measuring whether students could answer items correctly, now we are measuring whether students can answer items correctly and quickly.

TABLE 5 | The results of Bayesian model assessment in real data analysis.

	3PL	G3PLT0	G3PLT1	G3PLT2	G3PLT3
DIC	8392.374	9708.940	9217.431	8825.986	8561.295
LPML	−4197.832	−4792.301	−4565.769	−4395.082	−4275.351
	G3PLT4	**G3PLT5**	**G3PLT6**	**G3PLT7**	**G3PLT8**
DIC	8441.556	8398.835	8389.316	8391.581	8390.254
LPML	−4221.003	−4200.857	−4196.672	−4197.678	−4197.906

TABLE 6 | The results of item parameter estimation in real data analysis.

Parameter	EAP	SD	HPDI	Parameter	EAP	SD	HPDI
a_1	0.980	0.003	[0.873, 1.120]	a_9	1.199	0.003	[1.116, 1.312]
a_2	0.927	0.003	[0.824, 1.025]	a_{10}	0.821	0.004	[0.688, 0.946]
a_3	0.986	0.004	[0.857, 1.114]	a_{11}	1.059	0.006	[0.890, 1.200]
a_4	1.034	0.003	[0.928, 1.139]	a_{12}	1.004	0.007	[0.874, 1.195]
a_5	0.893	0.007	[0.723, 1.047]	a_{13}	1.037	0.006	[0.899, 1.198]
a_6	1.084	0.005	[0.965, 1.211]	a_{14}	1.011	0.005	[0.883, 1.137]
a_7	1.216	0.005	[1.062, 1.336]	a_{15}	0.986	0.006	[0.848, 1.190]
a_8	1.087	0.004	[0.974, 1.203]	a_{16}	0.803	0.002	[0.715, 0.917]
b_1	−0.065	0.009	[−0.240, 0.111]	b_9	0.839	0.007	[0.670, 0.995]
b_2	−1.405	0.014	[−1.617, −1.170]	b_{10}	0.065	0.020	[−0.186, 0.391]
b_3	−0.921	0.010	[−1.085, −0.693]	b_{11}	−0.147	0.016	[−0.374, 0.114]
b_4	−0.519	0.009	[−0.700, −0.321]	b_{12}	0.530	0.014	[0.324, 0.795]
b_5	−1.187	0.021	[−1.430, −0.849]	b_{13}	−1.608	0.015	[−1.846, −1.369]
b_6	−0.920	0.011	[−1.124, −0.730]	b_{14}	−0.083	0.012	[−0.280, 0.149]
b_7	0.900	0.007	[0.726, 1.069]	b_{15}	−1.145	0.016	[−1.429, −0.933]
b_8	1.085	0.007	[0.876, 1.236]	b_{16}	0.272	0.016	[0.062, 0.547]
c_1	0.065	0.000	[0.018, 0.120]	c_9	0.042	0.000	[0.016, 0.069]
c_2	0.098	0.001	[0.026, 0.189]	c_{10}	0.117	0.001	[0.029, 0.192]
c_3	0.079	0.001	[0.017, 0.156]	c_{11}	0.132	0.001	[0.056, 0.216]
c_4	0.079	0.001	[0.015, 0.143]	c_{12}	0.128	0.001	[0.071, 0.190]
c_5	0.107	0.002	[0.028, 0.199]	c_{13}	0.093	0.001	[0.028, 0.176]
c_6	0.092	0.001	[0.019, 0.158]	c_{14}	0.115	0.001	[0.034, 0.177]
c_7	0.026	0.000	[0.006, 0.045]	c_{15}	0.097	0.002	[0.022, 0.185]
c_8	0.032	0.000	[0.009, 0.056]	c_{16}	0.103	0.001	[0.035, 0.165]

6. CONCLUDING REMARKS

In this paper, we propose a new and flexible general three-parameter logistic model with response time (G3PLT), which is different from previous response time models, such as the hierarchical model framework proposed by van der Linden (2007), in which the response and the response time are considered in different measurement models, while a high-level model represents the correlation between latent ability and speed through a population distribution. However, our model integrates latent ability, time, and item difficulty into a item response model to comprehensively consider the impact on the probability of correct response. This approach to modeling is simpler and more intuitive. In addition, time weights are introduced in our model to investigate the influence of time intensity limited by different tests on the correct-response probability. When the time weight reaches 8, our model reduces to the traditional 3PL model, which indicates that the time factor has little influence on the correct-response probability. The examinees then answer each item correctly with the response probability given by the 3PL model.

However, the computational burden of the Bayesian algorithm becomes excessive when large numbers of examinees or items are considered or a large MCMC sample size is used. Therefore, it is desirable to develop a standalone R package associated with C++ or Fortran software for more extensive large-scale assessment programs.

Other issues should be investigated in the future. First of these is whether the G3PLT model can be combined with a multilevel structure model to analyze the influence of covariates on the latent ability at different levels, for example, to explore the influence of the time effect, gender, and socioeconomic status on latent ability. Second, although we have found that for different examinees with the same response framework, the ability estimates from the 3PL model is the same, those from the G3PLT model differ greatly. Examinees who take less time should be more proficient in answering items, and their ability should be higher than that of examinees who take longer. "Proficiency" is a latent skill that is not the same as latent ability. Whether we can connect proficiency and latent ability through a multidimensional 3PLT model to analyze their relationship is also an important topic for our future research. Third, our new model can also be used to detect various abnormal response behaviors, such as rapid guessing and cheating, with the aim of eliminating deviations in ability estimates caused by such behaviors.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://www.oecd.org/pisa/data/.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant no. 11571069) and Foundation for basic and applied basic research of Guangdong Province (Grant no. 2019A1515110448).

REFERENCES

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibb sampling. *J. Educ. Stat.* 17, 251–269. doi: 10.3102/10769986017003251

Baker, F. B., and Kim, S.-H. (2004). *Item Response Theory: Parameter Estimation Techniques, 2nd Edn.* New York, NY: Marcel Dekker.

Béguin, A. A., and Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika* 66, 541–561. doi: 10.1007/BF02296195

Birnbaum, A. (1968). "Some latent trait models and their use in inferring an examinee's ability," in *Statistical Theories of Mental Test Scores*, eds F. M. Lord and M. R. Novick (Reading, MA:Addison-Wesley), 397–479.

Bock, R. D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* 46, 443–459. doi: 10.1007/BF02293801

Bolsinova, M., and Tijmstra, J. (2018). Improving precision of ability estimation: getting more from response times. *Br. J. Math. Stat. Psychol.* 71, 13–38. doi: 10.1111/bmsp.12104

Bridgeman, B., and Cline, F. (2004). Effects of differentially time-consuming tests on computer adaptive test scores. *J. Educ. Measure.* 41, 137–148. doi: 10.1111/j.1745-3984.2004.tb01111.x

Brooks, S., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* 7, 434–455. doi: 10.1080/10618600.1998.10474787

Chang, H. (2004). "Computerized testing, E-rater, and generic algorithm: Psychometrics to support emerging technologies," in *Invited Symposium, 28th International Congress of Psychology* (Beijing).

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York, NY: Springer.

Chib, S., and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *Am. Stat.* 49, 327–335. doi: 10.1080/00031305.1995.10476177

Choe, E. M., Kern, J. L., and Chang, H-H. (2018). Optimizing the use of response times for item selection in computerized adaptive testing. *J. Educ. Behav. Stat.* 43, 135–158. doi: 10.3102/10769986177 23642

De Boeck, P., and Jeon, M. (2019). An overview of models for response times and processes in cognitive tests? *Front. Psychol.* 10:102. doi: 10.3389/fpsyg.2019.00102

Embretson, S. E, and Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Earlbaum Associates.

Fox, J.-P., Klein Entink, R. H., and van der Linden, W. J. (2007). Modeling of responses and response times with the package CIRT. *J. Stat. Softw.* 20, 1–14. doi: 10.18637/jss.v020.i07

Fox, J.-P., and Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivar. Behav. Res.* 51, 540–553. doi: 10.1080/00273171.2016.1171128

Geisser, S., and Eddy, W. F. (1979). A predictive approach to model selection. *J. Am. Stat. Assoc.* 74, 153–160. doi: 10.1080/01621459.1979.10481632

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). "Model determination using predictive distributions with implementation via sampling based methods (with discussion)," in *Bayesian Statistics*, Vol. 4, eds J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Oxford, UK: Oxford University Press), 147–167.

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–472. doi: 10.1214/ss/1177011136

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Han, K. T. (2012). *Fixing the c Parameter in the Three-Parameter Logistic Model.* Practical Assessment, Research and Evaluation, 17. Available online at: http://pareonline.net/getvn.asp?v=17andn=1

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97–109. doi: 10.1093/biomet/57.1.97

Hung, L.-F., and Wang, W.-C. (2012). The generalized multilevel facets model for longitudinal data. *J. Educ. Behav. Stat.* 37, 231–255. doi: 10.3102/1076998611402503

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. New York, NY: Springer.

Klein Entink, R. H., Fox, J.-P., and van der Linden, W. J. (2009a). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika* 74, 21–48. doi: 10.1007/s11336-008-9075-y

Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., and Fox, J.-P. (2009b). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychol. Methods* 14, 54–75. doi: 10.1037/a0014877

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: AddisonWesley.

Lu, J., Wang, C., Zhang, J., W. and Tao, J. (2019). A mixture model for responses and response times with a higher-order ability structure to detect rapid guessing behaviour. *Br. J. Math. Stat. Psychol.* doi: 10.1111/bmsp.12175

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., and Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *J. Educ. Behav. Stat.* 39, 426–451. doi: 10.3102/1076998614559412

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1092. doi: 10.1063/1.1699114

Patz, R. J., and Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *J. Educ. Behav. Stat.* 24, 146–178. doi: 10.3102/10769986024002146

Patz, R. J., and Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. *J. Educ. Behav. Stat.* 24, 342–366. doi: 10.3102/10769986024004342

Pokropek, A. (2016). Grade of membership response time model for detecting guessing behaviors. *J. Educ. Behav. Stat.* 41, 300–325. doi: 10.3102/1076998616636618

Qian, H., Staniewska, D., Reckase, M., and Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educ. Measure.* 35, 38–47. doi: 10.1111/emip.12102

Roskam, E.E. (1987). "Toward a psychometric theory of intelligence," in *Progress in Mathematical Psychology*, eds E. E. Roskam and R. Suck (Amsterdam: North-Holland), 151–174.

Roskam, E. E. (1997). "Models for speed and time-limit tests," in *Handbook of Modern Item Response Theory*, eds W. J. van der Linden and R. K. Hambleton (New York, NY: Springer), 187–208.

Rupp, A. A., Dey, D. K., and Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether to when: applications of Bayesian methodology to modeling. *Struct. Equat. Model.* 11, 424–451. doi: 10.1207/s15328007sem1103_7

Schnipke, D. L., and Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: a new method of measuring speededness. *J. Educ. Measure.* 34, 213–232. doi: 10.1111/j.1745-3984.1997.tb00516.x

Sinharay, S., and Johnson, M. S. (2019). The use of item scores and response times to detect examinees who may have benefited from item preknowledge. *Br. J. Math. Stat. Psychol.* doi: 10.1111/bmsp.12187. [Epub ahead of print].

Skorupski, W. P., and Wainer, H. (2017). "The case for Bayesian methods when investigating test fraud," in *Handbook of Detecting Cheating on Tests*, eds G. J. Cizek and J. A. Wollack (London: Routledge), 214–231.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and vander Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B* 64, 583–639. doi: 10.1111/1467-9868.00353

Swaminathan, H., and Gifford, J. A. (1979). *Estimation of Parameters in the Three-Parameter Latent Trait Model* (Report No. 90). Amherst: Laboratory of Psychometric and Evaluation Research; School of Education; University of Massachusetts.

Thissen, D. (1983). "Timed teting: An approach using item response theory," in *New Horizons in Testing*, ed D. J. Weiss (New York, NY: Academic Press), 179–203.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Stat.* 22, 1701–1762. doi: 10.1214/aos/1176325750

van der Linden, W. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* 72, 297–308. doi: 10.1007/s11336-006-1478-z

van der Linden, W. J. (2009). A bivariate lognormal response-time model for the detection of collusion between test takers. *J. Educ. Behav. Stat.* 34, 378–394. doi: 10.3102/1076998609332107

van der Linden, W. J., Breithaupt, K., Chuah, S. C., and Zhang, Y. (2007). Detecting differential speededness in multistage testing. *J. Educ. Measure.* 44, 117–130. doi: 10.1111/j.1745-3984.2007.00030.x

van der Linden, W. J., and Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika* 75, 120–139. doi: 10.1007/s11336-009-9129-9

van der Linden, W. J., and Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika* 73, 365–384. doi: 10.1007/s11336-007-9046-8

van der Linden, W. J., and Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory*. New York, NY: Springer-Verlag.

Verhelst, N. D., Verstralen, H. H. F. M., and Jansen, M. G. (1997). "A logistic model for time-limit tests," in *Handbook of Modern Item Response Theory*, eds W. J. van der Linden and R. K. Hambleton (New York, NY: Springer), 169–185.

von Davier, M. (2009). Is there need for the 3PL model? Guess what? *Measurement* 7, 110–114. doi: 10.1080/15366360903117079

Wang, C., Fan, Z., Chang, H.-H., and Douglas, J. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *J. Educ. Behav. Stat.* 38, 381–417. doi: 10.3102/1076998612461831

Wang, C., and Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *Br. J. Math. Stat. Psychol.* 68, 456–477. doi: 10.1111/bmsp.12054

Wang, C., Xu, G., and Shang, Z. (2018a). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika* 83, 223–254. doi: 10.1007/s11336-016-9525-x

Wang, C., Xu, G., Shang, Z., and Kuncel, N. (2018b). Detecting aberrant behavior and item preknowledge: a comparison of mixture modeling method and residual method. *J. Educ. Behav. Stat.* 43, 469–501. doi: 10.3102/1076998618767123

Wang, T., and Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Appl. Psychol. Measure.* 29, 323–339. doi: 10.1177/0146621605275984

Zhang, Z., Hamagami, F., Wang, L., Grimm, K. J., and Nesselroade, J. R. (2007). Bayesian analysis of longitudinal data using growth curve models. *Int. J. Behav. Dev.* 31, 374–383. doi: 10.1177/0165025407077764

Zhu, Z. M., Wang, C., and Tao, J. (2018). A two-parameter logistic extension model: an efficient variant of the three-parameter logistic model. *Appl. Psychol. Measure.* 21, 1–15. doi: 10.1177/0146621618800273

Zopluoglu, C. (2019). Detecting examinees with item preknowledge in large-scale testing using extreme gradient boosting (XGBoost). *Educ. Psychol. Measure.* 79, 931–961. doi: 10.1177/0013164419839439

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 06 August 2020
doi: 10.3389/fpsyg.2020.01987

Detecting Conditional Dependence Using Flexible Bayesian Latent Class Analysis

Jaehoon Lee[1], *Kwanghee Jung*[1]* *and Jungkyu Park*[2]*

[1] Department of Educational Psychology and Leadership, Texas Tech University, Lubbock, TX, United States, [2] Department of Psychology, Kyungpook National University, Daegu, South Korea

A fundamental assumption underlying latent class analysis (LCA) is that class indicators are conditionally independent of each other, given latent class membership. Bayesian LCA enables researchers to detect and accommodate violations of this assumption by estimating any number of correlations among indicators with proper prior distributions. However, little is known about how the choice of prior may affect the performance of Bayesian LCA. This article presents a Monte Carlo simulation study that investigates (1) the utility of priors in a range of prior variances (i.e., strongly non-informative to strongly informative priors) in terms of Type I error and power for detecting conditional dependence and (2) the influence of imposing approximate independence on model fit of Bayesian LCA. Simulation results favored the use of a weakly informative prior with large variance–model fit (posterior predictive p–value) was always satisfactory when the class indicators were either independent or dependent. Based on the current findings and the additional literature, this article offers methodological guidelines and suggestions for applied researchers.

Keywords: conditional dependence, Bayesian latent class analysis, approximate independence, prior variance, model fit

OPEN ACCESS

Edited by:
Pietro Cipresso,
Italian Auxological Institute (IRCCS),
Italy

Reviewed by:
Carlos Alberto De Bragança
Pereira,
University of São Paulo, Brazil
Philip L. H. Yu,
The University of Hong Kong,
Hong Kong
Peida Zhan,
Zhejiang Normal University, China

****Correspondence:***
Kwanghee Jung
kwanghee.jung@ttu.edu
Jungkyu Park
jkp@knu.ac.kr

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: *03 July 2019*
Accepted: *17 July 2020*
Published: *06 August 2020*

Citation:
Lee J, Jung K and Park J (2020)
Detecting Conditional Dependence
Using Flexible Bayesian Latent Class
Analysis. Front. Psychol. 11:1987.
doi: 10.3389/fpsyg.2020.01987

INTRODUCTION

Latent class analysis (LCA; Lazarsfeld and Henry, 1968) is a probability model–based tool that analyzes categorically scored data by introducing a latent variable. As the name suggests, the latent variable (usually) consists of a small number of levels, called "latent classes" that characterize the categories of a theoretical construct. The primary aim of LCA is to identify class members that are homogenous within the same class but distinct between different classes in terms of responses to a set of observed variables (i.e., latent class indicators). Once identified, the latent classes are compared with each other for auxiliary variables such as covariates and distal outcomes presumed to be antecedents or consequences of the classification (Asparouhov and Muthén, 2014; Vermunt, 2010).

LCA has been extended to accommodate various types of observed data–for example, latent profile analysis with continuous indicators, multilevel mixture models for clustered data, growth mixture models and latent transition models for longitudinal observations, and survival mixture models with time–censored indicators. Owing to such flexibility in data distribution that can be modeled, LCA and other mixture approaches recently have been increasingly adopted in a variety of disciplines, including cognitive diagnostic testing (Rupp et al., 2010) health and medicine (Schlattmann, 2010) genetics (McLachlan et al., 2004) machine learning (Yang and Ahuja, 2001) and economics (Geweke and Amisano, 2011). More recently, methodological advances have made it feasible to estimate LCA models within the Bayesian framework; see Li et al. (2018) and Asparouhov and Muthén (2010).

Assumption of Conditional Independence in LCA

Another key advantage of using LCA is that it does not require the rigid assumptions of traditional classification methods (Muthén, 2002; Muthén and Shedden, 1999; Magidson and Vermunt, 2004). Still, LCA assumes that class indicators are conditionally independent of each other, given class membership–i.e., class indicators are uncorrelated within each class. This implies that the associations among the indicators are accounted for only by the latent classes and there are no other latent variables influencing the indicators. A violation of this local independence assumption (i.e., *conditional dependence*) can lead to severe bias in estimating LCA parameters that include classification error, class probabilities, and posterior classification probabilities (Vacek, 1985; Torrance-Rynard and Walter, 1998; Albert and Dodd, 2004). An additional drawback that is common with conditional dependence is model misfit. That is, unmodeled dependence among indicators can induce poor model fit and incorrect values of information criteria [e.g., Akaike information criterion (AIC), Bayesian information criterion (BIC)], resulting in spurious latent classes (usually with an overestimated number of classes).

Current Methods for Handling Conditional Dependence

When conditional dependence is suspected, a viable option is to model the dependence "directly" (Uebersax, 1999; Hagenaars, 1988). The correlation between each pair of indicators is freely estimated; and a significant improvement in model fit supports relaxing the conditional independence assumption on locally dependent indicator pairs. Still, there is a caveat to this approach. Freeing the constrained parameters (correlations) often makes the model non–identifiable or results in highly unstable parameter estimates, because the dependencies captured by the latent classes are difficult to separate from nuisance local dependencies. An alternative option to deal with conditional dependence is employing a latent variable(s). *Factor* mixture modeling, for example, models conditional dependence by allowing for the indicators to be loaded on a continuous latent variable in addition to their loading on the discrete latent variable representing/forming classes (Lubke and Muthén, 2007). This approach is yet limited because in many applications indicators do not necessarily represent an interrelated dimension(s) of a generic construct and thus models can suffer from estimation challenges. Other applications of modeling conditional dependence can be found in Qu et al. (1996), Wang and Wilson (2005), Im (2017), Hansen et al. (2016), and Zhan et al. (2018).

Beyond handling conditional dependence, researchers may want to monitor and detect the sources of conditional dependence. Magidson and Vermunt (2004) proposed bivariate residual (BVR)–a high value of BVR for a pair of indicators reveals model misfit due to (residual) conditional dependence between the indicators. A drawback of BVR is that its distribution is unknown. Oberski et al. (2013) recommended using BVR with a bootstrapping procedure which approximates a chi–square distribution. They also showed that Lagrange multiplier test, also called modification index, performs well in identifying the sources of conditional dependence, showing adequate power and controlled Type I error.

New Method: Assumption of Conditional Independence in Bayesian LCA

In Bayesian statistics, the researcher's belief about the value of a parameter is formulated into a distribution, which is called *prior distribution* (often simply called *prior*). Data also inform about the parameter value, yielding a conditional distribution of the data given the parameter, which is called *likelihood*. The likelihood modifies the prior distribution into a *posterior distribution* (often simply called *posterior*). Finally, a parameter estimate is inferred through a sampling of 'plausible' values from the posterior.

A prior having small variance (i.e., a narrow prior distribution) represents the researcher's small uncertainty about the parameter value. This small–variance ("informative") prior makes relatively more contribution to constructing the posterior than does the likelihood. On the other hand, a prior having a large variance (i.e., a wide prior distribution) represents large uncertainty about the parameter value, and the large–variance ("non–informative") prior yields relatively less influence on the formation of the posterior than does the likelihood (Muthén and Asparouhov, 2012; MacCallum et al., 2012). Thus, the model would fit the data very closely if non–informative priors were specified for all model parameters, but the parameter estimates might be scientifically untenable (Gelman, 2002).

Recent methodological advances have made it possible to incorporate latent variable modeling in the framework of Bayesian statistics (O'sullivan, 2013; Silva and Ghahramani, 2009). Bayesian estimation in latent variable modeling is advocated particularly for avoiding the likely problem of a non–identifiable model or an improper solution in maximum–likelihood estimation. For instance, the researcher may replace the parameter specification of "exact zeros" with "approximate zeros" by imposing informative priors on the parameters that would have been fixed to 0 for hypothesis testing or scale setting in ML estimation (Muthén and Asparouhov, 2012). In Muthén and Asparouhov (2012) illustration of Bayesian structural equation modeling (BSEM), priors for factor loadings are specified to be normal with zero mean and infinity variance (i.e., non–informative priors), while priors for cross–loadings are specified to follow a normal distribution having zero mean and 0.01 variance (i.e., informative priors)–95% of the cross–loading values are between −0.2 and 0.2 in the prior distribution. A few real–data applications and Monte Carlo simulations showed that the "approximate zeros" strategy performs well for both measurement and structural models that involve cross–loadings, residual correlations, or latent regressions with respect to model fit testing, coverage for key parameters, and power to detect model misspecifications (Muthén and Asparouhov, 2012).

The idea of this "approximate zeros" approach is applicable for detecting and accommodating violations of the conditional independence assumption in LCA. Rather than fixing to 0, the researcher may freely estimate all or some tetrachoric (for binary class indicators) or polychoric (for polytomous class indicators) correlations among the indicators using informative priors with zero mean and small variance (i.e., *approximate independence*; Asparouhov and Muthén, 2011). Asparouhov and Muthén (2010, 2011) suggested Bayesian LCA that relaxes the conditional independence assumption to an assumption of approximate independence. This flexible Bayesian LCA can avoid a false class formation that is often caused by ignoring conditional dependence, or equivalently, neglecting (i.e., fixing) nonzero correlations among indicators.

To illustrate this method, a model was fitted to the data from Midlife in the United States (MIDUS), 2004—-2006, a national survey of 4,963 Americans aged 35 to 86. The class indicators were 10 binary items (*yes*/*no*) asking the main reasons for discrimination respondents experienced: age, gender, race, ethnicity or nationality, religion, height or weight, some other aspect of appearance, physical disability, sexual orientation, and some other reason (B1SP3A–B1SP3J). The model specified two classes and approximate independence between each pair of the indicators by imposing a prior on the tetrachoric correlation. The model fit was adequate, and the classification quality was excellent with an entropy value of 1 (35% of the sample in the first class and 65% in the second class). More important, the correlations among the indicators deviated from zero ranging from –0.28 to 0.42 and they were higher in the first class (M = 0.19, SD = 0.14) than in the second class (M = –0.05, SD = 0.15). These results suggest that the model of approximate independence could detect and properly model the fair amounts of conditional dependence. The M*plus* input for this analysis is shown in **Appendix** (B1SP3A-B1SP3J are renamed to U1-U10 for illustrative purpose).

More research is warranted to investigate the performance of this method as a tool for detecting conditional dependence and acknowledge potential consequences of incorporating approximate independence into LCA models. Another scientific inquiry is evaluating model fit of Bayesian LCA under the assumption of approximate independence. Asparouhov and Muthén (2011) argued that posterior predictive checking is needed to evaluate the fit of flexible Bayesian LCA models. In Bayesian statistics, one possible measure of model-data fit is *posterior predictive p-value* (PPP). The process of deriving PPP is quite technical (Gelman et al., 1996) but the key application is simple-a PPP around 0.50 suggests an excellent fit. Although there is no theoretical cutoff of alarming poor fit, Muthén and Asparouhov (2012) suggested that a PPP around 0.10, 0.05, or 0.01 would indicate a significantly ill-fitting model.

Purpose of Study

Although the literature has suggested flexible Bayesian LCA as an alternative approach to detect and accommodate conditional dependence between indicators, there is no specific guideline about to what degree a prior should be informative to properly model the conditional dependence (Ulbricht et al., 2018). To the authors' knowledge, only a strongly informative prior was empirically studied under limited conditions (Asparouhov and Muthén, 2011). Thus, this article presents a simulation study of which aim is to examine (1) the utility of priors in flexible Bayesian LCA with a wide range of variances (i.e., strongly non-informative to strongly informative priors) in terms of Type I error and power for detecting conditional dependence; and (2) the influence of imposing approximate independence on fit (PPP) of Bayesian LCA. The current investigation focuses on the simple case of binary indicators measured in cross-sectional research-that is, flexible Bayesian LCA as a beginning attempt to understand the performance of LCA under the assumption of approximate independence.

MATERIALS AND METHODS

This section illustrates the model specifications of LCA; and describes the simulated conditions and Monte Carlo procedure utilized to examine the performance of Bayesian LCA under the approximate independence assumption.

Bayesian LCA Models

Let Y be a full response vector for a set of J indicators, where $j = 1,\ldots, J$; and let X be a discrete latent variable consisting of M latent classes. A particular class is denoted by m. The probability of a particular response pattern on J indicators can be defined as follows:

$$P(Y) = \sum_{m=1}^{M} P(X = m) f(Y|X = m) \tag{1}$$

Let y_j denote a response on indicator j; then conditional density for indicator j $(f(y_j|X = m))$ is statistically independent of each other, given latent class membership m. Therefore, the conditional independence assumption can be represented as

$$P(Y) = \sum_{m=1}^{M} P(X = m) \prod_{j=1}^{J} f(y_j|X = m) \tag{2}$$

The conditional density $f(y_j|X = m)$depends on the assumed distribution of responses. Suppose the response vector $Y = (y_1, y_2, \ldots, y_J)^T$consists of J binary variables. Then the mth latent class density is given by $f(y_j|X = m) = \rho_{mj}^{y_j}(1 - \rho_{mj})^{1-y_j}$, where ρ_{mj} denotes the probability of endorsing item j for given latent class membership m $(f(y_j = 1|X = m))$.

This standard LCA model can be formulated in terms of a multivariate probit model with a continuous latent response variable y_j^* for indicator j:

$$y_j^*|X \sim N(\mu_{jm}, 1), \tag{3}$$

where$y_j = 0$, then$y_j^* < 0$; thus $\rho_{mj} = 0$, then $P(y_j^* < 0|X = m) = \Phi(\mu_{jm})$. A multivariate form of this model can be expressed as

$$y_j^*|X \sim N(\mu_{jm}, 1), \tag{4}$$

where $Y^* = (y_1^*, y_2^*, \ldots, y_J^*)^T$, $\mu_m = (\mu_{1m}, \mu_{2m}, \ldots, \mu_{Jm})^T$ and I is a correlation matrix that all the off-diagonal elements are equal to 0. The conditional dependence model with correlated indicators is to replace Eq. 4 with

$$Y^*|X \sim N(\mu_m, \Sigma_m), \tag{5}$$

where Σ_m represents an unrestricted correlation matrix. The off-diagonal elements in Σ_m are called a tetrachoric correlation for binary indicators that can be varied across classes. When indicators have more than two categories, the correlation is called a polychoric correlation. These correlations can be estimated by maximizing the log-likelihood of the multivariate normal distribution. The parameters of the conditional dependence LCA model can be estimated using the Markov chain Monte Carlo (MCMC) algorithm. The details about MCMC estimation with prior information on each parameter are provided by Asparouhov and Muthén (2011).

Population Model (Data Generation)

Two classes (Class 1 and Class 2) of equal size were simulated in the population model with 10 binary class indicators. The simulation conditions examined by Asparouhov and Muthén (2011) were included in the current study for the sake of comparability, along with some additional conditions. To vary the level of conditional dependence among the indicators, data were generated such that the tetrachoric correlation matrix in Class 1 had all zero values (no dependence); or three nonzero values $\rho_{1,12} = \rho_{1,39} = \rho_{1,57} = 0.20$, 0.50, or 0.80 (small, medium, and large, respectively) and zero values for all other elements. Here, $\rho_{m,jk}$ is the correlation between indicator j and indicator k in class m. The tetrachoric correlation matrix in Class 2 had all zero values; or all zero values except for $\rho_{2,46} = 0.20$, 0.50, or 0.80. The size of the nonzero correlations, if present, were matched to be equal between the two classes. In addition, the indicator thresholds $\mu_{m,k}$ were set to be equal within each class but opposite in sign between the two classes-$\mu_{1,k} = 1.00$ and $\mu_{2,k} = -1.00$, which yields a reasonable class separation at two standard deviations (Lubke and Muthén, 2007). Sample size was also simulated as $N = 50$, 75, 100, 500 in increments of 25, and $N = 1{,}000$.

Analysis Model

The analysis model specified a weakly informative prior for the indicator thresholds $\mu_{m,k}$ ($\sim N$ [0, 5]) and for the class threshold q_m ($\sim$ Dirichlet distribution D [10, 10]). They are the default priors in M*plus* 8 and not the focus of the current study. Another default prior, inverse Wishart distribution IW (I, f), where I is identity matrix and f is degrees of freedom (df), was specified for the tetrachoric correlations among the class indicators. To vary the variance of the priors, f was set to be 11, 52, 108, 408, or 4,000. In this way, the correlations were modeled as following a symmetric beta distribution on the interval [−1, 1] with mean zero and variance of 0.33, 0.02, 0.01, 0.003, or 0.00003-consequently, 95% confidence limits of the correlations approximately equal to ±1.13, ±0.30, ±0.20, ±0.10, or ±0.01, respectively (Barnard et al., 2000; Gill, 2008; Asparouhov and Muthén, 2011). The prior having the largest variance (0.33) was strongly non-informative because it indeed corresponds to a uniform distribution on the interval [−1, 1]. The prior having the second largest variance (0.02) was considered weakly non-informative, and the other three priors with relatively small variance were considered strongly informative (0.00003), informative (0.003), and weakly informative (0.01).

Monte Carlo Specifications

Two hundred samples were drawn from each of 400 simulation conditions (20 sample sizes × 4 levels of correlations among the indicators × 5 prior variances), yielding a total of 80,000 replications. In Bayesian estimation, two independent Markov chains created approximations to the posterior distributions, with a maximum of 50,000 iterations for each chain. The first half of each chain was discarded as being part of the burn-in phase. Convergence was assessed for each parameter using the Gelman-Rubin criterion (Gelman and Rubin, 1992) the convergence rate was 100% in all simulated conditions. The medians of the posterior distributions were reported as Bayesian point estimates, which is the default setting in M*plus* 8. Model fit (PPP) was calculated based on the chi-square discrepancy function (Scheines et al., 1999; Asparouhov and Muthén, 2010).

RESULTS

This section presents the results of the simulation study regarding (1) the effects of the condition factors (sample size, correlation size, prior variance) on Type I error and power of flexible Bayesian LCA for detecting conditional dependence; (2) the effects of the condition factors on fit (PPP) of the model; and (3) bias in Bayesian estimates of indicator correlations.

Type I Error and Power for Testing Conditional Dependence

Figure 1 depicts how often (zero or nonzero) correlations were detected to be significantly different from zero at the nominal alpha level of 0.05-"% significant." It should be noted that this figure summarizes the outcomes on a particular pair of indicators (the first and second indicators at Class 1), but the results are almost identical to those from other indicator pairs. Type I error, false positive on a true zero correlation, was well controlled in the flexible Bayesian LCA model. In fact, the average % significant, represented by the blue lines in **Figure 1**, was below 5% for any prior variance and for any sample size (see the bottom panel).

Power for detecting a nonzero correlation increased with the use of less informative priors (see the top three panels). As would be anticipated, a true nonzero correlation ($\rho_{1,12} = 0.20$, 0.50, or 0.80) was seldom estimated to be significantly different from zero if the strongly informative prior was imposed on this parameter. Also, power for detecting a true small correlation ($\rho_{1,12} = 0.20$) was always less than satisfactory (i.e., <80%) regardless of prior variance (see the bottom second panel). When the correlation was moderate ($\rho_{1,12} = 0.50$), a (either weakly or strongly) non-informative prior and a sample size of at least 500 were required to yield acceptable power. For the

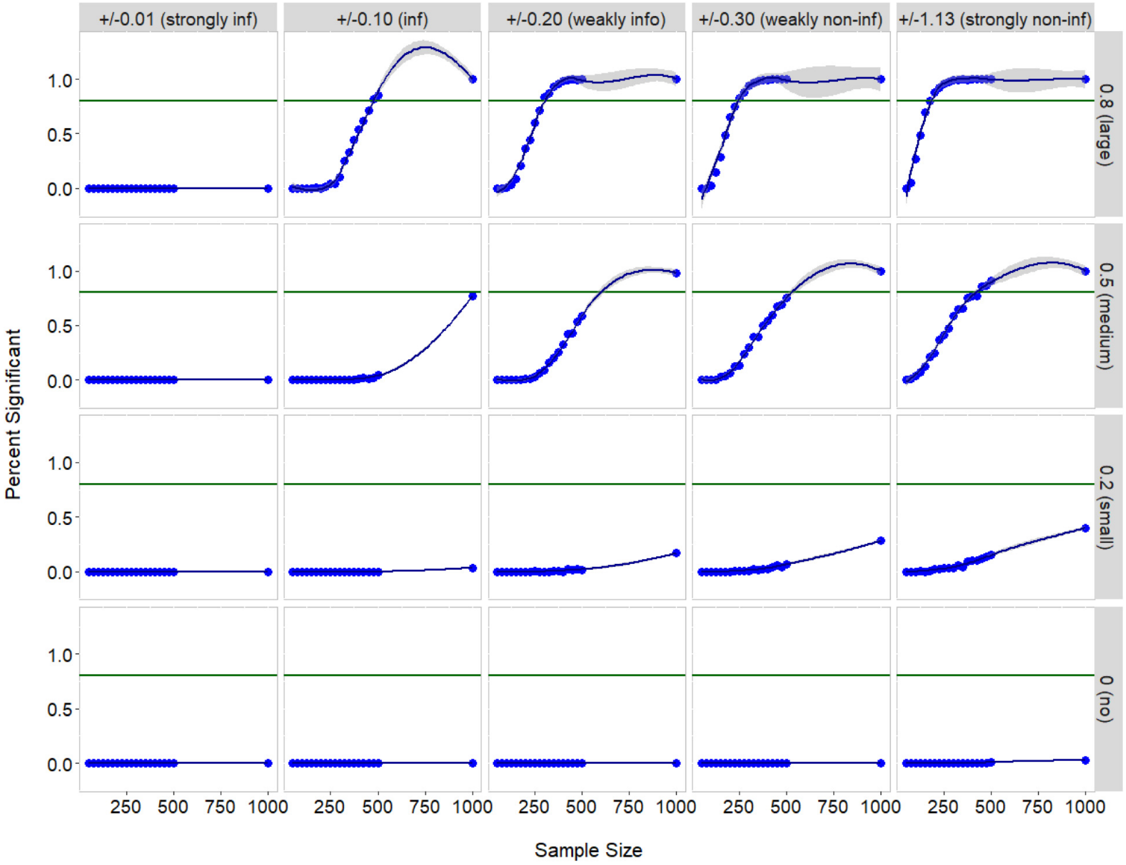


FIGURE 1 | Accuracy of testing zero and nonzero correlations by correlation size, sample size, and prior variance. The blue dots represent average % significant and the blue line represents smoothed conditional means of % significant. The green line indicates % significant of 0.80, a satisfactory power to detect a true nonzero correlation (top three panels). Note that the conditional means greater than 1 are not plausible values and merely indicate prediction artifact.

prior that allows for 95% of estimates to be within +0.10 and –0.10 (i.e., an informative prior), power was adequate (≥80%) when and only when a true large correlation ($\rho_{1,12} = 0.80$) was estimated from a sample greater than $N = 500$. For the priors having larger variance (weakly informative, non-informative, and strongly non-informative priors), power for detecting a true large correlation ($\rho_{1,12} = 0.80$) was satisfactory if the sample size was at least 300.

Model Fit of Flexible Bayesian LCA

Analysis of variance was conducted to identify which condition factors considerably influenced PPP. The estimated effect sizes (η^2) of the three condition factors and their interactions are provided in **Table 1**. Sample size had a negligible effect on PPP ($\eta^2 = 0.020$), which is similar to the findings for Bayesian confirmatory factor analysis in Muthén and Asparouhov (2012). PPP was largely influenced by the size of correlations (i.e., the magnitude of conditional dependence) among the indicators ($\eta^2 = 0.388$), as well as by the choice of prior variance for these parameters ($\eta^2 = 0.104$). Small to moderate effects were observed for the interactions between the condition factors ($\eta^2 = 0.014$–0.057).

TABLE 1 | Effects of simulation condition factors on posterior predictive *P*-value.

Condition factor	η^2
Sample size (N)	0.020
Correlation size (C)	0.388
Prior variance (P)	0.104
N × C	0.042
N × P	0.022
C × P	0.057
N × C × P	0.014

Figure 2 describes the (large) effects of correlation size and prior variance in a series of plots, in which the *y*-axis represents PPP and the *x*-axis represents sample size. Recall that the tetrachoric correlations between the indicators were simulated to be 0, 0.20, 0.50, or 0.80-no, small, medium, and large, respectively (in **Figure 2**, from the bottom to top panels). Also, recall that the prior (symmetric beta) distributions for these parameters were specified to have mean zero and variance of 0.00003, 0.003, 0.01, 0.02, or 0.33-strongly informative, informative, weakly informative, weakly non-informative, and

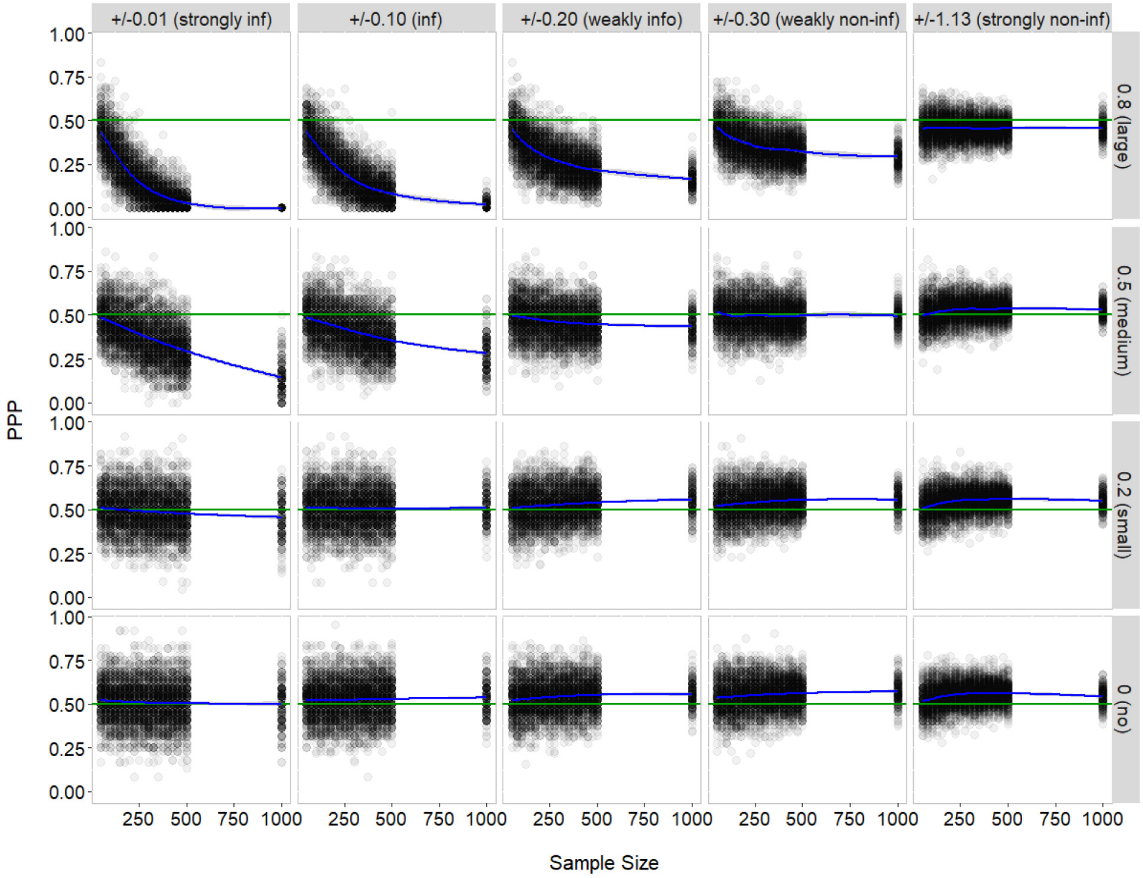


FIGURE 2 | Posterior predictive *P*-value by sample size, correlation size, and prior variance. The gray dots represent estimated PPP and the blue line represents smoothed conditional means of PPP. The green line indicates PPP of 0.50, an excellent fit.

strongly non-informative, respectively (in **Figure 2**, the panels from left to right). The average PPP, represented by the blue lines in **Figure 2**, was close to its expected value of 0.50 when the actual value of the correlations was zero-that is, when the indicators were conditionally independent within each class-regardless of different choices for prior variance. Deviations from 0.50 appeared with the use of less informative priors (i.e., larger prior variances), though the discrepancy was negligible (see the bottom panel). Similar results were found in the case of small correlations, or equivalently, small conditional dependence ($\rho_{m,jk}$ = 0.20; see the second bottom panel). In contrast, the average PPP decreased farther from 0.50 with the correlations greater than small ($\rho_{m,jk}$ = 0.50–0.80; see the top two panels), and more quickly when a more informative prior was chosen for the correlations-that is, interaction between correlation size and prior variance. Still, model fit was good if the strongly non-informative prior was specified for moderate and large correlations (in fact, correlations in any size).

The interactions of sample size with correlation size and prior variance are also exhibited in **Figure 2**. When the correlations between the indicators were less than moderate ($\rho_{m,jk}$ = 0–0.20; see the bottom two panels), sample size had no impact on model fit. When the correlations were rather moderate or large ($\rho_{m,jk}$ = 0.50–0.80; see the top two panels), PPP decreased as sample size increased; such deterioration in model fit became greater as the prior was more informative. In general, PPP was less variable as compared to the findings for Bayesian CFA (Muthén and Asparouhov, 2012).

Bias in Bayesian Estimates Due to the Presence of Conditional Dependence

Table 2 shows the Bayesian estimates of indicator correlations in Class 1 from the fitted flexible Bayesian LCA models. Recall that the correlation between the first and second indicators was simulated to be$\rho_{1,12}$ = 0, 0.20, 0.50, or 0.80, while the correlations between the first and third indicators and between the third and eighth indicators were always zero in the population ($\rho_{1,13} = \rho_{1,38} = 0$). These three indicators were purposefully selected to scrutinize the influence of conditional dependence on estimating correlations among other pairs of indicators. Similar to the findings for cross-loadings in Bayesian CFA (Asparouhov and Muthén, 2011) the estimate of a true zero correlation (i.e., conditional independence) was negatively biased by the presence of a true nonzero correlation(s) (i.e., conditional dependence). For example, both $\hat{\rho}_{1,13}$ and $\hat{\rho}_{1,38}$ deviated farther from their

TABLE 2 | Bayesian estimates of true zero correlations between class indicators.

$\rho_{1,12}$	Prior	$\hat{\rho}_{1,12}$	$\hat{\rho}_{1,13}$	$\hat{\rho}_{1,38}$
0.20	Strongly informative	0.0002	0.0000	0.0000
	Informative	0.0134	−0.0006	0.0008
	Weakly informative	0.0385	−0.0022	0.0019
	Weakly non-informative	0.0594	−0.0036	0.0027
	Strongly non-informative	0.1012	−0.0077	0.0052
0.50	Strongly informative	0.0004	0.0000	0.0000
	Informative	0.0370	−0.0006	0.0000
	Weakly informative	0.1023	−0.0024	−0.0002
	Weakly non-informative	0.1544	−0.0040	−0.0008
	Strongly non-informative	0.2543	−0.0073	−0.0023
0.80	Strongly informative	0.0009	0.0000	0.0000
	Informative	0.0711	−0.0025	−0.0009
	Weakly informative	0.1892	−0.0083	−0.0034
	Weakly non-informative	0.2769	−0.0139	−0.0061
	Strongly non-informative	0.4361	−0.0233	−0.0116

population value (0) in the negative direction as $\rho_{1,12}$ increased. Such bias became greater (i) when a less informative prior was specified for these parameters and (ii) when one (or both) of the two indicators had a nonzero correlation with other indicator(s) within the same class. Still, Type I error was well controlled-a true zero correlation ($\rho_{1,13} = \rho_{1,38} = 0$) was estimated as significantly different from zero in less than 5% of chances.

DISCUSSION

The central assumption of LCA is the conditional independence of indicators, given latent class membership. The current literature has shown that Bayesian LCA under the assumption of approximate independence provides an accessible alternative way of detecting violations of the conditional independence assumption (Asparouhov and Muthén, 2011). Unfortunately, little is known about how the performance of Bayesian LCA would be changed by a different choice of prior, even though the prior is the key element of Bayesian analysis. The current study, therefore, explores the utility of priors in a range of prior variances in terms of Type I error and power for detecting conditional dependence, model fit (PPP), and parameter bias. In doing so, the authors believe this article contributes to the methodology and applied communities by offering modeling guidance to be considered when researchers choose Bayesian LCA as a tool for analyzing highly correlated data.

Summary of Findings and Implications

The findings of the current simulation study show that Bayesian LCA could adequately control for the Type I error of falsely finding a true zero correlation as significantly different from zero. In fact, it is a somewhat rigorous test with Type I error smaller than 5% for all conditions examined. Power for detecting a nonzero correlation increased if a non-informative, rather than informative, prior was chosen. If the researcher secured a sample that included 300 or more, power for testing a large correlation would be satisfactory with any choice from weakly informative to strongly non-informative priors. For all priors examined, a true small correlation (i.e., 0.20) was significant in less than half of the replications. This finding implies that even when a correlation is estimated to have a positive value, the modeling may not produce enough power to establish significance for that correlation. Thus, a non-significant correlation should not be automatically discounted as being zero. Instead, the size of the estimated value should also be taken into account (Ulbricht et al., 2018). Unfortunately, another layer of complexity is that the Bayesian estimate may be biased downwardly by the presence of other nonzero correlations among the indicators, as observed in the current simulation. In many research settings, the true distribution of parameters is usually unknown and thus, researchers should be cautious about choosing extremely informative priors in either direction.

This study also found that model fit (PPP) of Bayesian LCA is susceptible to the magnitude of conditional dependence and the prior variance specified for the corresponding parameters (correlations). It is not surprising that in our simulation, approximate independence models fit well when the actual value of indicator correlations was equal to the prior mean (0). Rather, it is interesting that when the actual correlation value was different from the prior mean, model fit decreased as a more informative prior (i.e., smaller prior variance) was imposed on the correlations. A smaller variance may not let correlations escape from their zero prior mean, producing a worse PPP value. In a similar vein, model fit was acceptable when a non-informative prior was specified with a large prior variance regardless of the degree of dependence.

One should determine priors ahead of data collection in accordance with his/her substantive theory and/or previous findings from similar populations. In the context of cluster analysis, the researcher may consider either informative or non-informative priors when conditional independence has been confirmed a priori so that indicator correlations are nuisance parameters. More often, the nature of cluster analysis is rather exploratory (Lanza and Cooper, 2016) looking for or testing for correlated indicators. In such a case, the researcher may begin with non-informative priors reflecting large uncertainty on the parameter values. Otherwise, a range of priors will be equally inspired. Nevertheless, our simulation suggests that less informative priors, even strongly non-informative priors, would be a promising choice for running flexible Bayesian LCA.

Limitations and Future Research

Although a few important findings and implications were discussed in this article, the current simulation study has two notable limitations that need to be addressed in future research. First, one must assess the validity of the findings because any variation in Bayesian application may affect the trustworthiness of the simulation results. For instance, other BSEM fit measures-e.g., deviance information criteria (Spiegelhalter et al., 2002) widely applicable information criterion (Watanabe, 2010) and leave-one-out cross-validation statistics (Gelfand, 1996) and available significance tests for conditional independence (Andrade et al., 2014) should be considered to confirm the performance of flexible Bayesian LCA.

Second, caution should be paid to generalizing the findings beyond the conditions included in the study. The simulation considered only two latent classes and a relatively small number (10) of binary indicators. The number of latent classes and the number of their indicators are not expected to considerably affect Type I error and power of the analysis and bias in parameter estimation but may have an impact on model fit (PPP). In addition, only the default priors set by M*plus* 8 were analyzed in this study. Asymmetric, rather than symmetric, binomial distribution may be a better prior for correlations among class indicators because correlations are bounded by two values (−1 and 1). Mode or mean of posterior distribution, rather than median, can serve as a better point estimate for correlations particularly when the distribution is not normal. Because the exact distribution of a posterior is typically not known, it is recommended to plot the posterior distribution and choose the measure that best represents the sample. Taken together, further simulation work is encouraged to continue to increase the utility of Bayesian LCA for various models and data environments.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available upon request.

AUTHOR CONTRIBUTIONS

All authors substantially contributed to the conception or design of the work and analysis and interpretation of data for the work.

REFERENCES

Albert, P., and Dodd, L. (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* 60, 427–435. doi: 10.1111/j.0006-341x.2004.00187.x

Andrade, P. D., Stern, J. M., and Pereira, C. A. (2014). Bayesian test of significance for conditional independence: the multinomial model. *Entropy* 16, 1376–1395. doi: 10.3390/e16031376

Asparouhov, T., and Muthén, B. (2010). *Bayesian Analysis Of Latent Variable Models Using Mplus.* Technical report. Los Angeles, CA: Muthén & Muthén.

Asparouhov, T., and Muthén, B. (2011). Using Bayesian priors for more flexible latent class analysis. *Proc. Joint Statist. Meet.* 2011, 4979–4993.

Asparouhov, T., and Muthén, Â (2014). Auxiliary variables in mixture modeling: three-step approaches using Mplus. *Struct. Equ. Model.* 21, 329–341. doi: 10.1080/10705511.2014.915181

Barnard, J., McCulloch, R., and Meng, X. L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage. *Statist. Sin.* 10, 1281–1311.

Gelfand, A. E. (1996). "Model determination using sampling-based methods," in *Markov Chain Monte Carlo in Practice*, eds W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (London: Chapman & Hall), 145–162.

Gelman, A. (2002). "Prior distribution," in *Encyclopedia of Environmetrics*, Vol. 3, eds A. H. El-Shaarawi and W. W. Piegorsch (Chichester: John Wiley and Sons), 1634–1637.

Gelman, A., Meng, X. L., Stern, H. S., and Rubin, D. B. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sin.* 6, 733–807.

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.* 7, 457–511.

Geweke, J., and Amisano, G. (2011). Hierarchical Markov normal mixture models with applications to financial asset returns. *J. Appl. Econometr.* 26, 1–29. doi: 10.1002/jae.1119

Gill, J. (2008). *Bayesian Methods: A Social And Behavioral Sciences Approach*. New York, NY: Chapman & Hall.

Hagenaars, J. A. (1988). Latent structure models with direct effects between indicators local dependence models. *Sociol. Methods Res.* 16, 379–405. doi: 10.1177/0049124188016003002

Hansen, M., Cai, L., Monroe, S., and Li, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response models. *Br. J. Math. Statist. Psychol.* 69, 225–252. doi: 10.1111/bmsp.12074

Im, K. S. (2017). *The Hierarchical Testlet Response Time Model: Bayesian Analysis Of A Testlet Model For Item Responses And Response Times*. Doctoral dissertation, University of Kansas, Lawrence, KS.

Lanza, S. T., and Cooper, B. R. (2016). Latent class analysis for developmental research. *Child Dev. Perspect.* 10, 59–64. doi: 10.1111/cdep.12163

Lazarsfeld, P. F., and Henry, N. W. (1968). *Latent Structure Analysis*. Boston, MA: Houghton Mifflin.

Li, Y., Lord-Bessen, J., Shiyko, M., and Loeb, R. (2018). Bayesian Latent Class Analysis Tutorial. *Multivar. Behav. Res.* 53, 430–451. doi: 10.1080/00273171.2018.1428892

Lubke, G., and Muthén, B. (2007). Performance of factor mixture models as a function of model size, criterion measure effects, and class-specific parameters. *Struct. Equ. Model.* 14, 26–47. doi: 10.1080/10705510709336735

MacCallum, R. C., Edwards, M. C., and Cai, L. (2012). Hopes and cautions in implementing Bayesian structural equation modeling. *Psychol. Methods* 17, 340–345. doi: 10.1037/a0027131

Magidson, J., and Vermunt, J. K. (2004). "Latent class models," in *Handbook Of Quantitative Methodology For The Social Sciences*, ed. D. Kaplan (Newbury Park, CA: Sage), 175–198.

McLachlan, G. J., Do, K. A., and Ambroise, C. (2004). *Analyzing Microarray Gene Expression Data*. Hobokin, NJ: Wiley.

Muthén, B., and Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychol. Methods* 17, 313–335. doi: 10.1037/a0026802

Muthén, B. O. (2002). Beyond SEM: general latent variable modeling. *Behaviormetrika* 29, 81–117. doi: 10.2333/bhmk.29.81

Muthén, B. O., and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* 55, 463–469. doi: 10.1111/j.0006-341x.1999.00463.x

Oberski, D. L., van Kollenburg, G. H., and Vermunt, J. K. (2013). A monte carlo evaluation of three methods to detect local dependence in binary data latent class models. *Adv. Data Analy. Classif.* 7, 267–279. doi: 10.1007/s11634-013-0146-2

O'sullivan, A. M. (2013). *Bayesian Latent Variable Models With Applications*. Doctoral dissertation, Imperial College London, London.

Qu, Y., Tan, M., and Kutner, M. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* 52, 797–810.

Rupp, A., Templin, J., and Henson, R. (2010). *Diagnostic Measurement: Theory, Methods, And Applications*. New York, NY: Guilford Press.

Scheines, R., Hoijtink, H., and Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika* 64, 37–52. doi: 10.1007/bf02294318

Schlattmann, P. (2010). *Medical Applications Of Finite Mixture Models*. Berlin, Germany: Springer.

Silva, R., and Ghahramani, Z. (2009). The hidden life of latent variables: bayesian learning with mixed graph models. *J. Mach. Learn. Res.* 10, 1187–1238.

Spiegelhalter, D. J., Best, N. G., Carlin, B., and Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Statist. Soc. Ser. B* 64, 583–639. doi: 10.1111/1467-9868.00353

Torrance-Rynard, V., and Walter, S. (1998). Effects of dependent errors in the assessment of diagnostic test performance. *Statist. Med.* 16, 2157–2175. doi: 10.1002/(sici)1097-0258(19971015)16:19<2157::aid-sim653>3.0.co;2-x

Uebersax, J. S. (1999). Probit latent class analysis with dichotomous or ordered category measures: conditional independence/dependence models. *Appl. Psychol. Measur.* 23, 283–297. doi: 10.1177/01466219922031400

Ulbricht, C. M., Chrysanthopoulou, S. A., Levin, L., and Lapane, K. L. (2018). The use of latent class analysis for identifying subtypes of depression: a systematic review. *Psychiatr. Res.* 266, 228–246. doi: 10.1016/j.psychres.2018.03.003

Vacek, P. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics* 41, 959–968.

Vermunt, J. K. (2010). Latent class modeling with covariates: two improved three-step approaches. *Polit. Analy.* 18, 450–469. doi: 10.1093/pan/mpq025

Wang, W.-C., and Wilson, M. (2005). The Rasch testlet model. *Appl. Psychol. Measur.* 29, 126–149. doi: 10.1177/0146621604271053

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* 11, 3571–3594.

Yang, M.-H., and Ahuja, N. (2001). *Face Detection And Gesture Recognition For Human-Computer Interaction*. Boston, MA: Springer.

Zhan, P., Liao, M., and Bian, Y. (2018). Joint testlet cognitive diagnosis modeling for paired local item dependence in response times and response accuracy. *Front. Psychol.* 9:607. doi: 10.3389/fpsyg.2018.00607

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

```
TITLE:           Bayesian latent class analysis model of approximate independence
DATA:            FILE = example.dat;
                 !Names data set.
VARIABLE:        NAMES = u1-u10;
                 !Assigns names to the variables in the data set
                 CATEGORICAL = u1-u10;
                 !Specifies which dependent variables are treated as binary or ordered categorical variables in the model and its estimation. In this example,
                 all dependent variables are binary.
                 MISSING = .;
                 !Specifies the values or symbols in the data set that are treated as missing or invalid. In this example, dot (.) is the missing value flag.
                 CLASSES = C(2);!Assigns names to the categorical latent variables in the model and specifies the number of latent classes for each
                 categorical latent variable. In this example, the latent variable C has two classes.
ANALYSIS:        ESTIMATOR = BAYES;
                 !Activates the Bayesian estimator.
                 CHAINS = 2;
                 !Specifies using two Markov Chains for conducting the analysis.
                 PROCESSORS = 2;
                 !For use in multi-core systems, assigns two processors with one per chain;
                 TYPE = MIXTURE;
                 !Carries out a mixture analysis.
MODEL:           %OVERALL%
                 %C#1%
                 !In the first class,
                 [u1$1-u10$1*1]; !Specifies starting values (in this example, 1) for the thresholds of the dependent variables
                 u1-u10 WITH u1-u10*0 (p1-p45);
                 !Assigns labels to the correlations among the dependent variables
                 %C#2%
                 [u1$1-u10$1*1];
                 u1-u10 WITH u1-u10*0 (p46-p90);
MODEL PRIORS:    p1-p90 ∼ IW(0, 11);
                 !Assigns priors (in this example, inverse Wishart distributions) to the correlations among the dependent variables.
```

ORIGINAL RESEARCH
published: 26 October 2020
doi: 10.3389/fpsyg.2020.516424

Preliminary Validation of the CI-FRA Checklist: A Simple Screening Tool for Measuring the Early Signs of Reading and Spelling Disorders in Italian Primary Students

Sara Giovagnoli[1]*, *Luigi Marotta*[2], *Sara Magri*[1], *Michela Muccinelli*[3], *Alessandra Albani*[1], *Giulia Casu*[1], *Sara Garofalo*[1] *and Mariagrazia Benassi*[1]

[1] *Department of Psychology, University of Bologna, Bologna, Italy,* [2] *Bambino Gesù Children Hospital (IRCCS), Rome, Italy,* [3] *Azienda Unità Sanitaria Locale (AUSL) della Romagna, Emilia-Romagna, Italy*

OPEN ACCESS

Edited by:
Pietro Cipresso,
Istituto Auxologico Italiano (IRCCS),
Italy

Reviewed by:
Simone Aparecida Capellini,
São Paulo State University, Brazil
Turid Helland,
University of Bergen, Norway

***Correspondence:**
Sara Giovagnoli
sara.giovagnoli@unibo.it

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *30 November 2019*
Accepted: *30 September 2020*
Published: *26 October 2020*

Citation:
Giovagnoli S, Marotta L, Magri S, Muccinelli M, Albani A, Casu G, Garofalo S and Benassi M (2020) Preliminary Validation of the CI-FRA Checklist: A Simple Screening Tool for Measuring the Early Signs of Reading and Spelling Disorders in Italian Primary Students. Front. Psychol. 11:516424. doi: 10.3389/fpsyg.2020.516424

Although several screening tests for recognizing early signs of reading and spelling difficulties have been developed, brief and methodologically grounded tools for teachers are very limited. The present study aimed to lay the foundation for a new screening tool for teachers: the Checklist for early Indicators of risk Factors in Reading Ability (CI-FRA). The proposed checklist consists of 20 items, based on a 7-point Likert scale, and it investigates five domains: reading, writing, attention, and motor skills. Six hundred sixty-seven children were evaluated by 40 teachers during the first year of primary school and, longitudinally, in the second year. Exploratory factor analysis and confirmatory factor analysis (CFA) were applied to verify structural validity. Concurrent validity was assessed by Spearman correlation to analyze the link between CI-FRA and reading and spelling standardized tests and cognitive tests. Reliability was assessed by Cronbach α and interclass correlation coefficient. The CFA reported a three-factor structure as the optimal solution, including language (reading and writing), visuospatial attention, and fine motor skills subscales. Good reliability, good internal consistency, and acceptable test–retest indices were found. Concurrent validity was confirmed by significant correlations between CI-FRA total score and standardized reading and spelling test, as well as by correlations between CI-FRA subscales and neuropsychological standardized test scores. Preliminary evaluation of sensitivity by receiver operating characteristic curves showed that the CI-FRA score has particularly high sensitivity and specificity for word reading speed deficit. In conclusion, the results confirm that CI-FRA is a theoretically grounded and statistically valid tool that could help the teachers to screen for early signs of reading and spelling difficulties.

Keywords: reading, spelling, reading disorders, confirmatory factor analysis, checklist, early indicators

INTRODUCTION

Several studies show that early intervention is crucial to correct some of the adverse effects of reading difficulties (Torgesen et al., 2001; Poskiparta et al., 2003; Elbro and Scarborough, 2004; Torgesen, 2005). For this reason, the identification of early signs characterizing children with reading and spelling difficulties is essential.

Although many standardized tests are available for clinicians to assess learning disorders, little effort has been done in literature in terms of tools dedicated to teachers. Teachers are the first adults evaluating the daily signs of progress in children, therefore having the highest chance to recognize learning disorders at an early stage. The behavioral checklists currently available (Mash and Wolfe, 2002; Wagner, 2003) are global or broad-spectrum rating scales based on parent and teachers' ratings about the frequency and intensity of a wide range of behaviors. To our knowledge, none of these focuses on both precursors and current learning abilities as evaluated by the teacher. Furthermore, the need for new instruments dedicated to teachers has been largely demonstrated by the absence of theoretically grounded and statistically validated tools (Snowling, 2013; Catts et al., 2015). Indeed, screening tests for specific learning disorders have often proved to be inaccurate (Catts et al., 2009; Johnson et al., 2009; Compton et al., 2010). For example, longitudinal studies have shown how tests evaluating early predictors of reading difficulties resulted in high percentages of false-positive and false-negatives (Catts et al., 2015; Poulsen et al., 2017). Poulsen et al. (2017) stated that there is a methodological "dilemma": on the one hand, preschool screening would allow early intervention, but measures to detect reading difficulties are inaccurate; on the other hand, a school-based screening test would be more accurate but would delay the intervention. In line with Poulsen et al. (2017), the present study aimed to create a fast screening tool based on different domains of evaluation and including both early indicators of learning disorders and a measure of current reading and spelling abilities.

Learning disorders are related to a complex neuropsychological profile where multiple difficulties are traced within different cognitive domains (Pennington, 2006). The multifactorial theory suggests that the etiology of learning disorders is multifactorial; i.e., it involves the interaction of multiple risks and environmental factors that impact on multiple cognitive domains (Menghini et al., 2010). According to this model, both phonological and non-phonological abilities could be impaired in subjects having reading and spelling disorders; therefore, careful early evaluation of a wide range of cognitive abilities appears to be necessary for early detection of future reading and spelling disorders.

Language-related visual abilities (such as letter knowledge, phonological awareness, and rapid automatized naming) are regarded by some authors as one of the main precursors of future reading ability (Kirby et al., 2003; Elbro and Scarborough, 2004; Landerl and Wimmer, 2008; Puolakanaho et al., 2008; Catts et al., 2009; Lervåg et al., 2009; Landerl et al., 2013; Poulsen et al., 2015), arguing that the inconsistencies among studies are related to the differences among the orthographies investigated in each study (Ziegler et al., 2010). Yet, a study by Moll et al. (2014) including a large cohort of students from five European countries ordered by grapheme-phoneme consistency (from the lowest, i.e., English, to the highest consistency, i.e., Finnish) found comparable results for different orthographies. Phonological processing and rapid automatizing naming (RAN) were both reported as important indicators for reading and spelling development. RAN was the best reading speed indicator, whereas phonological processing was the best predictor of reading accuracy and spelling. The lower the consistency of the language, the better these indices worked as predictors. Indeed, English orthography, being the less consistent one, shows a stronger predictive effect of RAN and phonological processing than all other orthographies. Italian is instead characterized by high grapheme–phoneme consistency (transparent orthography) with no irregular words, no non-homographic homophones, and no alternative acceptable phonological ways of spelling words (Zoccolotti et al., 1999). Considering phonological abilities, a recent study demonstrated that phonological awareness is a strong predictor for word reading in Italian language (Holopainen et al., 2020).

Although there is general agreement on the role of phonological abilities as an early indicator of reading and spelling disorders, the role of other predictive factors is more controversial. A group of studies pointed out that visuospatial attention could be considered a general precursor of reading difficulties both in Italian and French cohorts (Facoetti and Molteni, 2001; Valdois et al., 2004; Bosse et al., 2007; Bosse and Valdois, 2009; Facoetti et al., 2010). Indeed, visuospatial attention was found to predict irregular word reading independently from phoneme awareness. However, recent studies found a double dissociation between dyslexia and visuospatial attention, thus opposing the importance of attention in predicting reading disorders (Lukov et al., 2015).

Further controversy is related to the predictive role of motor skills for children's school readiness (e.g., Grissmer et al., 2010), showing a link with intellectual skills. In particular, Grissmer et al. (2010) showed that both attention and fine motor skills measured at kindergarten are important developmental predictors of later academic achievement. Cameron et al. (2016) also emphasized how motor skills are implicated in children's self-regulation and their future reading, spelling, and numeracy. In a study by Roebers and Jäger (2014), fine motor skills were found to have a noticeable predictive power for school achievement and literacy, but strictly associated with executive functions. Other studies (e.g., Viholainen et al., 2002) confirmed that motor and language problems are often interconnected. Moreover, fine motor skills have been found to be associated to executive functions (Roebers and Jäger, 2014), and some studies demonstrated that they are predictors of written expression achievement (Carlson et al., 2013). Besides, some studies demonstrated that, in preschool and early elementary classrooms, motor coordination, executive functions, and visuospatial processes are combined with other skills to form the basis for children's successful learning (Pagani and Messier, 2012; Cameron et al., 2016).

Taken together, this evidence suggests that early evaluation of language, visuospatial attention, and fine motor skills should be considered for the early identification of children at risk of learning disorders (Gabrieli and Norton, 2012; Cameron et al., 2016).

Moving from this evidence and the multifactorial theory (Menghini et al., 2010), we aimed to lay the foundation for a new screening tool for teachers: the Checklist for early Indicators of risk Factors in Reading Ability (CI-FRA). This checklist is based on the evaluation of early precursors and current state of reading and spelling difficulties through the analysis of five domains (language, reading, spelling, attention, and motor skills).

Reading and spelling difficulties are knowingly overlapping in learning disorders. At least three plausible theoretical models can explain the heterogeneity of early signs of reading and spelling difficulties along a continuum of specificity. In line with the multifactorial theory, a first model includes altogether the learning difficulties evaluated as explained by a general factor expressing the severity of the learning disorder (minimum specificity); a second model distinguishes the specificity of each specific domain investigated with factors explaining the variability of each domain (maximum specificity); and a third model accomplishes the possible overlapping as well as the specificity of the domains investigated and includes a number of factors that are inferior to the number of domains, i.e., macrodimension (intermediate specificity).

Because no previous study validated such a theoretical model, a confirmatory factor analysis (CFA) was used to test the dimensionality of early signs of reading and spelling difficulties. Moreover, the validity, sensitivity, and reliability of the CI-FRA were statistically tested.

MATERIALS AND METHODS

Participants and Procedure

Participants were recruited through contact and direct agreement with the school managers of the 23 primary schools located in the province of Forlì-Cesena, in the Emilia Romagna region (North of Italy). Italian education is based on a state school system and follows the same rules and the same educational curriculum for all the regions. The primary school is the first compulsory school, and it is commonly preceded by 3 years of kindergarten. The primary school lasts 5 years, and the first year starts at 6 years. During the first 3 months of the first year, the pupils learn reading and spelling, and at the end of grade 2, they are expected to be proficient. All the data were collected between the end of February 2017 and June 2017 (after 5–6 months from the beginning of the academic year). A total of 667 children (310 females) attending grade 1 of the primary school [mean age is 6.64 years, standard deviation (SD) = 0.28 years] participated in the study.

Institutional review boards approved the study, and both parents gave written informed consent. In the case of single parent, we asked the responsible parent for the informed consent.

The exclusion criteria adopted were those recommended by the Consensus Conference on Specific Learning Disorders promoted by the Italian National Institute of Health (Lorusso et al., 2014) for diagnosis of developmental dyslexia. Participants with an IQ lower than 70 and having referred sensory disability were excluded from the study. In order to evaluate the predictive, concurrent validity, and sensitivity of CI-FRA, we selected from the total sample a subsample of 106 children (males = 64; mean age = 6.6 years, SD = 0.28 years) recruited from two schools who agreed to participate in the follow-up evaluation. This subsample was assessed in regard to specific cognitive measures (general cognitive functioning, visual attention, phonological skills) simultaneously with CI-FRA administration during the first grade and at the end of the second grade (i.e., September 2018) reevaluated for reading and spelling abilities with standardized tests. A group of expert psychologists was responsible for the assessment and the relationships with the teachers and parents. The standardized cognitive tests were chosen because they are considered the main indicators for Italian reading and spelling acquisition. These indications have been published in official public documents promoted by the Italian National Institute of Health as the Consensus Conference on Specific Learning Disorders for diagnosis of developmental dyslexia (Istituto Superiore di Sanità, 2011). Within the subsample, 86 children (males = 51; mean age = 6.6 years, SD = 0.29 years) had a second evaluation by their teachers at CI-FRA (first at the end of February 2017 and a second time at the end of May 2017) to measure test–retest reliability.

No significant differences were found between the total sample and the two subsamples for demographics characteristics or gender. As **Table 1** shows, no differences were found for maternal education level or for paternal ones between the three samples. The percentages of mothers and fathers who had school difficulties in the past were not significantly different between the three subsamples, as well as no differences were found for gender distribution (**Table 1**).

According to the Italian school general population, it is common to have a high percentage of bilingual pupils; therefore, we decided to include in the whole sample also the bilinguals and ask the teachers to have additional information about their exposure to the Italian language. In the total sample, 81.5% of the pupils were monolingual and used Italian as their language, whereas 18.5% of the total sample were bilingual. Within the bilinguals, 46.4% had the Italian language as L1, and 54.6% as L2. The languages more common after Italian were Arabic and Albanian (16.7 and 13.8%, respectively), whereas the most common languages as L2 after Italian were Arabic, Albanian, and Romanian (7.4, 9.5, and 9.3%). The 8.7% of bilinguals had been in Italy for over 3 years, and the 4.8% (only five children) had been in Italy for less than 3 years; all the others were born in Italy. Teachers reported that 87.8% of the total sample had a good oral comprehension ability, 11.7% showed a sufficient ability, whereas the 0.5% had difficulties (all those children are bilinguals). Most of the parents' participants (71%) had finished high school, college, or university, and 95.3% declared that they did not have difficulty at school. Finally, 87.3% stated that no one in the family presented specific learning disorders.

Instruments

CI-FRA Checklist

The CI-FRA checklist was created by a team of psychologists and speech therapists who worked in two different research teams and jointly collaborated to the project.

TABLE 1 | Demographic characteristics of the three samples used in the study.

			Entire sample	Cognitive measures sample	Test–retest sample	χ^2
Maternal Variables	Education (%)	Primary and secondary school	102 (25.0%)	14 (23.7%)	14 (23.7%)	0.079
		High school and university	306 (75.0%)	45 (76.3%)	45 (76.3%)	
	School difficulties (%)	Yes	19 (4.7%)	1 (1.7%)	2 (3.4%)	1.23
		No	387 (95.3%)	58 (98.3%)	56 (96.6%)	
Paternal Variables	Education (%)	Primary and secondary school	132 (33.1%)	15 (25.4%)	16 (27.6%)	1.88
		High school and university	267 (66.9%)	44 (74.6%)	42 (72.4%)	
	School difficulties (%)	Yes	19 (4.7%)	4 (6.9%)	4 (6.9%)	0.85
		No	382 (95.3%)	54 (93.1%)	54 (93.1%)	
Children variables	Gender	Males	357 (53.5%)	64 (60.4%)	51 (59.3%)	2.47
		females	310 (46.5%)	42 (39.6)%	35 (40.7%)	

Three steps were settled for the development of the final version of the checklist. In a first step, three psychologists and one speech therapist, two are coauthors in the present articles, prepared a list of possible items on the basis of multifactorial model of dyslexia (Pennington, 2006; Menghini et al., 2010; Ziegler et al., 2019). According to the multifactorial model of dyslexia, the phonological memory and phonological awareness, as well as visual attention functions and motor skills, were considered important predictive factors of reading and spelling difficulties. Therefore, for each one of these cognitive domains, we included a specific set of items. In a second step, three additional psychologists and one speech therapist revised the set of items by selecting the most relevant items for each one of the cognitive domains by ordering them for their importance. In the last step, a group of 60 teachers evaluated the adequacy of the checklist by excluding some items because of their redundancy and by ameliorating the terms used in some others items (the specialistic language used in some cases was not sufficiently clear for the teachers).

The final version of the CI-FRA checklist comprises 20 items that measure the student's learning disorders, as referred by the teacher's teaching experience (**Table 2**). The teacher (or the team of teachers) is asked to evaluate each student's difficulties in the different domains by comparing him/her to an ideal reference "average student" based on his/her teaching experience. The frequency of occurrence of each problematic behavior is measured using a seven-point Likert scale ranging from 1 (never observed) to 7 (often observed). The checklist has to be compiled by the teacher; as a first administration point, it is recommended to administer the CI-FRA 3 to 4 months after the beginning of the Academic year, then it can be used every 5 months to monitor the developmental changes.

The checklist has been developed to measure five dimensions or subscales (phonological awareness and verbal memory, reading, spelling, motor skills, attention): language and verbal memory subscale includes items 1 to 3 and item 20; reading subscale includes the items 4 to 8, fine motor skills is composed of items 9 to 12, spelling dimension includes items 13 to 15, and the last dimension regards attention abilities and is composed of items 16 to 19. The scores of the subscales are obtained as the mean of the items included in each subscale. The total score is obtained as the sum of all the items.

Moreover, the CI-FRA is accompanied with an interview in which the teacher is asked to specify the presence of bilingualism and the familiarity for learning disorders (e.g., the parental education levels and possible relatives with learning disorders).

The CI-FRA is not only available as a paper-and-pencil tool but is also in digital format that includes the formula to calculate the scores for each student and to show the changes in each area graphically. Moreover, the teacher could have a graphical representation of the entire class.

Standardized Cognitive Tests

Raven's colored progressive matrices (CPM)

The test evaluates the non-verbal intellectual abilities, such as logical ability, visuospatial components, and the ability to analyze abstract images according to similarity, dissimilarity, numerical progression, and size (Raven, 1994). The test consists of 36 items, and the subject is required to look at an incomplete figure and identify the missing piece between 4 and 8 alternatives. The subject total performance represents the subject total score.

Digit span (Wechsler intelligence scale for children-IV subtest)

This subtest measures short-term auditory memory and working memory's ability (Wechsler, 2005). The subject's task is to listen and repeat a sequence of numbers. The sequence increases in length at each trial. Forward and backward digit span abilities are tested. In the backward task, the participant has to recall the sequence in reverse order (working memory). Subject's total score is obtained by summing forward and backward correct responses.

RAN colors

The task measures automatization naming ability that is a competence related to language abilities (De Luca et al., 2005). "Colors" condition is composed of a sequence of colored dots, and the subject's task is to name the colors as fast as possible. Total time and total correct answers represent the subject's scores in speed and accuracy.

TABLE 2 | CI-FRA checklist.

	Italian version	English version
Item 1	L'alunno fatica ad esprimersi oralmente (le difficoltà possono riguardare gli aspetti fonologici, articolatori e/o la produzione morfosintattica)	The student has difficulty in oral expression (difficulties can concern phonological and articulatory aspects and/or morphosyntactic production).
Item 2	L'alunno per esprimersi utilizza poche parole e sempre le stesse (ampiezza del vocabolario limitata)	In oral communication, the student uses a limited number of words and tends to use always the same words (restricted vocabulary).
Item 3	L'alunno fatica a costruire una parola dai singoli fonemi (sintesi fonemica) o a individuare i fonemi che compongono la parola (segmentazione fonemica)	The student struggles to create a word starting from separated phonemes (phonemic synthesis) or to identify the phonemes that compose the word (phonemic segmentation).
Item 4	L'alunno legge più lentamente rispetto ai coetanei	The student has a lower reading speed compared to peers.
Item 5	L'alunno, quando legge, commette molti errori	When reading, the student makes many mistakes.
Item 6	L'alunno, quando legge, commette errori di confusione tra lettere che hanno un suono simile (es. p-b, c-g, f-v) o che sono visivamente simili (es. m-n, b-d, a-e)	When reading, the student makes confusion errors between letters that have a similar sound (e.g., p-b, c-g, f-v) or that are visually similar (e.g., m-n, b-d, a-e).
Item 7	L'alunno mostra difficoltà nella lettura delle parole bisillabiche piane	The student struggles in reading simple disyllabic words.
Item 8	L'alunno, quando legge, si affatica facilmente	When reading, the student gets tired quickly.
Item 9	La grafia dell'alunno risulta poco leggibile	The student's handwriting is difficult to read.
Item 10	L'alunno impugna la matita/penna con difficoltà o in modo inadeguato	The student holds the pencil/pen having difficulty or inadequately.
Item 11	L'alunno mostra difficoltà nella gestione del foglio (rispetto delle righe, dei quadretti, i margini)	The student shows difficulties in managing spaces in the paper (poor awareness of lines, squares, margins of the paper).
Item 12	L'alunno mostra difficoltà nella motricità fine (es. usare le forbici, allacciare bottoni)	The student shows difficulties in fine motor skills (e.g., using scissors, fastening buttons).
Item 13	L'alunno, quando scrive, commette errori di confusione tra lettere che hanno un suono simile (es. p-b, c-g, f-v) o che sono visivamente simili (es. m-n, b-d, a-e)	When writing, the student confuses letters that have a similar sound (e.g., p-b, c-g, f-v) or that are visually similar (e.g., m-n, b-d, a-e).
Item 14	L'alunno, quando scrive, tende ad invertire le lettere, ad esempio gli capita di scrivere "la" invece che "al"	When writing, the student tends to reverse the letters (e.g., he writes "fo" instead of "of").
Item 15	L'alunno mostra difficoltà nella scrittura delle parole bisillabiche piane	The student has difficulties in writing simple disyllabic words.
Item 16	L'alunno si distrae facilmente	The student is easily distracted.
Item 17	L'alunno si affatica facilmente	The student gets tired easily.
Item 18	L'alunno si muove molto sulla sedia mentre deve eseguire i compiti, giocherella con gli oggetti presenti sul tavolo, ecc. . .	The student moves a lot on the chair while doing homework, plays with the objects on the table, etc.
Item 19	L'alunno impiega molto più tempo degli altri per portare a termine le attività in classe	The student takes much longer than others to complete classroom activities.
Item 20	L'alunno fatica nei compiti che riguardano la memoria (esempio poesie, mesi, filastrocche)	The student has difficulties in memory tasks (poems, nursery rhymes, months).

Visual search (VS) objects

The task is used to evaluate visual attention ability (De Luca et al., 2005). "Objects" condition is composed of matrices of different figures (stars, pears, cows, trains, and hands) painted in a paper, and the subject's task is to identify and mark with a pen the target figure (star) as quickly as possible. Total time and total correct answers represent the subject's score in speed and accuracy.

Metaphonological skills (CMF)

This test allows for evaluating the development of metaphonological skills in children from 5 to 11 years (Marotta et al., 2008). Metaphonological abilities represent important prerequisites for adequate learning and development of reading and spelling abilities and are related to language competences. In this study, to evaluate the different types and levels of phonological awareness, we used the following subtests: *segmentation test*, in which it is required to say, in the correct sequence, the segmental units (syllables), which constitute the different words (segmentation); *phonemic synthesis test*, in which the word resulting from the fusion of a series of phonemes pronounced by the examiner in the correct sequence (synthesis); *deletion of the initial syllable test*, in which it is required to pronounce a word without the initial syllable (manipulation); FAS test (verbal fluency test with phonemic facilitation), in which it is required to say as many words as possible starting with the same letter/sound (classification). The sum of correct answers in each subtest represents the subject's scores.

Non-word repetition (NWR)

This task involves different processes, including phonological memory and speech production and evaluates the ability to listen and repeat unusual sound patterns (non-word) (Subtest of PROMEA test; Vicari, 2007). This test consists of 40 non-words, and non-words can have high or low resemblance according to the number of changed letters compared to an existing Italian word. The sum of correct answers represents the subject's score.

Reading and Spelling Standardized Tests

Standardized batteries for Italian reading and spelling ability (DDE-2, Sartori et al., 2007; MT, Cornoldi et al., 2018) were used to test the presence of specific reading disability and specific spelling disability. Within the battery, accuracy and speed in reading were evaluated by using a written text (MT, Cornoldi et al., 2018). Spelling abilities were tested as the accuracy in a test (DDE-2, Sartori et al., 2007) that requires the child to write non-words list. Fine motor skills were tested by a test for evaluating the speed and fluidity of handwriting in which the child is asked to produce specific graphemes (lelele) as much as he can and as quickly as possible (BVSCO, Tressoldi et al., 2013). Raw scores were transformed into *z* standardized scores according to normative data. For each test, the children having a score equal to or less than 1.5 SDs are considered as having a deficit in the specific learning domain (reading, spelling, or fine motor skills).

Statistical Analysis

To assess the CI-FRA structural validity, exploratory factor analyses (EFA) and CFA were conducted on two randomly created subsamples. The sample size was established *a priori* as to have a subject to an item ratio of 10:1 in the EFA (Nunnally, 1978) and at least 10 observations for each freely estimated model parameter in the CFA (Kline, 1998).

Exploratory factor analyses with principal axis factoring (PAF) and Promax rotation was performed on the first subsample (n = 200). PAF is an extraction method generally used when testing a theoretical model method (Tabachnick and Fidell, 2001) as in this study where we expected a model-structure fitting the multifactorial model theory (Pennington, 2006; Menghini et al., 2010). Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy test and Bartlett test of sphericity were used to check whether the data were adequate to apply factor analysis. Factors were extracted based on Kaiser's criterion Kaiser (1960) of eigenvalue higher than 1. Items with loadings greater than 0.40 and cross-loadings less than 0.10 were considered for inclusion in a factor.

Confirmatory factor analysis was performed on the second subsample (n = 467) to test the factor model that resulted from EFA. Model parameters were estimated using the robust maximum likelihood method. The closeness of the hypothesized model to the empirical data was evaluated through the following goodness-of-fit indices: χ^2, Satorra–Bentler scaled χ^2 statistic (S-B χ^2); root mean square error of approximation [RMSEA, cutoff < 0.10, upper bound of the 90% confidence interval (CI) ≤ 0.10]; standardized root mean square residual (cutoff < 0.10); and comparative fit index (CFI, cutoff > 0.90) (Weston and Gore, 2006).

Aiming to evaluate possible alternative models explaining specificity or overlapping between the investigated domains, the three-factor model obtained by EFA and confirmed by CFA was compared to a one-factor model solution and to a five-factor model solution by CFA.

The predictive validity and the concurrent validity were verified by correlation analysis between the CI-FRA and standardized measures of reading and spelling abilities, and between CI-FRA and standardized measures general cognitive functioning and phonological skills that served as prerequisites for reading–spelling in a subsample of 106 participants. Expecting correlations with a moderate effect size, this sample size was considered adequate to have approximately 95% power (α = 0.05, two-tailed) to reject the null hypothesis.

On the same subsample, test sensitivity was evaluated by receiver operating characteristic (ROC) curves applied for each learning disorder (reading and spelling), using as state variable the qualitative results of the standardized test for reading and spelling. For each learning disorder, different standardized clinical tests were used. Children scoring equal to or less than the clinical cutoff score at the specific reading and spelling tests (1 SD for reading text, 2 SD for word lists) were registered as "below the norms" score, and these subjects were considered in the "clinical group." The category clinical group was used as a reference category in the state variable, whereas the CI-FRA subscale scores were the test variables.

Internal consistency reliability was assessed by calculating Cronbach α (cutoff ≥ 0.70; Nunnally, 1978) and corrected item-total correlations (cutoff ≥ 0.30; Streiner and Norman, 2008). Test–retest reliability over a 3-month period was assessed in a subsample of 86 participants by calculating the intraclass correlation coefficient (ICC) with a two-way random-effects (absolute agreement) model (cutoff ≥ 0.70; Streiner and Norman, 2008). This sample size was established *a priori* to detect an expected large effect size with a power of 0.85 or greater and α = 0.05 (two-tailed).

Interpretation of results was based on both statistical significance (significant level set at $p < 0.05$) and measures of effect size, with Spearman ρ of 0.10 considered small, 0.30 medium, and 0.50 large, and Cohen d of 0.20 considered small, 0.50 medium, and 0.80 large (Cohen, 1988). Sample sizes were calculated *a priori* with the statistical software G*Power 3 (Faul et al., 2007). CFA was performed using LISREL 8.80 (Scientific Software International, Lincolnwood, IL, United States); all other analyses were performed with IBM SPSS 25 (SPSS Inc., Chicago, IL, United States).

RESULTS

Structural Validity

The EFA run on the first sample (n = 200) yielded three factors explaining 74.41% of the variance (**Table 3**). The KMO measure of sampling adequacy proved to be extremely good (KMO = 0.92; Hutcheson and Sofroniou, 1999), and Bartlett test of sphericity proved to be highly significant ($p < 0.001$). According to Kaiser criterion Kaiser (1960), three factors had an eigenvalue >1 and explained 74.4% of the total variance. All the items respect the inclusion criteria in a factor (item's loadings greater than 0.40 and a cross-loading less than 0.10). **Table 3** shows the items' loading for the 3-factor solution.

The first factor extracted is composed of 12 items and includes the items representing behaviors related to phonological abilities (1, 2, and 3), reading (4, 5, 6, 7, and 8), and spelling ability difficulties (13, 14, and 15). In addition, item 20 (created to be representative of behaviors related to verbal memory difficulties) was included in Factor 1. The second

TABLE 3 | EFA factor loadings (n = 200), and CFA goodness-of-fit indices (n = 467).

Item content	Mean ± SD	F1	F2	F3
7. The student struggles in reading simple disyllabic words.	1.21 ± 1.62	**0.98**	−0.02	−0.09
5. When reading, the student makes many mistakes.	1.55 ± 1.77	**0.98**	0.03	−0.07
6. When reading, the student makes confusion errors between letters that have a similar sound (e.g., p-b, c-g, f-v) or that are visually similar (e.g., m-n, b-d, a-e).	1.46 ± 1.70	**0.97**	0.03	−0.06
15. The student has difficulties in writing simple disyllabic words.	1.29 ± 1.70	**0.96**	−0.04	−0.06
4. The student has a lower reading speed compared to peers.	1.81 ± 1.90	**0.90**	0.13	−0.08
3. The student struggles to create a word starting from separated phonemes (phonemic synthesis) or to identify the phonemes that compose the word (phonemic segmentation).	1.35 ± 1.75	**0.87**	−0.18	0.20
13. When writing, the student confuses letters that have a similar sound (e.g., p-b, c-g, f-v) or that are visually similar (e.g., m-n, b-d, a-e).	1.68 ± 1.71	**0.86**	0.00	0.08
20. The student has difficulties in memory tasks (poems, nursery rhymes, months).	1.37 ± 1.70	**0.74**	0.14	0.00
1. The student has difficulty in oral expression (difficulties can concern phonological and articulatory aspects and/or morphosyntactic production)	1.42 ± 1.85	**0.65**	−0.21	0.39
8. When reading, the student gets tired quickly.	1.36 ± 1.77	**0.63**	0.34	−0.03
2. In oral communication, the student uses a limited number of words and tends to use always the same words (restricted vocabulary).	1.54 ± 1.89	**0.62**	−0.13	0.32
14. When writing, the student tends to reverse the letters (e.g., he writes “fo” instead of “of”).	1.27 ± 1.58	**0.59**	0.08	0.11
16. The student is easily distracted.	2.43 ± 1.93	−0.08	**0.88**	0.12
17. The student gets tired easily.	1.78 ± 1.86	0.17	**0.81**	−0.02
18. The student moves a lot on the chair while doing homework, plays with the objects on the table, etc.	2.18 ± 1.94	−0.24	**0.80**	0.24
19. The student takes much longer than others to complete classroom activities.	2.05 ± 2.12	0.33	**0.67**	−0.10
11. The student shows difficulties in managing spaces in the paper (poor awareness of lines, squares, margins of the paper).	1.39 ± 1.71	0.05	0.12	**0.79**
12. The student shows difficulties in fine motor skills (e.g., using scissors, fastening buttons).	1.32 ± 1.68	0.01	0.17	**0.79**
9. The student’s handwriting is difficult to read.	1.02 ± 1.42	0.04	0.15	**0.71**
10. The student holds the pencil/pen having difficulty or inadequately.	0.92 ± 1.39	0.04	0.01	**0.57**
Eigenvalues		11.98	2.02	0.88
Eigenvalues after Promax rotation		11.11	7.3	7.6
Cronbach α		0.97	0.91	0.89

Fit indices	χ^2 (167)	S-B χ^2 (167)	RMSEA [90% CI]	CFI
	1,534.07*	881.143*	0.096 [0.090, 0.10]	0.98

*$p < 0.001$. In bold the highest items’ loading for the 3-factor solution. S-B X^2, Satorra-Bentler scaled X^2; RMSEA, Root Mean Square Error of Approximation; CFI, Comparative Fit Index.

factor included the four items describing behaviors typically related to attention difficulties (16, 17, 18, and 19). Finally, the third factor included four items (9, 10, 11, and 12) representing potential fine motor skills difficulties. Skewness and kurtosis computed on the three-factor scores indicated approximately normal univariate distributions, being lower than |2| (Tabachnick and Fidell, 2006); skewness (SE = 0.17) was between 0.58 and 1.27, and kurtosis (SE = 0.34) between −0.11 and 0.95.

This three-factor model was tested on the second sample (n = 467) using CFA. Results indicated an acceptable fit to the data, with all indices close to the expected value (**Table 3**). Each item loaded highly (>0.70) and significantly (p < 0.001) on its designated factor, with factor loadings in the 0.71–0.97 range and

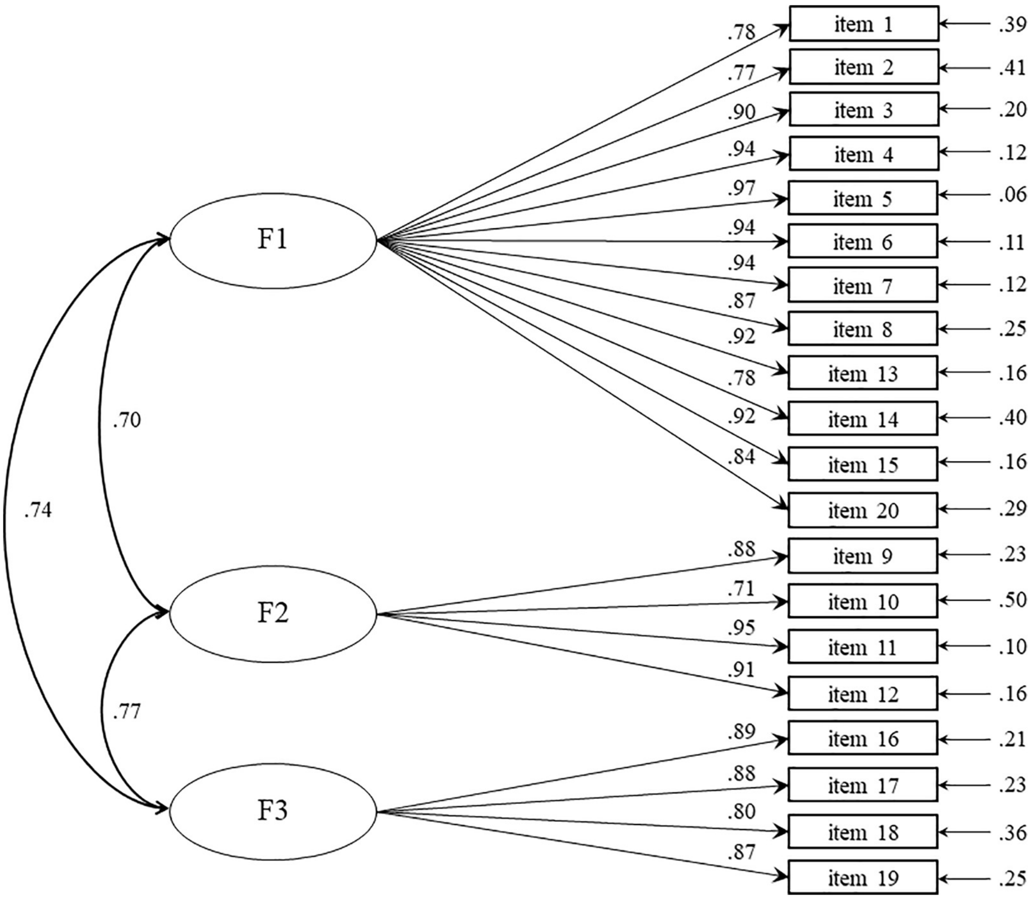


FIGURE 1 | Measurement model with standardized parameters (n = 467).

error variances in the 0.06–0.50 range (**Figure 1**). Latent variables were positively, strongly correlated ($p < 0.001$).

In order to evaluate possible alternative models explaining specificity or overlapping between the investigated domains, the three-factor model confirmed by CFA was compared to the one-factor model solution (representing unique general factor as an expression of the severity of the difficulty) and to a five-factor model solution (representing the specificity of each specific domain investigated) by CFA (**Table 4**).

All the models have shown significant χ^2 indices, suggesting that no good model fits. However, as suggested by many authors, the χ^2 test is widely recognized to be problematic (Jöreskog, 1969; Kim and Mueller, 1978; Bentler, 1990). It is sensitive to sample size, and it becomes more difficult to obtain a non-significant test as the number of cases increases. Analyzing the difference between the CFI in the three models, it is possible to see as the three- and five-factor models have very similar CFIs, and in both cases, CFI that resulted above the cutoff generally indicated a sign of good model fit (>0.95; Hu and Bentler, 1999). The CFI of the one-factor model is lower than the threshold of 0.95 and is significantly lower than those found for the three- and five-factor models. RMSEA indices resulted above the threshold of 0.06 commonly used to indicate a good model fit (Hu and Bentler, 1999) in all three models. However, according to the criteria suggested by MacCallum et al. (1996), a RMSEA of less than 0.10 can be considered as an indicator of a mediocre fit of the model as in the case of the three- and five-factor models, whereas, also using this criterion, the RMSEA for the one-factor model suggested a non-adequate fit model (with a RMSEA equal to 0.19). As regards Expected Cross-Validation Index (ECVI), no specific parameters for model acceptance or rejection exist for ECVI values; instead, this statistic assesses the likelihood that a model cross-validates across similar sized samples from the same population. In other words, the ECVI is used to compare competing models, with smaller values suggestive of greater generalizability (Byrne, 1998). When a model has a lower ECVI value, and when the ECVI value for a competing model is above the upper 90% confidence limit of the first model, it can be concluded with greater confidence that the first is the better of the two competing models (O'Rourke and Hatcher, 2013). In this case, the ECVI of the one-factor model was above both of the upper 90% CI of the three-factor model and the upper 90% CI of five-factor model; these results suggested preferring the three- and five-factor models to the one-factor one. As regards the parsimony fit indices, it is possible to note as the three-factor model showed the lowest

TABLE 4 | CFA goodness-of-fit indices ($n = 467$) for the one-factor model, the three-factor model, and the five-factor model.

Model	n. par	χ^2 (*df*)	CFI	RMSEA (CI)	ECVI (CI)	AIC	CAIC	BIC
One-factor	40	3,203.63* (170)	0.752	0.196 (0.190, 0.202)	7.046 (6.660 7.449)	3,283.63	3,489.48	3,449.48
Three-factor	63	1,534.07* (167)	0.981	0.096 (0.090, 0.10)	2.161 (1.927; 2.326)	1,007.14	1,331.36	1,268.36
Five-factor	70	1,444.53* (160)	0.981	0.100 (0.093, 0.105)	2.228 (1.990; 2.395)	1,038.07	1,398.31	1,328.31

*N. par, number of parameters estimated; CFI, Comparative Fit Index; RMSEA, Root Mean Square Error of Approximation; ECVI, Expected Cross-Validation Index; AIC, Akaike Information Criterion; CAIC, consistent Akaike Information Criterion; BIC, Bayesian Information Criterion. *p < 0.05.*

values for Akaike Information Criterion, Bayesian Information Criterion, and consistent Akaike Information Criterion, whereas the five-factor solution showed the highest but similar values, the one-factor model presented significantly highest values for all the indices. These results suggested preferring the three-factor model. Observing the number of estimated parameters, the most parsimonious model was the one-factor model, with 40 estimated parameters, followed by the three-factor model (63) and by the five-factor model (70).

Summarizing all the information on the model fit indices, both for the three-factor model and for the five-factor model, the fit to the data was acceptable; in fact, both the three- and five-factor models showed fit indices close to the expected value. The one-factor model showed the worst fit indices and resulted to be the model that less fit the data. Despite that the three- and five-factor models resulted to be very alike, obtaining very similar fit indices, the evaluation of the information criteria indices suggested to prefer the three-factor model to the five-factor model.

Finally, following the parsimony principle (essential because it helps discriminate the signal from the noise, allowing better prediction and generalization to new data; Vandekerckhove et al., 2015) and bearing in mind that the one-factor model has shown the worst fit indices with respect to the three- and five-factor models, the three-factor model, showing the best fit indices and having 63 parameters estimated (against the 70 of the five-factor model), turned out to be the model that fits the data better and that better explains the dimensionality of the analyzed data.

However, because latent variables in the three-factor model were highly correlated, a higher-order model with one second-order general factor and three first-order factors was also estimated. Such a second-order model was statistically equivalent to the model with three correlated factors, thus yielding exactly the same number of estimated parameters, fitted residuals and model fit statistics. The second-order factor loadings associated with the general factor were 0.82 for F1, 0.85 for F2, and 0.90 for F3 ($p < 0.001$). The unique first-order factor residual variances were all positive and significant, being 0.33 for F1, 0.27 for F2, and 0.19 for F3 ($p < 0.001$). Because the amount of variance associated with a first-order factor's residual decreases as the first-order factor loading onto the general factor increases, a statistically significant residual variance indicates that a dimension is, at least partly, unique or separable (Gignac and Kretzschmar, 2017). Altogether, results support the plausibility of both a general CI-FRA factor and three unique first-order factors, corresponding to phonological, reading and spelling abilities, attention competences, and fine motor skills. The three first-order factor scores provide information about teacher-assessed abilities in specific domains, whereas the global (second-order) CI-FRA score provides summary information on students' learning disorders as referred by the teacher.

Reliability

Reliability estimates were adequate. In the first ($n = 220$) and second ($n = 467$) subsamples, Cronbach α's were 0.97 and 0.98 for Factor 1, 0.89 and 0.92 for Factor 2, and 0.91 and 0.92 for Factor 3, respectively. All corrected item-total correlations were >0.50, being in the 0.74–0.94 range for Factor 1, 0.57–0.87 for Factor 2, and 0.74–0.89 for Factor 3. Test–retest reliability estimate over a 3-month period ($n = 68$) was acceptable for all the three factors. ICC of 0.73 (95% CI [0.66, 0.79]) was found for Factor 1, ICC of 0.69 (95% CI [0.61, 0.76]) for Factor 2, and an ICC of 0.67 (95% CI [0.59, 0.74]) for Factor 3.

Predictive and Concurrent Validity and Sensitivity

Spearman ρ correlation analysis evidenced significant positive correlations between CI-FRA and accuracy and speed in standardized reading test and between CI-FRA scores and accuracy in spelling (**Table 5**).

Spearman ρ correlation analysis between CI-FRA subscales and total score and the different cognitive measures showed significant correlation with phonological awareness scales (CMF), NWR, digit span, naming colors (RAN), VS, and Raven scale (CPM). As **Table 5** shows, all the CI-FRA subscales and the total CI-FRA score were correlated with the phonological awareness scale (CMF) subtests [phonemic synthesis, deletion, segmentation and verbal fluency test with phonemic facilitation (FAS)]. Significant positive correlations were found also between CI-FRA and Non-Word Repetition, as well as for CI-FRA and Digit Span test. Speed in naming colors (RAN) was significantly correlated with all the CI-FRA subtests and the CI-FRA total score except for language subscale, whereas the accuracy of RAN was not correlated with CI-FRA subscales and total score as well as visual search (VS) speed and accuracy. Raven score (CPM) was correlated with all the CI-FRA subscales and total score. Overall, the significant correlation index varied from 0.21 to 0.70, indicating small to moderate correlations between CI-FRA and the aforementioned cognitive measures.

The sensitivity and specificity of CI-FRA were analyzed by means of ROC curves. In the ROC curve, the state variable was created on the basis of standardized clinical measures for each specific learning disorder: those children scoring equal

TABLE 5 | Spearman ρ correlation analysis results to evaluate concurrent validity between CI-FRA subscales and total score and the different cognitive measures.

		CI-FRA					
		Language	**Reading**	**Spelling**	**Motor skills**	**Visuospatial attention**	**Total score**
Cognitive measures	CMF_synthesis	0.500**	0.617**	0.551**	0.388**	0.550**	0.590**
	CMF_deletion	0.491**	0.567**	0.473**	0.369**	0.492**	0.534**
	CMF_segmentation	0.461**	0.540**	0.497**	0.402**	0.455**	0.534**
	CMF_FAS	0.609**	0.701**	0.550**	0.550**	0.637**	0.671**
	NWR	0.470**	0.535**	0.493**	0.254**	0.446**	0.496**
	Digit Span	0.377**	0.449**	0.460**	0.257**	0.482**	0.475**
	RAN_speed	0.173	0.289**	0.287**	0.208*	0.270**	0.293**
	RAN_accuracy	0.052	0.010	0.024	0.030	0.066	0.041
	VS_speed	0.137	0.084	0.085	0.027	0.039	0.095
	VS_accuracy	0.090	0.119	0.089	0.162	0.111	0.116
	CPM	0.443**	0.461**	0.429**	0.360**	0.371**	0.469**
	Text reading_speed	0.333**	0.511**	0.437**	0.324**	0.388**	0.444**
	Text reading_accuracy	0.494**	0.580**	0.573**	0.474**	0.520**	0.574**
	Words reading_speed	0.336**	0.544**	0.447**	0.363**	0.442**	0.478**
	Words reading_accuracy	0.487**	0.562**	0.561**	0.345**	0.574**	0.570**
	Spelling_accuracy	0.411**	0.467**	0.444**	0.263**	0.434**	0.460**
	Handwriting_fluidity	0.280*	0.416**	0.264*	0.332**	0.399**	0.383**

*CMF_synthesis, Metaphonological skills_Phonemic Synthesis; CMF_deletion, Metaphonological skills_Deletion of the initial syllable; CMF_segmentation, Metaphonological skills_Segmentation test; CMF_FAS, Metaphonological skills_Verbal fluency test with phonemic facilitation; Digit Span, Digit Span, WISC-IV subtest; VS_speed, Visual Search Objects_speed; VS_accuracy, Visual Search Objects_accuracy; RAN_speed, Rapid Automatization Naming Colors_speed; RAN_accuracy, Rapid Automatization Naming Colors_accuracy; NWR, Non-Word Repetition; CPM, Raven's Colored Progressive Matrices; Text reading_speed, speed in reading a written text (MT, Cornoldi et al., 2018); Text reading_accuracy, accuracy in reading a written text (MT, Cornoldi et al., 2018); Words reading_speed, speed in reading a words list (DDE-2, Sartori et al., 2007); Words reading_accuracy, accuracy in reading a words list (DDE-2, Sartori et al., 2007); Spelling_accuracy, accuracy in writing non-words list (DDE-2, Sartori et al., 2007); Handwriting_fluidity, fluidity in handwrite productions (BVSCO, Tressoldi et al., 2013). **p < 0.01, *p < 0.05.*

to or less than the clinical cutoff score at the specific reading and spelling tests (1 SD for reading text, 2 SD for word lists) were considered as having learning disorders. This clinical assessment was collected by clinicians at the end of second grade. Considering the reading abilities; we used two gold standard tests for the reading and spelling disorder diagnosis: word list test and text reading test (including both accuracy and speed). The standardized tests evidenced that the 4% of children (4/102) had deficit in word reading speed, and the 5% (5/100) had spelling deficit in non-word spelling test. In details, the ROC curves' areas were all significant (Word List Reading Speed Area under the ROC curves = 0.97; p = 0.001; word list reading accuracy area under the ROC curves = 0.89; p = 0.001; text reading speed area under the ROC curves = 0.77; p = 0.001; text reading accuracy area under the ROC curves = 0.88; p < 0.001; spelling area under the ROC curves = 0.80; p = 0.020), indicating a promising applicability of CI-FRA as screening test for reading and spelling disorders. However, the highest sensitivity and specificity were obtained for the speed in word reading test (**Figure 2**). This result, even if preliminary, suggested using as a possible cutoff for speed in reading words a value at CI-FRA total score greater than 75 (having sensitivity = 0.99, and specificity = 0.95).

DISCUSSION

The main aim of the present study was to propose and test a new screening tool for teachers able to detect early signs of reading and spelling difficulties. To this aim, the dimensionality, validity, sensitivity, and reliability of the CI-FRA checklist were analyzed. The relation between the CI-FRA scores and

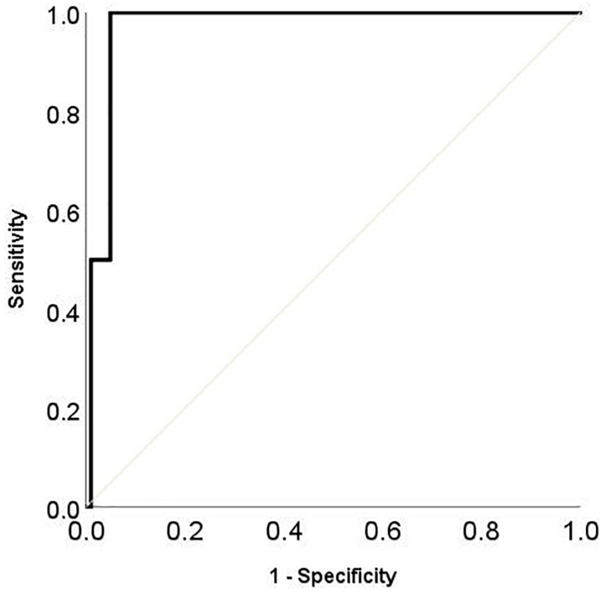


FIGURE 2 | ROC curves representing sensitivity and specificity of CI-FRA total score in predicting the speed in standardized word list reading test.

the results obtained by standardized Italian tests used to assess learning and general cognitive abilities were studied to check for concurrent validity. Preliminary results were obtained with ROC curves to confirm the sensitivity and specificity of the CI-FRA. Findings indicated that CI-FRA shows good internal validity, according to the multifactorial model of reading and spelling disorders. Moreover, high predictive validity, good test–retest reliability, acceptable concurrent validity, and overall adequate psychometric properties were reported.

Exploratory factor analyses provided a three-factor solution representing the dimension of language, phonological awareness, and reading–spelling competences (1), attention (2), and fine motor skills (3), which was confirmed as adequate with CFA. A comparison between the three theoretical models shown as the three-factor model is the one that best represents and fits the original data having all estimated indices close to the expected values. This result suggests that the difficulties emerging in the early stages of reading and spelling acquisition pertain to three different domains. In particular, the three-factor model accomplishes the possible overlapping of specific abilities (phonological domain: reading, spelling, phonological abilities), as well as other important skills investigated (attention and fine motor skills). The model, showing an intermediate specificity along this continuum, could explain the frequent overlapping of different cognitive difficulties in reading and spelling disorders.

The first factor included items related to early reading and spelling abilities (items 4–8 and items 13–15), language and phonological competences (items 1–3), and one item related to short-term verbal ability (item 20). Taken together, these items converge in the dimension of phonological awareness. It is well known that in an early phase, inadequate development of reading, spelling, and language abilities is the principal aspect of the evolution of learning skills. Moreover, verbal memory is strongly linked to linguistic and phonological skills (Adams and Gathercole, 2000) and relates to the difficulty in learning poems, nursery rhymes, and months.

The second factor included items concerning attentional difficulties, mainly related to the tendency to distraction and slight hyperactive attitude (items 16 to 19). Many studies demonstrated the relevant relation existing between attentional problems (in particular visuospatial attention) and learning disorders. Some authors stated that visuospatial attention is decisive in the initial processing of raw visual information that is a process necessary for the elaboration, synthesis, and reading of graphemes, and then a process necessary for the development of reading abilities (Facoetti and Molteni, 2001; Facoetti et al., 2010). Moreover, given the high frequency of visual attention span disorder in dyslexic children, visual attention deficit could be considered as a predictor of future reading difficulties (Valdois et al., 2004; Bosse et al., 2007; Bosse and Valdois, 2009).

The third factor included items related to fine motor skills abilities, specifically regarding difficulties in managing the space in the paper, in fine motricity, and handwriting ability (items 9–12). The predictive role of fine motor skills for the future development of learning abilities (i.e., reading and spelling) has been demonstrated by many studies (Grissmer et al., 2010; Roebers and Jäger, 2014; Cameron et al., 2016) showing that motor skills and language difficulties are often interconnected (Viholainen et al., 2002). Indeed, although motor skills impairment can overlap with executive function deficits (Roebers and Jäger, 2014), an early evaluation of these aspects may still represent a useful indicator of future reading and spelling disorders.

Significant correlations between the CI-FRA scales and scores obtained from standardized tests commonly used to evaluate developmental reading and spelling abilities (measured in the second grade) demonstrated good predictive validity. Moreover, CI-FRA total score and subscale scores correlated with scores obtained from standardized neuropsychological tests. In detail, abilities commonly considered as predictive factors of reading and spelling acquisition (metaphonological and phonological awareness, working memory, phonological memory, speech production) were significantly correlated with all CI-FRA subscales and total score. This confirmed that the CI-FRA scores are in line with those obtained with standardized tests. Even if the correlation analysis results showed moderate effect sizes, these results are in line with former literature investigating the link between different phonological abilities and reading ability (Moll et al., 2014). It is important to mention that, especially in the first 2 years of alphabetization, pupils are highly diverse in the development of reading and spelling abilities. It is therefore not surprising to find high variability in data collected from a primary school sample. Overall, the CI-FRA showed correlations both with domain-specific abilities (i.e., reading and spelling) and with other cognitive abilities (i.e., non-verbal intellectual abilities and motor skills).

Furthermore, even if the sample size was small, the ROC curves showed promising results. Crucially, the CI-FRA total score is highly sensitive for predicting the presence of word reading speed deficit, which is the most important parameter for distinguishing the reading deficit in Italian orthography (Zoccolotti et al., 1999).

LIMITATIONS OF THE STUDY

The present preliminary results indicate a promising value of the CI-FRA checklist for primary school teachers. Nevertheless, some limitations should be taken into account. The first limitation relates to the small sample size. The ROC curves are referred to a very small sample of 5 to 10 cases with specific reading or spelling disorder. A larger sample size is therefore required to confirm the sensitivity and specificity here reported. Moreover, a larger clinical sample will allow verifying the consistency of the CI-FRA cutoff score. A second limitation concerns the geographic origin of the sample. All the schools involved in the present study were from the Emilia-Romagna region, located in the north of Italy. Even if the representativeness of the sample compared to the general Italian school population is ensured (see Barbiero et al., 2019), samples from other Italian regions should be included to endure the representativeness of the sample.

CONCLUSION

The CI-FRA checklist is conceived as a brief screening tool for teachers for the evaluation of the early signs of reading and spelling disorders. The challenge of a fast tool is to be not only as simple as possible, but also methodologically well-founded. The preliminary evaluation of the psychometric properties of the CI-FRA confirmed that it could be considered a good screening tool for reading and spelling disorders. The CI-FRA includes a general score that could be used as a good indicator of reading and spelling difficulties, as well as specific subscales corresponding to more general abilities (i.e., attention, fine motor, and executive skills) that allow defining the profile of each pupil. The simplicity of the checklist and the reliability allow using the CI-FRA also for the evaluation of the evolution of the pupil's profile and of the overall class' composition. The accordance between CI-FRA and cognitive tests highlights the possibility to recognize not only a general fragility in the prerequisites of learning but also the specific early signs for reading and spelling developmental process (Catts et al., 2015).

Importantly, the results of the present research could be considered as preliminary evidence for the development of other checklists for the early screening of learning disorders. Such tools could help teachers to plan early intervention and eventually inform families and clinicians about the possible need for an in-depth evaluation. Crucially, such a tool could represent a significant advantage also for the National Health Service.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The research was carried out in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

SGi, MB, and LM designed the experiments. SGi, MB, and AA wrote the manuscript. SGi, AA, SM, and MM collected the data. SGi, MB, GC, and SGa performed the data analysis. All authors read and commented on the manuscript.

ACKNOWLEDGMENTS

Assistance in collecting data provided by Gioele Piobbico was greatly appreciated by the authors.

REFERENCES

Adams, A. M., and Gathercole, S. E. (2000). Limitations in working memory: implications for language development. *Int. J. Lang. Commun.Disord.* 35, 95–116. doi: 10.1080/136828200247278

Barbiero, C., Montico, M., Lonciari, I., Monasta, L., Penge, R., Vio, C., et al. (2019). The lost children: the underdiagnosis of dyslexia in Italy. A cross-sectional national study. *PLoS One* 14:e0210448. doi: 10.1371/journal.pone.0210448

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychol. Bull.* 107, 238–246. doi: 10.1037/0033-2909.107.2.238

Bosse, M. L., Tainturier, M. J., and Valdois, S. (2007). Developmental dyslexia: the visual attention span deficit hypothesis. *Cognition* 104, 198–230. doi: 10.1016/j.cognition.2006.05.009

Bosse, M. L., and Valdois, S. (2009). Influence of the visual attention span on child reading performance: a cross-sectional study. *J. Res. Read.* 32, 230–253. doi: 10.1111/j.1467-9817.2008.01387.x

Byrne, B. M. (1998). *Structural Equation Modeling With LISREL, PRELIS, and SIMPLIS: Basic Concepts, Applications, and Programming*. Mahwah, NJ: Lawrence Erlbaum.

Cameron, C. E., Cottone, E. A., Murrah, W. M., and Grissmer, D. W. (2016). How are motor skills linked to children's school performance and academic achievement? *Child Dev. Perspect.* 10, 93–98. doi: 10.1111/cdep.12168

Carlson, A. G., Rowe, E., and Curby, T. W. (2013). Disentangling Fine Motor Skills' relations to academic achievement: the relative contributions of visual-spatial integration and visual-motor coordination. *J. Genet. Psychol.* 174, 514–533. doi: 10.1080/00221325.2012.717122

Catts, H. W., Nielsen, D. C., Bridges, M. S., Liu, Y. S., and Bontempo, D. E. (2015). Early identification of reading disabilities within an RTI framework. *J. Learn. Disabil.* 48, 281–297. doi: 10.1177/0022219413498115

Catts, H. W., Petscher, Y., Schatschneider, C., Sittner Bridges, M., and Mendoza, K. (2009). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *J. Learn. Disabil.* 42, 163–176. doi: 10.1177/0022219408326219

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edn. New Jersey, NJ: Erlbaum.

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., et al. (2010). Selecting at-risk first-grade readers for early intervention: eliminating false positives and exploring the promise of a two-stage gated screening process. *J. Educ. Psychol.* 102, 327–340. doi: 10.1037/a0018448

Cornoldi, C., Colpo, G., and Carretti, B. (2018). *Prove MT. Kit Scuola. Classi 1-2 Primaria*. Firenze: Giunti Editore.

De Luca, M., Di Filippo, G., Judica, A., Spinelli, D., and Zoccolotti, P. (2005). *Test di Denominazione Rapida e Ricerca Visiva di Colori, Figure e Numeri*. Roma: IRCCS Fondazione Santa Lucia.

Elbro, C., and Scarborough, H. S. (2004). "Early identification," in *Handbook of Children's Literacy*, eds T. Nunes, and P. Bryant, (Dordrecht: Kluwer), 339–359.doi: 10.1007/978-94-017-1731-1_19

Facoetti, A., and Molteni, M. (2001). The gradient of visual attention in developmental dyslexia. *Neuropsychologia* 39, 352–357. doi: 10.1016/S0028-3932(00)00138-X

Facoetti, A., Trussardi, A. N., Ruffino, M., Lorusso, M. L., Cattaneo, C., Galli, R., et al. (2010). Multisensory spatial attention deficits are predictive of phonological decoding skills in developmental dyslexia. *J. Cogn. Neurosci.* 22, 1011–1025. doi: 10.1162/jocn.2009.21232

Faul, F., Erdfelder, E., Lang, A. G., and Buchner, A. (2007). G*Power 3: a fiEXIBLE statistical power analysis program for the social, behavioral and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/BF03193146

Gabrieli, J. D. E., and Norton, E. S. (2012). Reading abilities: importance of visual-spatial attention. *Curr. Biol.* 22, R298–R299. doi: 10.1016/j.cub.2012.03.041

Gignac, G. E., and Kretzschmar, A. (2017). Evaluating dimensional distinctness with correlated-factor models: limitations and suggestions. *Intelligence* 62, 138–147. doi: 10.1016/j.intell.2017.04.001

Grissmer, D., Grimm, K. J., Aiyer, S. M., Murrah, W. M., and Steele, J. S. (2010). Fine motor skills and early comprehension of the world: two new school readiness indicators. *Dev. Psychol.* 46, 1008–1017. doi: 10.1037/A0020104

Holopainen, L., Kofler, D., Koch, A., Hakkarainen, A., Bauer, K., and Taverna, L. (2020). Ci sono differenti predittori della lettura nelle lingue che hanno un'ortografia trasparente? Evidenze da uno studio longitudinale. *J. Educ. Cult. Psychol. Stud.* 21, 111–129. doi: 10.7358/ecps-2020-021-holo

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. *Struc. Equat. Modeling* 6, 1–55. doi: 10.1080/10705519909540118

Hutcheson, G., and Sofroniou, N. (1999). *The Multivariate Social Scientist: Introductory Statistics Using Generalized Linear Models*. Thousand Oaks, CA: Sage Publication, doi: 10.4135/9780857028075

Istituto Superiore di Sanità (2011). *Disturbi Specifici Dell'apprendimento Consensus Conference. sui Disturbi Specifici dell'Apprendimento*. Roma: Istituto Superiore di Sanità, 6–7.

Johnson, E. S., Jenkins, J. R., Petscher, Y., and Catts, H. W. (2009). How can we improve the accuracy of screening instruments? *Learn. Disabil. Res. Pract.* 24, 174–185. doi: 10.1111/j.1540-5826.2009.00291.x

Jöreskog, K. G. (1969). A general approach to confirmatory factor analysis. *Psychometrika* 34, 183–202. doi: 10.1007/BF02289343

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educ. Psychol. Meas.* 20, 141–151. doi: 10.1177/001316446002000116

Kim, J. O., and Mueller, C. W. (1978). *Factor Analysis: Statistical Methods and Practical Issues*. Beverly Hills, CA: Sage.

Kirby, J. R., Parrila, R. K., and Pfeiffer, S. L. (2003). Naming speed and phonological awareness as predictors of reading development. *J. Educ. Psychol.* 95, 453–464. doi: 10.1037/0022-0663.95.3.453

Kline, R. B. (1998). *Principles and Practice of Structural Equation Modeling*. New York, NY: Guilford.

Landerl, K., Ramus, F., Moll, K., Lyytinen, H., Leppänen, P. H., Lohvansuu, K., et al. (2013). Predictors of developmental dyslexia in European orthographies with varying complexity. *J. Child Psychol. Psychiatry* 54, 686–694. doi: 10.1111/jcpp.12029

Landerl, K., and Wimmer, H. (2008). Development of word reading fiuency and spelling in a consistent orthography: an 8-year follow-up. *J. Educ. Psychol.* 100, 150–161. doi: 10.1037/0022-0663.100.1.150

Lervåg, A., Bråten, I., and Hulme, C. (2009). The cognitive and linguistic foundations of early reading development: a Norwegian latent variable longitudinal study. *Dev. Psychol.* 45, 764–781. doi: 10.1037/a0014132

Lorusso, M. L., Vernice, M., Dieterich, M., Brizzolara, D., Mariani, E., De Masi, S., et al. (2014). The process and criteria for diagnosing specific learning disorders: indications from the consensus conference promoted by the italian national institute of health. *Ann. Super. Sanità* 50, 77–89. doi: 10.4415/ANN-14-01-12

Lukov, L., Friedmann, N., Shalev, L., Khentov-Kraus, L., Shalev, N., Lorber, R., et al. (2015). Dissociations between developmental dyslexias and attention deficits. *Front. Psychol.* 5:1501. doi: 10.3389/fpsyg.2014.01501

MacCallum, R. C., Browne, M. W., and Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychol. Methods* 2, 130–149. doi: 10.1037/1082-989X.1.2.130

Marotta, L., Ronchetti, C., Trasciani, M., and Vicari, S. (2008). *CMF: Valutazione delle Competenze Metafonologiche*. Trento: Erickson.

Mash, E. J., and Wolfe, D. A. (2002). *Abnormal Child Psychology*, 2nd Edn. Belmont, CA: Wadsworth.

Menghini, D., Finzi, A., Benassi, M., Bolzani, R., Facoetti, A., Giovagnoli, S., et al. (2010). Different underlying neurocognitive deficits in developmental dyslexia: a comparative study. *Neuropsychologia* 48, 863–872. doi: 10.1016/j.neuropsychologia.2009.11.003

Moll, K., Ramus, F., Bartling, J., Bruder, J., Kunze, S., Neuhoff, N., et al. (2014). Cognitive mechanisms underlying reading and spelling development in five European orthographies. *Learn. Instr.* 29, 65–77. doi: 10.1016/j.learninstruc.2013.09.003

Nunnally, J. C. (1978). *Psychometric theory*, 2nd Edn. New York, NY: McGraw Hill.

O'Rourke, N., and Hatcher, L. (2013). *A Step-by-Step Approach to Using SAS§for Factor Analysis and Structural Equation Modeling*, 2nd Edn. Cary, NC: SAS Institute Inc.

Pagani, L., and Messier, S. (2012). Links between motor skills and indicators of school readiness at kindergarten entry in urban disadvantaged children. *J. Educ. Dev. Psychol.* 2, 95–107. doi: 10.5539/jedp.v2n1p95

Pennington, B. F. (2006). From single to multiple deficit models of developmental disorders. *Cognition* 101, 385–413.

Poskiparta, E., Niemi, P., Lepola, J., Ahtola, A., and Laine, L. (2003). Motivational-emotional vulnerability anddifficulties in learning to read and spell. *Br. J. Educ. Psychol.* 73, 187–206. doi: 10.1348/00070990360626930

Poulsen, M., Juul, H., and Elbro, C. (2015). Multiple mediation analysis of the relationship between rapid naming and reading. *J. Res. Read.* 38, 124–140. doi: 10.1111/j.1467-9817.2012.01547.x

Poulsen, M., Nielsen, A. M. V., Juul, H., and Elbro, C. (2017). Early identification of reading difficulties: a screening strategy that adjusts the sensitivity to the level of prediction accuracy. *Dyslexia* 23, 251–267. doi: 10.1002/dys.1560

Puolakanaho, A., Ahonen, T., Aro, M., Eklund, K., Leppänen, P. H., Poikkeus, A. M., et al. (2008). Developmental links of very early phonological and language skills to second grade reading outcomes: strong to accuracy but only minor to fluency. *J. Learn. Disabil.* 41, 353–370. doi: 10.1177/0022219407311747

Raven, J. C. (1994). *CPM: Coloured Progressive Matrices: Serie A-AB-B*. Firenze: Organizzazioni Speciali.

Roebers, C. M., and Jäger, K. (2014). The relative importance of fine motor skills, intelligence, and executive functions for first grader's reading and spelling skills. *Perspect. Lang. Lit.* 40, 13–17.

Sartori, G., Job, R., and Tressoldi, P. E. (2007). *DDE-2. Batteria per la Valutazione della Dislessia e Disortografia Evolutiva-2*. Firenze: Organizzazioni Speciali.

Snowling, M. J. (2013). Early identification and interventions for dyslexia: a contemporary view. *J. Res. Special Educ. Needs* 13, 7–14. doi: 10.1111/j.1471-3802.2012.01262.x

Streiner, D. L., and Norman, G. R. (2008). *Health Measurement Scales: A Practical Guide to Their devElopment and use*, 4th Edn. Oxford: Oxford University Press.

Tabachnick, B. G., and Fidell, L. S. (2001). *Using Multivariate Statistics*. Boston, MA: Allyn and Bacon A Pearson Education Company.

Tabachnick, B. G., and Fidell, L. S. (2006). *Using Multivariate Statistics*, 5th Edn. Boston: Allyn & Bacon.

Torgesen, J., Rashotte, C., and Alexander, A. W. (2001). "Principles of fluency instruction in reading: relationships with established empirical outcomes," in *Dyslexia, Fluency, and the Brain*, ed. M. Wolf, (Timonium, MD: York Press), 333–355.

Torgesen, J. K. (2005). "Recent discoveries on remedial interventions for children with dyslexia," in *The Science of Reading*, eds M. J. Snowling, and C. Hulme, (Oxford: Blackwell Publishing), 521–537.doi: 10.1002/9780470757642.ch27

Tressoldi, P. E., Cornoldi, C., and Re, A. M. (2013). *BVSCO-2. Batteria per la Valutazione della Scrittura e della Competenza Ortografica – 2*. Firenze: Giunti OS.

Valdois, S., Bosse, M. L., and Tainturier, M. J. (2004). The cognitive deficits responsible for developmental dyslexia: review of evidence for a selective visual attentional disorder. *Dyslexia* 10, 339–363. doi: 10.1002/dys.284

Vandekerckhove, J., Matzke, D., and Wagenmakers, E. J. (2015). "Model comparison and the principle of parsimony," in *Oxford Handbook of Computational and Mathematical Psychology*, eds J. Busemeyer, J. Townsend, Z. J. Wang, and A. Eidels, (Oxford: Oxford University Press), 300319.doi: 10.1093/oxfordhb/9780199957996.013.14

Vicari, S. (2007). *PROMEA: Prove di Memoria e Apprendimento per l'Età Evolutiva*. Firenze: Giunti Psychometrics.

Viholainen, H., Ahonen, T., Cantell, M., Lyytinen, P., and Lyytinen, H. (2002). Development of early motor skills and language in children at risk for familial dyslexia. *Dev. Med. Child Neurol.* 44, 761–769. doi: 10.1111/j.1469-8749.2002.tb00283.x

Wagner, W. G. (2003). *Counseling, psychology, and children: A multidimensional approach to intervention*. Upper Saddle River, NJ, Pearson Education.

Wechsler, D. (2005). *WISC-IV: Wechsler Intelligence Scale for Children-IV*. Firenze: Giunti Psychometrics.

Weston, R., and Gore, P. A. Jr. (2006). A brief guide to structural equation modeling. *Couns. Psychol.* 34, 719–751. doi: 10.1177/0011000006286345

Ziegler, J. C., Bertrand, D., Tóth, D., Csépe, V., Reis, A., Faísca, L., et al. (2010). Orthographic depth and its impact on universal predictors of reading: a cross-language investigation. *Psychol. Sci.* 21, 551–559. doi: 10.1177/0956797610363406

Ziegler, J. C., Perry, C., and Zorzi, M. (2019). "Modeling the variability of developmental dyslexia," in *Developmental Dyslexia Across Languages and Writing Systems*, eds L. Verhoeven, C. Perfetti, and K. Pugh, (Cambridge: Cambridge University Press), 350–371. doi: 10.1017/9781108553377.016

Zoccolotti, P., De Luca, M., Di Pace, E., Judica, A., Orlandi, M., and Spinelli, D. (1999). Markers of developmental surface dyslexia in a language (Italian) with high grapheme–phoneme correspondence. *App. Psycholinguistics* 20, 191–216. doi: 10.1017/S0142716499002027

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 18 December 2020
doi: 10.3389/fpsyg.2020.500039

A Rasch Model and Rating System for Continuous Responses Collected in Large-Scale Learning Systems

Benjamin Deonovic[1]*, *Maria Bolsinova*[2], *Timo Bechger*[3] and *Gunter Maris*[3,4]

[1] ACT, Inc., Iowa City, IA, United States, [2] Department of Methodology and Statistics, Tilburg University, Tilburg, Netherlands, [3] ACT, Inc., Amsterdam, Netherlands, [4] Department of Psychological Methods, University of Amsterdam, Amsterdam, Netherlands

An extension to a rating system for tracking the evolution of parameters over time using continuous variables is introduced. The proposed rating system assumes a distribution for the continuous responses, which is agnostic to the origin of the continuous scores and thus can be used for applications as varied as continuous scores obtained from language testing to scores derived from accuracy and response time from elementary arithmetic learning systems. Large-scale, high-stakes, online, anywhere anytime learning and testing inherently comes with a number of unique problems that require new psychometric solutions. These include (1) the cold start problem, (2) problem of change, and (3) the problem of personalization and adaptation. We outline how our proposed method addresses each of these problems. Three simulations are carried out to demonstrate the utility of the proposed rating system.

Keywords: Rasch model, longitudinal data analysis, rating system, item response theory (IRT), learning and assessment system, continuous response measurement

OPEN ACCESS

Edited by:
Jason C. Immekus,
University of Louisville, United States

Reviewed by:
Oscar Lorenzo Olvera Astivia,
University of South Florida,
United States
Stefano Noventa,
University of Tübingen, Germany

****Correspondence:***
Benjamin Deonovic
bdeonovic@gmail.com

Specialty section:
This article was submitted to Quantitative Psychology and Measurement,
a section of the journal Frontiers in Psychology

Received: *23 September 2019*
Accepted: *19 November 2020*
Published: *18 December 2020*

Citation:
Deonovic B, Bolsinova M, Bechger T and Maris G (2020) A Rasch Model and Rating System for Continuous Responses Collected in Large-Scale Learning Systems.
Front. Psychol. 11:500039.
doi: 10.3389/fpsyg.2020.500039

1. INTRODUCTION

Large-scale, high-stakes, online, anywhere anytime learning and testing inherently comes with a number of unique problems that require new psychometric solutions. First, there is the *cold start problem*: the system needs to start without data. The traditional solution is to start with a large item bank calibrated to an appropriate *Item Response Theory (IRT) model*, which is expensive and challenging as it requires large numbers of representative test takers to respond to items under realistic testing conditions. Second, there is the *problem of change*: learner and item properties change as a cohort of learners progresses through its education. While such changes are intended, they are not easily handled by traditional psychometrics developed to assess student's ability at a single time point. Finally, there is *the problem of personalization and adaptation*: to optimally support learning, each learner follows her own path at her own pace. This will give rise to sparse, incomplete data that are not easily analyzed using likelihood-based methods. Moreover, online learning systems, such as Duolingo, for foreign languages, and Math Garden, for elementary arithmetic, generate large data sets with large number of item responses per learner as learners practice with many items over extended periods of time.

The urnings rating system was introduced by Bolsinova et al. (2020) to address these challenges, but its usefulness is limited by the fact that it assumes a Rasch model (or its generalization for polytomous data) and is tied to discrete item responses. In this paper, we extend the urnings rating system to continuous responses and illustrate its relevance for online learning systems using simulated data. Throughout, the Duolingo English Test (DET; Wagner and Kunnan, 2015; LaFlair and Settles, 2019; Maris, 2020), and Math Garden (Klinkenberg et al., 2011) will serve as motivating examples.

2. THE CONTINUOUS RASCH MODEL

Continuous responses can be obtained from a wide variety of data and functions of data. In the DET, item responses are continuous numbers between zero and one. In Math Garden, continuous responses come from a combination of accuracy and time. Other learning and assessment systems may ask users to provide their perceived certainty that the chosen response is correct (Finetti, 1965; Dirkzwager, 2003). In this paragraph, we consider a general measurement model for continuous responses. For expository purposes, we consider the responses to be between zero and one.

The model we consider is the direct extension of the Rasch model to continuous responses and we will refer it as *the continuous Rasch (CR) model*. Suppressing the person index, the CR model is defined by the following response probabilities:

$$\begin{aligned} f(\mathbf{x}|\theta) &= \prod_i f(x_i|\theta) && (1)\\ &= \prod_i \frac{\exp(x_i(\theta-\delta_i))}{\int_0^1 \exp(s(\theta-\delta_i))ds} && (2)\\ &= \prod_i \frac{(\theta-\delta_i)\exp(x_i(\theta-\delta_i))}{\exp(\theta-\delta_i)-1}, && (3)\end{aligned}$$

where θ represents learner ability and δ_i item difficulty. This is an exponential family IRT model where the sum $x_+ = \sum_i x_i$ is the sufficient statistic for ability. Note that the CR model is not new as it is equivalent[1] to the Signed Residual Time (SRT) model proposed by Maris and van der Maas (2012) and the Rasch model for continuous responses found in Verhelst (2019). The key insight is that the model can be used for any type of continuous responses. For illustration, **Figure 1** shows plots of the probability density, cumulative distribution, and expectation functions under the CR model.

For our present purpose, we will not analyze the continuous responses directly but a limited number of binary responses derived from them. We now explain how this works. If we define two new variables as follows

$$\begin{aligned} y_{i1} &= (x_i > 0.5) && (4)\\ x_{i1} &= \begin{cases} x_i - 0.5 & \text{if } y_{i1} = 1 \\ x_i & \text{if } y_{i1} = 0 \end{cases} && (5)\end{aligned}$$

we obtain conditionally independent sources of information on ability from which the original observations can be reconstructed; that is, $Y_{i1} \perp\!\!\!\perp X_{i1}|\theta$. Moreover, it is readily found that the implied measurement model for Y_{i1} is the Rasch model:

$$p(Y_{i1}=1|\theta) = p(X_i > 0.5|\theta) = \frac{\exp(0.5(\theta-\delta_i))}{1+\exp(0.5(\theta-\delta_i))} \qquad (6)$$

where the discrimination is equal to a half. The other variable, X_{i1}, is continuous with the following distribution over the interval 0 to 1/2:

$$f(x_{i1}|\theta) = \frac{(\theta-\delta_i)\exp(x_{i1}(\theta-\delta_i))}{\exp(0.5(\theta-\delta_i))-1} \qquad (7)$$

The distribution of X_{i1} and X_i thus belong to the same family, but with a different range for the values of the random variable. We can now continue to split up X_{i1} into two new variables and recursively transform the continuous response to a set of conditionally independent Rasch response variables with discriminations that halve in every step of the recursion.

If we denote the binary response variable obtained in the j-th step of the recursion by Y_{ij}, we obtain the (non-terminating) dyadic expansion (see e.g., Billingsley, 2013) of the continuous response variables into conditionally independent binary response variables, as depicted in **Figure 2**. Since the discriminations halve in every step, most of the statistical information about ability contained in the continuous response is recovered by a limited number of binary variables. If the CR model fits, then at the point where $\theta = \delta_i$ the information in the continuous response is $\frac{4}{3}$ times the information contained in Y_{i1} alone[2].

Other models have been developed for continuous responses. Notably the extensions by Samejima to the graded response models (Samejima, 1973, 1974), Müller's extension to Andrich's rating formulation (Müller, 1987), and more recently, a generalization of the SRT model (van Rijn and Ali, 2017). Estimation procedures developed for these models have all been likelihood based and quite infeasible in a learning setting where there are many people and items, and each person answers a different subset of items. For the CR model, we will therefore turn to estimation via the use of rating systems.

3. METHODS: THE URNINGS RATING SYSTEM

3.1. Classic Urnings

Adaptive online tests produce data sets with both a large number of test takers and a large number of items. Even when we analyze binary response variables, direct likelihood-based inference will not scale-up to handle these large amounts of data. We will therefore use a rating system. A rating system is a method to assess a player's strength in games of skill and track its evolution over time. Here, learners solving items are considered players competing against each other and the ratings represent the skill of the learner and the difficulty of the item.

Rating systems, such as the Elo rating system (Elo, 1978; Klinkenberg et al., 2011), originally developed for tracking ability in chess, are highly scalable but come with their own set of problems. Elo ratings, in particular, are known to have an inflated variance, and their statistical properties are not very well-understood (e.g., Brinkhuis and Maris, 2009). The urnings rating system overcomes both issues while it is still highly scalable with

[1] After re-scaling, if $X \sim \text{SRT}(\eta)$ then $Y = \frac{1}{2}(X-1) \sim \text{CR}(2\eta)$.

[2] The infinite sum $\frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \ldots$ is equal to $\frac{1}{3}$.

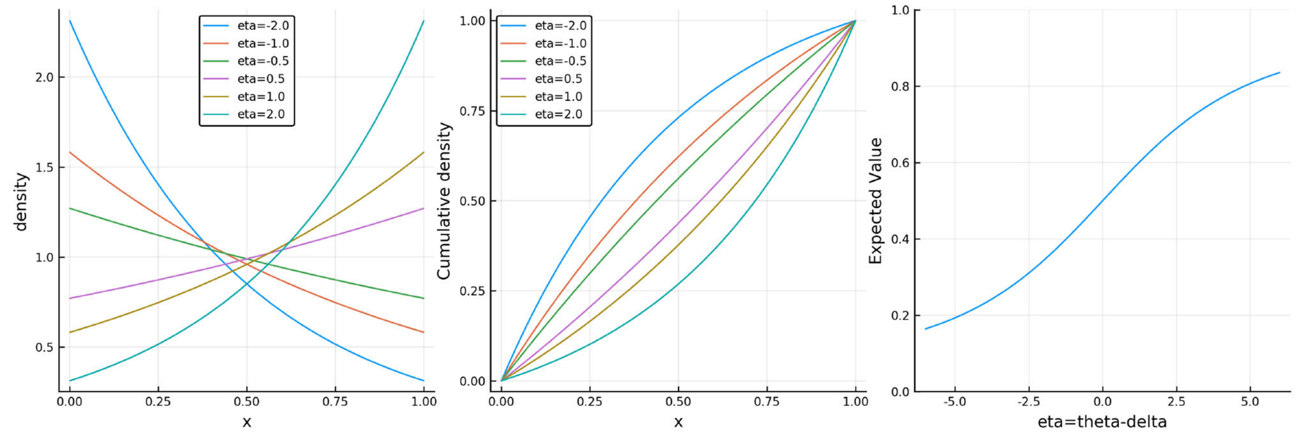
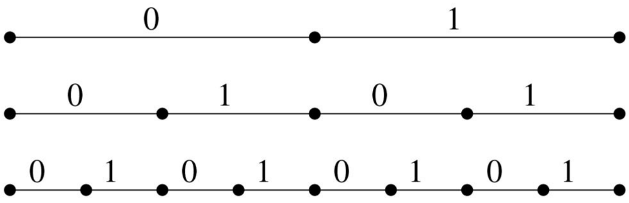


FIGURE 1 | (Left) The probability density function, (middle) the cumulative distribution function, and (right) the expectation of the continuous Rasch model where $\eta = \theta - \delta_i$.



$$P(Y_{i1} = 1|\theta) = \frac{\exp(0.5(\theta - \delta_i))}{1 + \exp(0.5(\theta - \delta_i))}$$

$$P(Y_{i2} = 1|\theta) = \frac{\exp(0.25(\theta - \delta_i))}{1 + \exp(0.25(\theta - \delta_i))}$$

$$P(Y_{i3} = 1|\theta) = \frac{\exp(0.125(\theta - \delta_i))}{1 + \exp(0.125(\theta - \delta_i))}$$

FIGURE 2 | The first three steps of a dyadic expansion of continuous responses into conditionally independent binary response variables. Each follows a Rasch model with a discrimination that halves at each subsequent step.

person and item ratings being updated after each response. In equilibrium, when neither learners nor items change, urnings are known to be binomially distributed variables, with the logits of the probability being the ability/difficulty in a Rasch model.

Urnings is a rating system where discrete parameters u_p and u_i, the "urnings," track the ability of a person and the difficulty of an item. Urnings assumes that the observed binary responses result from a game of chance played between persons and items matched-up with to probability $M_{pi}(u_p, u_i)$. The game proceeds with each player drawing a ball from an infinite urn containing red and green balls, the proportion of green balls being π_p in the person urn and π_i in the item urn. The game ends when the balls drawn are of different color and the player with the green ball wins. If the person wins, the item is solved and so the binary response corresponds to

$$X_{pi} = \begin{cases} 1 & \text{if } y_p^* = 1 \\ 0 & \text{if } y_i^* = 1 \end{cases}$$

where y_p^* and y_i^* indicate whether the green ball was drawn by the person or the item. An easy derivation shows that the observed responses follow a Rasch model:

$$\begin{aligned} p(X_{pi} &= 1) = p(y_p^* = 1, y_i^* = 0|\theta_p, \theta_i) \\ &= \frac{\pi_p(1-\pi_i)}{\pi_p(1-\pi_i) + (1-\pi_p)\pi_i} = \frac{\exp(\theta_p - \theta_i)}{1 + \exp(\theta_p - \theta_i)} \end{aligned} \quad (8)$$

where $\theta_p = \ln(\pi_p/(1-\pi_p))$ and similarly for θ_i.

The urnings rating system mimics this game using finite sized urns. For each "real" game that is played, a corresponding simulated game is played with finite urns containing, respectively u_p and u_i green balls out of n[3]. Let y_p and y_i denote the outcome of the simulated game. If the result of the simulated game does not match that of the real game, the balls drawn are replaced with the outcome of the real game. If person p lost the simulated game but solved item i, the proportion of green balls for p is thus increased while the proportion of green balls for i is decreased. This can be summarized with the updated equations

$$u_p^* = u_p + y_p^* - y_p \quad (9)$$

$$u_i^* = u_i + y_i^* - y_i \quad (10)$$

[3] Note that in practice the number of balls in the person urns and item urns don't have to be equal, but for notations sake we will keep them the same.

Match making: pair person p with item i with probability $M_{pi}(\mathbf{u})$

```
Reality:
    repeat
        $y_p^* \sim Bernoulli(\pi_p)$
        $y_i^* \sim Bernoulli(\pi_i)$
    until $y_p^* \neq y_i^*$
    return $(y_p^*, y_i^*)$

Urnings:
    repeat
        $y_p \sim Bernoulli(u_p/n)$
        $y_i \sim Bernoulli(u_i/n)$
    until $y_p \neq y_i$
    return $(y_p, y_i)$
```

Update:

$$u_p^* = u_p + y_p^* - y_p$$
$$u_i^* = u_i + y_i^* - y_i$$

Metropolis-Hastings: accept new urnings with probability:

$$\min\left(1, \frac{u_p(n-u_i)+(n-u_p)u_i}{u_p^*(n-u_i^*)+(n-u_p^*)u_i^*}\frac{M_{pi}(\mathbf{u}^*)}{M_{pi}(\mathbf{u})}\right)$$

FIGURE 3 | Urnings rating system.

where u_p^* and u_i^* are the proposed new configurations for the number of balls in each urn. This new configuration is then accepted or rejected using a Metropolis-Hastings acceptance probability to ensure that the ratings u_p/n and u_i/n converge to the proportions π_p and π_i when neither persons nor items change.

Figure 3 gives an overview of the urnings updating scheme. Bolsinova et al. (2020) prove that each of the urn proportions forms a constructed Markov-chain such that the invariant distribution of $\mathbf{u} = (u_p, u_i)^\intercal$ is a binomial distribution with parameters n and $\boldsymbol{\pi} = (\pi_p, \pi_i)^\intercal$. Note that the urn size n functions as a design parameter similar to the K-factor in Elo ratings. Larger urns mean that the system is more sensitive to change and the system converges more rapidly when the urns are smaller.

As the urnings rating system is designed to work with dichotomous response variables it is not directly applicable to the CR. However, through the use of the dyadic expansion, the continuous responses are transformed into a series of dichotomous responses. The urnings rating system can be applied directly to these dichotomous response variables that result from the dyadic expansion of the continuous responses. For a dyadic expansion of order k, we will use k urns for each person and k separate urns for each item. Due to the difference in discrimination, each person urn will be tracking $\theta_p/2^j$, where $j \in \{1, \ldots, k\}$ corresponds to the step in the dyadic expansion. Once the proportions in the urns are in equilibrium, one could combine them to get an overall estimate of θ_p. This will be similar for the item urns and item difficulty. In the simulation section below, we show how this multi-urn solution can be used to identify model misspecification.

In the next section we derive an extension to the classical urnings rating system, which tracks the θ_p using a single urn.

3.2. Extension to Urnings

Recall that the jth item in the dyadic expansion corresponds to the ability $\theta_p/2^j$. We shall see that the differences in discrimination that derive from the dyadic expansion of the continuous response variables in the CR model translate into differences in the stakes of the game. The *stakes* of the urnings algorithm correspond to how much the number of green balls can increase (or decrease). In the classic urnings algorithm, the stakes are always equal to 1. In the extended urnings algorithm we allow items with different discriminations to combine. For a dyadic expansion of order k we let the item with the lowest discrimination, the final expansion, have a stake of one. For each previous item, we double the stakes such that the jth item in the dyadic expansion has a stake of 2^{k-j}.

How does this impact the urnings update? **Figure 4** has a summary of the extended urnings rating system. The observed binary outcomes X_{pi} are now assumed to be generated by the following game of chance. The game is same as above for classic

Match making: pair person p with item i with probability $M_{pi}(\mathbf{u})$

```
Reality:
    repeat
        $y_p^* \sim Binomial(s, \pi_p)$
        $y_i^* \sim Binomial(s, \pi_i)$
    until $|y_p^* - y_i^*| = s$
    return $(y_p^*, y_i^*)$

Urnings:
    repeat
        $y_p \sim HyperGeometric(n, u_p, s)$
        $y_i \sim HyperGeometric(n, u_i, s)$
    until $|y_p - y_i| = s$
    return $(y_p, y_i)$
```

Update:

$$u_p^* = u_p + y_p^* - y_p$$

$$u_i^* = u_i + y_i^* - y_i$$

Metropolis-Hastings: accept new Urnings with probability:

$$\min\left(1, \frac{\binom{u_p}{s}\binom{n-u_i}{s} + \binom{n-u_p}{s}\binom{u_i}{s}}{\binom{u_p^*}{s}\binom{n-u_i^*}{s} + \binom{n-u_p^*}{s}\binom{u_i^*}{s}} \frac{M_{pi}(\mathbf{u}^*)}{M_{pi}(\mathbf{u})}\right)$$

FIGURE 4 | Extended Urnings rating system.

urnings, except now the game has stakes *s*. For a game with stakes *s*, the process to generate the observed outcome is to continue drawing *s* balls from both urns (y_p^* and y_i^*) until we get *s* green ones from the one urn and *s* red ones from the other. Thus

$$X_{pi} = \begin{cases} 1 & \text{if } y_p^* = s \\ 0 & \text{if } y_i^* = s \end{cases}$$

Similarly, a simulated game is played where balls are drawn (y_p and y_i) from finite urns until *s* have been drawn from one urn and none from the other (without replacement). We once again just replace these *s* balls by *s* of the color consistent with the real item response. That is, a learner stands to lose or gain *s* balls based on her response to this particular item. This is why we refer to the discriminations as stakes in this context. **Figure 4** has the updated Metropolis-Hastings acceptance probability, which is consistent with this extension. Theorem 1 provides the necessary theoretical justification for this correction. For a proof of the theorem see Appendix 1.

THEOREM 1. (Extension of Urnings Invariant Distribution) *If invariant distribution for the current configuration of balls is*

$$p(u_p, u_i) = \left(\frac{s!}{n!/(n-s)!}\right)^2 \frac{\binom{u_p}{s}\binom{n-u_i}{s} + \binom{n-u_p}{s}\binom{u_i}{s}}{\pi_p^s(1-\pi_i)^s + (1-\pi_p)^s\pi_i^s}\binom{n}{u_p} \pi_p^{u_p}(1-\pi_p)^{n-u_p}\binom{n}{u_i}\pi_i^{u_i}(1-\pi_i)^{n-u_i}$$

then the invariant distribution for the updated configuration of balls is the same, where s corresponds to the stakes.

4. SIMULATION STUDY

We provide three simulation studies to illustrate the benefits of the proposed method. Simulation 1 shows how the urnings algorithm can recover the true ability of the persons and is robust to misspecification of the model generating the continuous responses. Simulation 2 simulates a more realistic setting and aims to show how our proposed approach handles the problems inherent in learning and assessment specified in the introduction. Simulation 3 highlights the problems inherent in any model which tracks ability and difficulty: these quantities are not separately identified, and it is easy to be misled when this is not taken into account (Bechger and Maris, 2015).

4.1. Simulation 1

We simulate 1,000 persons with ability uniformly distributed between -4 and 4, $\theta_p \sim U(-4, 4)$ and 100 items with difficulty distributed between -4 and 4, $\delta_i \sim U(-4, 4)$. We simulate a total of 100 million person-item interactions in order to create a data set that is comparable to the large-scale learning system data that the model is built for. At each interaction, a randomly sampled person and item is picked. The person's response is then simulated from the CR model based on their ability and the item's difficulty. This continuous response is then expanded using the dyadic expansion of order 3 to create three dichotomous responses. These dichotomous responses are then tracked by the multi-urn system with learner urns having an urn size of 50 and the item urns having urn sizes of 100.

4.1.1. Tracking With Multiple Urns

The results of tracking the responses using the three urn system is in **Figures 5**, **6**. The colored lines in **Figure 5** correspond to the probability contours for the probability an item is answered correctly (from low probability given by purple to high probability given by red) given the urns for the person (horizontal axis) and the urns for the item (vertical axis). The smooth colored lines correspond to the expected probabilities while the noisier colored lines plotted on top correspond to the observed proportion of correct responses for every combination of Urnings from simulation 1. These plots show that there is good model fit, especially in the first urn. **Figure 6** shows the final urn proportions in the three urns plotted against the simulated ability values (on the inverse logit scale, which we call "expit"). In red is the implied 95% confidence ellipse. The blue points are within the 95% ellipse while the red ones are outside of it. Each plot in **Figure 6** also shows the correlation and the proportion of points within the ellipse (the coverage) in the plot title.

4.1.2. Model Misspecification

How robust is this approach to deviations from the assumptions? We investigate this through simulating from a different underlying model. The learning and assessment system Math Garden also has continuous responses and assumes the same distribution for the scores as we have. The scores in Math Garden are derived as a particular function of response accuracy, i.e., was the response correct or incorrect, and response time to produce the continuous item score in such a way that penalizes fast incorrect responses. Specifically, $S_i = (2Y_i - 1)(d - T_i)$ where Y_i indicates whether the response was correct or not and T_i is time when the time-limit for responding is set to d. However, the fact that time is, literally, monetized in Math Garden, may entice learners to employ a different, more economic utility-based rule. Students may value their time and thus the relationship between their response scores, accuracy, and time may be $S_i = Y_i - T_i$ in which a slow incorrect response has a large negative score. The question is can we detect that learners follow the alternative scoring rule rather than the intended one? The answer is yes. We will show this by means of a simulation.

We augment the first simulation. Rather than simulating from the CR model we will simulate from the distribution implied by the scoring rule $S_i = Y_i - T_i$. One can show that in order to simulate from this distribution we can do the following. We first simulate the response Y_i from the CR model, but if the response is <0.5, $Y_i < 0.5$, then we set the score to be $Y_i = 0.5 - Y_i$. One of the benefits of using three separate urns to track the ability is that model misfit can be detected by comparing the urns to each other. The relationship between the true urn proportions is a known function. Specifically, if θ_p are the true simulated abilities we can plot the inverse logit of $\theta_p/2$ against the inverse logit of $\theta_p/4$. If the observed own proportions don't follow this relationship there is model misfit.

Figure 7 shows the relationship between the urn proportions in urns 1 and 2 using the true generating model and the modified generating model. This figure shows that when the generating model is the modified one the model misspecification can be detected as the relationship between the urn proportions follows a U-shaped curve rather than the expected monotonic relationship.

4.2. Simulation 2

For Simulation 2 we consider a more realistic setting. Specifically, we deal with two problems in learning and assessment systems: *the problem of change* and *the problem of personalization and adaptation*. We allow the ability of the persons to change over time. Specifically, the ability changes as a function of time according to a generalized logistic function

$$\theta_p(t) = \theta_{p1} + \frac{\theta_{p2} - \theta_{p1}}{1 + \exp\left(-\alpha_p t\right)} \tag{11}$$

where t is the simulation index (from 1 to 10^8) mapped to the interval (−4,4), $\theta_{p1} \sim \mathrm{U}(-4, 4)$, $\theta_{p2} \sim \mathrm{U}(-4, 4)$, and $\alpha_p \sim \mathrm{Gamma}(1, 1)$. The item difficulty is simulated from the uniform again, $\delta_i \sim \mathrm{U}(-4, 4)$ and held constant. Once again, we simulate 10^8 responses from the continuous Rasch model where a person is (uniformly) randomly selected but now a random item is selected by choosing one with the following weights

$$\begin{aligned} M_{pi}(\mathbf{u}) \;=\; & \exp(-2(\ln{(u_p + 1)}/(n_p - u_p + 1)) \\ & -\; \ln{(u_i + 1)}/(n_i - u_i + 1))^2 \end{aligned} \tag{12}$$

where u_p corresponds to the selected person's urn proportion, u_i corresponds to item i's urn proportion, and n_p and n_i the person and item urn sizes, respectively. This results in items whose difficulty are closer to the selected person's ability being more likely selected. For this simulation we track the ability using a single urn with urn sizes of 420 for both the person and item urns.

Figure 8 shows the results for one person and one item in particular. In red is the true ability and difficulty of this person and item and the blue trace line is the urn proportion. These show that the extended Urnings rating system can track the change in ability well. We can increase the urn size if we wish to decrease the variance in the urn proportions. Another traceplot that can be generated is **Figure 9**. The leftmost plot in this figure is the probability that the response to the first dyadic expansion of a particular item is 1, the middle one is the 2nd dyadic expansion of the same person and item, and the rightmost plot is the third expansion. This also shows good fit to the simulated data. Along with increasing the urn size in order to decrease variance we can also keep track of a running mean. In **Figure 9** we also plot the average of the previous 2,000 probabilities at each new interaction which closely tracks the true probability.

4.3. Simulation 3

For the final simulation we explore the trouble with every measurement model, which relates ability to difficulty as the Rasch model does: the issue of unidentifiability of these parameters. In most assessment frameworks this issue is often

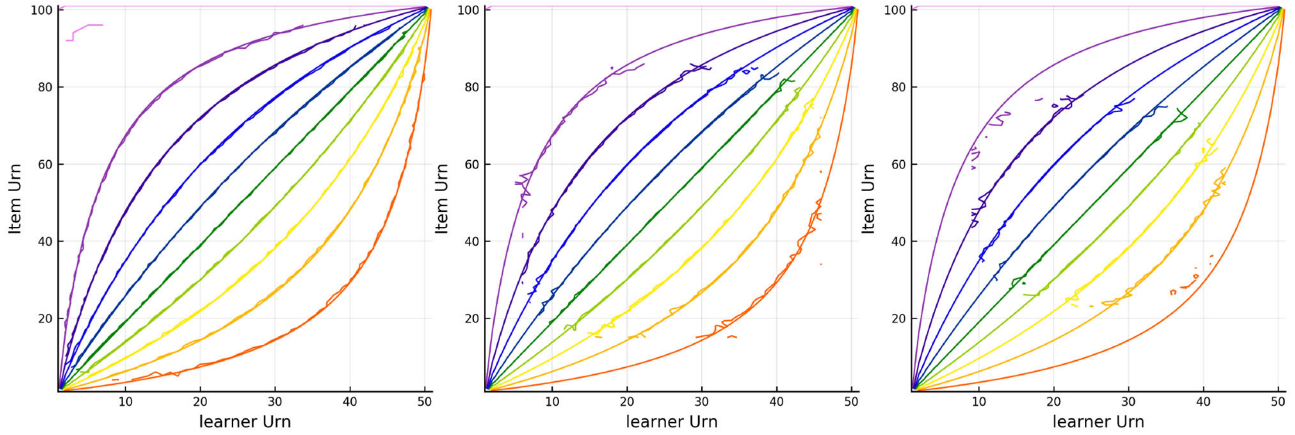


FIGURE 5 | Contours for the predicted and observed proportion of correct responses for every combination of Urnings from simulation 1. Plots from left to right correspond to the urn associated with the respective step in the dyadic expansion.

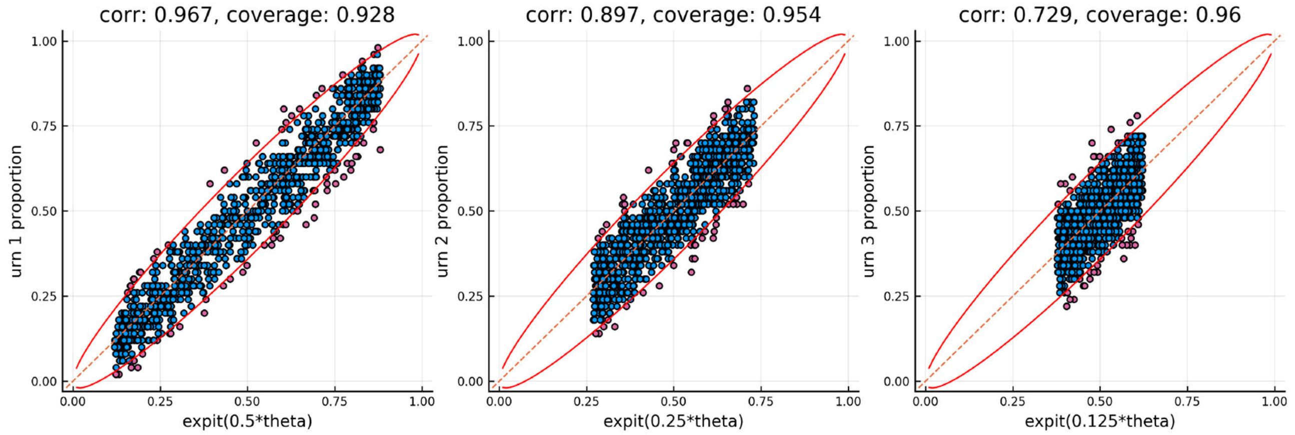


FIGURE 6 | Urn proportions of the three urns plotted against the expit of the scaled ability, $\theta_p/2^k$ where k indexes the dyadic expansion. A 95% confidence interval is included along with the correlation and coverage.

circumvented by several assumptions, such as the assumption that the abilities of the persons and the difficulties of the items are static and not changing. Additionally, some arbitrary zero point must be decided on, which is typically that the average difficulty of the population of items is equal to zero. In this final simulation, we challenge some of these assumptions as typically happens in real data, especially in learning systems.

As before, we allow the ability to change over time in the same was as we did in simulation 2. However, we restrict the change in ability to only be positive by sampling $\theta_{p1} \sim \mathrm{U}(-4, 0)$ and $\theta_{p2} \sim \mathrm{U}(0, 4)$ so that each person's ability increases. Furthermore, we allow the difficulty of the items to change over time. The item difficulties change in the same way as the person ability, but they all decrease over time. Specifically, the difficulty is

$$\delta_i(t) = \delta_{i1} + \frac{\delta_{i2} - \delta_{i1}}{1 + \exp(-2(t - t_0))} \tag{13}$$

where $\delta_{i1} \sim \mathrm{U}(0, 4)$ and $\delta_{i2} \sim \mathrm{U}(-4, 0)$. Additionally, we split the items into four groups such that the point, t_0 (at which the difficulty is half way between its starting difficulty, δ_1, to its ending difficulty, δ_{i2}) varies between groups. In the first group of items the mid-point is at the first quarter of the number of simulated interactions, the second group is half way through the simulated interactions (just like the person ability), the third group is three quarters of the way through the simulated interactions, and the last group does not change in ability. **Figure 10** plots the (true) change in item difficulty over the simulated interactions. In this way we simulate an experience that is close to a learning environment. Items whose relative difficulty decreases early on represent items related to skills which the persons learn early on in the learning environment. Just as in simulation 2, at each interaction we randomly pick a person and then select an item using the same weights as described in simulation 2. The single urn scheme is used to track the abilities and difficulties with urns of size 420 for both persons and items.

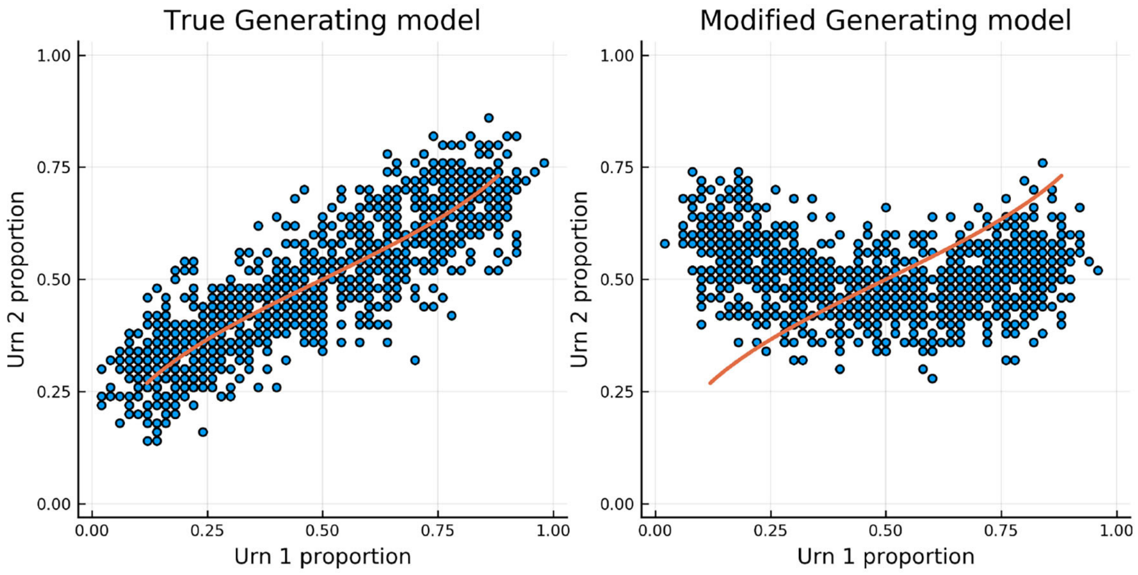


FIGURE 7 | Urn proportions in urn 1 plotted against urn proportions in urn 2 using the true generating model and the alternative model.

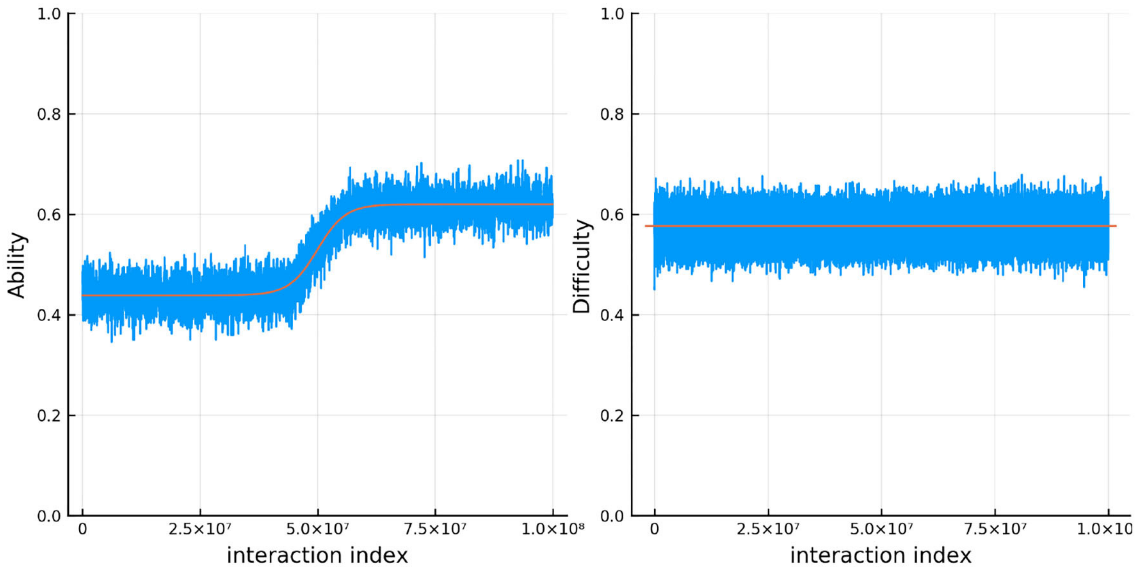


FIGURE 8 | The true (solid red line) and estimated (blue line) change in ability (left) for 1 specific person and item difficulty (right) for 1 specific item in simulation 2.

Figure 11 shows the true and estimated ability and difficulty for a particular person and a particular item. The true ability change is in red on the left and the true difficulty is in red on the right. In blue, the urn proportion for the ability on the left and the difficulty on the right. What is happening here? Clearly the urn proportions do not track the true values; this is most evident with the ability on the left. As the number of balls in the person and item urns is always fixed, if we allow the items to become easier over time and the person abilities to increase over time, the persons are literally stealing balls away from the items. This results in under-estimation of the person abilities and over-estimation of the item difficulties. In the previous simulation this effect was circumvented by allowing the distribution of ability (and difficulty) to be the same at the start of the simulation and at the end, by allowing some people's ability to increase and others to decrease (and the item difficulty was kept constant). This is not the case in this simulation. Only quantities that are properly contextualized can be accurately tracked, such as the probability that a person answers an item correctly. Consider **Figure 12**. As in the previous simulation, this figure plots the probability that a

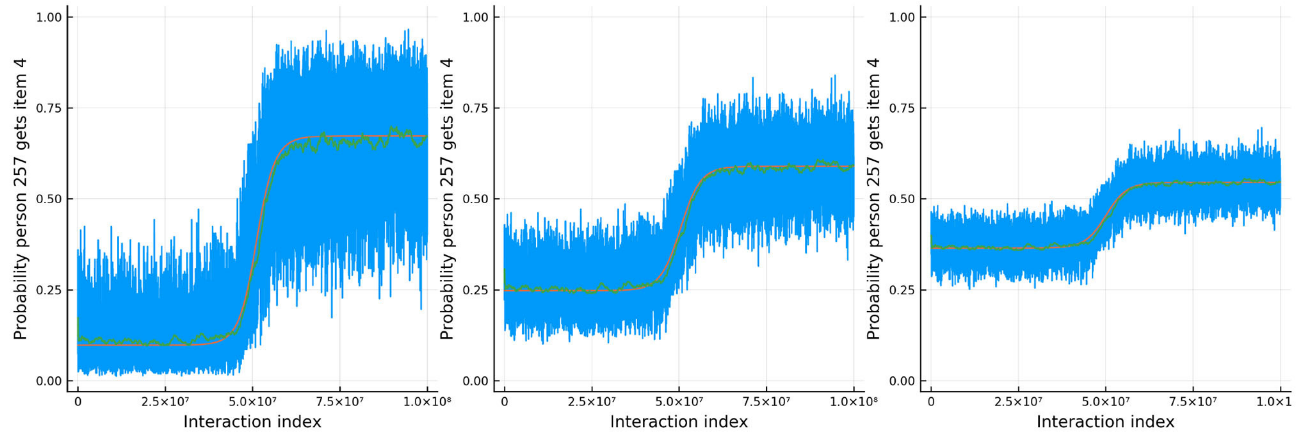


FIGURE 9 | The probability that a specific person answers the *d*th item in the dyadic expansion of a specific item correctly in simulation 2.

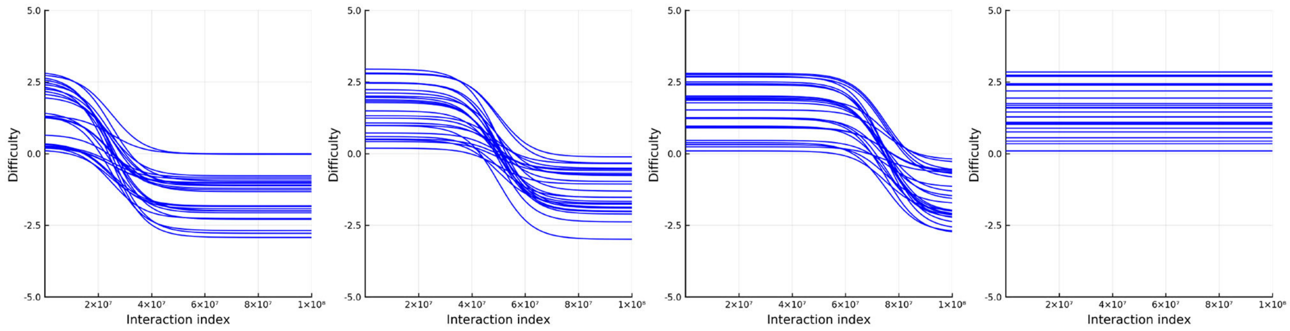


FIGURE 10 | True item difficulties in simulation 3.

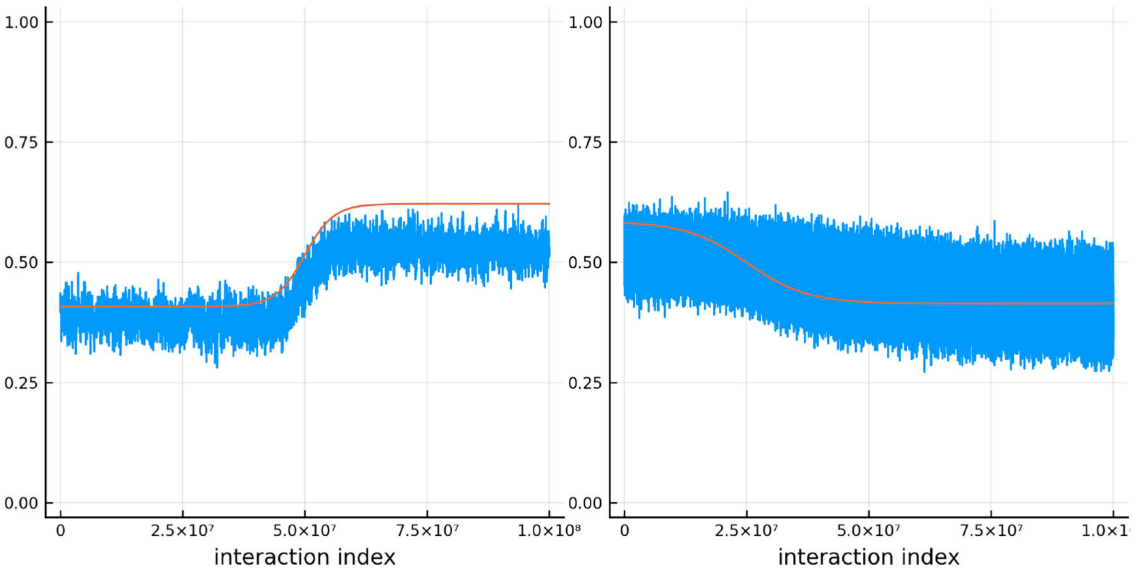


FIGURE 11 | The true (solid red line) and estimated (blue line) change in ability (left) for one specific person and item difficulty (right) for one specific item in simulation 3.

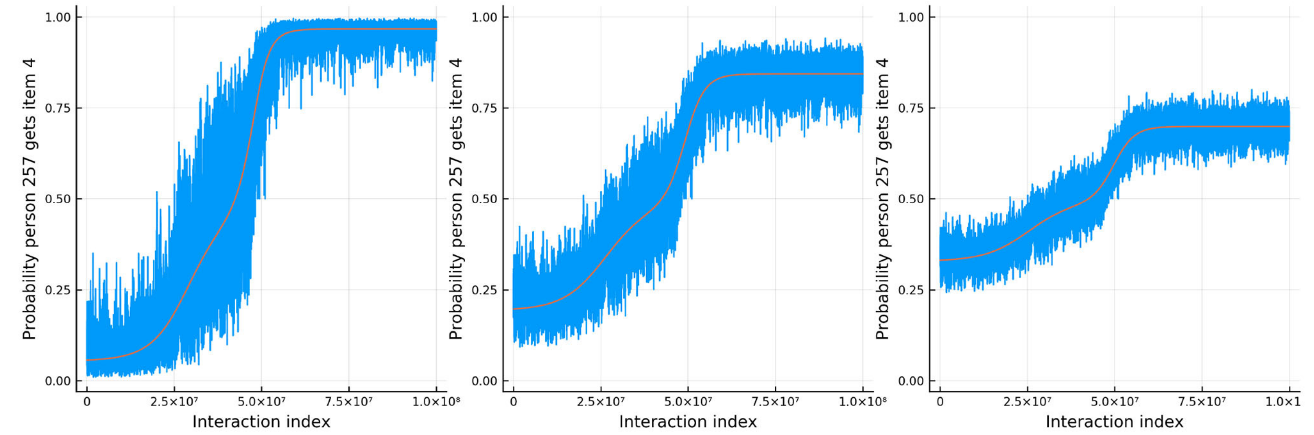


FIGURE 12 | The probability that a specific person answers the *d*th item in the dyadic expansion of a specific item correctly in simulation 3.

particular person gets one of the dyadic expansion items correct on a particular item.

5. DISCUSSION

In this article, we have proposed a new method to analyze data generated by massive online learning systems, such as DET or Math Garden, based on the CR model and the Urnings ratings system. We have demonstrated its feasibility using simulation.

The approach described here is new and based on three ingredients. First, we found that the SRT model is a special case of a Rasch model for continuous item responses. Second, we established that, if the CR model holds, continuous responses can be transformed to independent binary responses that follow the Rasch model and contain most of the information in the original responses. Of course, the Rasch model is known to not always fit the data, as it assumes each item discriminates equally well (Verhelst, 2019). We have discussed the topic of model misspecification (with regard to the misspecification of the scoring rule rather than the true data-generating process), but the focus of this paper has been on the use of the CR in the context of a learning system. Third, the urnings rating system can be applied to the binary responses to track both learners and items in real time.

In the introduction, three unique problems with large-scale, high-stakes, online, anywhere anytime learning and testing were identified. Having dealt with the problem of change and of personalization and adaptation we now briefly comment on the cold start problem. Having introduced the notion of stakes, as a way of dealing with differences in item discrimination, we can reuse the same idea for addressing the cold start problem. When a new person or item is added, we initially multiply their stakes by some number. This has the effect, similar to decreasing the urn size, of taking large(r) steps, and hence more rapidly converging to the "correct" value, but with a larger standard error. After some initial responses have been processed, the multiplier can decrease to one. Note that, in principle, the same approach can be used continuously to adjust the stakes depending on how fast or slow a person or item parameter is changing.

An extension of the urnings system was introduced in order to make use of the dichotomous responses with varying discriminations. It will be clear that we have only begun to explore the possibilities offered by the new method.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

GM developed the initial idea. BD, MB, TB, and GM were involved in further developments, writing, and critical revisions. BD and GM developed code and simulations. All authors contributed to the article and approved the submitted version.

FUNDING

MB was partially funded by the NAEd/SpencerFoundation Fellowship. All others were funded by employer, ACT, Inc. The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.500039/full#supplementary-material

REFERENCES

Bechger, T. M., and Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika* 80, 317–340. doi: 10.1007/s11336-014-9408-y

Billingsley, P. (2013). "Probability and measure," in *Wiley Series in Probability and Statistics* (Hoboken, NJ: Wiley).

Bolsinova, M., Maris, G., Hofman, A. D., van der Maas, H., and Brinkhuis, M. J. S. (2020). Urnings: a new method for tracking dynamically changing parameters in paired comparison systems. doi: 10.31219/osf.io/nep6a

Brinkhuis, M. J., and Maris, G. (2009). *Dynamic Parameter Estimation in Student Monitoring Systems*. Measurement and Research Department Reports (Rep. No. 2009-1). Arnhem: Cito.

Dirkzwager, A. (2003). Multiple evaluation: a new testing paradigm that exorcizes guessing. *Int. J. Test*. 3, 333–352. doi: 10.1207/S15327574IJT0304_3

Elo, A. E. (1978). *The Rating of Chess Players, Past and Present*. New York, NY: Arco Pub.

Finetti, B. D. (1965). Methods for discriminating levels of partial knowledge concerning a test item. *Br. J. Math. Stat. Psychol*. 18, 87–123. doi: 10.1111/j.2044-8317.1965.tb00695.x

Klinkenberg, S., Straatemeier, M., and van der Maas, H. L. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Comput. Educ*. 57, 1813–1824. doi: 10.1016/j.compedu.2011.02.003

LaFlair, G. T., and Settles, B. (2019). *Duolingo English Test: Technical Manual*. Pittsburgh, PA: DuoLingo, Inc.

Maris, G. (2020). *The Duolingo English Test: Psychometric Considerations*. Technical Report DRR-20-02, Duolingo.

Maris, G., and van der Maas, H. (2012). Speed-accuracy response models: scoring rules based on response time and accuracy. *Psychometrika* 77, 615–633. doi: 10.1007/s11336-012-9288-y

Müller, H. (1987). A rasch model for continuous ratings. *Psychometrika* 52, 165–181. doi: 10.1007/BF02294232

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika* 38, 203–219. doi: 10.1007/BF02291114

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika* 39, 111–121. doi: 10.1007/BF02291580

van Rijn, P. W., and Ali, U. S. (2017). A generalized speed–accuracy response model for dichotomous items. *Psychometrika* 83, 109–131. doi: 10.1007/s11336-017-9590-9

Verhelst, N. D. (2019). "Exponential family models for continuous responses," in *Theoretical and Practical Advances in Computer-based Educational Measurement*, eds. B. Veldkamp, and C. Sluijter (New York, NY: Springer), p. 135–160. doi: 10.1007/978-3-030-18480-3_7

Wagner, E., and Kunnan, A. J. (2015). The Duolingo English test. *Lang. Assess. Q*. 12, 320–331. doi: 10.1080/15434303.2015.1061530

Conflict of Interest: BD, TB, and GM work at ACT, Inc.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Psychology

ORIGINAL RESEARCH
published: 11 August 2021
doi: 10.3389/fpsyg.2021.509575

A Novel and Highly Effective Bayesian Sampling Algorithm Based on the Auxiliary Variables to Estimate the Testlet Effect Models

Jing Lu[1], *Jiwei Zhang*[2]*, *Zhaoyuan Zhang*[3]*†, *Bao Xu*[4] *and Jian Tao*[1]

[1] Key Laboratory of Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University, Changchun, China, [2] Key Lab of Statistical Modeling and Data Analysis of Yunnan Province, School of Mathematics and Statistics, Yunnan University, Kunming, China, [3] Department of Statistics, School of Mathematics and Statistics, Northeast Normal University, Changchun, China, [4] Institute of Mathematics, Jilin Normal University, Siping, China

OPEN ACCESS

Edited by:
Pietro Cipresso,
University of Turin, Italy

Reviewed by:
Holmes Finch,
Ball State University, United States
Peida Zhan,
Zhejiang Normal University, China

Correspondence:
Jiwei Zhang
zhangjw713@nenu.edu.cn
Zhaoyuan Zhang
zhangzy328@nenu.edu.cn

†Present address:
Zhaoyuan Zhang,
School of Mathematics and Statistics,
Yili Normal University, Yili, China

Specialty section:
This article was submitted to Quantitative Psychology and Measurement,
a section of the journal Frontiers in Psychology

Received: *02 November 2019*
Accepted: *06 July 2021*
Published: *11 August 2021*

Citation:
Lu J, Zhang J, Zhang Z, Xu B and Tao J (2021) A Novel and Highly Effective Bayesian Sampling Algorithm Based on the Auxiliary Variables to Estimate the Testlet Effect Models.
Front. Psychol. 12:509575.
doi: 10.3389/fpsyg.2021.509575

In this paper, a new two-parameter logistic testlet response theory model for dichotomous items is proposed by introducing testlet discrimination parameters to model the local dependence among items within a common testlet. In addition, a highly effective Bayesian sampling algorithm based on auxiliary variables is proposed to estimate the testlet effect models. The new algorithm not only avoids the Metropolis-Hastings algorithm boring adjustment the turning parameters to achieve an appropriate acceptance probability, but also overcomes the dependence of the Gibbs sampling algorithm on the conjugate prior distribution. Compared with the traditional Bayesian estimation methods, the advantages of the new algorithm are analyzed from the various types of prior distributions. Based on the Markov chain Monte Carlo (MCMC) output, two Bayesian model assessment methods are investigated concerning the goodness of fit between models. Finally, three simulation studies and an empirical example analysis are given to further illustrate the advantages of the new testlet effect model and Bayesian sampling algorithm.

Keywords: bayesian inference, deviance information criterion, logarithm of the pseudomarignal likelihood, item response theory, testlet effect models, slice-Gibbs sampling algorithm, Markov chain Monte Carlo

1. INTRODUCTION

In education and psychological tests, a testlet is defined as that a bundle of items share a common stimulus (a reading comprehension passage or a figure) (Wainer and Kiely, 1987). For example, in a reading comprehension test, a series of questions may be based on a common reading passage. The advantages of the testlet design are not only to allow for more complicated and interrelated set of items, but also to improve the testing efficiency (Thissen et al., 1989). Namely, with several items embedded in a testlet, test takers need not waste a considerable amount of time and energy in processing a long passage just to answer a single item. Despite their appealing features, this testing format poses a threat to item analysis because items within a testlet often violate the local independence assumption of item response theory (IRT). The traditional item response analysis tends to overestimate the precision of person ability obtained from testlets, and overestimate test reliability\information, and yields biased estimation for item difficulty and discrimination parameters (Sireci et al., 1991; Yen, 1993; Wang and Wilson, 2005a; Wainer et al., 2007; Eckes, 2014; Eckes and Baghaei, 2015).

In the face of these problems, two methods have been proposed to cope with the local item dependence. One method is to estimate a unidimensional model but treat items within a testlet as a single polytomous item (Sireci et al., 1991; Yen, 1993; Wainer, 1995; Cook et al., 1999) and then apply polytomous item response models such as the generalized partial-credit model (Muraki, 1992), the graded response models (Samejima, 1969), or the nominal response model (Bock, 1972). This method is appropriate when the local dependence between items within a testlet is moderate and the test contains a large proportion of independent items (Wainer, 1995), but it becomes impractical as the number of possible response patterns increases geometrically with the number of items in a testlet and thus is not frequently used (Thissen et al., 1989). An alternative method is testlet effects can be taken into account by incorporating specific dimensions in addition to the general dimension into the IRT models. Two such multidimensional IRT models are often used by researchers. That is, the bi-factor models (Gibbons and Hedeker, 1992) and the random-effects testlet models (Bradlow et al., 1999; Wainer et al., 2007). However, Li et al. (2006), Rijmen (2010), and Min and He (2014) find that the random-effects testlet models can be used as a special case of the bi-factor models. It is obtained by constraining the loadings on the specific dimension to be proportional to the loading on the general dimension within each testlet. In practice, researchers prefer to use simple random-effects testlet models if the two models are available and the model fit is not too much damage. Next, we discuss the specific forms of some commonly used testlet effect models.

Several literatures on testlet structure modeling have been proposed to capture the local item dependence from different perspectives for the past two decades. Bradlow et al. (1999) and Wainer et al. (2000) extend the traditional IRT models including a random effect parameter to explain the interaction between testlets and persons. The probit link function of the above model is formulated as $\Phi\left[a_j\left(\theta_i - b_j + \eta_{id(j)}\right)\right]$, where Φ is the normal cumulative distribution function, θ_i denotes the the ability for the ith examinee, a_j and b_j, respectively denote the discrimination parameter and difficulty parameter for the jth item, and $\eta_{id(j)}$ is a random effect that represents the interaction of examinee i with testlet $d\left(j\right)$ [$d\left(j\right)$ denotes the testlet d contains item j]. Further, Li et al. (2006) propose a general two parameter normal ogive testlet response theory (2PNOTRT) model from the perspective of multidimensionality. Each item response in the multidimensional model depends on both the primary dimension and the secondary testlet dimensions. Under the 2PNOTRT model, the basic form of probit link function is expressed as $\Phi\left[a_{j1}\theta_i - t_j + a_{j2}\eta_{id(j)}\right]$, where t_j is a threshold parameter related to the item difficulty. The latent traits underlying examinees' responses to items in testlets consist of general ability θ and several secondary dimensions, one for each testlet. Item parameters a_{j1} and a_{j2} indicate the discriminating power of an item with respect to the primary ability θ and the secondary dimension η_d, respectively. Because the secondary dimension $\eta_{id(j)}$ is a random effect that represents the interaction of examinee i with testlet $d\left(j\right)$, it is believed that the loading of the secondary dimensions η_d should be the discriminating power of the testlet with respect to it, and it should be related to the discrimination parameters of the items in the testlet with respect to the intended ability, θ. The above two testlet effect models are constructed in the framework of probit link function. On this basis, Zhan et al. (2014) propose the concept of within-item multidimensional testlet effect. In this paper, we introduce a new item parameter as a testlet discrimination parameter and propose a new two parameter logistic testlet model in the framework of logit link function for dichotomously scored items, as detailed in the next section. Moreover, testlet response theory modeling has also been extended to the other field of educational and psychological measurement such as large-scale language assessments (Rijmen, 2010; Zhang, 2010; Eckes, 2014), hierarchical data analysis (Jiao et al., 2005, 2013), cognitive diagnostic assessments (Zhan et al., 2015, 2018).

One of the most commonly used estimation methods for the above-mentioned testlet effect models is the marginal maximum likelihood method via the expectation-maximization (EM; Dempster et al., 1977) algorithm (Bock and Aitkin, 1981; Mislevy, 1986; Glas et al., 2000; Wang and Wilson, 2005b). The ability parameters and testlet effects are viewed as unobserved data (latent variables), and then we can find the maximum of a complete data likelihood (the responses and unobserved data) marginalized over unobserved data. However, the marginal maximum likelihood estimation of testlet models has been hampered by the fact that the computations often involve analytically intractable high dimensional integral and hence it is hard to find the maximum likelihood estimate of the parameters. More specifically, when the integrals over latent variable distributions are evaluated using Gaussian quadrature (Bock and Aitkin, 1981), the number of calculations involved increases exponentially with the number of latent variable dimensions. Even though the number of quadrature points per dimension can be reduced when using adaptive Gaussian quadrature (Pinheiro and Bates, 1995), the total number of points again increases exponentially with the number of dimensions. In addition, when the EM algorithm is employed to compute marginal maximum likelihood estimates with unobserved data, the convergence of EM algorithm can be very slow whenever there is a large fraction of unobserved data, and the estimated information matrix is not a direct by product of maximization.

An alternative method is to use a fully Bayesian formulation, coupled with a Markov Chain Monte Carlo (MCMC) procedure to estimate the testlet model parameters (e.g., Wainer et al., 2000, 2007). The Bayesian method, including Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970; Tierney, 1994; Chib and Greenberg, 1995; Chen et al., 2000) and Gibbs algorithm (Geman and Geman, 1984; Tanner and Wong, 1987; Albert, 1992), has some significant advantages over classical statistical analysis. It allows meaningful assessments in confidence regions, incorporates prior knowledge into the analysis, yields more precise estimators (provided the prior knowledge is accurate), and follows the likelihood and sufficiency principles. In this current study, an effective slice-Gibbs sampling algorithm (Lu et al., 2018) in the framework of Bayesian is used to estimate the model parameters. The slice-Gibbs sampling, as the name suggests, can be conceived of an extension of Gibbs algorithm. The sampling process consists of two parts. One part is the slice algorithm (Damien et al., 1999; Neal, 2003; Bishop, 2006; Lu et al., 2018), which samples the two parameter logistic testlet effect models from the truncated full conditional posterior distribution by introducing the auxiliary variables. The other part is Gibbs algorithm which updates variance parameters based on the sampled values from the two parameter logistic testlet effect models. The motivation for this sampling algorithm is manifold. First, the slice-Gibbs sampling algorithm is a fully Bayesian method, which averts to calculate multidimensional numerical integration compared with the marginal maximum likelihood method. Second, the slice algorithm has the advantage of a flexible prior distribution being introduced to obtain samples from the full conditional posterior distributions rather than being restricted to using the conjugate distributions, which is required in Gibbs sampling algorithm and limited using the normal ogive framework (Tanner and Wong, 1987; Albert, 1992; Bradlow et al., 1999; Wainer et al., 2000; Fox and Glas, 2001; Fox, 2010; Tao et al., 2013). The detailed discussions about the informative priors and non-informative priors of item parameters are shown in the simulation 2. Third, it is known that the Metropolis-Hasting algorithm (Metropolis et al., 1953; Hastings, 1970; Tierney, 1994; Chib and Greenberg, 1995; Chen et al., 2000) severely depends on the standard deviation (tuning parameter) of the proposal distributions, and it is sensitive to step size. More specifically, if the step size is too small random walk, the chain will take longer to traverse the support of the target density; If the step size is too large there is great inefficiency due to a high rejection rate. However, the slice algorithm automatically tunes the step size to match the local shape of the target density and draws the samples with acceptance probability equal to one. Thus, it is easier and more efficient to implement.

The remainder of this article is organized as follows. Section 2 describes the two parameter logistic testlet effect model, the prior assumptions and model identifications. A detailed description of the slice-Gibbs sampling algorithm and Bayesian model assessment criteria are presented in section 2. In section 3, three simulation studies are given, the first of which considers the performances of parameter recovery using the slice-Gibbs algorithm under different design conditions. In the second simulation, the prior sensitivity of the the slice-Gibbs sampling algorithm is assessed using the simulated data. In the third simulation, based on the Markov chain Monte Carlo (MCMC) output, two Bayesian model assessment methods are used to evaluate the model fit. In section 5, an empirical example is analyzed in detail to further demonstrate the applicability of the testlet structure models and the validity of the slice-Gibbs sampling algorithm. At last, we conclude with a few summary remarks in section 6.

2. THE NEW TWO PARAMETER LOGISTIC TESTLET MODEL AND PRIOR ASSUMPTIONS

The new two parameter logistic testlet model (N2PLTM):

$$
\begin{aligned}
p_{ij} &= p\left(y_{ij}=1 \,\middle|\, \theta_i, a_j, b_j, \eta_{id(j)}\right) \\
&= \frac{\exp\left[a_j\left(\theta_i - b_j\right) + \alpha_{d(j)}\eta_{id(j)}\right]}{1+\exp\left[a_j\left(\theta_i - b_j\right) + \alpha_{d(j)}\eta_{id(j)}\right]}, \qquad (1)
\end{aligned}
$$

In Equation (1), $i = 1, \ldots, n.$ indicates persons. Suppose a text contains J items, items in such tests are grouped into $K\,(1 \leq K \leq J)$ mutually exclusive and exhaustive testlets. Denote testlet d containing item by $d\left(j\right)$ and the size of each testlet by $n_k\left(1 \leq k \leq K\right)$ which can be written as with $d\,(1)$ and $d\,(J) = K$. y_{ij} represents the response of the ith examinee answering the jth item, and the correct response probability is expressed as p_{ij}. And θ_i denotes ability parameter for the ith examinee. a_j is the discrimination parameter of the item j. b_j denotes the difficulty parameter of the item j, and $\alpha_{d(j)} = \sum\limits_{j \in S_{d(j)}} \frac{a_j}{n_{d(j)}}$ is the testlet discrimination parameter where $n_{d(j)}$ is the numbers of items in testlet (testlet d contains item j) and $S_{d(j)}$ is the set of the serial numbers of item in the testlet. The purpose of using the testlet discrimination parameter is to consider the interaction between the discrimination parameters for all $S_{d(j)}$ items in the same testlet and the testlet effect, rather than just examining the influence of the jth item discrimination parameter on the testlet effect for the traditional testlet models. The random effect $\eta_{id(j)}$ represents the interaction of individual i with testlet $d\left(j\right)$. It can be interpreted as a random shift in individuals' ability or another ability dimension (Li et al., 2006). The following priors and hyper-priors are used to estimate the parameters of N2PLTM. The latent ability θ and the testlet effect η are assumed to be independently and normally distributed under the testlet model. That is, $\boldsymbol{\eta}^* = (\theta_i, \eta_{i1}, \ldots, \eta_{iK})'$ has a multivariate normal distribution $N\,(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is mean vector, $\boldsymbol{\Sigma}$ is a diagonal matrix, $\boldsymbol{\Sigma} = diag\left(\sigma^2_\theta, \sigma^2_{\eta_1}, \ldots, \sigma^2_{\eta_K}\right)$. The variances of $\eta_{ik}\left(k = 1, 2, \ldots, K\right)$, which can be allowed to vary across testlets, indicate the amount of local dependence in each testlet. If the variance of η_{ik} is zero, the items within the testlet can be considered conditionally independent. As the variance increases, the amount of local dependence increases. The priors to the discrimination parameters are set from truncated normal priors, $N\left(\mu_a, \sigma^2_a\right) \mathrm{I}\,(0, +\infty)$, where $\mathrm{I}\,(0, +\infty)$ denotes the indicator function that the values range from zero to infinity, and the difficulty parameters are assumed to follow the normal distribution, $b_k \sim N\left(\mu_b, \sigma^2_b\right)$. In addition, the hyper-priors for σ^2_a, σ^2_b and $\sigma^2_{\eta_k}\left(k = 1, 2, \ldots, K\right)$ are assumed to follow inverse Gamma distribution with shape parameter v and scale parameter τ. Let $\boldsymbol{\Omega} = (\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\eta})$ represents the collection of the unknown parameters in model (1), where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$, $\boldsymbol{a} = \left(a_1, \ldots, a_J\right)'$, $\boldsymbol{b} = \left(b_1, \ldots, b_J\right)'$ and $\boldsymbol{\eta} = \left(\eta_{1d(1)}, \ldots., \eta_{d(J)}\right)'$. The joint posterior distribution of $\mathbf{\Omega}$ given the data is represented by

$$\begin{aligned} p\left(\mathbf{\Omega}\,|\boldsymbol{Y}\right) \propto & \prod_{i=1}^{n}\prod_{j=1}^{J} p\left(y_{ij}\,\middle|\theta_i, a_j, b_j, \eta_{id(j)}\right) p\left(\theta_i\right) \\ & p\left(a_j\,|\mu_a, \sigma_a^2\right) \mathrm{I}\left(a_j > 0\right) p\left(b_j\,|\mu_b, \sigma_b^2\right) \\ & \times p\left(\sigma_a^2\right) p\left(\sigma_b^2\right) p\left(\eta_{id(j)}\,\middle|\mu_\eta, \sigma^2_{\eta_{d(j)}}\right) p\left(\sigma^2_{\eta_{d(j)}}\right) \\ \propto & \left\{\prod_{i=1}^{n}\prod_{j=1}^{J}\left[p_{ij}^{y_{ij}}\left(1-p_{ij}\right)^{1-y_{ij}}\right]\right\}\left[\prod_{i=1}^{n}\exp\left(-\frac{\theta_i^2}{2}\right)\right] \\ & \left(\sigma_a^2\sigma_b^2\right)^{-\frac{J}{2}}\prod_{j=1}^{J}\exp\left[-\frac{\left(a_j-\mu_a\right)^2}{2\sigma_a^2}\right] \\ & \times\exp\left[-\frac{\left(b_j-\mu_b\right)^2}{2\sigma_b^2}\right]\mathrm{I}\left(a_j>0\right)\left(\sigma_a^2\right)^{-(\nu_1+1)} \\ & \left(\sigma_b^2\right)^{-(\nu_2+1)}\exp\left[-\frac{\tau_1}{\sigma_a^2}-\frac{\tau_2}{\sigma_b^2}\right] \\ & \times\prod_{i=1}^{n}\prod_{j=1}^{J}\exp\left(-\frac{\eta^2_{id(j)}}{2\sigma^2_{\eta_{d(j)}}}\right)\left(\sigma^2_{\eta_{d(j)}}\right)^{-(\nu_3+1)} \\ & \exp\left(-\frac{\tau_3}{\sigma^2_{\eta_{d(j)}}}\right). \end{aligned} \tag{2}$$

2.1. Model Identifications

In Equation 1, the linear part of the testlet effect model, $a_j\left(\theta_i - b_j\right) + \alpha_{d(j)}\eta_{id(j)}$, can be rewritten as follows

$$a_j\left(\theta_i - b_j + \frac{\eta_{id(j)}}{n_{d(j)}}\right) + \frac{\sum\limits_{k\in S_{d(j)}-\{j\}} a_k}{n_{d(j)}}\eta_{id(j)},$$

where the testlet discrimination $\alpha_{d(j)}$ consists of the discrimination parameters a_j. That is, $\alpha_{d(j)} = \sum\limits_{j\in S_{d(j)}}\frac{a_j}{n_{d(j)}}$, and $k \in S_{d(j)} - \{j\}$ means that k belongs to the set $S_{d(j)}$ excluding the index j. To eliminate the trade offs among the ability θ, difficulty parameter b and testlet effect $\eta_{id(j)}$ in location, we fix the mean population level of ability to zero and restrict a item difficulty parameter to zero. Meanwhile, to eliminate the trade off between the ability θ and the discrimination parameter a in scale, we need restrict the variance population level of ability to one. However, $a_j b_j$, $a_j\frac{\eta_{id(j)}}{n_{d(j)}}$ and $\frac{\sum\limits_{k\in S_{d(j)}-\{j\}} a_k}{n_{d(j)}}\eta_{id(j)}$ still have the trade offs in scale. In fact, we only need fix a item discrimination parameter to one. In summary, the required identification conditions are as follows:

$$\theta_i \sim N\left(0,1\right),\; a_1 = 1 \text{ and } b_1 = 0.$$

Several identification restriction methods of two parameter IRT models have been widely used. The identification restrictions of our model are based on the following methods.

(1) To fix the mean population level of ability to zero and the variance population level of ability to one (Lord and Novick, 1968; Bock and Aitkin, 1981; Fox and Glas, 2001; Fox, 2010). That is, $\theta \sim N\left(0,1\right)$;

(2) To fix the item difficulty parameter to a specific value, most often zero, and restrict the discrimination parameter to a specific value, most often one (Fox and Glas, 2001; Fox, 2010). That is, $b_1 = 0$ and $a_1 = 1$.

3. BAYESIAN INFERENCES

3.1. Slice-Gibbs Algorithm to Estimate Model Parameters

The motivation for the slice-Gibbs sampling algorithm is that the inferred samples can easily be drawn from the full conditional distribution by introducing the auxiliary variables. Before giving the specific Bayesian sampling process, we give the definition of auxiliary and its role in the sampling process. Auxiliary variables are variables that can help to make estimates on incomplete data, while they are not part of the main analysis. Basically, the auxiliary variables are latent unknown parameters without any direct interpretation which are introduced for technical/simulation reasons or for the reason of making an analytically intractable distribution tractable. Within the Bayesian framework, in the method of auxiliary variables, realizations from a complicated distribution can be obtained by augmenting the variables of interest by one or more additional variables such that the full conditionals are tractable and easy to simulate from. The construction of sampling algorithms via the introduction of auxiliary variable received much attention since it resulted in both simple and fast algorithms (Tanner and Wong, 1987; Higdon, 1998; Meng and van Dyk, 1999; Fox, 2010).

For each of the response variable y_{ij}, we introduce two mutually independent random auxiliary variables λ_{ij} and φ_{ij}. The random variables λ_{ij} and φ_{ij} are assumed to follow a Uniform (0,1). The following two cases must be satisfied.

Case 1: When $y_{ij} = 1$, an equivalent condition for $y_{ij} = 1$ is the indicator function $\mathrm{I}\left(0 < \lambda_{ij} \le p_{ij}\right)$ must be equal to 1, as opposed to $\mathrm{I}\left(0 < \varphi_{ij} \le q_{ij}\right)$ is set to 0, where $q_{ij} = 1 - p_{ij}$. In addition, if the joint distribution (λ_{ij} and p_{ij}) integrate out the auxiliary variables λ_{ij}, the obtained marginal distribution is just equal to the correct response probability of the ith individual answering the jth item.

Case 2: Similarly, when $y_{ij} = 0$, an equivalent condition for $y_{ij} = 0$, that is, the indicator function $\mathrm{I}\left(0 < \varphi_{ij} \le q_{ij}\right)$ must be equal to 1, as opposed to is $\mathrm{I}\left(0 < \lambda_{ij} \le p_{ij}\right)$ set to 0.

Therefore, the joint posterior distribution based on the auxiliary variables is given by

$$\begin{aligned} p\left(\mathbf{\Omega}, \boldsymbol{\lambda}, \boldsymbol{\varphi}\ |\boldsymbol{Y}\right) \propto \prod_{i=1}^{n}\prod_{j=1}^{J}\big[&\mathrm{I}\left(y_{ij}=1\right)\mathrm{I}\left(0<\lambda_{ij}\le p_{ij}\right) \\ &+\mathrm{I}\left(y_{ij}=0\right)\mathrm{I}\left(0<\varphi_{ij}\le q_{ij}\right)\big] \end{aligned}$$

$$\times \left(\sigma_a^2\sigma_b^2\right)^{-\frac{J}{2}} \prod_{j=1}^{J} \exp\left[-\frac{\left(a_j-\mu_a\right)^2}{2\sigma_a^2} - \frac{\left(b_j-\mu_b\right)^2}{2\sigma_b^2}\right]$$

$$\mathrm{I}\left(a_j>0\right)\left[\prod_{i=1}^{n} \exp\left(-\frac{\theta_i^2}{2}\right)\right]$$

$$\times \left(\sigma_a^2\right)^{-(v_1+1)} \left(\sigma_b^2\right)^{-(v_2+1)} \exp\left[-\frac{\tau_1}{\sigma_a^2}-\frac{\tau_2}{\sigma_b^2}\right]$$

$$\times \prod_{i=1}^{n}\prod_{j=1}^{J} \exp\left(-\frac{\eta_{id(j)}^2}{2\sigma_{\eta_{d(j)}}^2}\right)\left(\sigma_{\eta_{d(j)}}^2\right)^{-(v_3+1)}$$

$$\exp\left(-\frac{\tau_3}{\sigma_{\eta_{d(j)}}^2}\right). \qquad (3)$$

We find that the Equation (2) can be obtained by taking expectations about the auxiliary variables for the Equation (3). Each step of the algorithm needs to satisfy the Equation (3). The detailed slice-Gibbs sampling algorithm is given by

Step 1: Sample the auxiliary variables λ_{ij} and φ_{ij} given the response variable $\boldsymbol{Y}$ and the parameters $\mathbf{\Omega}$. The full conditional posterior distributions can be written as

$$\begin{aligned} \lambda_{ij}\,|\boldsymbol{Y},\ \mathbf{\Omega} &\sim \text{Uniform}\left(0,\ p_{ij}\right), \quad \text{if } y_{ij}=1, \\ \varphi_{ij}\,|\boldsymbol{Y},\ \mathbf{\Omega} &\sim \text{Uniform}\left(0,\ q_{ij}\right), \quad \text{if } y_{ij}=0. \end{aligned} \qquad (4)$$

Step 2: Sample the discrimination parameter a_j. The prior of the discrimination parameters is $N\left(\mu_a, \sigma_a^2\right)\mathrm{I}\left(0,+\infty\right)$. According to the Equation (3), for all i, if $0 < \lambda_{ij} \le p_{ij}$, $\left(\theta_i - b_j + \frac{\eta_{id(j)}}{n_{d(j)}}\right) > 0$ or $0 < \varphi_{ij} \le q_{ij}$, $\left(\theta_i - b_j + \frac{\eta_{id(j)}}{n_{d(j)}}\right) < 0$. The following inequalities are established

$$a_j\left(\theta_i-b_j\right)+\alpha_{d(j)}\eta_{id(j)} \ge \log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right),$$

Or equivalently,

$$a_j \ge \left(\theta_i - b_j + \frac{\eta_{id(j)}}{n_{d(j)}}\right)^{-1}\left[\log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right) - \frac{\sum\limits_{k\in S_{d(j)}-\{j\}} a_k}{n_{d(j)}}\eta_{id(j)}\right],$$

And,

$$a_j\left(\theta_i-b_j\right)+\alpha_{d(j)}\eta_{id(j)} \ge \log\left(\frac{1-\varphi_{ij}}{\varphi_{ij}}\right),$$

Or equivalently,

$$a_j \ge \left(\theta_i - b_j + \frac{\eta_{id(j)}}{n_{d(j)}}\right)^{-1}\left[\log\left(\frac{1-\varphi_{ij}}{\varphi_{ij}}\right) - \frac{\sum\limits_{k\in S_{d(j)}-\{j\}} a_k}{n_{d(j)}}\eta_{id(j)}\right].$$

Similarly, for all i, if $0 < \lambda_{ij} \le p_{ij}$, $\left(\theta_i - b_j + \frac{\eta_{id(j)}}{n_{d(j)}}\right) < 0$ or $0 < \varphi_{ij} \le q_{ij}$, $\left(\theta_i - b_j + \frac{\eta_{id(j)}}{n_{d(j)}}\right) > 0$. The following inequalities are established

$$a_j\left(\theta_i-b_j\right)+\alpha_{d(j)}\eta_{id(j)} \ge \log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right),$$

Or equivalently,

$$a_j \le \left(\theta_i - b_j + \frac{\eta_{id(j)}}{n_{d(j)}}\right)^{-1}\left[\log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right) - \frac{\sum\limits_{k\in S_{d(j)}-\{j\}} a_k}{n_{d(j)}}\eta_{id(j)}\right],$$

And,

$$a_j\left(\theta_i-b_j\right)+\alpha_{d(j)}\eta_{id(j)} \ge \log\left(\frac{1-\varphi_{ij}}{\varphi_{ij}}\right),$$

Or equivalently,

$$a_j \le \left(\theta_i - b_j + \frac{\eta_{id(j)}}{n_{d(j)}}\right)^{-1}\left[\log\left(\frac{1-\varphi_{ij}}{\varphi_{ij}}\right) - \frac{\sum\limits_{k\in S_{d(j)}-\{j\}} a_k}{n_{d(j)}}\eta_{id(j)}\right].$$

Let

$$\Delta_j = \left\{ i \,\middle|\, 0 < \lambda_{ij} \le p_{ij}, \left(\theta_i - b_j + \frac{\eta_{id(j)}}{n_{d(j)}}\right) > 0 \right\},$$

$$G_j = \left\{ i \,\middle|\, 0 < \varphi_{ij} \le p_{ij}, \left(\theta_i - b_j + \frac{\eta_{id(j)}}{n_{d(j)}}\right) < 0 \right\},$$

$$\nabla_j = \left\{ i \,\middle|\, 0 < \lambda_{ij} \le p_{ij}, \left(\theta_i - b_j + \frac{\eta_{id(j)}}{n_{d(j)}}\right) < 0 \right\},$$

$$\Lambda_j = \left\{ i \,\middle|\, 0 < \varphi_{ij} \le p_{ij}, \left(\theta_i - b_j + \frac{\eta_{id(j)}}{n_{d(j)}}\right) > 0 \right\}.$$

When given the response variable $\boldsymbol{Y}$, the auxiliary variable $\boldsymbol{\lambda}$, $\boldsymbol{\varphi}$ and other parameters $\mathbf{\Omega}_1$ (all of the parameters except a_j), the full conditional distribution is represented by

$$a_j \,\middle|\, \boldsymbol{\lambda},\ \boldsymbol{\varphi},\ \mathbf{\Omega}_1 \sim N\left(\mu_a, \sigma_a^2\right)\mathrm{I}\left(0 < a_j^L \le a_j \le a_j^U\right). \qquad (5)$$

In Equation (5),

$$a_j^L = \max\left\{ \max_{i\in\Delta_j}\left(\theta_i - b_j + \frac{\eta_{id(j)}}{n_{d(j)}}\right)^{-1}\right.$$

$$\left[\log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right) - \frac{\sum\limits_{k\in S_{d(j)}-\{j\}} a_k}{n_{d(j)}}\eta_{id(j)}\right],$$

$$\max_{i\in G_j}\left(\theta_i-b_j+\frac{\eta_{id(j)}}{n_{d(j)}}\right)^{-1}\left[\log\left(\frac{1-\varphi_{ij}}{\varphi_{ij}}\right)-\frac{\sum\limits_{k\in S_{d(j)}-\{j\}}a_k}{n_{d(j)}}\eta_{id(j)}\right]\Bigg\}\Bigg\}.$$

And

$$a_j^U=\min\left\{\min_{i\in\nabla_j}\left(\theta_i-b_j+\frac{\eta_{id(j)}}{n_{d(j)}}\right)^{-1}\right.$$
$$\left[\log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right)-\frac{\sum\limits_{k\in S_{d(j)}-\{j\}}a_k}{n_{d(j)}}\eta_{id(j)}\right],$$
$$\left.\min_{i\in\Lambda_j}\left(\theta_i-b_j+\frac{\eta_{id(j)}}{n_{d(j)}}\right)^{-1}\left[\log\left(\frac{1-\varphi_{ij}}{\varphi_{ij}}\right)-\frac{\sum\limits_{k\in S_{d(j)}-\{j\}}a_k}{n_{d(j)}}\eta_{id(j)}\right]\right\}\Bigg\}.$$

Step 3: Sample the difficulty parameter b_j. The prior of the difficulty parameters is $N\left(\mu_b,\sigma_b^2\right)$. According to the Equation (3), for $\forall i$, if we have $0<\lambda_{ij}\le p_{ij}$, the following inequalities are established,

$$a_j\left(\theta_i-b_j\right)+\alpha_{d(j)}\eta_{id(j)}\ge\log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right),$$

Or equivalently,

$$b_j\le\theta_i-\frac{1}{a_j}\left[\log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right)-\alpha_{d(j)}\eta_{id(j)}\right].$$

Similarly, for all i, if $0<\varphi_{ij}\le q_{ij}$, the following inequalities are established

$$a_j\left(\theta_i-b_j\right)+\alpha_{d(j)}\eta_{id(j)}\ge\log\left(\frac{1-\varphi_{ij}}{\varphi_{ij}}\right),$$

Or equivalently,

$$b_j\le\theta_i-\frac{1}{a_j}\left[\log\left(\frac{1-\varphi_{ij}}{\varphi_{ij}}\right)-\alpha_{d(j)}\eta_{id(j)}\right].$$

Let $D_j=\left\{i\,\middle|y_{ij}=1,\ 0<\lambda_{ij}\le p_{ij}\right\}$, $E_j=\left\{i\,\middle|y_{ij}=0,\ 0<\varphi_{ij}\le q_{ij}\right\}$. Thus, given the response variable $\boldsymbol{Y}$, the auxiliary variable $\boldsymbol{\lambda}$, $\boldsymbol{\varphi}$ and other parameters $\boldsymbol{\Omega}_2$ (all of the parameters except b_j). The full conditional posterior distribution is given by

$$b_j\,\middle|\boldsymbol{\lambda},\ \boldsymbol{\varphi},\ \boldsymbol{\Omega}_2\sim N\left(\mu_b,\sigma_b^2\right)\mathrm{I}\left(b_j^L\le b_j\le b_j^U\right),\qquad(6)$$

In Equation (6),

$$b_j^L=\max_{i\in E_j}\left\{\theta_i-\frac{1}{a_j}\left[\log\left(\frac{1-\varphi_{ij}}{\varphi_{ij}}\right)-\alpha_{d(j)}\eta_{id(j)}\right]\right\},$$

And

$$b_j^U=\min_{i\in D_j}\left\{\theta_i-\frac{1}{a_j}\left[\log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right)-\alpha_{d(j)}\eta_{id(j)}\right]\right\}.$$

Step 4: Sample the latent ability θ_i, the prior of the latent ability is assumed to follow a normal distribution with mean μ_θ and variance σ_θ^2. Given the response variable $\boldsymbol{Y}$, the auxiliary variable $\boldsymbol{\lambda}$, $\boldsymbol{\varphi}$ and other parameters $\boldsymbol{\Omega}_3$ (all of the parameters except θ_i). The full conditional posterior distribution of θ_i is

$$\theta_i\,\middle|\boldsymbol{\lambda},\ \boldsymbol{\varphi},\ \boldsymbol{\Omega}_3,\boldsymbol{Y}\sim N\left(\mu_\theta,\sigma_\theta^2\right)\mathrm{I}\left(\theta_i^L\le\theta_i\le\theta_i^U\right),\qquad(7)$$

In Equation (7),

$$\theta_i^L=\max_{j\in C_i}\left\{\frac{1}{a_j}\left[\log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right)-\alpha_{d(j)}\eta_{id(j)}\right]+b_j\right\},$$
$$\text{where } C_i=\left\{j\,\middle|y_{ij}=1,\ 0<\lambda_{ij}\le p_{ij}\right\},$$

$$\theta_i^U=\min_{j\in B_i}\left\{\frac{1}{a_j}\left[\log\left(\frac{1-\varphi_{ij}}{\varphi_{ij}}\right)-\alpha_{d(j)}\eta_{id(j)}\right]+b_j\right\},$$
$$\text{where } B_i=\left\{j\,\middle|y_{ij}=0,\ 0<\varphi_{ij}\le q_{ij}\right\}.$$

Step 5: Sample the testlet random effect $\eta_{id(j)}$. Assuming that the jth term comes from the kth testlet [i.e., $d\left(j\right)=k$] and the order of the terms in the kth testlet is form j_k to n_k+j_k-1. Then, the joint posterior distribution can be rewritten as

$$p\left(\boldsymbol{\Omega},\boldsymbol{\lambda},\boldsymbol{\varphi}\ \middle|\boldsymbol{Y}\right)\propto\prod_{i=1}^{n}\prod_{k=1}^{K}\prod_{j=j_k}^{n_k+j_k-1}\left[\mathrm{I}\left(y_{ij}=1\right)\mathrm{I}\left(0<\lambda_{ij}\le p_{ij}^*\right)\right.$$
$$\left.+\mathrm{I}\left(y_{ij}=0\right)\mathrm{I}\left(0<\varphi_{ij}\le q_{ij}^*\right)\right]$$
$$\times\left(\sigma_a^2\sigma_b^2\right)^{-\frac{J}{2}}\prod_{j=1}^{J}\exp\left[-\frac{\left(a_j-\mu_a\right)^2}{2\sigma_a^2}\right.$$
$$\left.-\frac{\left(b_j-\mu_b\right)^2}{2\sigma_b^2}\right]\mathrm{I}\left(a_j>0\right)\left[\prod_{i=1}^{n}\exp\left(-\frac{\theta_i^2}{2}\right)\right]$$

$$\times\left(\sigma_a^2\right)^{-(v_1+1)}\left(\sigma_b^2\right)^{-(v_2+1)}\exp\left[-\frac{\tau_1}{\sigma_a^2}-\frac{\tau_2}{\sigma_b^2}\right]$$
$$\times\prod_{i=1}^{n}\prod_{j=1}^{J}\exp\left(-\frac{\eta_{ik}^2}{2\sigma_{\eta_k}^2}\right)\left(\sigma_{\eta_k}^2\right)^{-(v_3+1)}\exp\left(-\frac{\tau_3}{\sigma_{\eta_k}^2}\right).$$

where $p_{ij}^*=\frac{\exp\left[a_j(\theta_i-b_j)+\alpha_k\eta_{ik}\right]}{1+\exp\left[a_j(\theta_i-b_j)+\alpha_k\eta_{ik}\right]}$, $q_{ij}^*=1-p_{ij}^*$. The prior of the testlet random effect η_{ik} is assumed to follow a normal distribution with mean μ_η and variance σ_η^2. Given the response variable $\boldsymbol{Y}$, the auxiliary variable $\boldsymbol{\lambda}$, $\boldsymbol{\varphi}$ and other parameters $\boldsymbol{\Omega}_4$ (all of the parameters except η_{ik}). The full conditional distribution of η_{ik} is given by

$$\eta_{ik}\,\middle|\boldsymbol{\lambda},\ \boldsymbol{\varphi},\ \boldsymbol{\Omega}_4,\boldsymbol{Y}\sim N\left(\mu_\eta,\sigma_\eta^2\right)\mathrm{I}\left(\eta_{ik}^L\le\eta_{ik}\le\eta_{ik}^U\right),\qquad(8)$$

In Equation (8),

$$\eta_{ik}^{L}=\frac{1}{\alpha_k}\left[\log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right)-a_j\left(\theta_i-b_j\right)\right],\ \text{and}$$

$$\eta_{ik}^{U}=\frac{1}{\alpha_k}\left[\log\left(\frac{1-\varphi_{ij}}{\varphi_{ij}}\right)-a_j\left(\theta_i-b_j\right)\right].$$

Step 6: Sample the variance parameter σ_a^2, the variance is assumed to follow a Inverse-Gamma(ν_1,τ_1) hyper prior. Given the discrimination parameters $\boldsymbol{a}$, the hyper parameters ν_1 and τ_1. The full conditional posterior distribution of σ_a^2 is given by

$$\begin{aligned}p\left(\sigma_a^2\left|\boldsymbol{a},\nu_1,\tau_1\right.\right)&\propto p\left(\boldsymbol{a}\left|\mu_a,\sigma_a^2\right.\right)p\left(\sigma_a^2\right)\\&\propto\left|\sigma_a^2\right|^{-\frac{J}{2}}\exp\left\{-\frac{\sum_{j=1}^{J}\left(a_j-\mu_a\right)^2}{2\sigma_a^2}\right\}\\&\quad\left|\sigma_a^2\right|^{-(\nu_1+1)}\exp\left\{-\frac{\tau_1}{\sigma_a^2}\right\}.\end{aligned}$$

Thus,

$$\sigma_a^2\left|\boldsymbol{a},\nu_1,\tau_1\right.\sim\text{Inverse}-\text{Gamma}\left(\frac{J}{2}+\nu_1,\frac{\sum_{j=1}^{J}\left(a_j-\mu_a\right)^2}{2}+\tau_1\right).$$

Step 7: Sample the variance parameter σ_b^2, the variance is assumed to follow a Inverse-Gamma(ν_2,τ_2) hyper prior. Given the difficulty parameters $\boldsymbol{b}$, the hyper parameters ν_2 and τ_2. The full conditional posterior distribution of σ_b^2 is given by

$$\begin{aligned}p\left(\sigma_b^2\left|\boldsymbol{b},\nu_2,\tau_2\right.\right)&\propto p\left(\boldsymbol{b}\left|\mu_b,\sigma_b^2\right.\right)p\left(\sigma_b^2\right)\\&\propto\left|\sigma_b^2\right|^{-\frac{J}{2}}\exp\left\{-\frac{\sum_{j=1}^{J}\left(b_j-\mu_b\right)^2}{2\sigma_b^2}\right\}\\&\quad\left|\sigma_b^2\right|^{-(\nu_2+1)}\exp\left\{-\frac{\tau_2}{\sigma_b^2}\right\}.\end{aligned}$$

Thus,

$$\sigma_b^2\left|\boldsymbol{b},\nu_2,\tau_2\right.\sim\text{Inverse}-\text{Gamma}\left(\frac{J}{2}+\nu_2,\frac{\sum_{j=1}^{J}\left(b_j-\mu_b\right)^2}{2}+\tau_2\right).\tag{9}$$

Step 8: Sample the random effect variance parameter $\sigma_{\eta_k}^2$, the variance is assumed to follow a Inverse-Gamma (ν_3,τ_3) hyper prior. Given the random effect parameters $\boldsymbol{\eta}$, the hyper parameters ν_3 and τ_3. The full conditional posterior distribution of $\sigma_{\eta_k}^2$ is given by

$$\begin{aligned}p\left(\sigma_{\eta_k}^2\left|\boldsymbol{\eta},\nu_3,\tau_3\right.\right)&\propto p\left(\boldsymbol{\eta}\left|\mu_\eta,\sigma_{\eta_k}^2\right.\right)p\left(\sigma_{\eta_k}^2\right)\\&\propto\left|\sigma_{\eta_k}^2\right|^{-\frac{n}{2}}\exp\left\{-\frac{\sum_{i=1}^{n}\left(\eta_{ik}-\mu_\eta\right)^2}{2\sigma_{\eta_k}^2}\right\}\\&\quad\left|\sigma_{\eta_k}^2\right|^{-(\nu_3+1)}\exp\left\{-\frac{\tau_3}{\sigma_{\eta_k}^2}\right\}.\end{aligned}$$

Thus,

$$\sigma_{\eta_k}^2\left|\boldsymbol{\eta},\nu_3,\tau_3\right.\sim\text{Inverse}-\text{Gamma}\left(\frac{n}{2}+\nu_3,\frac{\sum_{i=1}^{N}\left(\eta_{ik}-\mu_\eta\right)^2}{2}+\tau_3\right).\tag{10}$$

3.2. Bayesian Model Assessment

Within the framework of Bayesian, Bayes factor has played a major role in assessing the goodness of fit of competing models (Kass and Wasserman, 1995; Gelfand, 1996). It is defined as the ratio of the posterior odds of model 1 to model 2 divided by the prior odds of model 1 to model 2

$$\text{BF}=\frac{p\left(M_1\left|\boldsymbol{y}\right.\right)/p\left(M_2\left|\boldsymbol{y}\right.\right)}{p\left(M_1\right)/p\left(M_2\right)}=\frac{p\left(\boldsymbol{y}\left|M_1\right.\right)}{p\left(\boldsymbol{y}\left|M_2\right.\right)},\tag{11}$$

In Equation (11), $\boldsymbol{y}$ denotes the observation data, $p\left(M_h\right)$ denotes the model prior likelihood, and $p\left(M_h\left|\boldsymbol{y}\right.\right)$ are the marginal likelihoods of the data matrix $\boldsymbol{y}$ for model h, $h=1,2$. The Bayes factor (BF) provide a summary of evidence for M_1 compared to M_2. M_1 is supported when BF>1, and M_2 is supported otherwise. A value of BF between 1 and 3 is considered as minimal evidence for M_1, a value between 3 and 12 as positive evidence for M_1, a value between 12 and 150 as strong evidence for M_1, and a value >150 as very strong evidence (Raftery, 1996). However, one of the obstacles to use of the Bayes factors is the difficulty associated with calculating them. As we known, while the candidate model with high-dimensional parameters are used to fit the data, it is not possible integrate out the all parameters of models to obtain the closed-form expression of marginal distribution. In addition, it are acutely sensitive to the choice of prior distributions. If the use of improper priors for the parameters in alternative models results in Bayes factors that are not well defined. However, numerous approaches have been proposed for model comparison with improper priors (Aitkin, 1991; Gelfand et al., 1992; Berger and Pericchi, 1996; Ando, 2011). In our article, Based on the noninformative priors, a "pseudo-Bayes factor" approach is implemented, which provides a type of approximation to the BF.

3.2.1. Pseudo-Bayes Factor

The pseudo-Bayes factor (PsBF) method (Geisser and Eddy, 1979) overcome BF sensitive to the choice of prior distributions. It can be obtained by calculating the cross-validation predictive densities. Considering $i = 1, \ldots, n$ individuals response to items. Let $\boldsymbol{y}_{-(ij)}$ be the observed data without the ijth observation and let Ξ denote all the parameters under the assumed model. The cross-validation predictive density (CVPD) can be defined by

$$p\left(y_{ij}\middle|\boldsymbol{y}_{-(ij)}\right) = \int p\left(y_{ij}\middle|\boldsymbol{y}_{-(ij)}, \Xi\right) p\left(\Xi\middle|\boldsymbol{y}_{-(ij)}\right) d\Xi, \quad (12)$$

In Equation (12), the density $p\left(y_{ij}\middle|\boldsymbol{y}_{-(ij)}\right)$ denotes supporting the possibility of values of y_{ij} when the model is fitted to observations except y_{ij}. According to conditional independence hypothesis, the equation $p\left(y_{ij}\middle|\boldsymbol{y}_{-(ij)}, \Xi\right) = p\left(y_{ij}\,|\,\Xi\right)$ can be established, the responses on the different items are independent given ability and the responses of the individuals are independent of one another. The Pseudo Bayes factor (PsBF) for comparing two models (M_1 and M_2) is expressed in terms of the product of cross-validation predictive densities and can be written as

$$\text{PsBF} = \prod_{i,j} \frac{p\left(y_{ij}\middle|\boldsymbol{y}_{-(ij)}, M_1\right)}{p\left(y_{ij}\middle|\boldsymbol{y}_{-(ij)}, M_2\right)}. \quad (13)$$

In practice, we can calculate the logarithm of the numerator and denominator of the PsBF and it can be used for comparing different models. The model with a larger PsBF has a better fit of the data. Gelfand and Dey (1994) and Newton and Raftery (1994) proposed an importance sampling to evaluate the marginal likelihood (CVPD) of the data. Given the sample size R, $r = 1, \ldots, R$, the samples $\Xi^{(m)}$ from the posterior distribution $p\left(\Xi\middle|\boldsymbol{y}_{-(ij)}\right)$ often easily obtained via an MCMC sampler. The estimated likelihood function is

$$\begin{aligned} \widehat{p\left(y_{ij}\middle|\boldsymbol{y}_{-(ij)}\right)} &= \left[\frac{1}{M}\sum_{m=1}^{M}\frac{1}{p\left(y_{ij}\,|\,\Xi^{(m)}\right)}\right]^{-1} \\ &= \left[\frac{1}{M}\sum_{m=1}^{M}\frac{1}{\left(p_{ij}^{(m)}\right)^{y_{ij}}\left(1-p_{ij}^{(m)}\right)^{1-y_{ij}}}\right]^{-1} \end{aligned} \quad (14)$$

3.2.2. The Deviance Information Criteria (DIC)

A model comparison method is often based on a measure of fit and some penalty function based on the number of free parameters for the complexity of the model. Two well-known criteria of model selection based on a deviance fit measure are the Bayesian information criterion (BIC; Schwarz, 1978) and Akaike's information criterion (AIC; Akaike, 1973). These criteria depend on the effective number of parameters in the model as a measure of model complexity. However, in Bayesian hierarchical models, it is not clear how to define the number of parameters due to the prior distribution imposes additional restrictions on the parameter space and reduces its effective dimension. Therefore, Spiegelhalter et al. (2002) proposed the deviance information criterion (DIC) for model comparison when the number of parameters is not clearly defined in hierarchical models. The DIC is defined as the sum of a deviance measure and a penalty term for the effective number of parameters based on a measure of model complexity. This term estimates the number of effective model parameters and equals

$$\begin{aligned} P_D &= E_{\Xi|\boldsymbol{y}}\left\{-2\log p\left(\boldsymbol{y}\,|\,\Xi\right)\right\} + 2\log p\left(\boldsymbol{y}\,\middle|\,\widehat{\Xi}\right) \\ &= \overline{D\left(\Xi\right)} - D\left(\widehat{\Xi}\right). \end{aligned} \quad (15)$$

The DIC can be defined as

$$\begin{aligned} \text{DIC} &= \overline{D\left(\Xi\right)} + P_D \\ &= \overline{D\left(\Xi\right)} + \left(\overline{D\left(\Xi\right)} - D\left(\widehat{\Xi}\right)\right). \end{aligned} \quad (16)$$

In Equation (15), Ξ is the parameter of interest in the model. The complexity is measured by the effective number of parameters, P_D. $\overline{D\left(\Xi\right)}$ is the posterior expectation of the deviance. It is calculated from the MCMC output by taking the sample mean of the simulated values of the deviance, $D\left(\widehat{\Xi}\right) = -2\log p\left(\boldsymbol{y}\,\middle|\,\widehat{\Xi}\right)$. That is defined as the deviance of the posterior estimation mean. Here $\widehat{\Xi}$ denotes the posterior means of the parameters. The model with a smaller DIC has a better fit of the data.

4. SIMULATION STUDY

4.1. Simulation 1

This simulation study is conducted to evaluate the recovery performance of the slice-Gibbs sampling algorithm under different simulation conditions.

The following design conditions are considered: (a) testlet type: 20 dichotomous items in 2 or 4 testlets ($J = 20$, each testlet has 10 or 5 dichotomous items); (b) number of examinees, $N = 500$ and 1,000; and (c) testlet effect: the variances of the testlet random effect are 0.25 and 1.00. That is, $\sigma^2_{\eta_{ik}} = 0.25$ or 1.00, where $i = 1, \ldots, N$, $k = 1, 2$, or $k = 1, 2, 3, 4$. The true values of item discrimination parameters a_js are generated from a truncated normal distribution, that is, $a_j \sim N(0, 1)\,\mathrm{I}(0, +\infty)$, and the item difficulty parameters b_js are generated from $N(0, 1)$. Ability parameters θ_is for $N = 500$ or 1,000 examinees are drawn from a standard normal distribution. The testlets random effect parameters η_{ik}s are also generated from a normal distribution. That is, $\eta_{ik} \sim N\left(0, \sigma^2_{\eta_{ik}}\right)$. Response data are simulated using the N2PLTM in Equation (1). The non-informative priors and hyper priors of parameters are considered as follows:

$$\begin{aligned} &a_j \sim N(0, 100)\,\mathrm{I}(0, +\infty),\; b_j \sim N(0, 100),\; j = 1, \ldots, J, \\ &\sigma_a^2 \sim \mathrm{IG}(0.001, 0.001),\; \sigma_b^2 \sim \mathrm{IG}(0.001, 0.001),\; \sigma^2_{\eta_{ik}} \\ &\quad \sim \mathrm{IG}(0.001, 0.001). \end{aligned}$$

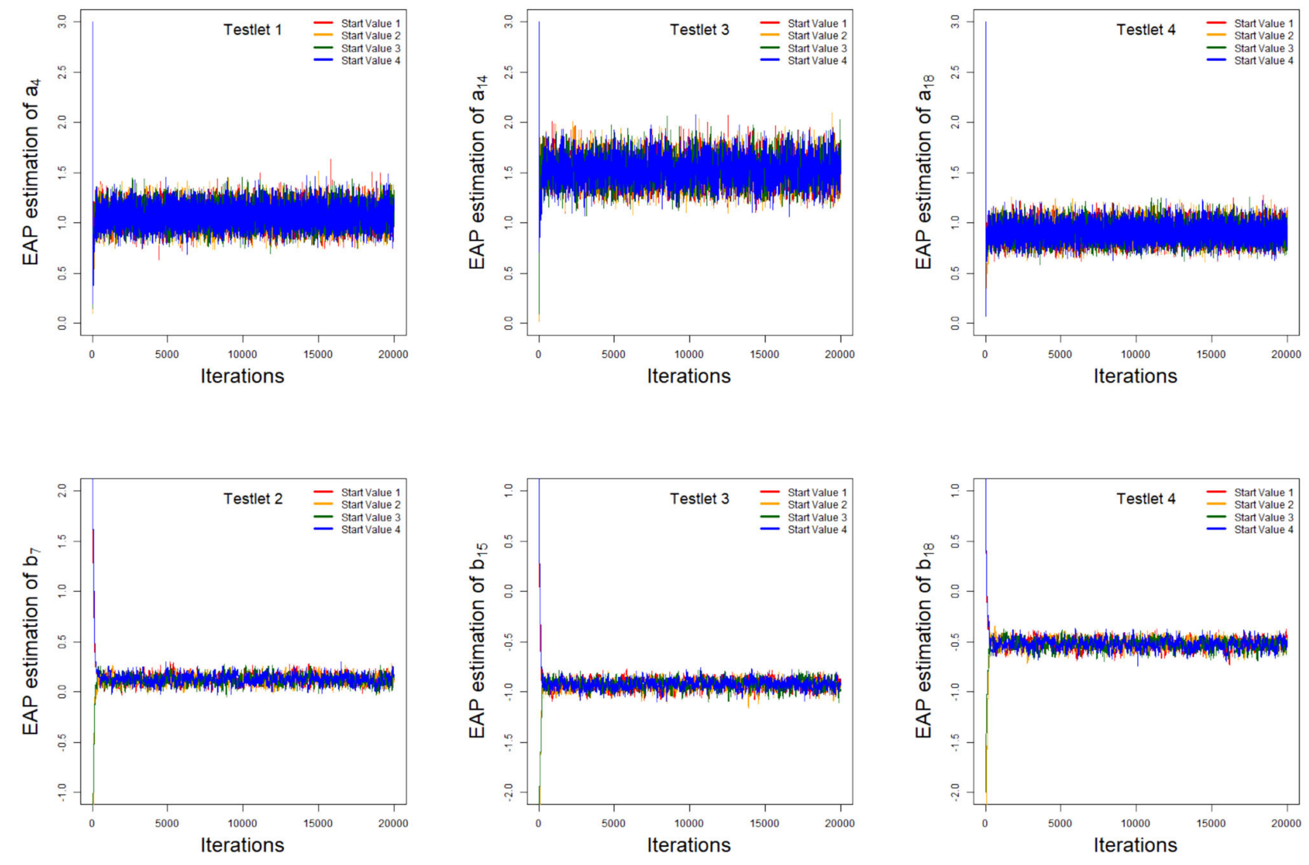


FIGURE 1 | The trace plots of the arbitrarily selected item parameters.

The non-informative priors and hyper priors are often used in many educational measurement studies (e.g., van der Linden, 2007; Wang et al., 2018). In this paper, the prior specification will be uninformative enough for the data to dominate the priors, so that the influence of the priors on the results will be minimal.

4.1.1. Convergence Diagnostic for Slice-Gibbs Algorithm

As an illustration, we only consider the convergence in the case of 20 dichotomous items in 4 testlets, the number of individuals is 500, and the variance of the random testlet variables is 0.25. Two methods are used to check the convergence of our algorithm. One is the “eyeball” method to monitor the convergence by visually inspecting the history plots of the generated sequences (Zhang et al., 2007), and another method is to use the Gelman-Rubin method (Gelman and Rubin, 1992; Brooks and Gelman, 1998) to check the convergence of the parameters. Bayesian computation procedure is implemented by R software. The convergence of slice-Gibbs algorithm algorithm is checked by monitoring the trace plots of the parameters for consecutive sequences of 20,000 iterations. We set the first 10,000 iterations as the burn-in period. Four chains started at overdispersed starting values are run for each replication. The trace plots of item parameters randomly selected are shown in **Figure 1**. In addition, we find the potential scale reduction factor (PSRF; Brooks and Gelman, 1998) values of all parameters are <1.1, which ensures that all chains converge as expected. As an illustration, the PSRF values of all item parameters are shown in **Figure 2**. On a desktop computer [AMD EPYC 7542 32-Core Processor] with 2.90 GHz dual core processor and 1TB of RAM memory, the average convergence times for our new algorithm and the traditional Metropolis-Hastings algorithm based on 50 replications, are shown in **Table 1**.

4.1.2. The Accuracy Evaluation of Parameter Estimation

The accuracy of the parameter estimates is measured by two evaluation methods, namely, Bias and mean squared error (MSE). The recovery results are based on the 50 replications in each simulation condition. The number of replication we choose is based on the previous research in educational psychological assessments. For example, Wang et al. (2013) proposed a semi-parametric approach, specifically, the Cox proportional hazards model with a latent speed covariate to analyze the response time data. In their simulation study, 10 replications (Page 15, section 4.1) are used for each simulation condition. Zhan et al. (2017) proposed joint modeling of attributes and response speed using item responses and response times simultaneously for cognitive diagnosis to provide more refined diagnostic feedback with collateral information in item response times. In their

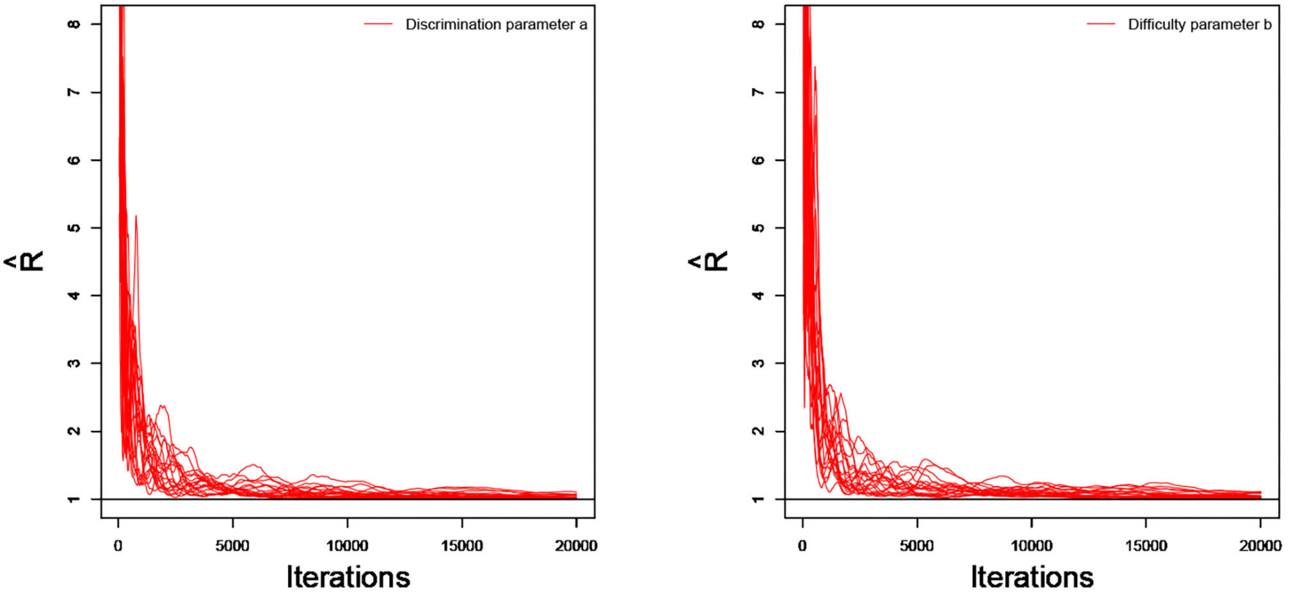


FIGURE 2 | The trace plots of $\widehat{R}$ for the simulation study 1.

TABLE 1 | Convergence times for all 8 simulation conditions in simulation study 1.

Sample size	Variance of	Time for convergence (Hours)	
× testlet type	testlet effect	Slice-Gibbs algorithm	MH algorithm
500 × 2		0.2624	0.3182
500 × 4	0.25	0.4428	0.5864
1,000×2		0.3261	0.4639
1,000×4		0.6354	0.7882
500 × 2		0.2781	0.3325
500 × 4	1	0.6262	0.7691
1,000×2		0.4045	0.5952
1,000×4		0.8827	1.1201

MH denotes the Metropolis-Hastings.

simulation study, they used 30 replications (Page 276) in each condition to reduce the random error. Lu et al. (2020) proposed a new mixture model for responses and response times with a hierarchical ability structure, which incorporates auxiliary information from other subtests and the correlation structure of the abilities to detect examinees' rapid guessing behavior. The recovery of the estimates was based on 20 replications (Page 14, section 5). Lu and Wang (2020) proposed to use an innovative item response time model as a cohesive missing data model to account for the two most common item nonresponses: not-reached items and omitted items. They considered 20 replications (Page 21) for each simulation condition. Therefore, based on the previous empirical conclusions, we adopt 50 replications in our simulation studies. If we consider a large number of replications, it is impossible to check the $\widehat{R}$ values (potential scale reduction factor; PSRF, Brooks and Gelman, 1998) calculated from each simulated dataset (replication) to ensure the parameter convergence. It will be a huge work when the simulated conditions increase. Let ϑ be the parameter of interest. $S = 50$ data sets are generated. Also, let $\widehat{\vartheta}^{(s)}$ denotes the posterior mean obtained from the sth simulated data set for $s = 1, \ldots, S$.

The Bias for parameter ϑ is defined as

$$\text{Bias}(\vartheta) = \frac{1}{S}\sum_{s=1}^{S}\left(\widehat{\vartheta}^{(s)} - \vartheta\right), \tag{17}$$

and the mean squared error (MSE) for parameter ϑ is defined as

$$\text{MSE}(\vartheta) = \frac{1}{S}\sum_{s=1}^{S}\left(\widehat{\vartheta}^{(s)} - \vartheta\right)^2. \tag{18}$$

From **Tables 2–4**, the Bias is between −0.3267 and 0.2769 for the discrimination parameters, between −0.2259 and 0.2071 for the difficulty parameters, between −0.0132 and 0.0161 for the variance parameters of $\boldsymbol{a}$, between −0.0219 and 0.1303 for the variance parameters of $\boldsymbol{b}$, between −0.2932 and 0.0332 for the variance parameter of testlet effect $\boldsymbol{\eta}$. the MSE is between 0.0000 and 0.1162 for the discrimination parameters, between 0.0000 and 0.0552 for the difficulty parameters, between 0.0002 and 0.0005 for the variance parameters of $\boldsymbol{a}$, between 0.0002 and 0.0449 for the variance parameters of $\boldsymbol{b}$, between 0.0000 and 0.1848 for the variance parameter of testlet effect $\boldsymbol{\eta}$. In summary, the slice-Gibbs algorithm provides accurate estimates of the parameters in term of various numbers of examinees and items.

TABLE 2 | Evaluating accuracy of the item parameter estimates based on different simulation conditions in the simulation study 1.

The testlet effect with small variance ($\sigma^2_{\eta_k} = 0.25$)											
		Two testlets ($k = 2$)						Four testlets ($k = 4$)			
		N = 500		N = 1,000				N = 500		N = 1,000	
Testlets	Para.	Bias	MSE	Bias	MSE	Testlet	Para.	Bias	MSE	Bias	MSE
1	a_1	0^*	0^*	0^*	0^*	1	a_1	0^*	0^*	0^*	0^*
	a_2	–0.0220	0.0122	–0.0596	0.0085		a_2	–0.0901	0.0320	–0.0331	0.0036
	a_3	0.1079	0.0259	0.0371	0.0053		a_3	–0.0437	0.0172	–0.1163	0.0299
	a_4	0.1293	0.0269	–0.0194	0.0100		a_4	–0.0517	0.0116	–0.0217	0.0046
	a_5	0.1430	0.0340	0.0201	0.0029		a_5	0.0375	0.0080	0.0209	0.0030
	a_6	0.0735	0.0211	0.0969	0.0236		b_1	0^*	0^*	0^*	0^*
	a_7	0.0296	0.0156	–0.0170	0.0058		b_2	–0.0229	0.0012	–0.1338	0.0194
	a_8	0.1060	0.0238	0.1418	0.0414		b_3	–0.0100	0.0016	–0.0489	0.0027
	a_9	0.0043	0.0119	–0.1767	0.0418		b_4	0.0678	0.0059	0.0084	0.0013
	a_{10}	0.0044	0.0162	0.0155	0.0050		b_5	–0.0338	0.0043	0.1382	0.0216
	b_1	0^*	0^*	0^*	0^*	2	a_6	0.0013	0.0055	-0.0099	0.0043
	b_2	0.0784	0.0066	0.0595	0.0046		a_7	–0.0321	0.0080	–0.0526	0.0121
	b_3	–0.0999	0.0121	0.1838	0.0346		a_8	–0.1421	0.0314	–0.0682	0.0195
	b_4	–0.1049	0.0120	–0.0586	0.0043		a_9	–0.1936	0.0484	–0.1320	0.02678
	b_5	0.0572	0.0064	0.0648	0.0081		a_{10}	–0.0459	0.0107	0.0698	0.0067
	b_6	–0.0441	0.0030	–0.1098	0.0125		b_6	0.0621	0.0088	–0.0551	0.0041
	b_7	0.0233	0.0021	0.0139	0.0018		b_7	–0.0227	0.0049	0.0557	0.0034
	b_8	–0.0780	0.0078	–0.0950	0.0093		b_8	0.0470	0.0042	0.0461	0.0024
	b_9	0.0061	0.0007	–0.0145	0.0007		b_9	–0.0519	0.0039	–0.1125	0.0129
	b_{10}	0.0309	0.0018	0.0711	0.0073		b_{10}	–0.0754	0.0105	0.1889	0.0382
2	a_{11}	–0.0930	0.0273	–0.0404	0.0079	3	a_{11}	0.0132	0.0080	–0.0040	0.0064
	a_{12}	–0.0566	0.0188	–0.0543	0.0109		a_{12}	–0.0766	0.0253	–0.0105	0.0100
	a_{13}	–0.0092	0.0112	0.0431	0.0266		a_{13}	–0.0444	0.0111	0.0010	0.0077
	a_{14}	0.0824	0.0223	–0.1066	0.0241		a_{14}	–0.0838	0.0255	0.0694	0.0086
	a_{15}	0.0670	0.0154	0.1983	0.0461		a_{15}	–0.1910	0.0489	–0.0047	0.0060
	a_{16}	0.0681	0.0201	–0.0650	0.0170		b_{11}	–0.0746	0.0069	0.0572	0.0039
	a_{17}	–0.0427	0.0116	0.2769	0.1023		b_{12}	–0.0766	0.0064	0.0149	0.0006
	a_{18}	0.0872	0.0183	0.1844	0.0403		b_{13}	0.0983	0.0128	0.0247	0.0015
	a_{19}	–0.0731	0.0164	–0.0246	0.0078		b_{14}	–0.0384	0.0020	0.1116	0.0140
	a_{20}	0.0856	0.0149	–0.1472	0.0302		b_{15}	0.1051	0.0121	–0.0203	0.0012
	b_{11}	0.0018	0.0008	–0.1063	0.0120	4	a_{16}	–0.1907	0.0522	–0.0602	0.0071
	b_{12}	0.0254	0.0018	0.0042	0.0005		a_{17}	0.0069	0.0057	–0.0596	0.0064
	b_{13}	0.0404	0.0029	–0.1164	0.0137		a_{18}	–0.0233	0.0084	–0.0467	0.0069
	b_{14}	0.0545	0.0082	–0.0481	0.0032		a_{19}	–0.1432	0.0368	–0.0512	0.0088
	b_{15}	0.0118	0.0029	0.1903	0.0365		a_{20}	–0.0780	0.0157	–0.1109	0.0276
	b_{16}	–0.0168	0.0064	–0.0048	0.0006		b_{16}	0.0351	0.0020	0.0784	0.0071
	b_{17}	–0.0871	0.0084	0.1171	0.0139		b_{17}	–0.1779	0.0372	–0.1403	0.0213
	b_{18}	0.1374	0.0203	0.2071	0.0437		b_{18}	0.0465	0.0052	–0.0353	0.0023
	b_{19}	0.0175	0.0015	–0.0419	0.0030		b_{19}	–0.0441	0.0029	–0.0976	0.0115
	b_{20}	–0.0676	0.0091	–0.0582	0.0038		b_{20}	0.0672	0.0057	0.0706	0.0054

Asterisks (*) indicates the constraints for model identifications. In fact, we need fix an item discrimination and difficulty parameters to one and zero due to model identifiability limitations. That is, a_1 =1, b_1 =0. In Bayesian estimation process, the Bias and MSE for the discrimination parameter a_1 are blackened 0. Similarly, the Bias and MSE for the difficulty parameter b_1 are also blackened 0.

4.2. Simulation 2

This simulation study is designed to show that the slice-Gibbs sampling algorithm is sufficiently flexible to recover various prior distributions of the item parameters and address the sensitivity of our slice-Gibbs algorithm with different priors.

TABLE 3 | Evaluating accuracy of the item parameter estimates based on different simulation conditions in the simulation study 1.

The testlet effect with large variance $\left(\sigma^2_{\eta_k} = 1.00\right)$											
		Two testlets $(k = 2)$						**Four testlets $(k = 4)$**			
		***N* = 500**		***N* = 1,000**				***N* = 500**		***N* = 1,000**	
Testlets	**Para.**	**Bias**	**MSE**	**Bias**	**MSE**	**Testlet**	**Para.**	**Bias**	**MSE**	**Bias**	**MSE**
1	a_1	0*	0*	0*	0*	1	a_1	0*	0*	0*	0*
	a_2	0.1068	0.0532	–0.0423	0.0060		a_2	–0.0206	0.0128	0.0762	0.0109
	a_3	0.0399	0.0122	0.0120	0.0023		a_3	0.0562	0.0201	–0.0674	0.0210
	a_4	0.0665	0.0164	0.0684	0.0130		a_4	0.0447	0.0137	0.0751	0.0137
	a_5	0.0898	0.0185	0.0541	0.0096		a_5	0.1790	0.0411	0.0915	0.0118
	a_6	–0.0190	0.0139	0.1984	0.0573		b_1	0*	0*	0*	0*
	a_7	–0.0810	0.0258	0.0352	0.0063		b_2	–0.0045	0.0008	–0.1139	0.0138
	a_8	0.0113	0.0150	0.2475	0.0768		b_3	0.0020	0.0011	–0.0247	0.0011
	a_9	–0.1398	0.0369	–0.0888	0.0217		b_4	0.0832	0.0079	0.0365	0.0024
	a_{10}	–0.1216	0.0358	0.0595	0.0061		b_5	–0.0402	0.0040	0.1794	0.0338
	b_1	0*	0*	0*	0*	2	a_6	0.0562	0.0087	0.0709	0.0109
	b_2	0.0777	0.0065	0.0795	0.0071		a_7	0.0629	0.0155	–0.0408	0.0133
	b_3	–0.0727	0.0086	0.1899	0.0367		a_8	–0.1050	0.0237	–0.0317	0.0139
	b_4	–0.0751	0.0063	–0.0479	0.0029		a_9	–0.1127	0.0269	–0.0780	0.0225
	b_5	0.0535	0.0067	0.1047	0.0136		a_{10}	0.0696	0.0128	0.1520	0.0259
	b_6	–0.0293	0.0017	–0.1021	0.0107		b_6	0.1359	0.0237	–0.0591	0.0045
	b_7	0.0236	0.0020	0.0503	0.0042		b_7	0.0162	0.0028	0.0435	0.0022
	b_8	–0.0498	0.0039	–0.0962	0.0094		b_8	0.0954	0.0110	0.0344	0.0016
	b_9	0.0044	0.0009	0.0047	0.0004		b_9	–0.0048	0.0007	–0.0918	0.0086
	b_{10}	0.0291	0.0020	0.1053	0.0130		b_{10}	–0.0405	0.0045	0.1919	0.0398
2	a_{11}	–0.1291	0.0416	–0.0248	0.0064	3	a_{11}	0.2072	0.0521	0.1561	0.0371
	a_{12}	–0.0855	0.0248	–0.0099	0.0092		a_{12}	0.0261	0.0241	0.1212	0.0288
	a_{13}	–0.0509	0.0204	0.0114	0.0120		a_{13}	0.0070	0.0086	0.1183	0.0262
	a_{14}	0.0745	0.0147	–0.0630	0.0124		a_{14}	0.0525	0.0187	0.2235	0.0569
	a_{15}	0.0388	0.0098	0.2199	0.0528		a_{15}	–0.3267	0.1162	0.1419	0.0311
	a_{16}	0.0719	0.0139	–0.0337	0.0127		b_{11}	–0.1127	0.0143	0.0245	0.0011
	a_{17}	0.0412	0.0331	0.2466	0.0734		b_{12}	–0.0932	0.0093	–0.0246	0.0011
	a_{18}	0.1039	0.0226	0.2060	0.0462		b_{13}	0.1460	0.0230	–0.0192	0.0018
	a_{19}	–0.1304	0.0333	0.0110	0.0102		b_{14}	–0.0334	0.0020	0.0751	0.0066
	a_{20}	0.0585	0.0105	–0.1228	0.0251		b_{15}	0.1157	0.0152	–0.0727	0.0059
	b_{11}	–0.0149	0.0015	–0.1035	0.0117	4	a_{16}	–0.1712	0.0499	0.0534	0.0091
	b_{12}	0.0055	0.0014	–0.0064	0.0005		a_{17}	0.1437	0.0265	0.0320	0.0052
	b_{13}	0.0277	0.0024	–0.0992	0.0100		a_{18}	0.0859	0.0176	0.0934	0.0141
	b_{14}	0.0286	0.0064	–0.0508	0.0032		a_{19}	–0.1100	0.0306	0.0515	0.0080
	b_{15}	–0.0027	0.0033	0.1773	0.03176		a_{20}	–0.0396	0.0180	–0.1562	0.0377
	b_{16}	–0.0326	0.0062	–0.0109	0.0006		b_{16}	0.0542	0.0037	0.1187	0.0151
	b_{17}	–0.0887	0.0087	0.1086	0.0121		b_{17}	–0.2259	0.0552	–0.1822	0.0344
	b_{18}	0.1242	0.0168	0.1821	0.0336		b_{18}	0.0843	0.0099	–0.0397	0.0023
	b_{19}	0.0057	0.0015	–0.0529	0.0040		b_{19}	–0.0275	0.0020	–0.1136	0.0137
	b_{20}	–0.0580	0.0073	–0.0641	0.0046		b_{20}	0.1055	0.0123	0.0684	0.0050

Asterisks (*) indicates the constraints for model identifications. In fact, we need fix an item discrimination and difficulty parameters to one and zero due to model identifiability limitations. That is, a_1 =1, b_1 =0. In Bayesian estimation process, the Bias and MSE for the discrimination parameter a_1 are blackened 0. Similarly, the Bias and MSE for the difficulty parameter b_1 are also blackened 0.

Response pattern with 500 examinees and 4 testlets (5 items per testlet) is generated by N2PLTM as given by Equation (1). The true values of item parameters and ability parameters are generated same as in simulation 1. The true value of the testlet effect variance is set equal to 0.25. The specified types of item parameter priors are given by the following:

TABLE 4 | Evaluating accuracy of the variance parameter estimates.

The testlet effect with small variance ($\sigma^2_{\eta_k} = 0.25$)									
Two testlets ($k = 2$)					**Four Testlets ($k = 4$)**				
	***N* = 500**		***N* = 1,000**			***N* = 500**		***N* = 1,000**	
Para.	**Bias**	**MSE**	**Bias**	**MSE**	**Para.**	**Bias**	**MSE**	**Bias**	**MSE**
σ^2_a	0.0161	0.0005	0.0080	0.0003	σ^2_a	0.0079	0.0002	–0.0092	0.0002
σ^2_b	–0.0219	0.0005	0.2119	0.0449	σ^2_b	0.0572	0.0033	0.1303	0.0170
$\sigma^2_{\eta_1}$	0.0283	0.0008	0.0209	0.0004	$\sigma^2_{\eta_1}$	–0.0051	0.0000	–0.0029	0.0000
$\sigma^2_{\eta_2}$	0.0234	0.0005	0.0332	0.0011	$\sigma^2_{\eta_2}$	–0.0021	0.0000	–0.0024	0.0000
					$\sigma^2_{\eta_3}$	–0.0102	0.0001	–0.0054	0.0000
					$\sigma^2_{\eta_4}$	–0.0059	0.0000	–0.0092	0.0000
The testlet effect with large variance ($\sigma^2_{\eta_k} = 1.00$)									
Two testlets ($k = 2$)					**Four testlets ($k = 4$)**				
	***N* = 500**		***N* = 1,000**			***N* = 500**		***N* = 1,000**	
Para.	**Bias**	**MSE**	**Bias**	**MSE**	**Para.**	**Bias**	**MSE**	**Bias**	**MSE**
σ^2_a	0.0106	0.0005	0.0094	0.0002	σ^2_a	0.0053	0.0002	–0.0132	0.0003
σ^2_b	–0.0135	0.0002	0.2181	0.0475	σ^2_b	0.0398	0.0016	0.1336	0.0178
$\sigma^2_{\eta_1}$	–0.1955	0.0382	–0.1953	0.0382	$\sigma^2_{\eta_1}$	–0.2333	0.1112	–0.2104	0.0964
$\sigma^2_{\eta_2}$	–0.2254	0.0509	–0.2014	0.0405	$\sigma^2_{\eta_2}$	–0.2932	0.0863	–0.2241	0.1051
					$\sigma^2_{\eta_3}$	–0.2194	0.1760	–0.2298	0.1848
					$\sigma^2_{\eta_4}$	–0.2024	0.1622	–0.2177	0.1745

Type I: Informative priors, $a_j \sim N(0, 1)\,\mathrm{I}(0, +\infty)$ and $b_j \sim N(0, 1)$;
Type II: Noninformative priors, $a_j \sim N(0, 100)\,\mathrm{I}(0, +\infty)$ and $b_j \sim N(0, 100)$;
Type III: Noninformative priors, $a_j \sim \text{Uniform}(0, 100)$ and $b_j \sim \text{Uniform}(0, 100)$.

Prior specifications for the other parameters are identical to the simulation study 1. To implement the MCMC sampling algorithm, chains of length 20,000 with an initial burn-in period 10,000 are chosen, and the PSRF values of all parameters are <1.1. Based on 25 replications, the average times for all parameters to converge in Type I, Type II and Type III are 0.4597, 0.4428, and 0.4506 h, respectively.

The average Bias and MSE for item parameters based on 50 replication are shown in **Table 5**. We find that the average Bias and MSE for item parameters are relatively unchanged under the three different prior distributions. The slice-Gibbs sampling algorithm allows for informative (Type I) or non-informative (Type II, Type III) priors of the item parameters and is not sensitive to the specification of priors. Moreover, a wider range of prior distributions is also appealing.

4.3. Simulation 3

In this simulation study, we will investigate the power of the model assessment methods. Namely, whether the Bayesian model comparison criteria based on the MCMC output could identify the true model from which the data are generated. The simulation design is as follows.

A data set with 500 examinees from standard normal distribution and four testlets (five items per testlet) is generated from the N2PLTM model. For the true values of parameters, the discrimination parameters $a_j s$ are generated from the truncated normal distribution, that is, $a_j \sim N(0, 1)\,\mathrm{I}(0, +\infty)$. The difficulty parameters $b_j s$ are generated from normal distribution, that is, $b_j \sim N(0, 1)$. The independent-items model as Model 1 is used to model assessment in which the random effects are set to zero. Model 1 is known as two parameter logistic model (2PLM; Birnbaum, 1957). In addition, the testlets random effect parameters $\eta_{ik}s$ are generated from a normal distribution. That is, $\eta_{ik} \sim N(0, 0.25)$, $k = 1, 2, 3, 4$. Model 2 is the traditional two parameter logistic testlet model (T2PLTM; Bradlow et al., 1999), which is give by

$$p_{ij} = p\left(y_{ij} = 1 \,\middle|\, \theta_i, a_j, b_j, \eta_{id(j)}\right) = \frac{\exp\left[a_j\left(\theta_i - b_j + \eta_{id(j)}\right)\right]}{1 + \exp\left[a_j\left(\theta_i - b_j + \eta_{id(j)}\right)\right]}. \tag{19}$$

Model 3 is the N2PLTM in Equation (1). The parameter priors are identical to the simulation study 1. The parameters are estimated based on 20,000 iterations after a 10,000 burn-in period, and the PSRF values of all parameters are <1.1. Two Bayesian model assessment methods are used to model fitting. That is, DIC and

TABLE 5 | Average Bias and MSE for the item parameter estimates using three prior distributions in the simulation study 2.

	Type I		Type II		Type III	
Parameter	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**
Discrimination ***a***	–0.0757	0.0250	–0.0641	0.0245	–0.0695	0.0260
Difficulty ***b***	–0.0039	0.0064	–0.0038	0.0064	–0.0038	0.0065

TABLE 6 | The results of Bayesian model assessment in the simulation 3.

		Fitted model		Model 1 (2PL)	Model 2 (T2PLT)	Model 3 (N2PLT)
True	Model 3		Q_1	11380.77	11124.27	11065.03
model	(N2PLT)	DIC	Median	11412.16	11153.87	11098.49
			Q_3	11488.77	11226.28	11159.71
			IQR	107.99	102.01	94.67
			Q_1	–5-903.97	–5658.31	–5634.16
		log-PsBF	Median	–5870.39	–5620.26	–5595.36
			Q_3	–5856.31	–5604.20	–5590.11
			IQR	47.65	54.11	44.05

log-PsBF. The results of Bayesian model assessment based on 50 replications are shown in **Table 6**.

From **Table 6**, we find that when the Model 3 (N2PLTM model) is the true model, the Model 3 is chosen as the best-fitting model according to the results of the DIC and log-PsBF, which is what we expect to see. The medians of DIC and log-PsBF are respectively 11098.49 and −5595.36. The Model 2 (T2PLTM model) is the second best fitting model, which is attributed to the fact that the Model 2 with testlet random effect as well as the Model 3 also can capture the dependency structure between items. The differences between Model 3 and Model 2 in the median of DIC and log-PsBF are −55.38 and 24.9, respectively. However, compared the T2PLTM model, the N2PLTM model with the testlet discrimination parameter α is more flexible and the fitting is more sufficient. The Model 1 (2PL model) is worst-fitting model. The medians of DIC and log-PsBF are respectively 11412.16 and −5870.39. The differences between Model 3 model and Model 1 in the median of DIC and log-PsBF are −313.67 and 275.03, respectively. This is because the Model 1 do not consider the complicated and interrelated sets of items, thus it can not improve the model fitting for the testlet item response data. In summary, the Bayesian assessment criteria is effective for identifying the true models and it can be used in the subsequent empirical example analysis.

TABLE 7 | The results of Bayesian model assessment in the real data.

Model	DIC	log-PsBF
2PLM	44179.93	–22021.39
T2PLTM	40796.35	–20794.23
N2PLTM	**40632.52**	**–20708.47**

The meaning of the bold values is the best fitting model.

5. EMPIRICAL EXAMPLE

To illustrate the applicability of the testlet IRT modeling method to large-scale test assessments, we consider a data set of students' English reading comprehension test for Maryland university (Tao et al., 2013). A total of 1,289 students take part in the test and answer 28 items. The 28 items consist of 4 testlets. Testlet 1 is formed by Items 1 to 8, that is, $d(1) = \cdots = d(8) = 1$; Testlet 2 by Items 9 to 15, that is, $d(9) = \cdots = d(15) = 2$; Testlet 3 by Items 16 to 23, that is, $d(16) = \cdots = d(23) = 3$; and Testlet 4 by Items 24–28, that is, $d(24) = \cdots = d(28) = 4$. The following prior distributions are used to analyze the data. That is,

$$a_j \sim N(0, 100)\,\mathrm{I}(0, +\infty), \; b_j \sim N(0, 100), \; j = 1, \ldots, 28,$$
$$\theta_i \sim N(0, 1), \; \eta_{id(j)} \sim N(0, 1), \; i = 1, \ldots, 1289, \; j = 1, \ldots, 28.$$

We consider three models to fit the real data. The three models are 2PLM, T2PLTM and N2PLTM, respectively. The slice-Gibbs algorithm is applied to estimate the parameters of the three models. The slice-Gibbs sampling is iterated 20,000 iterations, with a burn-in period of 10,000 iterations. The convergence of the chains is checked by PSRF, which are <1.1. The item parameters of the N2PLTM are estimated and the item parameter estimators and the corresponding standard deviations are provided in **Table 7**. In the Bayesian frame work, the 95% highest posterior density intervals (HPDI) are calculated as confidence regions for the item parameters and are given in the columns labeled HPDI$_a$ and HPDI$_b$ in **Table 8**.

Based on the results of Bayesian model selection form **Table 7**, we find that the N2PLTM is the best fitting model compared to the other two models. The DIC and log-PsBF are respectively 40632.52 and −20708.47. The second best fitting model is T2PLTM. The differences between N2PLTM and T2PLTM in the DIC and log-PsBF are −163.83 and 85.76, respectively. The

TABLE 8 | The estimation results of item parameter for the real data.

Testlets	Para.		EAP		SD		HPDI	
	a	b	$\hat{a}$	$\hat{b}$	SD$_a$	SD$_b$	HPDI$_a$	HPDI$_b$
1	a_1	b_1	1.0000	0.0000	0.0000	0.0000	[1.0000, 1.0000]	[0.0000, 0.0000]
1	a_2	b_2	1.6319	0.2606	0.0116	0.0001	[1.4281, 1.8411]	[0.2308, 0.2845]
1	a_3	b_3	0.7215	0.7808	0.0053	0.0017	[0.5837, 0.8673]	[0.6971, 0.8575]
1	a_4	b_4	0.6302	-0.2913	0.0033	0.0015	[0.5278, 0.7525]	[−0.3747, −0.2197]
1	a_5	b_5	0.8039	0.6052	0.0062	0.0007	[0.6385, 0.9471]	[0.5509, 0.6577]
1	a_6	b_6	0.7998	0.6283	0.0046	0.0010	[0.6667, 0.9380]	[0.5528, 0.6832]
1	a_7	b_7	1.1367	0.2697	0.0066	0.0004	[0.9717, 1.2945]	[0.2261, 0.3114]
1	a_8	b_8	1.1849	–0.0253	0.0053	0.0006	[1.0291, 1.3164]	[−0.0760, 0.0236]
2	a_9	b_9	0.8047	–0.7197	0.0018	0.0013	[0.7168, 0.8845]	[−0.7981, −0.6511]
2	a_{10}	b_{10}	0.6128	–0.7850	0.0016	0.0030	[0.5314, 0.6864]	[−0.8908, −0.6853]
2	a_{11}	b_{11}	1.6674	–0.0463	0.0069	0.0002	[1.5081, 1.8327]	[−0.0772, −0.0140]
2	a_{12}	b_{12}	1.0907	–0.2133	0.0076	0.0024	[0.9463, 1.2035]	[−0.3290, −0.1994]
2	a_{13}	b_{13}	1.7084	0.0546	0.0099	0.0001	[1.5124, 1.9014]	[0.0292, 0.0800]
2	a_{14}	b_{14}	1.0951	–0.0775	0.0047	0.0007	[0.9635, 1.2267]	[−0.1271, −0.0213]
2	a_{15}	b_{15}	0.9024	–0.1817	0.0042	0.0013	[0.7719, 1.0226]	[−0.2476, −0.1093]
3	a_{16}	b_{16}	0.6347	0.5639	0.0057	0.0011	[0.4895, 0.7859]	[0.4997, 0.6370]
3	a_{17}	b_{17}	0.7751	0.1933	0.0058	0.0011	[0.6331, 0.9295]	[0.1275, 0.2588]
3	a_{18}	b_{18}	1.5116	–0.6624	0.0045	0.0004	[1.3786, 1.6420]	[−0.7092, −0.6226]
3	a_{19}	b_{19}	0.4526	0.5646	0.0040	0.0023	[0.3234, 0.5688]	[0.4703, 0.6521]
3	a_{20}	b_{20}	0.6325	0.7146	0.0054	0.0017	[0.4886, 0.7769]	[0.6321, 0.7972]
3	a_{21}	b_{21}	0.9391	–0.7392	0.0024	0.0011	[0.8374, 1.0301]	[−0.8025, −0.6775]
3	a_{22}	b_{22}	1.0175	–0.2715	0.0036	0.0008	[0.8983, 1.1347]	[−0.3266, −0.2105]
3	a_{23}	b_{23}	1.0722	–0.3727	0.0037	0.0009	[0.9526, 1.1831]	[−0.4389, −0.3178]
4	a_{24}	b_{24}	2.0055	–0.0069	0.0116	0.0002	[1.7917, 2.2080]	[−0.0349, 0.0216]
4	a_{25}	b_{25}	0.7821	0.4765	0.0052	0.0011	[0.6391, 0.9178]	[0.4068, 0.5391]
4	a_{26}	b_{26}	1.5236	0.2656	0.0103	0.0002	[1.3277, 1.7270]	[0.2388, 0.2969]
4	a_{27}	b_{27}	1.1934	0.3662	0.0084	0.0003	[1.0189, 1.3794]	[0.3316, 0.4050]
4	a_{28}	b_{28}	0.6847	–0.1442	0.0045	0.0016	[0.5563, 0.8153]	[−0.2222, −0.0667]

Para. denotes the interest parameters. EAP denotes the expected a priori estimation. SD denotes the standard deviation. HPDI denotes the 95% highest posterior density intervals.

2PL model is worst-fitting model. The DIC and log-PsBF are respectively 44179.93 and –22021.39.

From **Table 8**, we find that for each testlet, the four items with highest discrimination are 2, 13, 18, and item 24, respectively. The expected a posteriori (EAP) estimations for the four item discrimination parameters are 1.6319, 1.7084, 1.5116, and 2.0055. The four most difficult items in each testlet are 3, 13, 20, and item 25 in turn. The EAP estimations for the four item difficulty parameters are 0.7808, 0.0546, 0.7146, and 0.4765. Compared to the items in the other three testlets, the items in the testlet 2 are relatively easy because the EAP estimates of the difficulty parameters (b_9, b_{10}, b_{11}, b_{12}, b_{14}, and b_{15}) are <0. In addition, the SD is between 0.0000 and 0.0116 for the discrimination parameters, between 0.0000 and 0.0030 for the difficulty parameters.

6. CONCLUDING REMARKS

To explore the relations between items with dependent structure, this current study proposes a N2PLTM and presents a effective Bayesian sampling algorithm. More specifically, an improved Gibbs sampling algorithm based on auxiliary variables is developed for estimating N2PLTM. The slice-Gibbs sampling algorithm overcomes the traditional Gibbs sampling algorithm's dependence on the conjugate prior for complex IRT model, and avoids some shortcomings of the Metropolis algorithm (such as sensitivity to step size, severe dependency on the candidate function or tuning parameter). Based on different simulation conditions, we find that the slice-Gibbs sampling algorithm can provide accurate parameter estimates in the sense of having small Bias and MSE values. In addition, the average Bias and MSE for item parameters are relatively unchanged under the three different prior distributions. The slice-Gibbs sampling algorithm allows for informative or non-informative priors of the item parameters and is not sensitive to the specification of priors. In summary, the algorithm is effective and can be used to analyze the empirical example.

However, the computational burden of the slice-Gibbs sampling algorithm becomes intensive especially when a large number of examinees or the items is considered, or a large number of the MCMC sample size is used. Therefore, it is desirable to develop a standing-alone R package associated with C++ or Fortran software for more extensive large-scale assessment program.

In addition, the new algorithm based on auxiliary variables can be extended to estimate some more complex item response and response time models, e.g., graded response model, Weibull response time model and so on.

DATA AVAILABILITY STATEMENT

The datasets analyzed in this manuscript are not publicly available. Requests to access the datasets should be directed to Bao Xu, xubao97@163.com.

AUTHOR CONTRIBUTIONS

JL and JZ completed the writing of the article, original thoughts, and provided key technical support. JL and ZZ provided key technical support. BX provided the data. JT and JL completed the article revisions. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (grant no. 12001091) and China Postdoctoral Science Foundations (grant nos. 2021M690587 and 2021T140108). In addition, this work was also supported by the Fundamental Research Funds for the Central Universities of China (grant no. 2412020QD025).

REFERENCES

Aitkin, M. (1991). Posterior bayes factor (with discussion). *J. R. Stat. Soc. B* 53, 111–142. doi: 10.1111/j.2517-6161.1991.tb01812.x

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle," in *Second International Symposium on Information Theory*, eds B. N. Petrov and F. Csaki (Budapest: Academiai Kiado), 267–281.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *J. Educ. Stat.* 17, 251–269. doi: 10.3102/10769986017003251

Ando, T. (2011). Predictive bayesian model selection. *Am. J. Math. Manag. Sci.* 31, 13–38. doi: 10.1080/01966324.2011.10737798

Berger, J. O., and Pericchi, L. R. (1996). The intrinsic Bayes factor for linear models. in *Bayesian Statistics 5*, eds J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Oxford: Oxford University Press), 25–44.

Birnbaum, A. (1957). *Efficient Design and Use of Tests of a Mental Ability For Various Decision Making Problems*. Series Report No. 58-16. Randolph Air Force Base, TX: USAF School of Aviation Medicine.

Bishop, C. (2006). *Slice Sampling. Pattern Recognition and Machine Learning*. New York, NY: Springer.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* 37, 29–51.

Bock, R. D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. *Psychometrika* 46, 443–459.

Bradlow, E. T., Wainer, H., and Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika* 64, 153–168.

Brooks, S. P., and Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *J. Comput. Graphical Stat.* 7, 434–455.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York, NY: Springer.

Chib, S., and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *Am. Stat.* 49, 327–335.

Cook, K. F., Dodd, B. G., and Fitzpatrick, S. J. (1999). A comparison of three polytomous item response theory models in the context of testlet scoring. *J. Outcome Meas.* 3, 1–20.

Damien, P., Wakefield, J., and Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by auxiliary variables, *J. R. Stat. Soc. B* 61, 331–344.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.* 39, 1–38.

Eckes, T. (2014). Examining testlets effects in the TestDaF listening section: a testlet response theory modeling approach. *Lang. Test.* 31, 39–61. doi: 10.1177/0265532213492969

Eckes, T., and Baghaei, P. (2015). Using testlet response theory to examine local dependence in C-Testes. *Appl. Meas. Educ.* 28, 1–14. doi: 10.1080/08957347.2014.1002919

Fox, J.-P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. New York, NY: Springer.

Fox, J.-P., and Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika* 66, 269–286. doi: 10.1007/BF02294839

Geisser, S., and Eddy, W. (1979). A predictive approach to model selection. *J. Am. Stat. Assoc.* 74, 153–160.

Gelfand, A. E. (1996). "Model determination using sampling-based methods," in W. R. Gilks, S. Richardson, and D. J. Spiegelhalter *Markov Chain Monte Carlo in Practice* (London: Chapman-Hall), 145–161.

Gelfand, A. E., and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *J. R. Stat. Soc. Ser. B* 56, 501–514.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). "Model determination using predictive distributions with implementation via sampling-based methods (with discussion)," in *Bayesian Statistics 4*, eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (Oxford: Oxford University Press), 147–167.

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–472.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* 6, 721–741.

Gibbons, R. D., and Hedeker, D. R. (1992). Full-information bi-factor analysis. *Psychometrika* 57, 423–436.

Glas, C. A. W., Wainer, H., and Bradlow, E. T. (2000). "Maximum marginal likelihood and expected a posteriori estimation in testlet-based adaptive testing," in *Computerized Adaptive Testing, Theory and Practice*, eds W. J. van der Linden, and C. A. W. Glas (Boston, MA: Kluwer-Nijhoff), 271–288.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97–109.

Higdon, D. M. (1998). Auxiliary variable methods for Markov Chain Monte Carlo with applications. *J. Am. Stat. Soc.* 93, 585–595.

Jiao, H., Wang, S., and He, W. (2013). Estimation methods for one-parameter testlet models. *J. Educ. Meas.* 50, 186–203. doi: 10.1111/jedm.12010

Jiao, H., Wang, S., and Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *J. Appl. Meas.* 6, 311–321.

Kass, R. E., and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Stat. Assoc.* 90, 928–934.

Li, Y., Bolt, D. M., and Fu, J. (2006). A comparison of alternative models for testlets. *Appl. Psychol. Meas.* 30, 3–21. doi: 10.1177/0146621605275414

Lord, F. M., and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Lu, J., and Wang, C. (2020). A response time process model for not reached and omitted items. *J. Educ. Meas.* 57, 584–620. doi: 10.1111/jedm.12270

Lu, J., Wang, C., Zhang, J., and Tao, J. (2020). A mixture model for responses and response times with a higher-order ability structure to detect rapid guessing behaviour. *Br. J. Math. Stat. Psychol.* 73, 262–288. doi: 10.1111/bmsp.12175

Lu, J., Zhang, J. W., and Tao, J. (2018). Slice-Gibbs sampling algorithm for estimating the parameters of a multilevel item response model. *J. Math. Psychol.* 82, 12–25. doi: 10.1016/j.jmp.2017.10.005

Meng, X.-L., and van Dyk, D. A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika* 86, 301–320. doi: 10.1093/biomet/86.2.301

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1092. doi: 10.1063/1.1699114

Min, S. C., and He, L. Z. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Lang. Test.* 31, 453–477. doi: 10.1177/0265532214527277

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika* 51, 177–195. doi: 10.1007/BF02293979

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Appl. Psychol. Meas.* 16, 159–176. doi: 10.1177/014662169201600206

Neal, R. (2003). Slice sampling. *Ann. Stat.* 31, 705–767. doi: 10.1214/aos/1056562461

Newton, M. A., and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *J. R. Stat. Soc. Ser. B* 56, 3–48.

Pinheiro, P. C., and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J. Comput. Graphical Stat.* 4, 12–35.

Raftery, A. E. (1996). "Hypothesis testing and model selection," in *Markov Chain Monte Carlo in Practice*, eds W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Washington, DC: Chapman & Hall), 163–187.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *J. Educ. Meas.* 47, 361–372. doi: 10.1111/j.1745-3984.2010.00118.x

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monogr. Suppl.* 17, 1–100.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464.

Sireci, S. G., Thissen, D., and Wainer, H. (1991). On the reliability of testlet-based tests. *J. Educ. Meas.* 28, 237–247.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. R. Stat. Soc. Ser. B* 64, 583–639. doi: 10.1111/1467-9868.00353

Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* 82, 528–550.

Tao, J., Xu, B., Shi, N.-Z., and Jiao, H. (2013). Refining the two-parameter testlet response model by introducing testlet discrimination sparameters. *Jpn. Psychol. Res.* 55, 284–291. doi: 10.1111/jpr.12002

Thissen, D., Steinberg, L., and Mooney, J. A. (1989). Trace lines for testlets: A use of multiple categorical response models. *J. Educ. Meas.* 26, 247–260.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussions). *Ann. Stat.* 22, 1701–1762.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy. *Psychometrika* 72, 287–308. doi: 10.1007/s11336-006-1478-z

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: the 1991 law school admissions test as an example. *Appl. Meas. Educ.* 8, 157–186.

Wainer, H., Bradlow, E. T., and Du, Z. (2000). "Testlet response theory: an analog for the 3PL model useful in testlet-based adaptive testing. in *Computerized Adaptive Testing: Theory and Practice*, eds W. J. van der Linden and C. A. W. Glas (Dordrecht; Boston; London: Kluwer Academic), 245–269.

Wainer, H., Bradlow, E. T., and Wang, X. (2007). *Testlet Response Theory and Its Applications*. Cambridge: Cambridge University Press.

Wainer, H., and Kiely, G. L. (1987). Item clusters and computerized adaptive testing: a case for testlet. *J. Educ. Meas.* 24, 185–201.

Wang, C., Fan, Z., Chang, H.-H., and Douglas, J. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *J. Educ. Behav. Stat.* 38, 381–417. doi: 10.3102/1076998612461831

Wang, C., Xu, G., and Shang, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika* 83, 223–254. doi: 10.1007/s11336-016-9525-x

Wang, W.-C., and Wilson, M. (2005a). Exploring local item dependence using a random-effects facet model. *Appl. Psychol. Meas.* 29, 296–318. doi: 10.1177/0146621605276281

Wang, W.-C., and Wilson, M. (2005b). The Rasch testlet model. *Appl. Psychol. Meas.* 29, 126–149. doi: 10.1177/0146621604271053

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *J. Educ. Meas.* 30, 187–213.

Zhan, P., Jiao, H., and Liao, D. (2017). Cognitive diagnosis modelling incorporating item response times. *Br. J. Math. Stat. Psychol.* 71, 262–286. doi: 10.1111/bmsp.12114

Zhan, P., Li, X., Wang, W.-C., Bian, Y., and Wang, L. (2015). The multidimensional testlet effect cognitive diagnostic models. *Acta Psychol. Sin.* 47, 689–701. doi: 10.3724/SP.J.1041.2015.00689

Zhan, P., Liao, M., and Bian, Y. (2018). Joint testlet cognitive diagnosis modeling for paired local item dependence in response times and response accuracy. *Front. Psychol.* 9:607. doi: 10.3389/fpsyg.2018.00607

Zhan, P., Wang, W.-C., Wang, L., and Li, X. (2014). The multidimensional testlet-effect Rasch model. *Acta Psychol. Sin.* 46, 1208–1222. doi: 10.3724/SP.J.1041.2014.01208

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Lang. Test.* 27, 119–140. doi: 10.1177/0265532209347363

Zhang, Z., Hamagami, F., Wang, L., Grimm, K. J., and Nesselroade, J. R. (2007). Bayesian analysis of longitudinal data using growth curve models. *Int. J. Behav. Dev.* 31, 374–383. doi: 10.1177/0165025407077764

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Psychology

ORIGINAL RESEARCH
published: 13 January 2022
doi: 10.3389/fpsyg.2021.498824

Self-Efficacy Beliefs of University Students: Examining Factor Validity and Measurement Invariance of the New Academic Self-Efficacy Scale

Andrea Greco[1]*, Chiara Annovazzi[2]*†, Nicola Palena[1], Elisabetta Camussi[2], Germano Rossi[2] and Patrizia Steca[2]

[1] Department of Human and Social Sciences, University of Bergamo, Bergamo, Italy, [2] Department of Psychology, University of Milano-Bicocca, Milan, Italy

OPEN ACCESS

Edited by:
Pietro Cipresso,
University of Turin, Italy

Reviewed by:
Silvia Testa,
University of Aosta Valley, Italy
José M. Pêgo,
University of Minho, Portugal

***Correspondence:**
Andrea Greco
andrea.greco@unibg.it
Chiara Annovazzi
chiara.annovazzi@unimib.it;
c.annovazzi@univda.it

†Present address:
Chiara Annovazzi,
Department of Human and Social
Sciences, University of Valle d'Aosta,
Aosta, Italy

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: 17 September 2019
Accepted: 14 July 2021
Published: 13 January 2022

Citation:
Greco A, Annovazzi C, Palena N,
Camussi E, Rossi G and Steca P
(2022) Self-Efficacy Beliefs of
University Students: Examining Factor
Validity and Measurement Invariance
of the New Academic Self-Efficacy
Scale. Front. Psychol. 12:498824.
doi: 10.3389/fpsyg.2021.498824

Academic self-efficacy beliefs influence students' academic and career choices, as well as motivational factors and learning strategies promoting effective academic success. Nevertheless, few studies have focused on the academic self-efficacy of university students in comparison to students at other levels. Furthermore, extant measures present several limitations. The first aim of this study was to develop a reliable and valid scale assessing university students' self-efficacy beliefs in managing academic tasks. The second aim was to investigate differences in academic self-efficacy due to gender, years of enrollment, and student status. The study involved 831 students (age *M* = 21.09 years; *SD* = 1.34 years; 66.3% women) enrolled in undergraduate programs. Indicators of academic experiences and performance (i.e., number of exams passed and average exam rating) were collected. A new scale measuring students' academic self-efficacy beliefs was administered. Results from a preliminary Exploratory Factor Analysis were consistently supported by findings from a Confirmatory Factor Analysis. Multigroup CFA supported the presence of measurement invariance. Analyses revealed that the new scale has eight factors: "Planning Academic Activities," "Learning Strategies," "Information Retrieval," "Working in Groups," "Management of Relationships with Teachers," "Managing Lessons," "Stress Management," and "Thesis Work." Self-efficacy dimensions showed significant relations with academic experiences and students' performance indicators, as well as differences due to gender, years of enrollment, and student status. Findings are discussed in terms of practical implications for the implementation of intervention programs aimed at fostering self-efficacy beliefs and academic success.

Keywords: academic self-efficacy beliefs, scale development and validation, measurement invariance, university students, academic experiences, students' performance

INTRODUCTION

Perceived self-efficacy refers to personal beliefs on the ability to maintain established goals and perform successful actions (Bandura, 1997), particularly in difficult moments (McGeown et al., 2014). Self-efficacy could concern a general or a specific belief: the first refers to a general perceived ability to face stressful conditions, while the second refers to a particular context or situation. This paper focuses on specific self-efficacy beliefs related to the academic field defined as students' perceived abilities to successfully master different curricular areas, to self-regulate learning activities, and to manage relationships with teachers and peers (Bandura et al., 1996; Bassi et al., 2007). Academic self-efficacy has a significant and strong relationship with academic achievement (e.g., Pajares and Urdan, 2005; Ferla et al., 2009; Brausch, 2011), as cognitive and learning skills are necessary but not always sufficient (Bandura, 1997). Effective functioning requires two components, skills and efficacy beliefs to execute them appropriately, that act upon one another in a reciprocal fashion. Bandura referred this as a "reciprocal causation" in which the functioning of one component depends, in part, upon the functioning of the other (Bandura, 1997). In this way, students with high levels of self-efficacy can transform troubles into opportunities, think strategically to solve their difficulties and feel in control of a majority of stressors in their lives (Bandura, 1997). Academic self-efficacy is particularly salient when students have to cope with performance adversity or failure (Bong and Skaalvik, 2003). Research has also indicated that there is a positive relationship between academic self-efficacy beliefs and motivation (Ommundsen et al., 2005), in particular with intrinsic motivation (Walker et al., 2006). Self-efficacy beliefs contribute to motivation in several ways: determining the goals people set for themselves, how much effort they expend, how long they persevere in the face of difficulties, and their resilience to failures (Bandura, 1994). Moreover, Bandura (1997) declared that people derive information to evaluate efficacy beliefs from four primary sources: (1) mastery experiences; (2) vicarious experiences; (3) forms of persuasion, both verbal and otherwise; and (4) "physiological and affective states from which people partly judge their capableness, strength, and vulnerability to dysfunction." Furthermore, self-efficacy beliefs foster positive social and supportive relationships (Bandura et al., 1996) that may help to reduce anxiety and to improve stress management (Mayer et al., 2002), especially in challenging contexts, such as the academic environment.

Studies on university students are less numerous compared to those on younger students. Nevertheless, they clearly show that those who feel more competent are more self-determined, demonstrating more effective self-regulation strategies and higher persistence to maintain their academic goals (Ryan and Deci, 2006). Conversely, students with low levels of academic self-efficacy have less motivation and are more passive and disengaged (Vallerand, 2000; Komarraju and Dial, 2014). Academic self-efficacy is also a significant predictor of university students' course selection (Britner and Pajares, 2006; Komarraju and Dial, 2014), academic continuance, and achievement (Britner and Pajares, 2006). In particular, a study conducted by Amini (2002) showed that 21% of academic achievement was explained by students' academic self-efficacy, whereas other studies showed a relationship with academic persistence (Robbins et al., 2004; Gore, 2006), and final GPA (Robbins et al., 2004).

In literature, it was also possible to find studies about self-efficacy in specific domains, such as in the health, sports, and educational fields. For instance, the Cardiovascular Management Self-efficacy by Steca et al. (2015) is an instrument to monitor differences during interventions to improve good disease management. About sports, Feltz and Lirgg (2001) made a literature review of individual beliefs, team beliefs, and coaches' and leaders' beliefs in sports, to better understand the dynamics of teams. Gould et al. (1999) found that personal and team self-efficacies were one of the most important elements to influence Olympic performances. In the educational field, Caprara et al. (2006) analyzed the Teachers' self-efficacy beliefs as determinants of their job satisfaction and students' academic achievement.

Specific self-efficacy beliefs are not stable, but, in line with Bandura's thought, they are workable and flexible as they are strongly influenced by multiple sources (Bandura, 1997). A wide variety of educational, psychological, and pedagogical interventions are aimed at improving students' self-efficacy beliefs for their beneficial effects on numerous outcomes (Lane et al., 2004). Indeed, it is critical to validly and reliably measure academic self-efficacy beliefs to set and evaluate interventions. A multifaced instrument that is specifically designed to measure numerous areas related to the activities of undergraduate students (e.g., individual effort and self-management skills, learning strategies, social, leisure and extracurricular activities, interaction with peers and teachers) would further facilitate substantive research in this area (Bandura et al., 1996; Cheung and Kwok, 1998; Amenkhienan and Kogan, 2004). Unfortunately, scales currently used to measure self-efficacy beliefs in university students present several limitations. For instance, Advance Care Planning Self-Efficacy focuses only on one aspect of academic self-efficacy beliefs, namely students' concern about planning and their ability to do so. Furthermore, as underlined by the authors, it can be used only for a specific sample: medical doctors or those who need to initiate an Advance Care Planning (ACP) conversation (Baughman et al., 2015). Similar limitations characterize the Self-Efficacy Scale in Academic Behaviors in Students of Social Science by Blanco et al. (2013) and the Engineering Self-Efficacy Scales by Mamaril (2014). Scales referring to more than one aspect are often very short, like the Student Self-Report of Academic Self-Efficacy Scale by Hoover-Dempsey and Sandler (2005) composed of only three items, which are unable to cover the complexity of academic self-efficacy beliefs. As a consequence, more than one instrument has to be used to have a full measurement of self-efficacy beliefs in the university field. This could present difficulties because students consider answering several questionnaires as burdensome and not inherent to their academic path. Moreover, apart the Engineering Self-Efficacy Scales by Mamaril (2014), the other instruments presented lack of a psychometrics validation. Therefore, a significant limitation in this area is the missingness of a valid and reliable instrument that assesses the multidimensional nature of self-efficacy in the context of an academic setting. It is imperative that a self-report scale that reflects the various facets of self-efficacy be available. Even at the detriment of a lower specificity compared to other scales, only a multifaceted scale may guarantee the possibility to compare students from different degree programs adding information on differences in self-efficacy beliefs and on the differential effectiveness that interventions may have on students from different curricula and conditions.

Current Study

Given the limitations of extant scales and considering the need for a multifaceted instrument, the main aim of this study was to develop a reliable and valid scale assessing university students' self-efficacy beliefs in managing academic tasks; moreover, we also aimed to investigate differences in self-efficacy beliefs due to gender, years of enrollment in an undergraduate degree course, and supplementary-year student status. Finally, the final aim was to show how to use the scale to develop students' profiles.

In relation to the first aim, we tested the dimensionality of the scale. Various researchers suggested developing a scale starting with Exploratory Factor Analysis (EFA) to assess factor structure and to refine the item pool; EFA should be followed by Confirmatory Factor Analysis (CFA) using a different sample to confirm the measure's factor-structure and psychometric properties (Costello and Osborne, 2005; Henson and Roberts, 2006; Worthington and Whittaker, 2006; Cabrera-Nguyen, 2010). Worthington and Whittaker (2006) highlighted that EFA followed by CFA is the most common approach to scale development and validation. Our sample size was sufficient to utilize EFA in a random split-half of the sample (named "development sample"); the results were then verified using CFA in the second split-half (named "validation sample"). Further, we tested measurement invariance on the whole sample by a mean of Multi-group CFA. We then explored the internal consistency of the scale and examined the associations of self-efficacy factors with indicators of academic experiences and students' performance, namely the number of exams passed and average exam rating. We hypothesized that self-efficacy factors were significantly and negatively associated with negative experiences during academic life as students feeling more able to manage academic tasks should have better management of their academic lives. Furthermore, we hypothesized significant and positive relations between self-efficacy beliefs and indicators of students' performance. These relations were tested considering students at their different years of the undergraduate program separately to examine these associations more carefully. In line with Bandura's previous studies (1997), we expected that self-efficacy beliefs would be more predictive of the students' career achievements. In that sense, one of our aims was to investigate which particular domain of self-efficacy belief is the best support for the personal competences useful in academia. In line with the Life Design approach (Savickas et al., 2009) and with the study of Azizli et al. (2015), we hypothesized that the ability to plan activities could be one of the most helpful self-efficacy beliefs to achieve career goals.

In relation to the second aim, related to the investigation of differences in self-efficacy beliefs due to gender, years of enrollment in an undergraduate degree course, and supplementary-year student status, we firstly conducted the relative tests for measurement invariance. In line with Ceci et al. (2014) and Voyer and Voyer (2014), we hypothesized that women would show better competences helpful for academia. We also considered it interesting to analyze the presence or absence of gender differences in other competences, in particular with stress management, given women generally show higher levels of stress (Cohen and Janicki-Deverts, 2012). Lastly, we were interested in exploring to what degree self-efficacy competence is present in students in line with their exam schedule vs. supplementary-year students, given that no study exists yet on this issue. We expected that students in line with the exam schedule had a higher general level of self-efficacy beliefs than the supplementary-year students. In addition, we hypothesized that the university and the academic context played an important role in improving the level of self-efficacy, establishing higher and higher demands as the degree course progresses; as suggested by Bandura (1997), in fact, mastery experience or, in other words, performing a task successfully, is the most effective way to strength self-efficacy beliefs. We hypothesized a higher level of self-efficacy competence in the second or third enrollment year.

The final aim of this study was to show how to use the scale to develop students' profiles consisting of perceived "strengths" and "weaknesses" that match their self-efficacy beliefs, which correspond to the areas in which students deem themselves more or less able to behave effectively. The profiles may be used in intervention programs aimed at fostering self-efficacy beliefs and academic success, starting from a precise and reliable assessment.

MATERIALS AND METHODS

Participants and Study Design

The participants were undergraduate students recruited from 24 Italian universities. The inclusion criteria were as follows: (a) no other previous degree, (b) age under 26 years, and (c) fluent in the Italian language. Eligible students received written information about the study and signed an informed consent form; participation was voluntary and provided no remuneration. The students filled the instruments during the weeks of teaching, in a 20-min session during a lesson. The study has a cross-sectional design and was approved by the Ethics Committees of the university that conducted the research.

We recruited 831 students from 13 faculties or departments and 73 courses. Participants were mostly women ($n = 551$, 66.3%), with a mean age of 21.09 years (SD = 1.34; range: 19–25 years). The majority ($n = 369$, 44.4%) were psychology students, followed by economics (n = 135, 16.2%) and engineering ones (n = 79, 9.5%); the rest (n = 248, 29.9%) were from other 10 faculties or departments. Two hundred and thirty-six students (28.4%) were enrolled in the first year of the degree courses, 108 in the second (13.0%) and 407 in the third (49.1%), while 78 (9.4%) were supplementary-year students (2 missing data). Finally, 380 students (45.7%) declared that they were preparing their theses.

Variables and Instruments

The Academic Self-Efficacy Scale

By following Bandura's guidelines exactly (Bandura, 2006), we conducted a preliminary study in one of the faculties participating in the research; we involved a teacher and nine volunteer students enrolled in a 3-year undergraduate degree program with three students for each year of the degree course. In this phase, participants had to answer open questions related to tasks and problems that students could encounter in managing academic demands [i.e., What are the tasks or activities that a student, as you or someone like you, have to do to fulfill to successfully manage academic demands? What are the problems that could encounter in managing academic demands? What are the ways (tasks or activities to do) out of such problems?]; the teacher had to answer these questions from their point of view considering tasks and problems that students could encounter in managing academic demands. In the same phase, they were asked to imagine how they could face these tasks and problems. This procedure allowed us to identify activities and situations that students frequently have to manage in their academic lives, as well as successful behaviors.

The behaviors that emerged were then transformed into items to measure students' self-efficacy beliefs. These items theoretically measure: "planning," namely students' beliefs that they can carefully plan and organize tasks, activities, and goals to achieve about academic demands; "information retrieval," reflecting students' perceptions of their ability to regularly collect information about the course of study; "learning strategies," namely students' perceptions of their abilities to strictly comply with study responsibilities and rework the concepts of the field of study; "relationships," namely students' beliefs that they are able to work in groups using appropriate study strategies; "stress management," reflecting students' perceptions of their abilities to adequately control negative emotions about exam-taking; "thesis work," namely students' perceptions regarding their abilities to strictly comply with thesis writing. After this step, a teacher and three volunteer students in the same department verified the comprehensibility of the items. The teacher and students involved in this phase were different from those involved in the previous one. They reported some suggestions for the items to be more easily understood by other students. Some of these changes required inserting specific examples referring to information retrieval (e.g., "opening times, ways of contacting offices") or to learning strategy (e.g., "relating concepts together, making outlines, exam review"). At the end of these phases, 44 items were developed: 37 general items for all students enrolled in an undergraduate program, and 7 for students involved in the final thesis preparation. For each item, participants rated the strength of their beliefs on a 5-point response format ranging from 1 (perceived inability) to 5 (complete self-assurance in one's ability).

None of the students involved in the construction of the scale took part in the subsequent phase of the study.

Academic Experiences

A pool of 24 questions developed by the authors was used to measure four kinds of experiences relating to academic experiences: planning experiences (11 items, referring to how many times the respondent could have problems passing exams because of several reasons, α = 0.78, example item "How many times have you failed an exam because you did not sort what you had to study in the time you had left?); finding information experiences (seven items, referring to how many times the respondent encountered problems because of several reasons such as for example not paying attention to warnings displayed on the bulletin board, α = 0.68, example item "How many times have you had problems because you did not find out about the exam format ahead of time?"); learning experiences (three items, referring to how many times the respondent could have problems getting to the exam unprepared because of several reasons such as not using appropriate learning strategies or focusing on less relevant concepts of a field of study, α = 0.64, example item "How many times have you focused on less relevant concepts in what you were studying and overlooked more important ones?); stress (three items, referring to how many times the respondent could have difficulty taking an exam because of several reasons such as being overwhelmed by anxiety, α = 0.79, example item "How many times have you skipped an exam because you were overcome with anxiety?"). All the items were rated using a 5-point Likert scale, ranging from "never" (1) to "very often" (5); the scores were calculated as mean item scores, where higher scores indicate more negative academic experiences.

Students' Performance Indicators

Indicators of students' performance were collected for each participant and are relative to the number of exams passed, the number of exams required each year by rules of the degree course, and the average exam rating. Given there are different rules for different degree courses (i.e., the number of exams for each degree course, the number of exams due each year for each degree course), the number of exams passed and the number of exams required each year by rules of the degree course were used to calculate a proportion of exams passed per participant; this new variable was used in the subsequent analyses. Students were also asked to indicate information about the year of the degree course in which they were enrolled and their status (in line with the exam schedule vs. supplementary-year students).

Data Analysis

The items of the new scale were preliminarily submitted to analyses to check the normal distribution by calculating mean, standard deviation, and indices of skewness and kurtosis; West et al. (1995) recommend concern if skewness > |2| and kurtosis > | 7|.

Students Not Involved in the Thesis Work

For students not involved in their thesis work, the total sample was later randomly divided into two halves. The first sample was used to perform an EFA (DEVELOPMENT SAMPLE, n = 414) and the second was used to perform a CFA for validating the EFA symptom structure (VALIDATION SAMPLE, n = 417). To avoid problems with missing data, the 7 items developed for students involved in thesis work were excluded from these analyses, because these items were filled out only by students in the situation proposed.

On DEVELOPMENT SAMPLE, the Kaiser Meyer Olkin (KMO) and Bartlett's test of sphericity were run to be sure that the correlation matrix could be subjected to analyses (KMO should be >0.5; Bartlett's test of sphericity should be significant). Horn's method of parallel analysis was used to identify the number of factors to be extracted using EFA (Horn, 1965). Horn's method was chosen because of its merits as an objective test for identifying the dimensionality of multivariate data (Hubbard and Allen, 1987). Horn's method is, in fact, more accurate than the Cattell scree test or the Kaiser-Guttman criteria: judging the elbow of a scree plot could reflect a sampling error, while an eigenvalue greater than one tends to retain too many factors (Hubbard and Allen, 1987; Netemeyer et al., 2003). EFA with the Promax oblique rotation was used to analyze the items on the Academic Self-Efficacy Scale. Oblique rotation was used because the factors extracted from the Academic Self-efficacy Scale are likely to correlate with each other. In the first step, all 37 general items were included. Subsequent factor analyses were conducted in a stepwise fashion to eliminate items until a stable factor solution emerged. Items that had a factor loading < 0.32 were excluded, and, after the first step, items that loaded at >0.32 on more than one factor were excluded. Loadings in the 0.32 range or above are generally considered the cut-off on substantial loadings (Comrey and Lee, 1992).

On VALIDATION SAMPLE, CFA was conducted and Maximum Likelihood (ML) was used as an estimation method. Hu and Bentler's guidelines for various fit (1999) indices were used to determine whether the expected model fits the data. The chi-square test statistic was used but considering the sensitivity of the chi-square statistic to the sample size other goodness of fit indices were considered as the root-mean square error of approximation (RMSEA) and the standardized root-mean-square residual (SRMR). RMSEA and SRMR ≤ 0.08 were interpreted as a reasonable fit. Moreover, it would be desirable to additionally report the comparative fit index (CFI) and the Tucker Lewis index (TLI). However, cases where the RMSEA of the null model is <0.158 render the CFI and the TLI non-interpretable (Kenny, 2020). Hence, such incremental indices were considered only when the null model RMSEA was above.158. CFI and TLI ≥ 0.90 were interpreted as reasonable.

Students Involved in the Thesis Work

With the same procedure indicated above, a separate EFA in a random subsample followed by a CFA in the other subsample were performed to test the dimensionality of the seven items developed for students involved in thesis work. Moreover, on the total subsample of students involved in thesis preparation, an overall CFA was performed to test the model resulted from the analyses on the whole set of 37 items, adding the seven items developed for students involved in thesis work. For both students involved and those not involved in thesis work, Cronbach's alpha, McDonald's (1999) omega, and the items' inter-correlations coefficients were performed on the total sample to examine internal consistency. Cronbach's Alpha and McDonald's omega below 0.60 are unacceptable (Nunnally and Bernstein, 1994), whereas the items' inter-correlations coefficients that are higher than 0.30 are adequate (Nunnally and Bernstein, 1994).

Validity, Measurement Invariance and Group Comparisons

To investigate the validity of the self-efficacy scale, we conducted correlations using all the scale scores computed as average item scores. For convergent validity, the relations among self-efficacy beliefs and academic experiences was assessed via Pearson correlation. Further, the relations among self-efficacy beliefs and students' performance indicators were also tested. Students at different years of the undergraduate program were tested separately to consider these associations more carefully. Following guidelines by Cohen (1988), we interpreted correlations as measures of effect size. Correlations were considered weak ($|0.10| < r < |0.29|$), moderate ($|0.30| < r < |0.49|$), or strong ($|0.50| < r < |1|$).

Furthermore, multi-group CFA were conducted on the whole sample to assess measurement-invariance (Blunch, 2012) for each of the three variables of interest: gender, status, and year of enrollment. Multi-group CFA were also conducted on the sample of 380 students preparing their thesis. Three different models were obtained and compared: (i) configural invariance, which served as a baseline model and where the structure is assumed to be the same in the various groups being compared (e.g., males vs. females); (ii) metric (or weak) invariance, where loadings are fixed to being equal across groups, and; (iii) scalar (or strong) invariance, where loadings and intercepts are fixed to be equal across groups. We considered metric and/or scalar invariance to be present when the corresponding models (ii and/or iii) fit the data, as well as model i (configural invariance), did. To compare the three models, we focused on the changes in RMSEA and SRMR (see also Lu et al., 2018; Zhao et al., 2019; Ma, 2020), since the $\chi 2$ difference test is too sensitive for the assessment of invariance with large samples ($N > 300$, Chen, 2007). Following Cheung and Rensvold (2002) and Chen (2007), we considered measurement invariance to be present when ΔRMSEA < 0.015 and ΔSRMR < 0.030. ΔCFI and ΔTLI were only reported if the null model RMSEA was < 0.158. Bayesian Information Criterion (BIC) values were also compared, with lower values indicating better fit and evidence of invariance (Cheung and Rensvold, 2002; Zhou et al., 2019). If both ii and iii forms of invariance were attained, we concluded that meaningful comparisons in the scores of the Academic Self-Efficacy Scale could be made for gender, and/or status, and/or year of enrollment. For such cases where invariance was assured, *t*-tests and univariate ANOVA were used to test the difference among profiles of the Academic Self-Efficacy Scale due to gender, year of enrollment in an undergraduate degree course, and students in line with the exam schedule vs. supplementary-year students (status).

Data analyses related to the normal distribution, EFA, Cronbach's alpha, items' inter-correlations, correlations, *t*-tests, and univariate ANOVA were performed using IBM SPSS Statistics (Version 22). Parallel analysis, CFA, and McDonald's omega were performed using MPlus software (Version 7) (Muthén and Muthén, 1998-2010). Multi-group CFA were performed with R (Version 4.0.3) (R Core Team, 2020) and R studio (Version 1.3.1093) (RStudio Team, 2020) using the R package lavaan (Rosseel, 2012). Missing values were treated via listwise deletion in SPSS and full information ML estimation in Mplus and R.

RESULTS

Preliminary Analysis

The average scores of the responses to the 44 items from all participants ranged from 2.32 to 4.32 ($SD_{\mathrm{MIN}} = 0.77$-$SD_{\mathrm{MAX}} = 1.13$). Furthermore, in line with recommendations by West et al. (1995), all the items showed an acceptable distribution; skewness and kurtosis showed no non-normally distributed items ($\text{Skewness}_{\text{MIN}}$ = –1.25-$\text{Skewness}_{\text{MAX}}$ = 0.64; $\text{Kurtosis}_{\text{MIN}}$ = –0.73-$\text{Kurtosis}_{\text{MAX}}$ = 1.12).

Factor Structure of the Academic Self-Efficacy Scale. Exploratory Factor Analysis

Data from Development Sample and 37 general items were used in these analyses. The Bartlett's sphericity test (χ^2 = 3628.64, $p < 0.001$) and the KMO = 0.85 have ensured that the correlation matrix could be subjected to factor analysis. The parallel analysis indicated that a seven-factor solution was the most appropriate. EFA was then conducted, with seven factors extracted. The initial pool of 37 general items, after subsequent factor analyses conducted in a stepwise fashion, was reduced to 30 (items are present in the **Supplementary Materials**). Four items were excluded because their loadings were lower than 0.32: "How well can you make friends with students who are stimulating for your degree course?"; "How good are you about consulting your representatives to find out about your rights?"; "How good are you at getting useful study advice by asking students who have already taken the tests?"; "How well can you gather useful study information by being present at other students' exams?". Three items were excluded because their loadings were above 0.32 on more than one factor: "How well can you critically judge the information other students give you?"; "How well can you get the materials you need on time to study for tests?"; "How well can you take advantage of appropriate and effective learning strategies (e.g., "relating concepts together, making outlines, exam review, etc.)?".

The pattern of factor loadings from the seven-factor exploratory measurement model for the self-efficacy scale's 30 items is given in **Table 1**.

The first extracted factor explains 9.32% of the variance. It showed loadings from six items assessing students' beliefs regarding their ability to carefully organize time, plan the number of exams, sort the study material, maintain a steady pace of study, and establish achievable goals concerning academic demands. This factor can be called "Planning Academic Activities." The second extracted factor explains 7.56% of the variance. It showed strong loadings from six items assessing students' beliefs regarding their ability to strictly comply with study tasks such as focus primarily on core concepts, create connections, enhance exam preparation, adequately reprocess and explain the study material. This factor can be called "Learning Strategies." The third extracted factor explains 7.29% of the variance. It showed loadings from six items assessing students' beliefs regarding their ability to regularly collect information about the course of study and the various examinations through the different sources available such as notice boards, administrative offices, and websites. This factor can be called "Information Retrieval." The fourth extracted factor explains 6.25% of the variance. It showed strong loadings from three items assessing students' beliefs regarding their ability to be good to create study groups and use adequate and productive strategies in this context. This factor can be called "Working in Groups." The fifth extracted factor explains 4.78% of the variance. It showed loadings from three items assessing students' beliefs regarding their ability to take an active role in classroom discussion, and refer to teachers for more information and clarification about courses and lessons. This factor can be called "Management of Relationships with Teachers." The sixth extracted factor explains 4.28% of the variance. It showed strong loadings from four items assessing students' beliefs regarding their ability to attend classes, keep focused even in challenging circumstances, take clear and helpful notes, and reprocess the main parts of a lesson. This factor can be called "Skills for lessons." The seventh and final extracted factor explains 3.51% of the variance. It consisted chiefly of two items assessing students' beliefs regarding their ability to adequately control exam-related anxiety, and discouragement after a failed exam. An appropriate name for this factor might be "Stress Management." The total variance explained by the seven factors extracted was 43.00%.

As shown in **Table 1**, no item displays a loading lower than 0.32. The extent of cross-loading between factors was moderate; the size of this secondary loading was usually small, below 0.32.

Factor Structure of the Academic Self-Efficacy Scale. Confirmatory Factor Analysis

Confirmatory factor analysis was conducted separately on data from Validation Sample using the 30 items; item selection to load on CFA factors was based on EFA loadings. **Table 1** presents the standardized factor loadings in Validation Sample. The fit of the CFA model to the data from the 417 students was acceptable [$\chi^2(384)$ = 930.206, $p < 0.001$; RMSEA = 0.058; SRMR = 0.067]; we therefore examined the RMSEA of the null model and found RMSEA null = 0.145. Therefore, we refrained from reporting the CFI or other incremental fit indices. Loadings from the CFA were comparable with those found in the EFA, identifying the seven factors.

Academic Self-Efficacy Scale Related to Thesis Work

Data from the 380 students that filled out the seven items developed for those involved in thesis preparation were used in these analyses. The sample was randomly split into two subsamples.

The first subsample (n = 190) was used to perform an EFA to test the dimensionality of the scale. The Bartlett's sphericity test (χ^2 = 678.46, $p < 0.001$) and the KMO = 0.87 have ensured that the correlation matrix could be subjected to factor analysis.

The pattern of factor loadings from the one-factor exploratory measurement model for the self-efficacy scale's 7 items is given in **Table 2**. The extracted factor explains 53.51% of the variance. It showed loadings from seven items assessing students' beliefs regarding their ability to strictly meet all graduation deadlines, to design, find, organize and regularly work to complete a good project for the thesis. This factor might be called "Thesis

TABLE 1 | Item percentage of response frequency and factors loadings from the Exploratory Factor Analysis in DEVELOPMENT SAMPLE and Confirmatory Factor Analysis in VALIDATION SAMPLE.

How well can you. . .	***Development Sample***								***Validation Sample***	
	% response	***PAA***	***LS***	***IR***	***WG***	***MRT***	***SL***	***SM***	***% response***	***Loadings^a***
. keep with the study schedule you set up	99.52	**0.81**	– 0.08	– 0.08	0.05	0.00	0.00	– 0.02	100	**0.77*****
. sort what you have to study in the time you have left to prepare for an exam	100	**0.79**	0.04	– 0.01	– 0.02	– 0.01	– 0.18	– 0.02	100	**0.71*****
. keep up continuous study habits throughout the school year	99.76	**0.70**	– 0.15	– 0.02	– 0.04	0.13	0.14	– 0.13	99.76	**0.73*****
. organize your time in order to finish a paper by the deadline	99.03	**0.70**	0.05	0.07	– 0.05	– 0.09	– 0.01	0.08	99.04	**0.71*****
. plan the number of exams you will take in each session based on how difficult they are	100	**0.53**	0.06	0.10	– 0.03	– 0.07	0.05	0.15	99.52	**0.59*****
. set achievable goals by knowing your abilities and your limitations	99.52	**0.39**	0.22	0.08	0.10	– 0.04	0.01	0.11	99.76	**0.54*****
. make connections, analogies and distinctions among the various subjects you are taking	100	– 0.08	**0.63**	0.08	– 0.04	0.01	0.00	0.03	100	**0.57*****
. at the exam, convey in writing what you'd studied	99.28	0.06	**0.62**	– 0.08	– 0.04	– 0.01	0.08–	–0.09	99.52	**0.65*****
. enhance your exam preparation with personalized, in-depth study	99.52	– 0.04	**0.56**	0.05	0.05	0.08	– 0.01	– 0.11	99.52	**0.47*****
. adjust your way of expressing yourself according to the situation and the person you're talking to	99.76	– 0.07	**0.56**	– 0.04	– 0.02	– 0.07	0.12	0.04	99.52	**0.55*****
. demonstrate your knowledge of that you've studied in an oral exam	98.79	0.11	**0.51**	0.00	0.11	– 0.09	0.03	0.00	98.80	**0.63*****
. focus on the main points of what you are studying	100	0.01	**0.51**	– 0.01	0.01	0.05	– 0.04	0.13	99.52	**0.55*****
. get the information you need about administrative offices (opening times, how to contact them.)	100	– 0.09	– 0.12	**0.76**	– 0.05	– 0.02	0.11	0.03	100	**0.67*****
. get information from the university website	99.76	– 0.03	0.10	**0.71**	– 0.06	– 0.01	– 0.13	0.06	99.76	**0.69*****
. regularly check the departmental notice board to get information about your degree course	100	0.10	0.00	**0.64**	0.00	0.11	– 0.07	– 0.10	100	**0.64*****
. get information on exam formats ahead of time	99.52	0.01	0.15	**0.46**	– 0.07	0.00	0.10	– 0.06	99.52	**0.61*****
. sign up for exams within the established timeline	99.76	0.04	– 0.23	**0.42**	0.08	– 0.13	0.31	0.11	99.76	**0.47*****
. find out ahead of time if there are any prerequisite exams to take in your degree course before beginning other courses	99.52	0.07	0.08	**0.42**	0.03	0.05	– 0.04	– 0.11	98.80	**0.47*****
. start efficient study groups	99.03	0.02	– 0.07	– 0.07	**0.82**	– 0.01	0.08	0.00	100	**0.76*****
. use good group study strategies (quiz each other, etc.)	99.52	0.00	0.09	– 0.01	**0.80**	– 0.07	– 0.08	– 0.06	99.76	**0.69*****
. work together productively by defining specific goals and tasks	99.03	– 0.04	0.00	– 0.01	**0.67**	0.13	0.03	0.06	99.76	**0.82*****
. raise your hand to ask the professor to explain parts of the lesson that you don't understand	99.76	– 0.03	– 0.01	0.06	– 0.05	**0.84**	0.00	0.02	99.76	**0.82*****
. participate actively in in-class discussion	99.76	0.00	0.00	– 0.08	0.05	**0.71**	0.00	0.10	99.76	**0.77*****
. go to your professors to get useful information on courses	98.79	0.06	0.03	0.16	0.16	**0.36**	0.05	– 0.04	99.52	**0.43*****
. stay focused in class even when is is noisy or crowded	99.76	0.10	0.09	– 0.09	– 0.18	0.11	**0.55**	– 0.01	99.52	**0.61*****
. attend class regularly even when the exam session approaches	100	– 0.18	0.08	0.15	0.11	– 0.06	**0.51**	0.02	99.76	**0.43*****
. take clear, useful notes in class	99.76	0.13	0.06	0.01	0.09	– 0.03	**0.49**	– 0.11	100	**0.70*****
. glean and reprocess the essential points in a lecture	99.76	– 0.02	0.30	– 0.06	– 0.01	0.08	**0.44**	0.06	99.52	**0.74*****
. keep exam anxiety under control	100	0.01	0.02	– 0.07	– 0.08	0.01	0.07	**0.69**	100	**0.87*****
. avoid getting discouraged when you fail an exam	98.79	0.03	– 0.02	0.03	0.07	0.10	– 0.12	**0.67**	99.04	**0.61*****

****$p < 0.001$. PAA, "Planning Academic Activities"; LS, "Learning Strategies"; IR, "Information Retrieval"; WG, "Working in Groups"; MRT, "Management of Relationships with Teachers"; SL, "Skills for Lessons"; SM, "Stress Management." [a]Items selected to load on CFA factors are based on EFA loadings. Bold items indicate factor membership.*

TABLE 2 | Item percentage of response frequency and factors loadings from the Exploratory Factor Analysis in a random subsample and Confirmatory Factor Analysis in the other subsample for the seven items related to the preparation of the thesis.

How well can you. . .	***EFA SUBSAMPLE***		***CFA SUBSAMPLE***	
	% response	***Loadings***	***% response***	***Loadings***
. select what is useful from all your research to write your thesis	99.47	0.80	98.95	0.74***
. use a clear and coherent structure to organize your research material for the thesis	98.95	0.76	98.42	0.69***
. devise a good project for your thesis	99.47	0.66	100	0.65***
. make good use of your advisor's suggestions to write your thesis	100	0.65	98.95	0.51***
. work continually in order to finish your thesis in time	100	0.65	98.95	0.65***
. use library resources to find materials for your thesis	100	0.62	98.42	0.68***
. respect all graduation deadlines (getting a thesis advisor, graduation application, handing in documents.)	100	0.61	98.95	0.60***

****$p < 0.001$.*

Work." As shown in **Table 2**, all items display adequate loadings, higher than 0.32.

Confirmatory factor analysis was conducted separately on the other subsample. **Table 2** presents the standardized factor loadings in these subsamples. The fit of the CFA model to the data from the 190 students was acceptable [$\chi^2(14) = 30.137$, $p < 0.01$; CFI = 0.96, TLI = 0.94; RMSEA = 0.078; SRMR = 0.038]. Loadings from the CFA were comparable with those found in the EFA, identifying one factor.

Factor Structure of the Academic Self-Efficacy Scale and the Academic Self-Efficacy Scale Related to Thesis Work

On the data from the 380 students that filled out the seven items developed for those involved in thesis preparation, an overall CFA was performed to test a model with eight factors, seven from the analyses on the whole set of 30 items, adding the "Thesis Work" factor. The fit of the CFA model to the data was acceptable [$\chi^2(601) = 1280.146$, $p < 0.001$; RMSEA = 0.055; SRMR = 0.066]. We therefore examined the RMSEA of the null model and found RMSEA null = 0.133. Therefore, we refrained from reporting the CFI or other incremental fit indices. Loadings from the CFA were comparable with those found in the previous CFA.

Reliability of the Academic Self-Efficacy Scale and Correlations Among Subscales

For each subscale, the score was calculated by computing the average score across items within a subscale (ranging from 1 to 5). All the factor scores showed an acceptable distribution; skewness and kurtosis showed normal distribution ($\text{Skewness}_{\text{MIN}} = -0.16$-$\text{Skewness}_{\text{MAX}} = 0.54$; $\text{Kurtosis}_{\text{MIN}} = -0.39$-$\text{Kurtosis}_{\text{MAX}} = 0.75$).

The analysis of reliability performed on the data collected from all participants (831 students for the Academic Self-efficacy Scale and 380 students for the Academic Self-efficacy Scale Related to Thesis Work) showed that the scale has adequate internal consistency for all factors. All Cronbach's alpha and McDonald's omega were adequate: "Planning Academic Activities" = $\alpha = 0.83$, $\omega = 0.83$; "Learning Strategies" = $\alpha = 0.75$, $\omega = 0.75$; "Information Retrieval" = $\alpha = 0.76$, $\omega = 0.76$; "Working in Groups" = $\alpha = 0.80$, $\omega = 0.80$; "Management of Relationships with Teachers" = $\alpha = 0.71$, $\omega = 0.73$; "Skills for lessons" = $\alpha = 0.68$, $\omega = 0.70$; "Stress Management" = $\alpha = 0.65$, $\omega = 0.65$; "Thesis Work" = $\alpha = 0.84$, $\omega = 0.85$. Moreover, the inter-correlations coefficients of items were all larger than.37, indicating adequate internal consistency.

As shown in **Table 3**, the self-efficacy factors were all positively and significantly correlated apart from "Information Retrieval" and "Stress Management," which were shown to be uncorrelated.

Measurement Invariance of the Academic Self-Efficacy Scale

Multigroup Confirmatory Factor analyses to test for measurement invariance showed that for both the whole sample (seven factors) and the sample of 380 students preparing their thesis (eight factors), measurement invariance could be deemed as present. Indeed, as **Table 4** shows, changes in RMSEA never exceeded 0.011, SRMR never exceeded 0.007, and that the BIC of the most parsimonious model (e.g., scalar invariance vs. metric invariance) were always the lowest. Hence, all comparisons for gender, status, and year of enrollment can be made (see below).

Correlations of Academic Self-Efficacy Factors With Indicators of Academic Experiences and Performance

We examined the correlations of the Academic Self-efficacy subscales with academic experiences and performance. As shown in **Table 3**, "Planning Academic Activities" was

TABLE 3 | Pearson correlations among Academic Self-efficacy factors and indicators of students' academic experiences and performance.

	PAA	LS	IR	WG	MRT	SL	SM	MTW
LS	0.41***	1						
IR	0.34***	0.23***	1					
WG	0.22***	0.23***	0.10**	1				
MRT	0.28***	0.40***	0.22***	0.32***	1			
SL	0.44***	0.47***	0.33***	0.12**	0.33***	1		
SM	0.13***	0.27***	0.00	0.13***	0.20***	0.11**	1	
MTW	0.47***	0.45***	0.35***	0.25***	0.41***	0.38***	0.17**	1
Planning experiences	−0.53**	−0.31**	−0.19*	− 0.12	−0.20*	−0.38**	− 0.05	− 0.16
Finding information experiences	−0.32**	−0.25**	−0.41**	0.17	− 0.15	−0.38**	0.17*	−0.38*
Learning experiences	−0.50**	−0.34**	− 0.14	0.07	− 0.08	−0.37**	− 0.02	− 0.18
Stress experiences	−0.42**	− 0.17	−0.20*	−0.24**	− 0.12	− 0.08	−0.37**	0.04
First year Proportion of exams passed	0.27**	0.05	0.12	0.09	0.14*	0.13*	0.01	–
Average exam rating	0.30**	0.22**	0.05	0.09	0.07	0.14*	0.00	–
Second year Proportion of exams passed	0.23*	0.34**	0.07	0.08	0.02	0.13	0.02	–
Average exam rating	0.21*	0.32**	− 0.03	0.14	0.23*	0.23*	− 0.09	–
Third year Proportion of exams passed	0.35**	0.19**	0.10	0.11**	0.23**	0.13*	0.06	0.28***
Average exam rating	0.40**	0.35**	0.13**	− 0.09	0.23**	0.27**	− 0.03	0.21***

**p* < 0.05; ***p* < 0.01; ****p* < 0.001. *PAA, "Planning Academic Activities"; LS, "Learning Strategies"; IR, "Information Retrieval"; WG, "Working in Groups"; MRT, "Management of Relationships with Teachers"; SL, "Skills for Lessons"; SM, "Stress Management"; MTW, "Management of Thesis Work."*

TABLE 4 | Fit indices for the assessment of measurement invariance.

	chisq	df	rmsea	srmr	bic
Configural (gender – 7 factors)	1600.091	768	0.051095	0.061165	61964.070
Metric (gender – 7 factors)	1633.939	791	0.050674	0.062409	61843.330
Scalar (gender – 7 factors)	1762.909	814	0.053	0.06438	61817.710
Configural (status – 7 factors)	1752.759	768	0.055619	0.061611	62318.070
Metric (status – 7 factors)	1797.327	791	0.055401	0.06253	62208.070
Scalar (status – 7 factors)	1837.998	814	0.05509	0.062632	62094.180
Configural (year – 7 factors)	2077.993	1152	0.058156	0.069731	53771.700
Metric (year – 7 factors)	2175.298	1198	0.058587	0.073902	53566.810
Scalar (year – 7 factors)	2331.973	1244	0.060662	0.07578	53421.290
Configural (gender – 8 factors)	2030.765	1202	0.06024	0.075723	35528.910
Metric (gender – 8 factors)	2067.303	1231	0.059797	0.077872	35393.180
Scalar (gender – 8 factors)	2153.694	1260	0.061099	0.079856	35307.310
Configural (status – 8 factors)	2389.350	1202	0.072295	0.076482	35504.680
Metric (status – 8 factors)	2451.060	1231	0.072415	0.078845	35394.280
Scalar (status – 8 factors)	2486.359	1260	0.071762	0.07892	35257.470

strongly and negatively correlated to negative experiences in planning and learning, and moderately and negatively associated with negative experiences in finding information and stress. “Learning Strategies” was moderately and negatively correlated to negative experiences in planning and learning, and weakly and negatively associated with negative experiences in finding information. “Information Retrieval” was moderately and negatively correlated to negative experiences in finding information, and weakly and negatively associated with negative experiences in planning and stress. “Working in Groups” was weakly and negatively correlated to negative experiences in stress. “Management of Relationships with Teachers” was weakly and negatively correlated to negative experiences in planning. “Skills for lessons” was moderately and negatively correlated to negative experiences in planning, finding information, and learning. “Stress Management” was moderately and negatively correlated to negative experiences in stress, while it was weakly and positively associated with negative experiences in finding information. “Management of Thesis Work” was moderately and negatively correlated to negative experiences in finding information.

Moreover, we examined the correlations of the Academic Self-efficacy subscales with indicators of students’ performance. The correlations were tested considering students at different years of the undergraduate program separately. As shown in **Table 3**, for the group of first year undergraduates, the proportion of exams passed was positively and weakly correlated to the “Planning Academic Activities,” “Management of Relationships with Teachers,” and “Skills for Lessons” subscales. The average exam rating was positively and moderately correlated to “Planning Academic Activities,” and weakly to the “Learning Strategies,” and “Skills for Lessons” subscales. For the group of second year undergraduates, the proportion of exams passed was positively and moderately correlated to “Learning Strategies,” and weakly to “Planning Academic Activities.” The average exam rating was positively and moderately correlated to “Learning Strategies,” and weakly to the “Planning Academic Activities,” “Management of Relationships with Teachers,” and “Skills for Lessons” subscales. For the group of third year undergraduates, the proportion of exams passed was positively and moderately correlated to “Planning Academic Activities,” and weakly to the “Learning Strategies,” “Working in Groups,” “Management of Relationships with Teachers,” “Skills for Lessons,” and “Management of Thesis Work” subscales. The average exam rating was positively and moderately correlated to “Planning Academic Activities” and “Learning Strategies,” and weakly associated with the “Information Retrieval,” “Management of Relationships with Teachers,” “Skills for Lessons,” and “Management of Thesis Work” subscales.

Academic Self-Efficacy Scale in Measuring Strengths vs. Weaknesses

Assessing self-efficacy beliefs allows us to develop profiles consisting of subjectively defined “strengths” and “weaknesses,” which reflect the areas in which students consider themselves more or less able to act effectively. **Figure 1** shows mean values of Academic Self-efficacy for the 831 students for the Academic Self-efficacy Scale and 380 students for the Academic Self-efficacy Scale Related to Thesis Work divided by gender. Both genders showed strengths in “Information Retrieval,” but weaknesses in “Management of Relationships with Teachers” and in “Working in Groups.” Furthermore, the results of the *t*-test showed a meaningful difference between male and female students in their levels of “Planning Academic Activities” [$t(\mathrm{df} = 828) = -2.64$, $p < 0.01$, Cohen’s $d = 0.19$],“Information Retrieval” [$t(\mathrm{df} = 828) = -4.31$, $p < 0.001$, Cohen’s $d = 0.31$],“Management of Relationships with Teachers” [$t(\mathrm{df} = 826) = 3.29$, $p < 0.01$, Cohen’s $d = 0.24$], “Skills for Lessons” [$t(\mathrm{df} = 828) = -5.07$, $p < 0.001$, Cohen’s $d = 0.36$], and “Stress Management” [$t(\mathrm{df} = 828) = 10.49$, $p < 0.001$, Cohen’s $d = 0.76$].

Figure 2 reports mean values of the Academic Self-efficacy factors separately for each year of enrollment in an undergraduate degree course. The three groups showed strengths in “Information Retrieval,” but weaknesses in “Working in Groups” and “Management of Relationships with Teachers.” Furthermore, the results of the univariate ANOVA and *post hoc* comparison based upon Tukey test showed a meaningful difference between first year students and second and third year students [$F(\mathrm{df} = 2, 751) = 9.46$, $p < 0.001$, $\eta^2 = 0.024$] in their

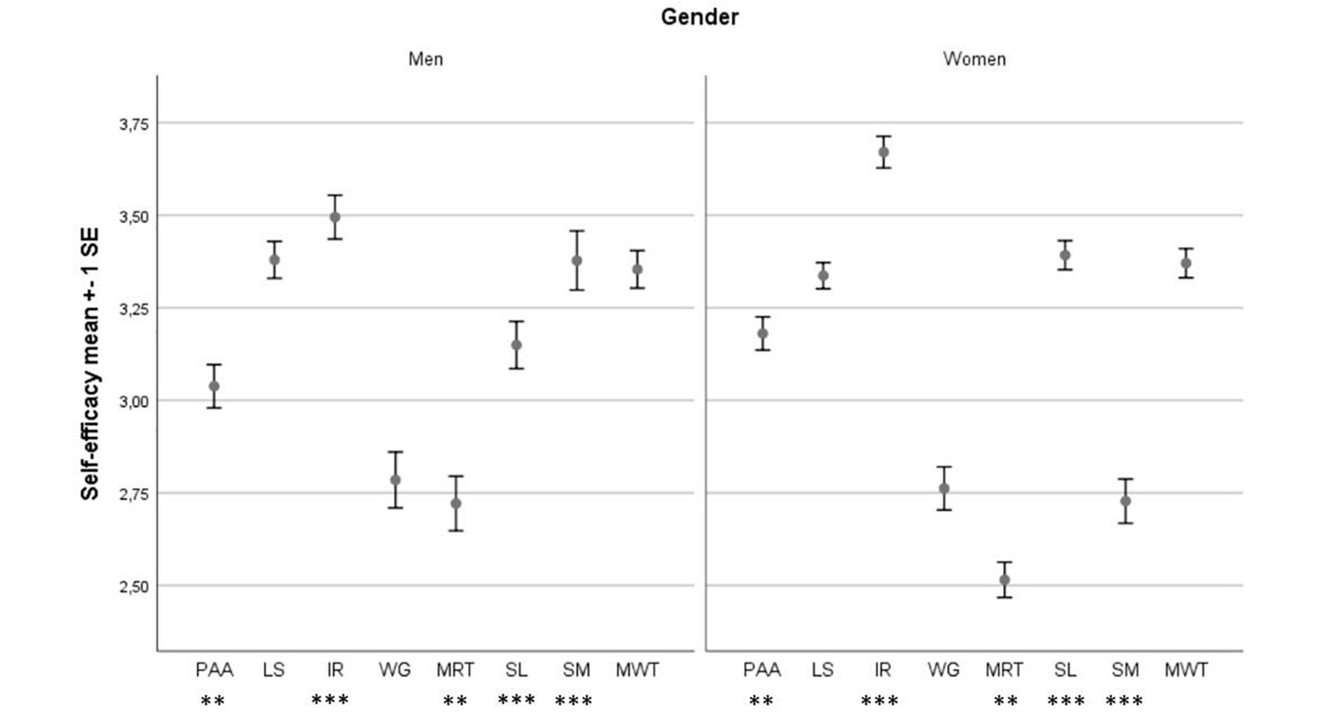


FIGURE 1 | Mean levels of the Academic Self-efficacy factors for men and women and *t*-test results. $^{**}p < 0.01$; $^{***}p < 0.001$; PAA, "Planning Academic Activities"; LS, "Learning Strategies"; IR, "Information Retrieval"; WG, "Working in Groups"; MRT, "Management of Relationships with Teachers"; SL, "Skills for Lessons"; SM, "Stress Management"; MWT, "Management of Thesis Work."

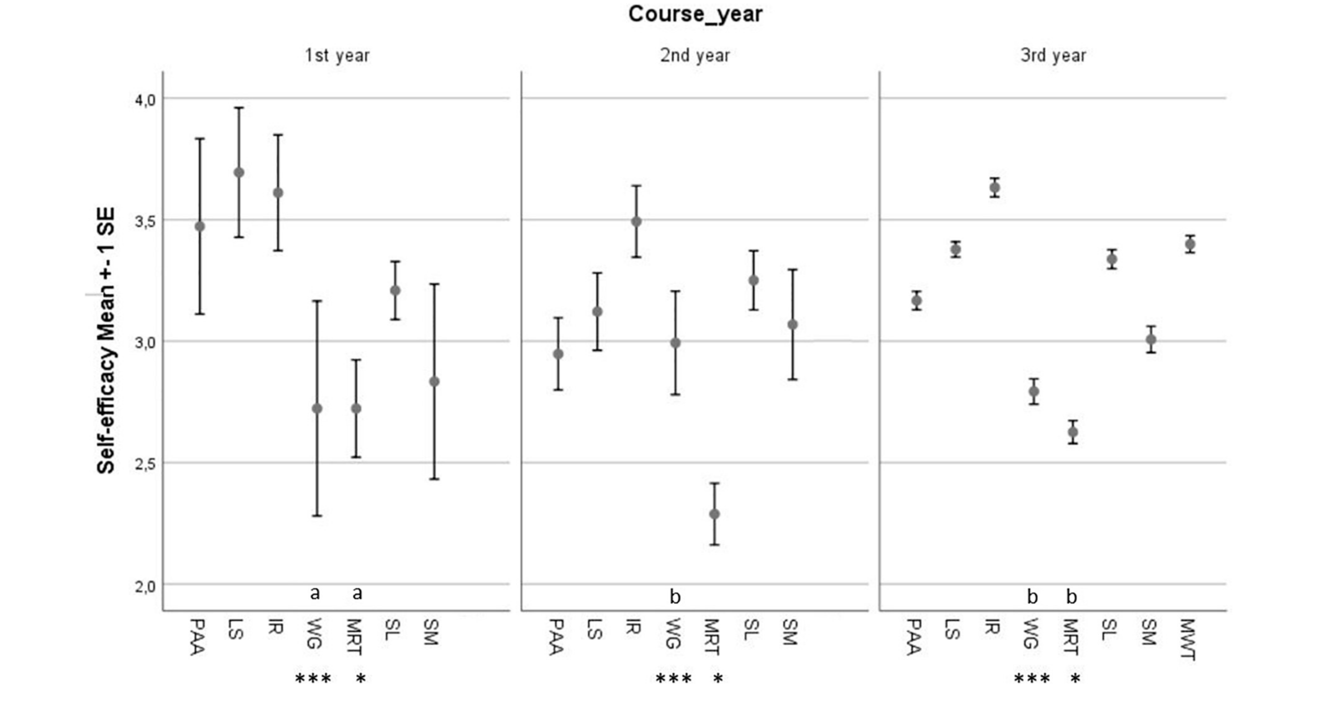


FIGURE 2 | Mean levels of the Academic Self-efficacy factors for students at the first, second, and third year of the undergraduate degree course and results from univariate analysis of variance and *post hoc* comparisons based upon Tukey tests. $^{*}p < 0.05$; $^{***}p < 0.001$; Different letters indicate significant differences among groups; PAA, "Planning Academic Activities"; LS, "Learning Strategies"; IR, "Information Retrieval"; WG, "Working in Groups"; MRT, "Management of Relationships with Teachers"; SL, "Skills for Lessons"; SM, "Stress Management."

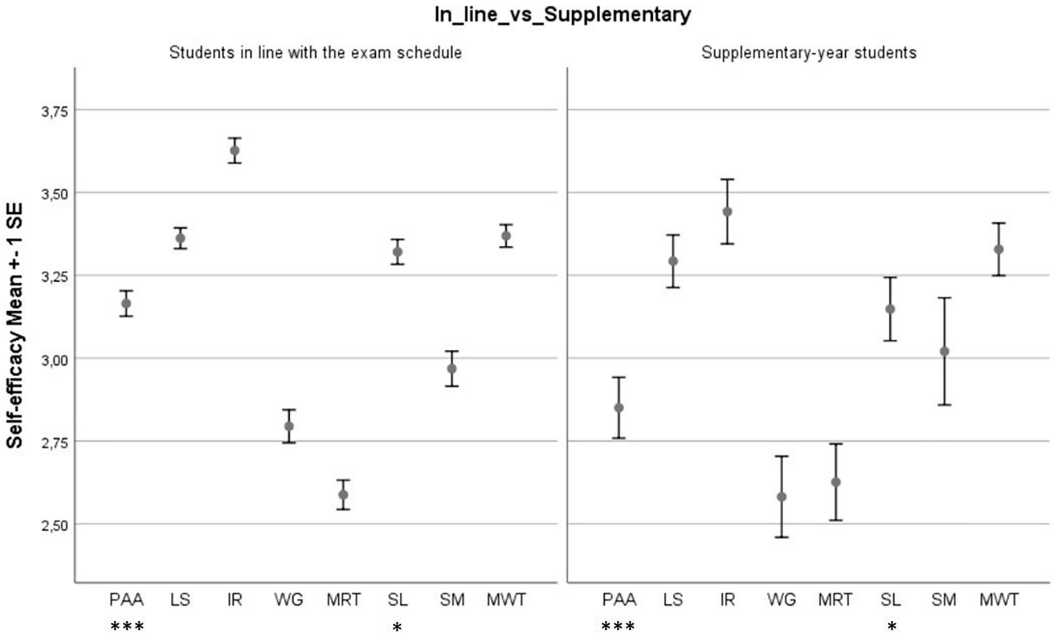


FIGURE 3 | Mean levels of the Academic Self-efficacy factors for students in line with the exam schedule vs. supplementary-year students and *t*-test results. **p* < 0.05; ****p* < 0.001; PAA, "Planning Academic Activities"; LS, "Learning Strategies"; IR, "Information Retrieval"; WG, "Working in Groups"; MRT, "Management of Relationships with Teachers"; SL, "Skills for Lessons"; SM, "Stress Management"; MTW, "Management of Thesis Work."

levels of "Working in Groups." Additionally, results showed a significant difference between first year students and second and third year students [$F(\text{df} = 2, 751) = 3.79$, $p < 0.05$, $\eta^2 = 0.010$] in their levels of "Management of Relationships with Teachers."

Finally, **Figure 3** shows mean values of Academic Self-efficacy for students in line with the exam schedule vs. supplementary-year students. Both groups showed strengths in "Information Retrieval," but weaknesses in "Management of Relationships with Teachers" and "Working in Groups." Students in line with the exam schedule showed strengths in "Planning Academic Activities," while supplementary-year students showed weakness in this factor. Furthermore, the results of the *t*-test showed a meaningful difference between students in line with the exam schedule and supplementary-year students in their levels of "Planning Academic Activities" [$t(\text{df} = 827) = 3.63$, $p < 0.001$, Cohen's $d = 0.43$], and "Skills for Lessons" [$t(\text{df} = 827) = 2.14$, $p < 0.05$, Cohen's $d = 0.25$].

DISCUSSION

The present study aimed to present the Academic Self-Efficacy Scale, a new multifaceted tool designed to measure self-efficacy beliefs in managing academic tasks among university students. The new scale presents adequate psychometric properties, the presence of measurement invariance, and associations with academic performance and experiences, and a remarkable discriminative validity. Analyses exploring the structure of the scale showed that it is made up of eight factors referring to the students' perceived abilities to manage tasks and situations that are crucial for their successful academic path, namely planning activities to be done, implementing effective learning strategies during lessons and at home, retrieving information, working with peers, managing relationships with teachers, managing negative emotions and stress, and thesis work. All these factors are in line with the self-efficacy features found in the literature (Bandura, 1997; Cheung and Kwok, 1998; Amenkhienan and Kogan, 2004) and cover a wide variety of the efficacy beliefs related to the academic context. In particular, results showed the crucial role of "Planning Academic Activities," related to the proportion of exams passed, the average exam rating for all the students independent of the enrollment year, the ability to manage stress, and with the ability to stay in line with academic achievement. Results are similar to the Life Design approach (Savickas et al., 2009), which underlined that the ability to plan personal aims and the next career steps is fundamental to career construction and career development. Findings from our study seem to suggest to focus on planning ability to develop intervention activities to support undergraduate students. In addition, our results showed that "Learning Strategies" and "Skills for Lessons" have relationships with the proportion of exams passed and the average exam rating for most of the students, even though they are not as strong as the ability to plan activities. In that sense, our results have highlighted that it is important to develop good study strategies and learn directly from class lectures, even though these are secondary to the ability to plan career steps. Finally, a strong negative correlation arose between "Working in Groups" and stress-related difficulties, showing how important it is to focus on the peer group to manage stress. In part, our study confirms findings present in the literature: self-efficacy assumes a key role in career planning and academic achievement. Our results showed that students with higher levels of self-efficacy are the ones who are in line with a traditional academic path and with academic goals. This particular result seems crucial to create specific interventions and promoting different levels of self-efficacy beliefs for different steps in the university career.

Differences between male and female students were found in their levels of "Planning Academic Activities," "Information Retrieval," and "Skills for Lessons"; these results confirm a female advantage in academic as stated by Eurostat data on European population and by previous researches (Ceci et al., 2014; Voyer and Voyer, 2014). Furthermore, our results showed higher levels of "Stress Management" in male students; these findings are in line with researches that show that female students are more likely to be influenced by academic stress and that perceived themselves are less able to manage it than male students (Ye et al., 2018). Surprisingly, our results showed higher levels of "Management of Relationships with Teachers" in male students; previous studies established closer and less conflictual relationships between teachers and girls than boys (Baker, 2006; Spilt et al., 2012). This result should be investigated further, considering also the possible influence due to teacher gender and its interaction with students' gender. Moreover, the study confirms previous results (Bandura, 1997): students in line with the traditional path at the university have higher levels of self-efficacy, particularly in the ability to plan and use information from the classroom in a formative way. Even in this case, our study suggests that self-efficacy beliefs support personal competences and that they are fundamental to developing personal and professional skills, useful for the academic context, but also for future planning, as highlighted in previous research (Pajares and Urdan, 2005; Brausch, 2011; Azizli et al., 2015).

Despite its strengths, our study has some limitations. First, since this study was conducted on Italian university students and considering the possible variation among the different university systems, additional work is needed to confirm the generalizability of the scale to other cultural contexts. Activities and tasks required to the students, and following related self-efficacy beliefs, may be different if, as in the Italian system, there are no penalties after an exam failed several times compared to other university systems in which the maximum number of exam failures is limited. Considered possible differences among university systems, future research could explore the structure of the scale in different languages and other countries. Further, although we tested convergent validity by exploring the relationship between the academic self-efficacy and the academic experiences scales, future studies should explore convergent validity in more detail. This will not be easy since the available scales only focus on specific aspects on self-efficacy or are limited to specific disciplines. Yet, future research is needed on this aspect. Finally, even if the use of self-reported academic grades is widely accepted in the social sciences (Stone et al., 1999; Kuncel et al., 2005; Baumeister et al., 2007; Sticca et al., 2017), further studies could explore the role of self-efficacy on different outcomes. Other methods would be useful to assess the truthfulness of participants' reported information, as data from university administrations concerning students' performance indicators or a proxy assessment of self-efficacy. In this way, the amount of missingness in the variables collected would be less.

The development of the Academic Self-efficacy Scale could be a significant contribution to the literature and to intervention in vocational guidance. Measuring self-efficacy beliefs has important implications for school counselors, career counselors, and psychologists working in the academic field. The scale could be used on two levels: preventing academic failure or dropout, and helping struggling students. The results showed that the scale could be a good instrument to identify the students' features and to intercept students with low levels of self-efficacy beliefs and those with more self-concern. The aim could be to create particular interventions for individuals, small groups (Koen et al., 2012) or large groups (Camussi et al., 2017), depending on the specific courses, the level of self-efficacy beliefs or their particular academic paths, to co-construct a new perception about their abilities. Moreover, the scale could be an instrument to build specific interventions and actions dedicated to sustaining the co-construction of academic motivation (Vallerand and Bissonnette, 1992; Vallerand et al., 1992; Vallerand, 2000) and general academic wellbeing. Starting from self-efficacy beliefs, the counselors could encourage the students to experiment with new strategies, promoting a new vision of their abilities, specifically in the new academic context, but expandable in the future working world. In that sense, the Academic Self-efficacy Scale could be a specific and brief instrument, helpful for working in synergy to implement a new representation of students and their abilities, to sustain academic and career success.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Milano-Bicocca. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AG and PS contributed conception and design of the study. AG and CA organized the database and wrote the first draft of the manuscript. AG and NP performed the statistical analysis. NP, EC, GR, and PS wrote sections of the manuscript. All authors contributed to manuscript revision, read and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.498824/full#supplementary-material

REFERENCES

Amenkhienan, C. A., and Kogan, L. R. (2004). Engineering students' perceptions of academic activities and support services: Factors that influence their academic performance. *Coll. Stud. J.* 38, 523–541.

Amini, S. H. (2002). *The role of self-efficacy, self-regulation and self-esteem in students' academic achievement in junior high school experimental sciences course* [dissertation]. Tehran: Tarbiat Moallem University.

Azizli, N., Atkinson, B. E., Baughman, H. M., and Giammarco, E. A. (2015). Relationships between general Self-Efficacy, planning for the future, and life satisfaction. *Personal. Indiv. Diff.* 82, 58–60. doi: 10.1186/s12913-016-1423-5

Baker, J. A. (2006). Contributions of teacher–child relationships to positive school adjustment during elementary school. *J. Sch. Psychol.* 44, 211–229. doi: 10.1111/1467-8624.00364

Bandura, A. (1994). "Self-efficacy," in *Encyclopedia of human behavior*, Vol. 4, ed. V. S. Ramachaudran (New York, NY: Academic Press), 71–81.

Bandura, A. (1997). *Self-Efficacy: The exercise of control*. New York, NY: wH Freeman, 3–604.

Bandura, A. (2006). Guide for constructing Self-Efficacy scales. *Self-Effic. Beliefs Adol.* 5, 307–337.

Bandura, A., Barbaranelli, C., Caprara, G. V., and Pastorelli, C. (1996). Multifaceted impact of self-efficacy beliefs on academic functioning. *Child Dev.* 67, 1206–1222. doi: 10.1111/j.1467-8624.1996.tb01791.x

Bassi, M., Steca, P., Delle Fave, A., and Caprara, G. V. (2007). Academic self-efficacy beliefs and quality of experience in learning. *J. Youth Adol.* 36, 301–312. doi: 10.1007/s10964-006-9069-y

Baughman, K. R., Ludwick, R., Palmisano, B., Hazelett, S., and Sanders, M. (2015). The relationship between organizational characteristics and advance care planning practices. *Am. J. Hosp. Pall. Med.* 32, 510–515. doi: 10.1177/1049909114530039

Baumeister, R. F., Vohs, K. D., and Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspect. Psychol. Sci.* 2, 396–403. doi: 10.1111/j.1745-6916.2007.00051.x

Blanco, V. H., Ornelas, C. M., Rueda, V. M. B., and Martinez, M. M. (2013). Factor structure of the Self-Efficacy scale in academic behaviors of university students of social sciences. *Rev. Mex. De Psicologia* 30, 79–88.

Blunch, N. (2012). *Introduction to structural equation modeling using IBM SPSS statistics and AMOS*, 2nd Edn. London, UK: Sage.

Bong, M., and Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educ. Psychol. Rev.* 15, 1–40.

Brausch, B. D. (2011). The Role of Mindfulness in Academic Stress, Self-Efficacy, and Achievement in College Students. *Masters Theses* 2011:147.

Britner, S. L., and Pajares, F. (2006). Sources of science self-efficacy beliefs of middle school students. *J. Res. Sci. Teach.* 43, 485–499.

Cabrera-Nguyen, P. (2010). Author guidelines for reporting scale development and validation results in the Journal of the Society for Social Work and Research. *J. Soc. Soc. Work Res.* 1, 99–103.

Camussi, E., Annovazzi, C., Montali, L., and Ginevra, M. C. (2017). "Gender-sensitive career counseling: An innovative approach," in *Counseling and Coaching in Times of Crisis and Transitions: from Research to Practice*. Abingdon, eds L. Nota and S. Soresi (Oxford: Routledge), 139–150. doi: 10.4324/9781315266596-12

Caprara, G. V., Barbaranelli, C., Steca, P., and Malone, P. S. (2006). Teachers' Self-Efficacy beliefs as determinants of job satisfaction and students' academic achievement: a study at the school level. *J. Sch. Psychol.* 44, 473–490. doi: 10.1016/j.jsp.2006.09.001

Ceci, S. J., Ginther, D. K., Kahn, S., and Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychol. Sci. Public Int.* 15, 75–141. doi: 10.1177/1529100614541236

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Struct. Equ. Model.* 14, 464–504. doi: 10.1080/10705510701301834

Cheung, C. K., and Kwok, S. T. (1998). Activities and academic achievement among college students. *J. Genet. Psychol.* 159, 147–162. doi: 10.1080/00221329809596142

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Struct. Equ. Model.* 9, 233–255. doi: 10.1207/S15328007SEM0902_5

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edn. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cohen, S., and Janicki-Deverts, D. (2012). Who's stressed? Distributions of psychological stress in the United States in probability samples from 1983, 2006, and 2009 1. *J. Appl. Soc. Psychol.* 42, 1320–1334. doi: 10.1111/j.1559-1816.2012.00900.x

Comrey, A. L., and Lee, H. B. (1992). "Interpretation and application of factor analytic results," in *A first course in factor analysis*, Vol. 2, eds A. L. Comrey and H. B. Lee (Hillsdale, NJ: Lawrence Eribaum Associates), 1992.

Costello, A. B., and Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Pract. Asses. Res. Eval.* 10:7.

Feltz, D. L., and Lirgg, C. D. (2001). Self-efficacy beliefs of athletes, teams, and coaches. *Handbook Sport Psychol.* 2, 340–361.

Ferla, J., Valcke, M., and Schuyten, G. (2009). Student models of learning and their impact on study strategies. *Stud. High. Educ.* 34, 185–202. doi: 10.1080/03075070802528288

Gore, P. A. Jr. (2006). Academic Self-Efficacy as a predictor of college outcomes: Two incremental validity studies. *J. Car. Asses* 14, 92–115.

Gould, D., Guinan, D., Greenleaf, C., Medbery, R., and Peterson, K. (1999). Factors affecting Olympic performance: Perceptions of athletes and coaches from more and less successful teams. *Sport Psychol.* 13, 371–394.

Henson, R., and Roberts, J. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educ. Psychol. Meas.* 66, 393–416. doi: 10.1177/0013164405282485

Hoover-Dempsey, K. V., and Sandler, H. M. (2005). Final Performance Report for OERI Grant # R305T010673: The Social Context of Parental Involvement: A Path to Enhanced Achievement. Available online at: https://discoverarchive.vanderbilt.edu/bitstream/handle/1803/7595/OERIIESfinalreport032205.pdf?sequence=1&isAllowed=y (accessed February 11, 2021).

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika* 30, 179–185. doi: 10.1007/BF02289447

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Stru. Equ. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Hubbard, R., and Allen, S. J. (1987). An empirical comparison of alternative methods for principal component extraction. *J. Business Res.* 15, 173–190. doi: 10.1016/j.jneumeth.2014.09.027

Kenny, D. (2020). *Measuring model fit*. Available from: http://davidakenny.net/cm/fit.htm [Accessed September 13, 2020].

Koen, J., Klehe, U. C., and Van Vianen, A. E. (2012). Training career adaptability to facilitate a successful school-to-work transition. *J. Vocat. Behav.* 81, 395–408. doi: 10.1016/j.jvb.2012.10.003

Komarraju, M., and Dial, C. (2014). Academic identity, Self-Efficacy, and self-esteem predict self-determined motivation and goals. *Learn. Indiv. Diff.* 32, 1–8. doi: 10.1016/j.lindif.2014.02.004

Kuncel, N. R., Credé, M., and Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Rev. Educ. Res.* 75, 63–82. doi: 10.3102/00346543075001063

Lane, J., Lane, A., and Kyprianou, A. (2004). Self-efficacy, self-esteem and their impact on academic performance. *Soc. Behav. Person.* 32, 247–256. doi: 10.2224/sbp.2004.32.3.247

Lu, S., Hu, S., Guan, Y., Xiao, J., Cai, D., Gao, Z., et al. (2018). Measurement Invariance of the Depression Anxiety Stress Scales-21 Across Gender in a Sample of Chinese University Students. *Front. Psychol.* 9:2064. doi: 10.3389/fpsyg.2018.02064

Ma, C. M. S. (2020). Measurement Invariance of the Multidimensional Scale of Perceived Social Support Among Chinese and South Asian Ethnic Minority Adolescents in Hong Kong. *Front. Psychol.* 11:3386. doi: 10.3389/fpsyg.2020.596737

Mamaril, N. J. A. (2014). Measuring undergraduate students' engineering Self-Efficacy: a scale validation study. *Theses Dissert.* 2014, 19.

Mayer, J. D., Salovey, P., and Caruso, D. R. (2002). Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT) Item Booklet. Toronto, Ontario: *UNH Personality Lab*, 26.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

McGeown, S. P., Putwain, D., Simpson, E. G., Boffey, E., Markham, J., and Vince, A. (2014). Predictors of adolescents' academic motivation: Personality, Self-Efficacy and adolescents' characteristics. *Learn. Indiv. Diff.* 32, 278–286. doi: 10.1016/j.lindif.2014.03.022

Muthén, L. K., and Muthén, B. O. (1998-2010). *Mplus User's Guide. Sixth Edition*. Los Angeles, CA: Muthén & Muthén.

Netemeyer, R. G., Bearden, W. O., and Sharma, S. (2003). *Scaling procedures: Issues and applications*. London: Sage Publications.

Nunnally, J. C., and Bernstein, I. H. (1994). *Psychometric Theory*, 3rd Edn. New York, NY: McGraw-Hill.

Ommundsen, Y., Haugen, R., and Lund, T. (2005). Academic self-concept, implicit theories of ability, and self-regulation strategies. *Scand. J. Educ. Res.* 49, 461–474. doi: 10.1080/00313830500267838

Pajares, F., and Urdan, T. (2005). *Self-Efficacy beliefs of adolescents*. Greenwich, CT: Information Age Publishers, 339–367.

R Core Team (2020). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., and Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychol. Bull.* 130:261. doi: 10.1037/0033-2909.130.2.261

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *J. Stat. Soft.* 48, 1–36. doi: 10.1002/9781119579038.ch1

RStudio Team (2020). *RStudio: Integrated Development for R*. Boston, MA: RStudio, PBC.

Ryan, R. M., and Deci, E. L. (2006). Self-regulation and the problem of human autonomy: Does psychology need choice, self-determination, and will? *J. Personal.* 74, 1557–1586. doi: 10.1111/j.1467-6494.2006.00420.x

Savickas, M. L., Nota, L., Rossier, J., Dauwalder, J. P., Duarte, M. E., Guichard, J., et al. (2009). Life designing: a paradigm for career construction in the 21st century. *J. Vocat. Behav.* 75, 239–250. doi: 10.1016/j.jvb.2009.04.004

Spilt, J. L., Koomen, H. M., and Jak, S. (2012). Are boys better off with male and girls with female teachers? A multilevel investigation of measurement invariance and gender match in teacher–student relationship quality. *J. Sch. Psychol.* 50, 363–378. doi: 10.1016/j.jsp.2011.12.002

Steca, P., Greco, A., Cappelletti, E., D'addario, M., Monzani, D., Pancani, L., et al. (2015). Cardiovascular management Self-Efficacy: Psychometric properties of a new scale and its usefulness in a rehabilitation context. *Ann. Behav. Med.* 49, 660–674. doi: 10.1007/s12160-015-9698-z

Sticca, F., Goetz, T., Bieg, M., Hall, N. C., Eberle, F., and Haag, L. (2017). Examining the accuracy of students' self-reported academic grades from a correlational and a discrepancy perspective: Evidence from a longitudinal study. *PLoS One* 12:e0187367. doi: 10.1371/journal.pone.0187367

Stone, A. A., Bachrach, C. A., Jobe, J. B., Kurtzman, H. S., and Cain, V. S. (eds) (1999). *The science of self-report: Implications for research and practice*. Mahwah, NJ: Psychology Press. doi: 10.4324/9781410601261

Vallerand, R. J. (2000). Deci and Ryan's self-determination theory: A view from the hierarchical model of intrinsic and extrinsic motivation. *Psychol. Inq.* 11, 312–318.

Vallerand, R. J., and Bissonnette, R. (1992). Intrinsic, extrinsic, and amotivational styles as predictors of behavior: A prospective study. *J. Person.* 60, 599–620. doi: 10.1111/j.1467-6494.1992.tb00922.x

Vallerand, R. J., Pelletier, L. G., Blais, M. R., Briere, N. M., Senecal, C., and Vallieres, E. F. (1992). The Academic Motivation Scale: a measure of intrinsic, extrinsic, and amotivation in education. *Educ. Psychol. Meas.* 52, 1003–1017.

Voyer, D., and Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psycholog. Bull.* 140:1174. doi: 10.1037/a0036620

Walker, C. O., Greene, B. A., and Mansell, R. A. (2006). Identification with academics, intrinsic/extrinsic motivation, and Self-Efficacy as predictors of cognitive engagement. *Learn. Indiv. Diff.* 16, 1–12. doi: 10.1016/j.lindif.2005.06.004

West, S. G., Finch, J. F., and Curran, P. J. (1995). "Structural equation models with nonnormal variables: Problems and remedies," in *Structural equation modeling: Concepts, issues, and applications*, ed. R. H. Hoyle (Thousand Oaks, CA: Sage Publications, Inc), 56–75.

Worthington, R., and Whittaker, T. (2006). Scale development research: A content analysis and recommendations for best practices. *Counsel. Psychol.* 34, 806–838.

Ye, L., Posada, A., and Liu, Y. (2018). The moderating effects of gender on the relationship between academic stress and academic self-efficacy. *Internat. J. Stress Manag.* 25:56. doi: 10.3390/ijerph16010048

Zhao, H., He, J., Yi, J., and Yao, S. (2019). Factor Structure and Measurement Invariance Across Gender Groups of the 15-Item Geriatric Depression Scale Among Chinese Elders. *Front. Psychol.* 10:1360. doi: 10.3389/fpsyg.2019.01360

Zhou, H., Liu, W., Fan, J., Xia, J., Zhu, J., and Zhu, X. (2019). The Temporal Experience of Pleasure Scale (TEPS): Measurement Invariance Across Gender in Chinese University Students. *Front. Psychol.* 10:2130. doi: 10.3389/fpsyg.2019.02130

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership