frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 27 March 2013
doi: 10.3389/fpsyg.2013.00109

On the relevance of assumptions associated with classical factor analytic approaches†

Daniel Kasper and Ali Ünlü*

Chair for Methods in Empirical Educational Research, TUM School of Education and Centre for International Student Assessment, Technische Universität München, Munich, Germany

Edited by:
Jason W. Osborne, Old Dominion University, USA

Reviewed by:
Andrew Jones, American Board of Surgery, USA
Evgueni Borokhovski, Concordia University, Canada

***Correspondence:**
Ali Ünlü, Chair for Methods in Empirical Educational Research, TUM School of Education and Centre for International Student Assessment, Technische Universität München, Lothstrasse 17, 80335 Munich, Germany.
e-mail: ali.uenlue@tum.de

†*The research reported in this paper is based on the dissertation thesis by Kasper (2012).*

A personal trait, for example a person's cognitive ability, represents a theoretical concept postulated to explain behavior. Interesting constructs are latent, that is, they cannot be observed. Latent variable modeling constitutes a methodology to deal with hypothetical constructs. Constructs are modeled as random variables and become components of a statistical model. As random variables, they possess a probability distribution in the population of reference. In applications, this distribution is typically assumed to be the normal distribution. The normality assumption may be reasonable in many cases, but there are situations where it cannot be justified. For example, this is true for criterion-referenced tests or for background characteristics of students in large scale assessment studies. Nevertheless, the normal procedures in combination with the classical factor analytic methods are frequently pursued, despite the effects of violating this "implicit" assumption are not clear in general. In a simulation study, we investigate whether classical factor analytic approaches can be instrumental in estimating the factorial structure and properties of the population distribution of a latent personal trait from educational test data, when violations of classical assumptions as the aforementioned are present. The results indicate that having a latent non-normal distribution clearly affects the estimation of the distribution of the factor scores and properties thereof. Thus, when the population distribution of a personal trait is assumed to be non-symmetric, we recommend avoiding those factor analytic approaches for estimation of a person's factor score, even though the number of extracted factors and the estimated loading matrix may not be strongly affected. An application to the Progress in International Reading Literacy Study (PIRLS) is given. Comments on possible implications for the Programme for International Student Assessment (PISA) complete the presentation.

Keywords: factor analysis, latent variable model, normality assumption, factorial structure, criterion-referenced test, large scale educational assessment, Programme for International Student Assessment, Progress in International Reading Literacy Study

1. INTRODUCTION

Educational research is concerned with the study of processes of learning and teaching. Typically, the investigated processes are not observable, and to unveil these, manifest human behavior in test situations is recorded. According to Lienert and Raatz (1998, p. 1) "a test [. . .] is a routine procedure for the investigation of one or more empirically definable personality traits" (translated by the authors), and to satisfy a minimum of quality criteria, a test is required to be objective, reliable, and valid.

In this paper we deal with factor analytic methods for assessing construct validity of a test, in the sense of its factorial validity (e.g., Cronbach and Meehl, 1955; Lienert and Raatz, 1998). Factorial validity refers to the factorial structure of the test, that is, to the number (and interpretation) of underlying factors, the correlation structure among the factors, and the correlations of each test item with the factors. There are a number of latent variable models that may be used to analyze the factorial structure of a test – for generalized latent variable modeling covering a plethora of models as special cases of a much broader framework, see Bartholomew et al. (2011) and Skrondal and Rabe-Hesketh (2004). This paper focuses on classical factor analytic approaches, and it examines how accurately different methods of classical factor analysis can estimate the factorial structure of test data, if assumptions associated with the classical approaches are not satisfied. The methods of classical factor analysis will include principal component analysis (PCA; Pearson, 1901; Hotelling, 1933a,b; Kelley, 1935), exploratory factor analysis (EFA; Spearman, 1904; Burt, 1909; Thurstone, 1931, 1965), and principal axis analysis (PAA; Thurstone, 1931, 1965). More recent works on factor analysis and related methods are Harman (1976), McDonald (1985), Cudeck and MacCallum (2007), and Mulaik (2009). Further references, to more specific topics in factor analysis, are given below, later in the text.[1]

[1]For the sake of simplicity and for the purpose and analysis of this paper, we want to refer to all of these approaches (PCA, EFA, PAA) collectively as classical factor analysis/analytic methods. Albeit it is known that PCA differs from factor analysis in important aspects, and that PAA rather represents an alternative estimation procedure for EFA. PCA and EFA are different technically and conceptually. PCA

A second objective of this paper is to examine the scope of these classical methods for estimating the probability distribution of latent ability values or properties thereof postulated in a population under investigation, especially when this distribution is skewed (and not normal). In applied educational contexts, for instance, that is not seldom the practice. Therefore a critical evaluation of this usage of classical factor analytic methods for estimating distributional properties of ability is important, as we do present with our simulation study in this paper, in which metric scale (i.e., at least interval scale; not dichotomous) items are used.

The results of the simulation study indicate that having a non-normal distribution for latent variables does not strongly affect the number of extracted factors and the estimation of the loading matrix. However, as shown in this paper, it clearly affects the estimation of the latent factor score distribution and properties thereof (e.g., skewness).

More precisely, the "estimation accuracy" for factorial structure of these models is shown to be worse when the assumption of interval-scaled data is not met or item statistics are skewed. This corroborates related findings published in other works, which we briefly review later in this paper. More importantly, the empirical distribution of estimated latent ability values is biased compared to the true distribution (i.e., estimates deviate from the true values) when population abilities are skewly distributed. It seems therefore that classical factor analytic procedures, even though they are performed with metric (instead of non-metric) scale indicator variables, are not appropriate approaches to ability estimation when skewly distributed population ability values are to be estimated.

Why should that be of interest? In large scale assessment studies such as the Programme for International Student Assessment (PISA)[2] latent person-related background (conditioning) variables such as sex or socioeconomic status are obtained as well by principal component analysis, and that "covariate" information is part of the PISA procedure that assigns to students their literacy or plausible values (OECD, 2012; see also Section 3.1 in the present paper). Now, if it is assumed that the distribution of latent background information conducted through questionnaires at the students, schools, or parents levels (the true latent variable distribution) is skewed, based on the simulation study of this paper we can expect that the empirical distribution of estimated background information (the "empirical" distribution of the calculated component scores) is biased compared to the true distribution (and is most likely skewed as well). In other words, estimated background values do deviate from their corresponding true values they ought to approximate, and so the inferred students' plausible values may be biased. Further research is necessary in order to investigate the effects and possible implications of potentially biased estimates of latent background information on students' assigned literacy values and competence levels, based on which the PISA rankings of OECD countries are reported. For an analysis of empirical large scale assessment (Progress in International Reading Literacy Study; PIRLS) data, see Section 6.

The paper is structured as follows. We introduce the considered classical factor analysis models in Section 2 and discuss the relevance of the assumptions associated with these models in Section 3. We describe the simulation study in Section 4 and present the results of it in Section 5. We give an empirical data analysis example in Section 6. In Section 7, we conclude with a summary of the main findings and an outlook on possible implications and further research.

2. CLASSICAL FACTOR ANALYSIS METHODS

We consider the method of principal component analysis on the one hand, and the method of exploratory factor and principal axis analysis on the other. At this point recall Footnote 1, where we clarified that, strictly speaking, principal component analysis is not factor analysis and that principal axis analysis is a specific method for estimating the exploratory factor analysis model. Despite this, for the sake of simplicity and for our purposes and analyses, we call these approaches collectively factor analysis/analytic methods or even models. For a more detailed discussion of these methods, see Bartholomew et al. (2011).

Our study shows, amongst others, that the purely computational dimensionality reduction method PCA performs surprisingly well, as compared to the results obtained based on the latent variable models EFA and PAA. This is important, because applied researchers often use PCA in situations where factor analysis more closely matches their purpose of analysis. In general, such computational procedures as PCA are easy to use. Moreover, the comparison of EFA (based on ML) with PAA (eigenstructure of the reduced correlation matrix based on communality estimates) in this paper represents an evaluation of different estimation procedures for the classical factor analysis model. This comparison of the two estimation procedures seems to be justified and interesting, as the (manifest) normality assumption in the observed indicators for the ML procedure is violated, both in the simulation study and empirical large scale assessment PIRLS application. At this point, see also Footnote 1.

seeks to create composite scores of observed variables while EFA assumes latent variables. There is no latent variable in PCA. PCA is not a model and instead is simply a re-expression of variables based on the eigenstructure of their correlation matrix. A statistical model, as is for EFA, is a simplification of observed data that necessarily does not perfectly reproduce the data, leading to the inclusion of an error term. This point is well-established in the methodological literature (e.g., Velicer and Jackson, 1990; Widaman, 2007). Correlation matrix is usually used in EFA, and the models for EFA and PAA are the same. There are several methods to fit EFA such as unweighted least squares (ULS), generalized least squares (GLS), or maximum likelihood (ML). PAA is just one of the various methods to fit EFA. PAA is a method of estimating the model of EFA that does not rely on a discrepancy function such as for ULS, GLS, or ML. This point is made clear, for instance, in MacCallum (2009). In fact, PAA with iterative communality estimation is asymptotically equivalent to ULS estimation. Applied researchers often use PCA in situations where factor analysis more closely matches the purpose of their analysis. This is why we want to include PCA in our present study with latent variables, to examine how well PCA results may approximate a factor analysis model. Such practice is frequently pursued, for example in empirical educational research, as we tried to criticize for the large scale assessment PISA study (e.g., OECD, 2005, 2012). Moreover, the comparison of EFA (based on ML) with PAA in this paper seems to be justified and interesting, as the (manifest) normality assumption in the observed indicator variables for the ML procedure is violated in the simulation study and empirical large scale assessment PIRLS application.

[2]PISA is an international large scale assessment study funded by the Organisation for Economic Co-operation and Development (OECD), which aims to evaluate education systems worldwide by assessing 15-year-old students' competencies in reading, mathematics, and science. For comprehensive and detailed information, see www.pisa.oecd.org.

2.1. PRINCIPAL COMPONENT ANALYSIS

The model of principal component analysis (PCA) is

$$\boldsymbol{Z} = \boldsymbol{FL'},$$

where $\boldsymbol{Z}$ is a $n \times p$ matrix of standardized test results of n persons on p items, $\boldsymbol{F}$ is a $n \times p$ matrix of p principal components ("factors"), and $\boldsymbol{L}$ is a $p \times p$ loading matrix.[3] In the estimation (computation) procedure $\boldsymbol{F}$ and $\boldsymbol{L}$ are determined as $\boldsymbol{F} = \boldsymbol{ZC\Lambda}^{-1/2}$ and $\boldsymbol{L} = \boldsymbol{C\Lambda}^{1/2}$ with a $p \times p$ matrix $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_p\}$, where λ_l are the eigenvalues of the empirical correlation matrix $\boldsymbol{R} = \boldsymbol{Z'Z}$, and with a $p \times p$ matrix $\boldsymbol{C} = (\boldsymbol{c_1}, \ldots, \boldsymbol{c_p})$ of corresponding eigenvectors $\boldsymbol{c_l}$.

In principal component analysis we assume that $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$, $\boldsymbol{F} \in \mathbb{R}^{n \times p}$, and $\boldsymbol{L} \in \mathbb{R}^{n \times p}$ and that empirical moments of the manifest variables exist such that, for any manifest variable $j = 1, \ldots, p$, its empirical variance is not zero ($s_j^2 \neq 0$). Moreover we assume that $\text{rk}(\boldsymbol{Z}) = \text{rk}(\boldsymbol{R}) = p$ (rk, the matrix rank) and that $\boldsymbol{Z}$, $\boldsymbol{F}$, and $\boldsymbol{L}$ are interval-scaled (at the least).

The relevance of the assumption of interval-scaled variables for classical factor analytic approaches is the subject matter of various research works, which we briefly discuss later in this paper.

2.2. EXPLORATORY FACTOR ANALYSIS

The model of exploratory factor analysis (EFA) is

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{Lf} + \boldsymbol{e},$$

where $\boldsymbol{y}$ is a $p \times 1$ vector of responses on p items, $\boldsymbol{\mu}$ is the $p \times 1$ vector of means of the p items, $\boldsymbol{L}$ is a $p \times k$ matrix of factor loadings, $\boldsymbol{f}$ is a $k \times 1$ vector of ability values (of factor scores) on k latent continua (on factors), and $\boldsymbol{e}$ is a $p \times 1$ vector subsuming remaining item specific effects or measurement errors.

In exploratory factor analysis, we assume that

$$\begin{aligned}
&\boldsymbol{y} \in \mathbb{R}^{p \times 1}, \boldsymbol{\mu} \in \mathbb{R}^{p \times 1}, \boldsymbol{L} \in \mathbb{R}^{p \times k}, \boldsymbol{f} \in \mathbb{R}^{k \times 1}, \quad \text{and} \quad \boldsymbol{e} \in \mathbb{R}^{p \times 1},\\
&\boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{L}, \boldsymbol{f}, \text{and } \boldsymbol{e} \text{ are interval-scaled (at the least)},\\
&\text{E}(\boldsymbol{f}) = \boldsymbol{0},\\
&\text{E}(\boldsymbol{e}) = \boldsymbol{0},\\
&\text{cov}(\boldsymbol{e},\boldsymbol{e}) = \text{E}(\boldsymbol{ee'}) = \boldsymbol{D} = \text{diag}\{v_1, \ldots, v_p\},\\
&\text{cov}(\boldsymbol{f},\boldsymbol{e}) = \text{E}(\boldsymbol{fe'}) = \boldsymbol{0},
\end{aligned}$$

where v_i are the variances of e_i ($i = 1, \ldots, p$). If the factors are not correlated, we call this the orthogonal factor model; otherwise it is called the oblique factor model. In this paper, we investigate the sensitivity of the classical factor analysis model against violated assumptions only for the orthogonal case (with $\text{cov}(\boldsymbol{f}, \boldsymbol{f}) = \text{E}(\boldsymbol{ff'}) = \boldsymbol{I} = \text{diag}\{1, \ldots, 1\}$).

Under this orthogonal factor model, $\boldsymbol{\Sigma}$ can be decomposed as follows:

$$\boldsymbol{\Sigma} = \text{E}\left[(\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})'\right] = \text{E}\left[(\boldsymbol{Lf} + \boldsymbol{e})(\boldsymbol{Lf} + \boldsymbol{e})'\right] = \boldsymbol{LL'} + \boldsymbol{D}.$$

This decomposition is utilized by the methods of unweighted least squares (ULS), generalized least squares (GLS), or maximum likelihood (ML) for the estimation of $\boldsymbol{L}$ and $\boldsymbol{D}$. For ULS and GLS, the corresponding discrepancy function is minimized with respect to $\boldsymbol{L}$ and $\boldsymbol{D}$ (Browne, 1974). ML estimation is performed based on the partial derivatives of the logarithm of the Wishart (W) density function of the empirical covariance matrix $\boldsymbol{S}$, with $(n-1)\,\boldsymbol{S} \sim \text{W}\,(\boldsymbol{\Sigma}, n-1)$ (Jöreskog, 1967). After estimates for $\boldsymbol{\mu}$, k, $\boldsymbol{L}$, and $\boldsymbol{D}$ are obtained, the vector $\boldsymbol{f}$ can be estimated by $\hat{\boldsymbol{f}} = (\boldsymbol{L'D}^{-1}\boldsymbol{L})^{-1}\boldsymbol{L'D}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$.

When applying this exploratory factor analysis, $\boldsymbol{y}$ is typically assumed to be normally distributed, and hence $\text{rk}(\boldsymbol{\Sigma}) = p$, where $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{y}$. For instance, one condition required for ULS or GLS estimation is that the fourth cumulants of $\boldsymbol{y}$ must be zero, which is the case, for example, if $\boldsymbol{y}$ follows a multivariate normal distribution (for this and other conditions, see Browne, 1974). For ML estimation note that $(n-1)\boldsymbol{S} \sim \text{W}(\boldsymbol{\Sigma}, n-1)$ if $\boldsymbol{y} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Another possibility of estimation for the EFA model is principal axis analysis (PAA). The model of PAA is

$$\boldsymbol{Z} = \boldsymbol{FL'} + \boldsymbol{E},$$

where $\boldsymbol{Z}$ is a $n \times p$ matrix of standardized test results, $\boldsymbol{F}$ is a $n \times p$ matrix of factor scores, $\boldsymbol{L}$ is a $p \times p$ matrix of factor loadings, and $\boldsymbol{E}$ is a $n \times p$ matrix of error terms. For estimation of $\boldsymbol{F}$ and $\boldsymbol{L}$ based on the representation $\boldsymbol{Z'Z} = \boldsymbol{R} = \boldsymbol{LL'} + \boldsymbol{D}$ the principal components transformation is applied. However, the eigenvalue decomposition is not based on $\boldsymbol{R}$, but is based on the reduced correlation matrix $\boldsymbol{R}_h = \boldsymbol{R} - \hat{\boldsymbol{D}}$, where $\hat{\boldsymbol{D}}$ is an estimate for $\boldsymbol{D}$. An estimate $\hat{\boldsymbol{D}}$ is derived using $h_j^2 = 1 - v_j$ and estimating the communalities h_j (for methods for estimating the communalities, see Harman, 1976).

The assumptions of principal axis analysis are

$$\begin{aligned}
&\boldsymbol{Z} \in \mathbb{R}^{n \times p}, \boldsymbol{L} \in \mathbb{R}^{p \times p}, \boldsymbol{F} \in \mathbb{R}^{n \times p}, \quad \text{and} \quad \boldsymbol{E} \in \mathbb{R}^{n \times p},\\
&\text{E}(\boldsymbol{f}) = \boldsymbol{0},\\
&\text{E}(\boldsymbol{e}) = \boldsymbol{0},\\
&\text{cov}(\boldsymbol{e},\boldsymbol{e}) = \text{E}(\boldsymbol{ee'}) = \boldsymbol{D} = \text{diag}\{v_1, \ldots, v_p\},\\
&\text{cov}(\boldsymbol{f},\boldsymbol{e}) = \text{E}(\boldsymbol{fe'}) = \boldsymbol{0},\\
&\text{cov}(\boldsymbol{f},\boldsymbol{f}) = \text{E}(\boldsymbol{ff'}) = \boldsymbol{I},
\end{aligned}$$

and empirical moments of the manifest variables are assumed to exist such that, for any manifest variable $j = 1, \ldots, p$, its empirical variance is not zero ($s_j^2 \neq 0$). Moreover, we assume that $\text{rk}(\boldsymbol{Z}) = \text{rk}(\boldsymbol{R}) = p$ and that the matrices $\boldsymbol{Z}$, $\boldsymbol{F}$, $\boldsymbol{L}$, and $\boldsymbol{E}$ are interval-scaled (at the least).

2.3. GENERAL REMARKS

Two remarks are important before we discuss the assumptions associated with the classical factor models in the next section.

First, it can be shown that $\boldsymbol{L}$ is unique up to an orthogonal transformation. As different orthogonal transformations may yield different correlation patterns, a specific orthogonal transformation must be taken into account (and fixed) before the estimation

[3] For the sake of simplicity and without ambiguity, in this paper we want to refer to component scores from PCA as "factor scores" or "ability values," albeit components conceptually may not be viewed as latent variables or factors. See also Footnote 1.

accuracies of the factor models can be compared. This is known as "rotational indeterminacy" in the factor analysis approach (e.g., see Maraun, 1996). For more information, the reader is also referred to Footnote 8 and Section 7.

Second, the criterion used to determine the number of factors extracted from the data must be distinguished as well. In practice, not all k or p but instead $\hat{k} < k$ or p factors with the $\hat{k}$ largest eigenvalues are extracted. Various procedures are available to determine $\hat{k}$. Commonly used criteria in educational research are the Kaiser-Guttman criterion (Guttman, 1954; Kaiser and Dickman, 1959), the scree test (Cattell, 1966), and the method of parallel analysis (Horn, 1965).

3. ASSUMPTIONS ASSOCIATED WITH THE CLASSICAL FACTOR MODELS

The three models described in the previous section in particular assume interval-scaled data and full rank covariance or correlation matrices for the manifest variables. Typically in the exploratory factor analysis model, the manifest variables $\boldsymbol{y}$ or the standardized variables $\boldsymbol{z}$ are assumed to be normally distributed. For the PCA and PAA models, we additionally want to presuppose – for computational reasons – that the variances of the manifest variables are substantially large. The EFA and PAA models assume uncorrelated factor terms and uncorrelated error terms (which can be relaxed in the framework of structural equation models; e.g., Jöreskog, 1966), uncorrelatedness between the error and latent ability variables, and expected values of zero for the errors as well as latent ability variables.

The question now arises whether the assumptions are critical when it comes to educational tests or survey data?[4]

3.1. CRITERION-REFERENCED TESTS AND PISA QUESTIONNAIRE DATA

From the perspective of applying these models to data of criterion-referenced tests, the last three of the above mentioned assumptions are less problematic. For a criterion-referenced test, it is important that all items of the test are valid for the investigated content. As such, the usual way of excluding items from the analysis when the covariance or correlation matrices are not of full rank does not work for criterion-referenced tests, because this can reduce content validity of a test. A similar argument applies to the assumption of substantially large variances of the manifest variables. As Klauer (1987) suggested and Sturzbecher et al. (2008) have shown for the driving license test in Germany, the variances of the manifest variables of criterion-referenced tests are seldom high, and in general the data obtained from those tests may lead to extracting too few dimensions. However, for the analysis of criterion-referenced tests, the assumption of interval-scaled data and the assumption of normality of the manifest test and latent ability scores are even more problematic. Data from criterion-referenced tests are rarely interval-scaled – instead the items of criterion-referenced tests are often dichotomous (Klauer, 1987). For criterion-referenced tests, it is plausible to have skewed (non-symmetric) test and ability score distributions, because criterion-referenced tests are constructed to assess whether a desired and excessive teaching goal has been achieved or not. In other words, the tested population is explicitly and intensively trained regarding the evaluated ability, and so it is rather likely that most people will have high values on the manifest test score as well as latent ability (e.g., see the German driving license test; Sturzbecher et al., 2008).

The assumption of interval-scaled data and the normality assumption for the manifest test and latent ability scores may also be crucial for the scaling of cognitive data in PISA (OECD, 2012; Chap. 9 therein). In PISA, the generated students' scores are plausible values. These are randomly drawn realizations basically from a multivariate normal distribution (as the prior) of latent ability values (person ability is modeled as a random effect, a latent variable), in correspondence to a fitted item response theory model (Adams et al., 1997) giving the estimated parameters of the normal distribution. The mean of the multivariate normal distribution is expressed as linear regression of various direct manifest regressors (e.g., administered test booklet, gender) and indirect "latent" or complex regressors obtained by aggregating over manifest and latent context or background variables (e.g., indicators for economic, social, and cultural status) in a principal component analysis. The component scores used in the scaling model as the indirect "latent" regressors are extracted, in the purely computational sense, to account for approximately 95% of the total variance in all the original variables. The background variables may be categorical or dummy-coded and may not be measured at an interval scale (nor be normally distributed). So as we said before, if one can assume that the distribution of latent background information revealed through questionnaires is skewed, we can expect that the empirical distribution of background information computed by principal component analysis is likely to be biased compared to the true distribution. This is suggested by the results of our simulation study. The bias of the empirical distribution in turn may result in biasing the regression expression for the mean. Therefore, special caution has to be taken regarding possible violations of those assumptions, and a minimum of related sensitivity analyses are required and necessary in order to control for their potential effects.

3.2. HISTORICAL REMARKS

The primary aim is to review results of previous studies focusing on the impact of violations of model assumptions. As to our knowledge, such studies did not systematically vary the distributions of the factors (in the case of continuous data as well) and

[4]Note that, at the latent level, there is no formal assumption that the latent factors (what we synonymously also want to call "person abilities") are normally distributed. At the manifest level, maximum likelihood estimation (EFA) assumes that the observed variables are normal; ULS and GLS (EFA), PAA (EFA), and PCA do not. The latter two methods only require a non-singular correlation matrix (e.g., see MacCallum, 2009). However, in applications, for example in empirical educational research, one often assumes that the latent ability values follow a normal distribution in the population of reference. Moreover, Mattson (1997)'s method described in Section 4.1 states that there is a connection between the manifest and latent distributions in factor analysis. Hence the question is what implications one can expect if this "implicit assumption" may not be justified. Related to the study and evaluation of the underlying assumptions associated with these classical factor models, this paper, amongst others, shows that the data re-expression method PCA performs surprisingly well if compared to the results obtained based on the latent variable approaches EFA and PAA. Moreover, ML and PAA estimation procedures for EFA are compared with one another, for different degrees of violating the normality assumption at the manifest or latent levels.

primarily investigated the impact of categorical data (however, not varying the latent distributions for the factors). Reviewing results of previous simulation studies based on continuous indicator variables that have compared different estimation methods (including PCA) and have compared different methods for determining the number of factors, as to our knowledge, would have not constituted reviewing relevant literature focusing primarily on the violations of the assumptions associated with those models.

Literature on classical factor models has in particular investigated violations of the assumption of interval-scaled data. In classical factor analysis, Green (1983) simulated dichotomous data based on the 3PL (three parameter logistic) model (Birnbaum, 1968) and applied PCA and PAA to the data, whereat Cattell's scree test and Horn's parallel analysis were used as extraction criteria. Although both methods were applied to the same data, the results regarding the extracted factors obtained from the analyses differed, and the true dimensionality was not detected. In general, the models extracted too many factors. These findings are in line with expectations. Green (1983) used the phi-coefficient ϕ as the input data, and according to Ferguson (1941), the maximum value of ϕ depends on the difficulty parameters of the items. Dependence of ϕ on item difficulty can in extreme cases lead to factors being extracted solely due to the difficulties of the items. Roznowski et al. (1991) referred to such factors as difficulty factors.

Carroll (1945) recommended to use the tetrachoric correlation ρ_{tet} for factor analysis of dichotomous data. The coefficient ρ_{tet} is an estimate of the dependency between two dichotomous items based on the assumption that the items measure a latent continuous ability – an assumption that corresponds to the factor analysis approach. Although one would expect that ρ_{tet} leads to less biased results as compared to ϕ, Collins et al. (1986) were able to show that ϕ was much better suited to capture the true dimensionality than ρ_{tet}. In simulations, they compared the two correlation coefficients within the principal component analysis, using a version of the scree test as extraction criterion. The simulated data followed the 2PL model with three latent dimensions, and in addition to item discrimination (moderate, high, very high), the item difficulty and its distribution were varied (easy, moderate, difficult, and extreme difficult item parameters; distributed normal, low frequency, rectangular, and bimodal). The coefficient ρ_{tet} led to better results when the distribution of item difficulty was rectangular. In all other cases, ϕ was superior to ρ_{tet}. But with neither of the two methods it was possible to detect the true number of factors in more than 45% of the simulated data sets. See Roznowski et al. (1991) for another study illustrating the superiority of the coefficient ϕ to the coefficient ρ_{tet}.

Clarification for findings in Green (1983), Collins et al. (1986), and Roznowski et al. (1991) was provided by Weng and Cheng (2005). Weng and Cheng varied the number of items, the factor loadings and difficulties of the items, and sample size. The authors used the parallel analysis extraction method to determine the number of factors. However, the eigenvalues of the correlation matrices were computed using a different algorithm, which in a comparative study proved to be more reliable (Wang, 2001). With this algorithm, ϕ and ρ_{tet} performed equally well and misjudged true unidimensionality only when the factor loadings or sample sizes were small, or when the items were easy. This means that it was not the correlation coefficient *per se* that led to inadequate estimation of the number of factors but the extraction method that was used.

Muthén (1978, 1983, 1984), Muthén and Christoffersson (1981), Dolan (1994), Gorsuch (1997), Bolt (2005), Maydeu-Olivares (2005), and Wirth and Edwards (2007) present alternative or more sophisticated ways for dealing with categorical variables in factor analysis or structural equation modeling. Muthén (1989), Muthén and Kaplan (1992), and Ferguson and Cox (1993) compared the performances of factor analytic methods under conditions of (manifest) non-normality for the observed indicator variables.

We will add to and extend this literature and investigate in this paper whether the classical factor analysis models can reasonably unveil the factorial structure or properties of the population latent ability distribution in educational test data (e.g., obtained from criterion-referenced tests) when the assumption of normality in the latency may not be justified. None of the studies mentioned above has investigated the "true distribution impact" in these problems.

4. SIMULATION STUDY

A simulation study is used to evaluate the performances of the classical factor analytic approaches when the latent variables are not normally distributed.

True factorial structures under the exploratory factor analysis model are simulated, that is, the values of n, k, $\boldsymbol{L}$, $\boldsymbol{f}$, and $\boldsymbol{e}$ are varied.[5] On the basis of the constructed factorial structures, the matrices of the manifest variables are computed. These matrices are used as input data and analyzed with classical factor analytic methods. The estimates (or computed values) $\hat{k}$, $\hat{\boldsymbol{L}}$, and $\hat{\boldsymbol{f}}$ (or $\hat{\boldsymbol{F}}$) are then compared to the underlying population values. As criteria for "estimation accuracy" we use the number of extracted factors (as compared to true dimensionality), the skewness of the estimated latent ability distribution, and the discrepancy between estimated and true loading matrix. Shapiro-Wilk tests for normality of the ability estimates are presented and distributions of the estimated and true factor scores are compared as well.

Note that in the simulation study metric scale, not dichotomous, items are analyzed. This can be viewed as a baseline informative for the dichotomous indicator case as well (cf. Section 6). The results of the simulation study can serve as a reference also for situations where violations of normality for latent and manifest variables and metric scale data are present. One may expect the reported results to become worse when, in addition to (latent) non-normality of person ability, data are discretized or item statistics are skewed (manifest non-normality).

4.1. MOTIVATION AND PRELIMINARIES

The present simulation study particularly aims at analyzing and answering such questions as:

[5]Obviously, PCA as introduced in this paper cannot be used as a data generating probability model underlying the population. However, the simulation study shows that PCA results can approximate a factor analysis (cf. also Footnote 1).

- To what extent does the estimation accuracy for factorial structure of the classical factor analysis models depend on the skewness of the population latent ability distribution?
- Are there specific aspects of the factorial structure or latent ability distribution with respect to which the classical factor analysis models are more or less robust in estimation when true ability values are skewed?
- Given a skewed population ability distribution does the estimation accuracy for factorial structure of the classical factor analysis models depend on the extraction criterion applied for determining the number of factors from the data?
- Can person ability scores estimated under classical factor analytic approaches be representative of the true ability distribution or properties thereof when this distribution is skewed?

Mattson (1997)'s method can be used for specifying the parameter settings for the simulation study (cf. Section 4.2). We briefly describe this method (for details, see Mattson, 1997). Assume the standardized manifest variables are expressed as $\boldsymbol{z}=\boldsymbol{A}\boldsymbol{v}$, where $\boldsymbol{v}$ is the vector of latent variables and $\boldsymbol{A}$ is the matrix of model parameters. Moreover, assume that $\boldsymbol{v}=\boldsymbol{T}\boldsymbol{\omega}$, where $\boldsymbol{T}$ is a lower triangular square matrix such that each component of $\boldsymbol{v}$ is a linear combination of at most two components of $\boldsymbol{\omega}$, $E(\boldsymbol{v}\boldsymbol{v}')=\boldsymbol{\Sigma}_{\boldsymbol{v}}=\boldsymbol{T}\boldsymbol{T}'$, and $\boldsymbol{\omega}$ is a vector of mutually independent standardized random variables ω_i with finite central moments μ_{1i}, μ_{2i}, μ_{3i}, and μ_{4i}, of order up to four. Then

$$E(z)=\boldsymbol{A}\boldsymbol{T}E(\boldsymbol{\omega})=\boldsymbol{0}$$

and

$$E(\boldsymbol{z}\boldsymbol{z}')(=\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{v}}\boldsymbol{A}')=\boldsymbol{A}\boldsymbol{T}E(\boldsymbol{\omega}\boldsymbol{\omega}')\boldsymbol{T}'\boldsymbol{A}'=\boldsymbol{A}\boldsymbol{T}\boldsymbol{T}'\boldsymbol{A}'.$$

Or equivalently, $E(z_iz_j)=\boldsymbol{\gamma}_{\boldsymbol{i}}'\boldsymbol{\gamma}_{\boldsymbol{j}}$, where $\boldsymbol{\gamma}_{\boldsymbol{i}}=(\boldsymbol{a}_{\boldsymbol{i}}'\boldsymbol{T})'$ and $\boldsymbol{a}_{\boldsymbol{i}}'$ is the i-th row of $\boldsymbol{A}$. Under these conditions the third and fourth order central moments of z_i are given by

$$E(z_i^3)=\sum_m \gamma_{im}^3\mu_{3m} \qquad \text{and}$$

$$E(z_i^4)=\sum_m \gamma_{im}^4\mu_{4m}+6\sum_{m\geqslant 2}\sum_{o=1}^{m-1}\gamma_{im}^2\gamma_{io}^2.$$

Hence the univariate skewness $\sqrt{\beta_{1i}}$ and kurtosis β_{2i} of any z_i can be calculated by

$$\sqrt{\beta_{1i}}=\frac{E\left(z_i^3\right)}{\left[E\left(z_i^2\right)\right]^{3/2}} \qquad \text{and} \qquad \beta_{2i}=\frac{E\left(z_i^4\right)}{\left[E\left(z_i^2\right)\right]^2}.$$

In the simulation study, the exploratory factor analysis model with orthogonal factors ($\text{cov}(\boldsymbol{f},\boldsymbol{f})=\boldsymbol{I}$) and error variables assumed to be uncorrelated and unit normal (with standardized manifest variables) is used as the data generating model. Let $\boldsymbol{A}:=(\boldsymbol{L},\boldsymbol{I_p})$ be the concatenated matrix of dimension $p\times(k+p)$, where $\boldsymbol{I_p}$ is the unit matrix of order $p\times p$, and let $\boldsymbol{v}:=(\boldsymbol{f}',\boldsymbol{e}')'$ be the concatenated vector of length $k+p$. Then we have $\boldsymbol{z}=\boldsymbol{A}\boldsymbol{v}$ for the simulation factor model. Let $\boldsymbol{T}:=\boldsymbol{I_{(k+p)\times(k+p)}}$ and $\boldsymbol{\omega}:=\boldsymbol{v}$, then $\boldsymbol{T}$ and $\boldsymbol{\omega}$ satisfy the required assumptions afore mentioned. Hence the skewness and kurtosis of any z_i are given by, respectively,

$$\sqrt{\beta_{1i}}=\frac{\sum_{m=1}^{k+p}a_{im}^3\mu_{3m}}{[\boldsymbol{a}_{\boldsymbol{i}}'\boldsymbol{a}_{\boldsymbol{i}}]^{3/2}} \qquad \text{and}$$

$$\beta_{2i}=\frac{\sum_{m=1}^{k+p}a_{im}^4\mu_{4m}+6\sum_{m=2}^{k+p}\sum_{o=1}^{m-1}a_{im}^2a_{io}^2}{[\boldsymbol{a}_{\boldsymbol{i}}'\boldsymbol{a}_{\boldsymbol{i}}]^2}.$$

Mattson's method is used to specify such settings for the simulation study as they may be observed in large scale assessment data. The next section describes this in detail.

4.2. DESIGN OF THE SIMULATION STUDY

The number of manifest variables was fixed to $p=24$ throughout the simulation study. For the number of factors, we used numbers typically found in large scale assessment studies such as the Progress in International Reading Literacy Study (PIRLS, e.g., Mullis et al., 2006) or PISA (e.g., OECD, 2005). According to the assessment framework of PIRLS 2006 the number of dimensions for reading literacy was four, in PISA 2003 the scaling model had seven dimensions. We decided to use a simple loading structure for $\boldsymbol{L}$, in the sense that every manifest variable was assumed to load on only one factor (within-item unidimensionality) and that each factor was measured by the same number of manifest variables. In reliance on PIRLS and PISA in our simulation study, the numbers of factors were assumed to be four or eight. We assumed that some of the factors were well explained by their indicators while others were not, with upper rows (variables) of the loading matrix generally having higher factor loadings than lower rows (variables). Thus, the loading matrices employed in our study for the four and eight dimensional simulation models were, respectively,

$$\boldsymbol{L}=\begin{pmatrix}
0.9&0&0&0\\0.8&0&0&0\\0.7&0&0&0\\0.6&0&0&0\\0.5&0&0&0\\0.4&0&0&0\\
0&0.8&0&0\\0&0.7&0&0\\0&0.6&0&0\\0&0.5&0&0\\0&0.4&0&0\\0&0.3&0&0\\
0&0&0.6&0\\0&0&0.6&0\\0&0&0.5&0\\0&0&0.4&0\\0&0&0.4&0\\0&0&0.3&0\\
0&0&0&0.6\\0&0&0&0.5\\0&0&0&0.5\\0&0&0&0.4\\0&0&0&0.3\\0&0&0&0.3
\end{pmatrix} \quad \text{and} \quad \boldsymbol{L}=\begin{pmatrix}
0.9&0&0&0&0&0&0&0\\0.8&0&0&0&0&0&0&0\\0.7&0&0&0&0&0&0&0\\
0&0.8&0&0&0&0&0&0\\0&0.8&0&0&0&0&0&0\\0&0.7&0&0&0&0&0&0\\
0&0&0.8&0&0&0&0&0\\0&0&0.7&0&0&0&0&0\\0&0&0.6&0&0&0&0&0\\
0&0&0&0.7&0&0&0&0\\0&0&0&0.7&0&0&0&0\\0&0&0&0.7&0&0&0&0\\
0&0&0&0&0.7&0&0&0\\0&0&0&0&0.6&0&0&0\\0&0&0&0&0.6&0&0&0\\
0&0&0&0&0&0.6&0&0\\0&0&0&0&0&0.6&0&0\\0&0&0&0&0&0.5&0&0\\
0&0&0&0&0&0&0.5&0\\0&0&0&0&0&0&0.4&0\\0&0&0&0&0&0&0.4&0\\
0&0&0&0&0&0&0&0.4\\0&0&0&0&0&0&0&0.4\\0&0&0&0&0&0&0&0.3
\end{pmatrix}.$$

We decided to analyze released items of the PIRLS 2006 study (IEA, 2007) to have an empirical basis for the selection of skewness values for $\boldsymbol{\omega}(=\boldsymbol{\nu})$. We used a data set of dichotomously scored responses of 7,899 German students to 125 test items. **Figure 1** displays the distribution of the PIRLS items' (empirical) skewness values.[6]

We decided to simulate under three conditions for the distributions of $\boldsymbol{\omega}$. Under the first condition, ω_m $(m=1, \ldots, k)$ are normal with $\mu_{1m}=0$, $\mu_{2m}=1$, $\mu_{3m}=0$, and $\mu_{4m}=3$. Under the second condition, ω_m $(m=1, \ldots, k)$ are slightly skewed with $\mu_{1m}=0$, $\mu_{2m}=1$, $\mu_{3m}=-0.20$, and $\mu_{4m}=3$. Under the third condition, ω_m $(m=1, \ldots, k)$ are strongly skewed with $\mu_{1m}=0$, $\mu_{2m}=1$, $\mu_{3m}=-2$, and $\mu_{4m}=9$. The error terms were assumed to be unit normal, that is, we specified $\mu_{1h}=0$, $\mu_{2h}=1$, $\mu_{3h}=0$, and $\mu_{4m}=3$ for ω_h $(h=k+1, \ldots, k+p)$. Skewness and kurtosis of any z_i under each of the three conditions were computed using Mattson's method (Section 4.1). The values are reported in **Tables 1** and **2** for the four and eight dimensional factor spaces, respectively.

Under the slightly skewed distribution condition, the theoretical values of skewness for the manifest variables range between -0.060 and -0.005, a condition that captured approximately 20% of the considered PIRLS test items. Under the strongly skewed distribution condition, the theoretical values of skewness lie between -0.599 and -0.047, a condition that covered circa 30% of the PIRLS items (cf. **Figure 1**). Based on these theoretical skewness and kurtosis statistics, we can see to what extent under these model specifications the distributions of the manifest variables deviate from the normal distribution.

How to generate variates ω_i $(i=1, \ldots, k+p)$ such that they possess predetermined moments μ_{1i}, μ_{2i}, μ_{3i}, and μ_{4i}? To simulate values for ω_i with predetermined moments, we used the generalized lambda distribution (Ramberg et al., 1979)

$$\omega_i = \lambda_1 + \frac{u^{\lambda_3} - (1-u)^{\lambda_4}}{\lambda_2},$$

where u is uniform (0, 1), λ_1 is a location parameter, λ_2 a scale parameter, and λ_3 and λ_4 are shape parameters. To realize the desired distribution conditions for the simulation study (normal, slightly skewed, strongly skewed) using this general distribution its parameters λ_1, λ_2, λ_3, and λ_4 had to be specified accordingly. Ramberg et al. (1979) tabulate the required values for the λ parameters for different values of μ. In particular, for a (more or less) normal distribution with $\mu_1=0$, $\mu_2=1$, $\mu_3=0$, and $\mu_4=3$ the corresponding values are $\lambda_1=0$, $\lambda_2=0.197$, $\lambda_3=0.135$, and $\lambda_4=0.135$. For a slightly skewed distribution with $\mu_1=0$, $\mu_2=1$, $\mu_3=-0.20$, and $\mu_4=3$, the values are $\lambda_1=0.237$, $\lambda_2=0.193$, $\lambda_3=0.167$, and $\lambda_4=0.107$. For a strongly skewed distribution with $\mu_1=0$, $\mu_2=1$, $\mu_3=-2$, and $\mu_4=9$, the parameter values are given by $\lambda_1=0.993, \lambda_2=-0.108\cdot10^{-2}, \lambda_3=-0.108\cdot10^{-2}$, and $\lambda_4=-0.041\cdot10^{-3}$.

Remark. Indeed, various distributions are possible (see Mattson, 1997); however, the generalized lambda distribution proves to be special. It performs very well in comparison to other distributions, when theoretical moments calculated according to the Mattson formulae are compared to their corresponding empirical moments computed from data simulated under a factor model (based on that distribution). For details, see Reinartz et al. (2002). These authors have also studied the effects of the use of different (pseudo) random number generators for realizing the uniform distribution in such a comparison study. Out of three compared random number generators – RANUNI from SAS, URAND from PRELIS, and RANDOM from Mathematica – the generator RANUNI performed relatively well or better. In this paper, we used the SAS program for our simulation study.[7]

Besides the number of factors and the distributions of the latent variables, sample size was varied. In the small sample case, every z_i consisted of $n=200$ observations, and in the large sample case z_i contained $n=600$ observations. **Table 3** summarizes the design of the simulation study. Overall there

[6]All figures of this paper were produced using the R statistical computing environment (R Development Core Team, 2011; www.r-project.org). The source files are freely available from the authors.

[7]For the factor analyses in this paper, we used the SAS program and its PROC FACTOR implementation of the methods PCA, EFA, and PAA. More precisely, variation of the PROC FACTOR statements, run in their default settings, yields the performed procedures PCA, EFA, and PAA (e.g., EFA if METHOD = ML).

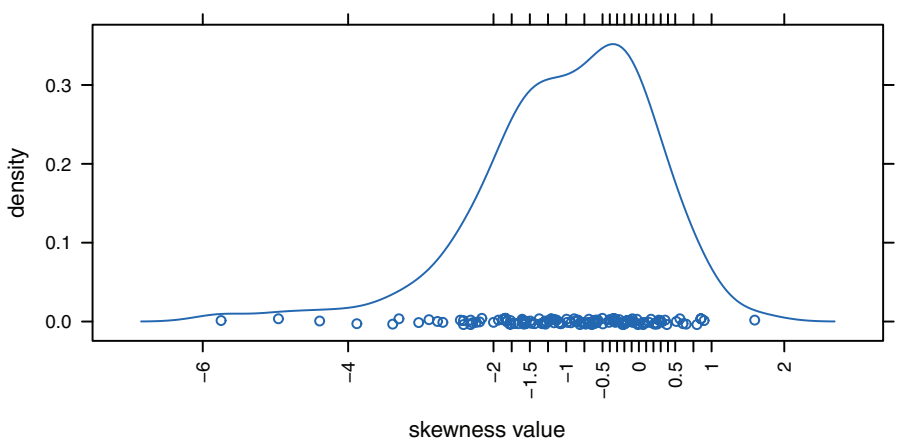


FIGURE 1 | Distribution of the skewness values for the 125 PIRLS test items.

Table 1 | Theoretical values of skewness and kurtosis for z_i (four factors).

z_i	Latent variable					
	Normal[a]		Slightly skewed[b]		Strongly skewed[c]	
	$\sqrt{\beta_{1i}}$	β_{2i}	$\sqrt{\beta_{1i}}$	β_{2i}	$\sqrt{\beta_{1i}}$	β_{2i}
z_1	0	3	−0.060	3	−0.599	4.202
z_2	0	3	−0.049	3	−0.488	3.914
z_3	0	3	−0.038	3	−0.377	3.649
z_4	0	3	−0.027	3	−0.272	3.420
z_5	0	3	−0.018	3	−0.179	3.240
z_6	0	3	−0.010	3	−0.102	3.114
z_7	0	3	−0.049	3	−0.488	3.914
z_8	0	3	−0.038	3	−0.377	3.649
z_9	0	3	−0.027	3	−0.272	3.420
z_{10}	0	3	−0.018	3	−0.179	3.240
z_{11}	0	3	−0.010	3	−0.102	3.114
z_{12}	0	3	−0.005	3	−0.047	3.041
z_{13}	0	3	−0.027	3	−0.272	3.420
z_{14}	0	3	−0.027	3	−0.272	3.420
z_{15}	0	3	−0.018	3	−0.179	3.240
z_{16}	0	3	−0.010	3	−0.102	3.114
z_{17}	0	3	−0.010	3	−0.102	3.114
z_{18}	0	3	−0.005	3	−0.047	3.041
z_{19}	0	3	−0.027	3	−0.272	3.420
z_{20}	0	3	−0.018	3	−0.179	3.240
z_{21}	0	3	−0.018	3	−0.179	3.240
z_{22}	0	3	−0.010	3	−0.102	3.114
z_{23}	0	3	−0.005	3	−0.047	3.041
z_{24}	0	3	−0.005	3	−0.047	3.041

[a] $\mu_{1m} = 0$, $\mu_{2m} = 1$, $\mu_{3m} = 0$, and $\mu_{4m} = 3$.
[b] $\mu_{1m} = 0$, $\mu_{2m} = 1$, $\mu_{3m} = -0.20$, and $\mu_{4m} = 3$.
[c] $\mu_{1m} = 0$, $\mu_{2m} = 1$, $\mu_{3m} = -2$, and $\mu_{4m} = 9$.

Table 2 | Theoretical values of skewness and kurtosis for z_i (eight factors).

z_i	Latent variable					
	Normal[a]		Slightly skewed[b]		Strongly skewed[c]	
	$\sqrt{\beta_{1i}}$	β_{2i}	$\sqrt{\beta_{1i}}$	β_{2i}	$\sqrt{\beta_{1i}}$	β_{2i}
z_1	0	3	−0.060	3	−0.599	4.202
z_2	0	3	−0.049	3	−0.488	3.914
z_3	0	3	−0.038	3	−0.377	3.649
z_4	0	3	−0.049	3	−0.488	3.914
z_5	0	3	−0.049	3	−0.488	3.914
z_6	0	3	−0.038	3	−0.377	3.649
z_7	0	3	−0.049	3	−0.488	3.914
z_8	0	3	−0.038	3	−0.377	3.649
z_9	0	3	−0.027	3	−0.272	3.420
z_{10}	0	3	−0.038	3	−0.377	3.649
z_{11}	0	3	−0.038	3	−0.377	3.649
z_{12}	0	3	−0.038	3	−0.377	3.649
z_{13}	0	3	−0.038	3	−0.377	3.649
z_{14}	0	3	−0.027	3	−0.272	3.420
z_{15}	0	3	−0.027	3	−0.272	3.420
z_{16}	0	3	−0.027	3	−0.272	3.420
z_{17}	0	3	−0.027	3	−0.272	3.420
z_{18}	0	3	−0.018	3	−0.179	3.240
z_{19}	0	3	−0.018	3	−0.179	3.240
z_{20}	0	3	−0.010	3	−0.102	3.114
z_{21}	0	3	−0.010	3	−0.102	3.114
z_{22}	0	3	−0.010	3	−0.102	3.114
z_{23}	0	3	−0.010	3	−0.102	3.114
z_{24}	0	3	−0.005	3	−0.047	3.041

[a] $\mu_{1m} = 0$, $\mu_{2m} = 1$, $\mu_{3m} = 0$, and $\mu_{4m} = 3$.
[b] $\mu_{1m} = 0$, $\mu_{2m} = 1$, $\mu_{3m} = -0.20$, and $\mu_{4m} = 3$.
[c] $\mu_{1m} = 0$, $\mu_{2m} = 1$, $\mu_{3m} = -2$, and $\mu_{4m} = 9$.

are 12 conditions and for every condition 100 data sets were simulated.

Each of the generated 1,200 data sets were analyzed using all of the models of principal component analysis, exploratory factor analysis (ML estimation), and principal axis analysis altogether with a varimax rotation (Kaiser, 1958).[8] For any data set under each model, the factors, and hence, the numbers of retained factors were determined by applying the following three extraction criteria or approaches: the Kaiser-Guttman criterion, the scree test, and the parallel analysis procedure.[9]

Table 3 | Summary of the simulation design and number of generated data sets.

Sample size	Number of factors	Latent variable distribution		
		Normal	Slightly skewed	Strongly skewed
200	4	100	100	100
	8	100	100	100
600	4	100	100	100
	8	100	100	100

4.3. EVALUATION CRITERIA

The criteria for evaluating the performance of the classical factor models are the number of extracted factors (as compared to true dimensionality), the skewness of the estimated latent ability

[8] Because of rotational indeterminacy in the factor analysis approach (e.g., Maraun, 1996), the results are as much an evaluation of varimax rotation as they are an evaluation of the manipulated variables in the study. For more information, see Section 7.

[9] The Kaiser-Guttman criterion is a poor way to determine the number of factors. However, due to the fact that none of the existing studies has investigated the estimation accuracy of this criterion when the latent ability distribution is skewed, we have decided to include the Kaiser-Guttman criterion in our study. This criterion may also be viewed as a "worst performing" baseline criterion, which other extraction methods need to outperform, as best as possible.

distribution, and the discrepancy between the estimated and the true loading matrix. The latter two criteria are computed using the true number of factors. Furthermore, Shapiro-Wilk tests for assessing normality of the ability estimates are presented and distributions of the estimated and true factor scores are compared.

For the skewness criterion, under a factor model and a simulation condition, for any data set the factor scores on a factor were computed and their empirical skewness was the value for this data set that was used and plotted. For the discrepancy criterion, under a factor model and a simulation condition, for any data set $i = 1, \ldots, 100$ a discrepancy measure D_i was calculated,

$$D_i = \frac{\sum_{x=1}^{p} \sum_{y=1}^{k} |\hat{l}_{i;xy} - l_{xy}|}{kp},$$

where $\hat{l}_{i;xy}$ and l_{xy} represent the entries of the estimated (varimax rotated, for data set $i = 1, \ldots, 100$) and true loading matrices, respectively. It gives the averaged sum of the absolute differences between the estimated and true factor loadings. We also report the average and variance (or standard deviation) of these discrepancy measures, over all simulated data sets,

$$\overline{D} = \frac{1}{100} \sum_{i=1}^{100} D_i \qquad \text{and} \qquad s^2 = \frac{1}{100 - 1} \sum_{i=1}^{100} (D_i - \overline{D})^2.$$

In addition to calculating estimated factor score skewness values, we also tested for univariate normality of the estimated factor scores. We used the Shapiro-Wilk test statistic W (Shapiro and Wilk, 1965). In comparison to other univariate normality tests, the Shapiro-Wilk test seems to have relatively high power (Seier, 2002). In our study, under a factor model and a simulation condition, for any data set the Shapiro-Wilk test statistic's p-value was calculated for the estimated factor scores on a factor and the distribution of the p-values obtained from 100 simulated data sets was plotted.

5. RESULTS

We present the results of our simulation study.

5.1. NUMBER OF EXTRACTED FACTORS

Figure 2 shows the relative frequencies of the numbers of extracted factors for sample size $n = 200$ and $k = 4$ as true number of factors. If the Kaiser-Guttman criterion is used, the number of extracted factors is overestimated (for PCA) or tends to be underestimated (for EFA and PAA). With the scree test, four dimensions where extracted in the majority of cases, but variation of the numbers of extracted factors over the different data sets is high. High variation in this case can be explained by the ambiguous and hence difficult to interpret eigenvalue graphics that one needs to visually inspect for the scree test. Applying the parallel analysis method, variation of the numbers of extracted factors can be reduced and the true number of factors is estimated very well (e.g., for PCA). There does not seem to be a relationship between the number of extracted factors and the underlying distribution (normal, slightly skewed, strongly skewed) of the latent ability values.

When sample size is increased to $n = 600$, variation of the estimated numbers of factors decreases substantially under many conditions (see **Figure 3**). Compared to small sample sizes, the scree test and the parallel analysis method perform very well. The Kaiser-Guttman criterion still leads to a biased estimation of the true number of factors. Once again, there seems to be no relationship between the distribution of the latent ability values and the number of extracted factors.

Figure 4, for a sample size of $n = 200$, shows the case when there are $k = 8$ factors underlying the data. The Kaiser-Guttman criterion again leads to overestimation or underestimation of the true number of factors. The extraction results for the scree test have very high variation, and estimation of the true number of factors is least biased when the parallel analysis method is used.

Increasing sample size from $n = 200$ to 600 results in a significant reduction of variation (**Figure 5**). However, the true number of factors can be estimated without bias only when the parallel analysis method is used as extraction criterion. A possible relationship between the distribution of the latent ability values and the number of extracted factors once again does not seem to be apparent.

To sum up, we suppose that the "number of factors extracted" is relatively robust against the extent the latent ability values may be skewed. Another observation is that the parallel analysis method seems to outperform the scree test and the Kaiser-Guttman criterion when it comes to detecting the number of underlying factors.

5.2. SKEWNESS OF THE ESTIMATED LATENT ABILITY DISTRIBUTION

Figure 6A shows the distributions of the estimated factor score skewness values, for $n = 200$, $k = 4$, and $\mu_{3m} = 0$. The majority of the skewness values lies in close vicinity of 0. In other words, for a true normal latent ability distribution with skewness $\mu_3 = 0$, under the classical factor models the estimated latent ability scores most likely seem to have skewness values of approximately 0. An impact of the factor model used for the analysis of the data on the skewness of the estimated latent ability values cannot be seen under this simulation condition. However, the standard deviations of the skewness values clearly decrease from the first to the fourth factor. In other words, the true skewness of the latent ability distribution may be more precisely estimated for the fourth factor than for the first.

When true latent ability values are slightly negative skewed, $\mu_3 = -0.20$, in our simulation study this skewness may only be properly estimated for the first and second extracted factors (**Figure 6B**). The estimated latent ability values of the third and fourth extracted factors more give skewness values of approximately 0. The true value of skewness for these factors hence may likely to be overestimated.

If true latent ability values are strongly negative skewed, $\mu_3 = -2$, unbiased estimation of true skewness may not be possible (**Figure 6C**). Even in the case of the first and second factors, the estimation is biased now. True skewness of the latent ability distribution may be overestimated regardless of the used factor model or factor position.

To sum up, under the classical factor models, the concept of "skewness of the estimated latent ability distribution" seems to be sensitive with respect to the extent the latent ability values may be skewed. It seems that, the more the true latent ability values are skewed, the greater is overestimation of true skewness. In other

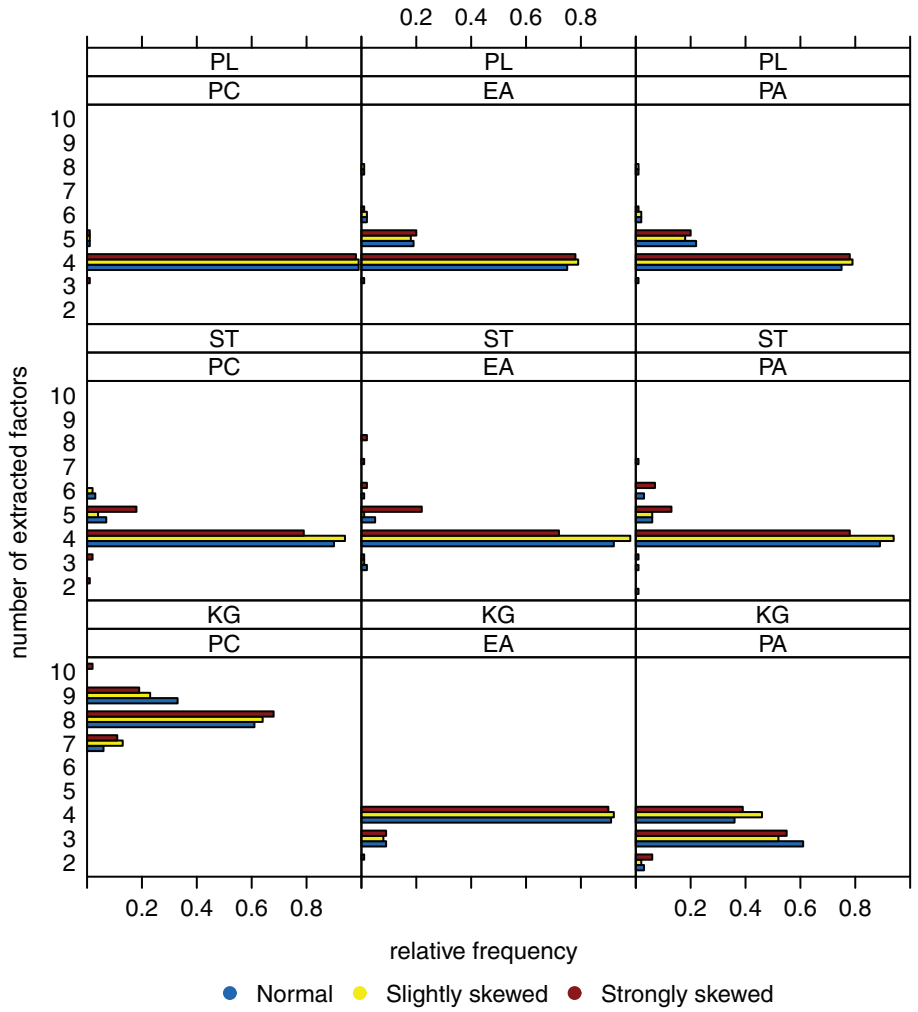


FIGURE 2 | Relative frequencies of the numbers of extracted factors, for $n = 200$ and $k = 4$. Factor models are principal component analysis (PCA, or PC), exploratory factor analysis (EFA, or EA), and principal axis analysis (PAA, or PA). Kaiser-Guttman criterion (KG), scree test (ST), and parallel analysis (PL) serve as factor extraction criteria.

words, strongly negative skewed distributions may not be estimated without bias based on the classical factor models. Increasing sample size, for example from $n = 200$ to 600, or changing the number of underlying factors, say from $k = 4$ to 8, did not alter this observation considerably. For that reason, the corresponding plots at this point of the paper are omitted and can be found in Kasper (2012).

We performed Shapiro-Wilk tests for univariate normality of the estimated factor scores. As can be seen from **Figure 7A**, under normally distributed true latent ability scores nearly all values of W are statistically non-significant. In these cases, the null hypothesis cannot be rejected.

A similar conclusion can be drawn when the true latent ability values are not normally distributed but instead follow a slightly skewed distribution (**Figure 7B**). Nearly all Shapiro-Wilk test statistic values are statistically non-significant. In other words, the null hypothesis stating normally distributed latent ability values is seldom rejected although the true latent distribution is skewed and not normal. No relationship between the p-values and the used factor model or factor position may be apparent (disregarding the observation that the p-values for the fourth factor are generally lower than for the other factors).

The case of a strongly skewed factor score distribution is depicted in **Figure 7C**. Virtually all values of W are statistically significant and the null hypothesis of normality of factor scores is rejected. Similar conclusions or observations may be drawn for increased sample size or factor space dimension and we do omit presenting plots thereof.

Finally, **Figure 8** shows the distribution of the estimated factor scores on the fourth factor (for $k = 4$) in comparison to the true strongly skewed ability distribution under the exploratory factor analysis model for a sample size of $n = 1{,}000$. The unit normal distribution is plotted as a reference. The estimated factor scores have a skewness value of -0.47 compared to true skewness -2. The estimated distribution deviates from the true distribution and does not approximate it acceptably well.

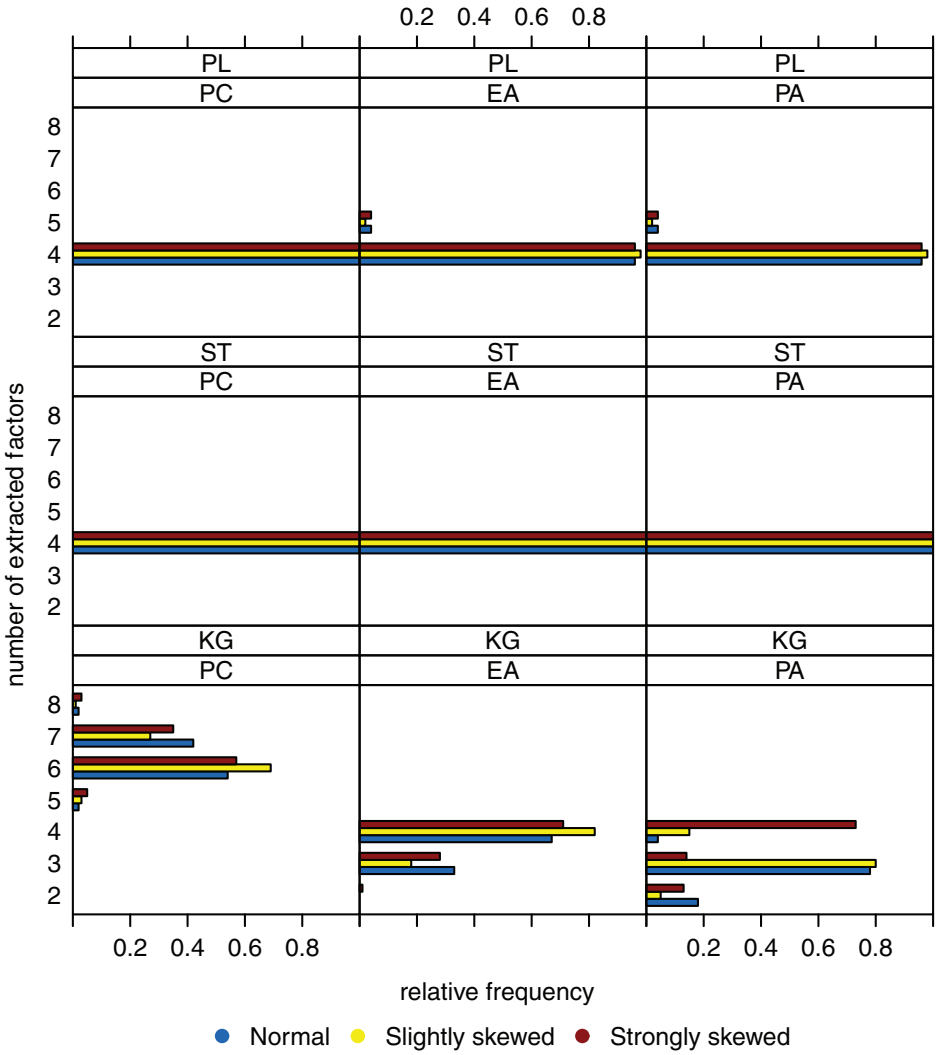


FIGURE 3 | Relative frequencies of the numbers of extracted factors, for $n = 600$ and $k = 4$. Factor models are principal component analysis (PCA, or PC), exploratory factor analysis (EFA, or EA), and principal axis analysis (PAA, or PA). Kaiser-Guttman criterion (KG), scree test (ST), and parallel analysis (PL) serve as factor extraction criteria.

5.3. DISCREPANCY BETWEEN THE ESTIMATED AND THE TRUE LOADING MATRIX

In **Table 4**, the average and standard deviation coefficients $\bar{D}$ and s for the discrepancies are reported. The largest average discrepancy values are obtained for the condition $n = 200$, $k = 8$, and the strongly skewed latent ability distribution: 0.173, 0.157, and 0.143 for PCA, EFA, and PAA, respectively. Under this condition, the true factor loadings are, mostly or clearly, overestimated or underestimated. Minor differences between the estimated and true factor loadings are obtained for $n = 600$, $k = 4$, and the normal latent ability distribution: with average discrepancies 0.076, 0.063, and 0.066 for PCA, EFA, and PAA, respectively.

Deviations of the estimated loading matrix from the true loading matrix can also be quantified and visualized at the level of individual absolute differences $|\hat{l}_{i;xy} - l_{xy}|$. In this way not only overall discrepancy averages can be studied but also the distribution of absolute differences at the individual entry level. **Figure 9** shows the distributions of the absolute differences $|\hat{l}_{i;xy} - l_{xy}|$ for the different sample sizes and numbers of underlying factors. In each panel, $100pk$ absolute differences are plotted.

The majority of the absolute differences lies in the range from 0 to circa 0.20. Larger absolute differences between the estimated and true factor loadings occurred rather rarely. It is also apparent that the 36 distributions hardly differ. This observation suggests that the effects or impacts of sample size, true number of factors, and the latent ability distribution on the accuracy of the classical factor models for estimating the factor loadings are rather weak. In that sense, estimation of the loading matrix seems to be robust overall. In our simulation study, we were not able to see a clear relationship between the distribution of the latent ability values and the discrepancy between the estimated and the true loading matrix.

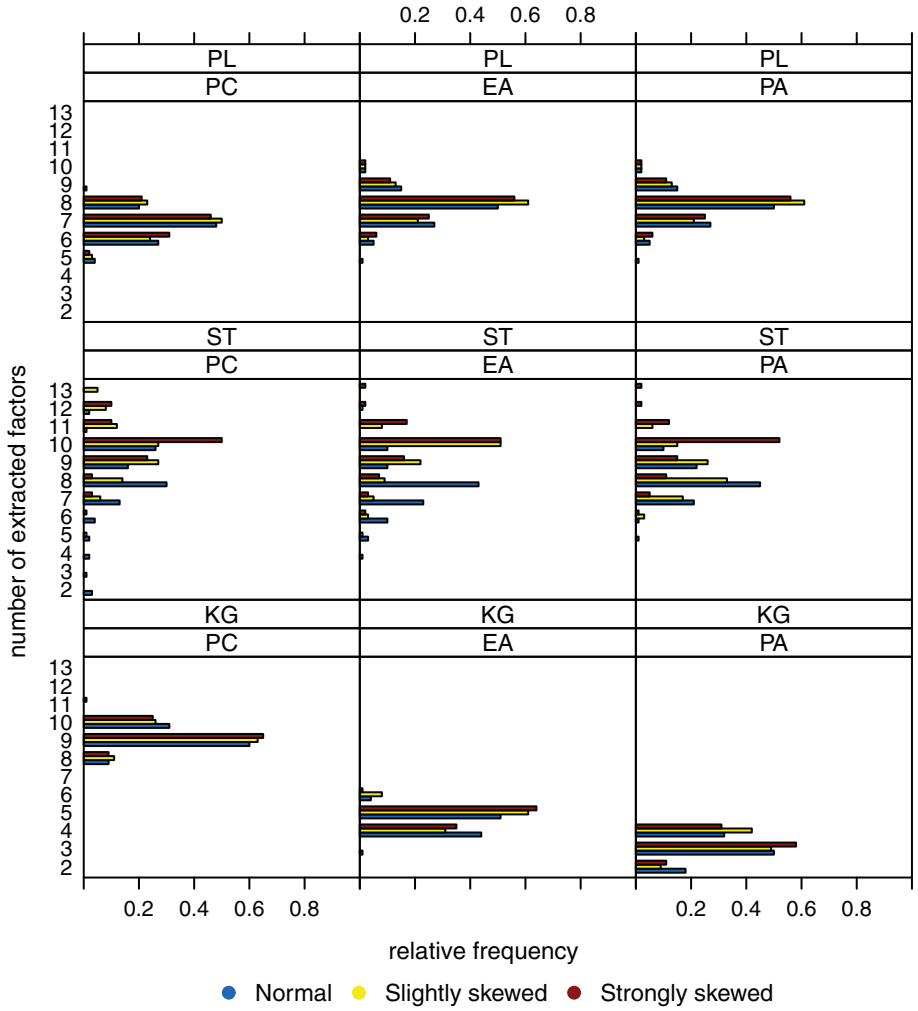


FIGURE 4 | Relative frequencies of the numbers of extracted factors, for $n = 200$ and $k = 8$. Factor models are principal component analysis (PCA, or PC), exploratory factor analysis (EFA, or EA), and principal axis analysis (PAA, or PA). Kaiser-Guttman criterion (KG), scree test (ST), and parallel analysis (PL) serve as factor extraction criteria.

6. ANALYSIS OF PIRLS 2006 DATA

In addition to the simulation study, the classical factor analytic approaches are also compared on the part of PIRLS 2006 data that we presented in Section 4.2. The booklet design in PIRLS implies that only a selection of the items has been administered to each student, depending on booklet approximately 23–26 test items per student (Mullis et al., 2006). As a consequence, the covariance or correlation matrices required for the factor models can only be computed for the items of a particular test booklet. Since analysis of all thirteen booklets of the PIRLS 2006 study is out of the scope of this paper, we decided to analyze booklet number 4. This booklet contains 23 items, and nine of these items (circa 40% of all items) have skewness values in the range of −0.6 to 0. This skewness range corresponds to the values considered in the simulation study, and no other test booklet had a comparably high percentage of items with skewness values in this range.

Note that in the empirical application dichotomized multicategory items are analyzed. In practice, large scale assessment data are discrete and not continuous. Yet, the metric scale indicator case considered in the simulation study can serve as an informative baseline; for instance (issue of polychoric approximation) to the extent that a product-moment correlation is a valid representation of bivariate relationships among interval-scaled variables (e.g., Flora et al., 2012). In our paper, the simulation results and the results obtained for the empirical large scale assessment application are, more or less, comparable.

In PIRLS 2006, four sorts of items were constructed and used for assigning "plausible values" to students (for details, see Martin et al., 2007). Any item loads on exactly one of the two dimensions "Literacy Experience" (L) and "Acquire and Use Information" (A) and also measures either the dimension "Retrieving and Straightforward Inferencing" (R) or the dimension "Interpreting, Integrating, and Evaluating" (I). Moreover, all of these items are assumed to be indicators for the postulated higher dimension "Overall Reading." In other words, PIRLS 2006 items may be assumed to be one-dimensional if the "uncorrelated" factor "Overall Reading" is

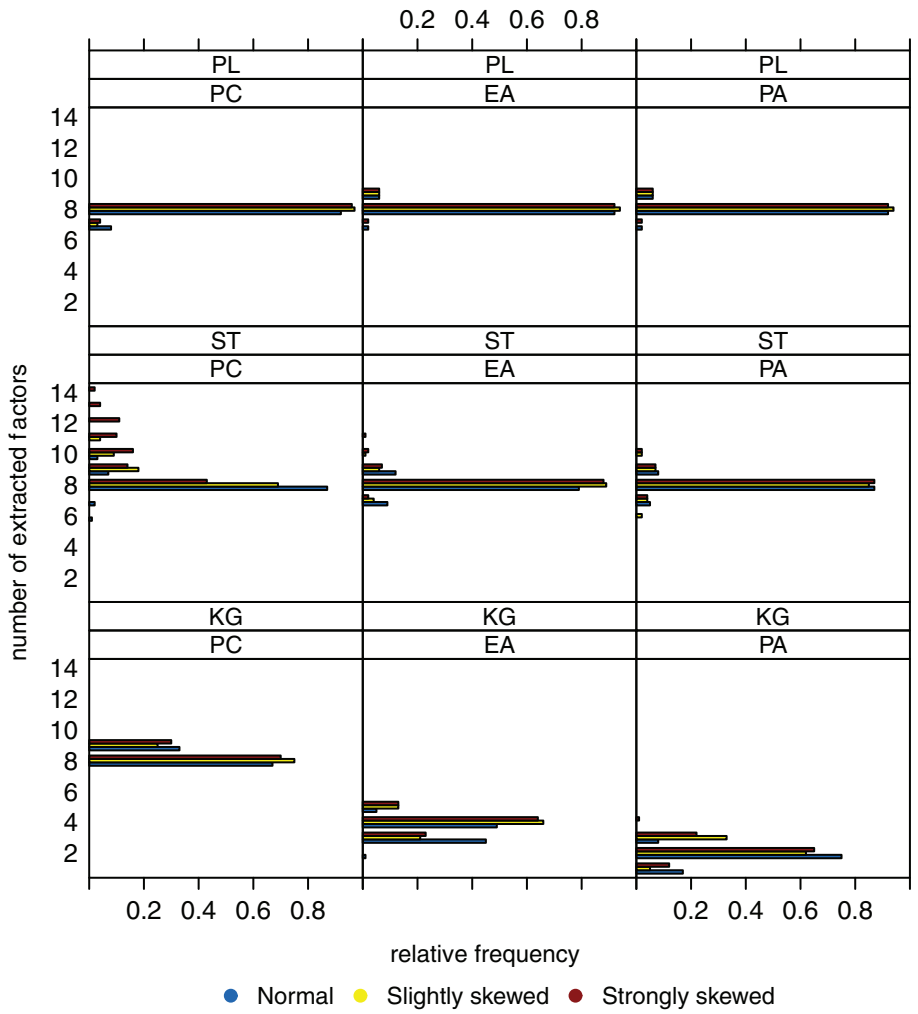


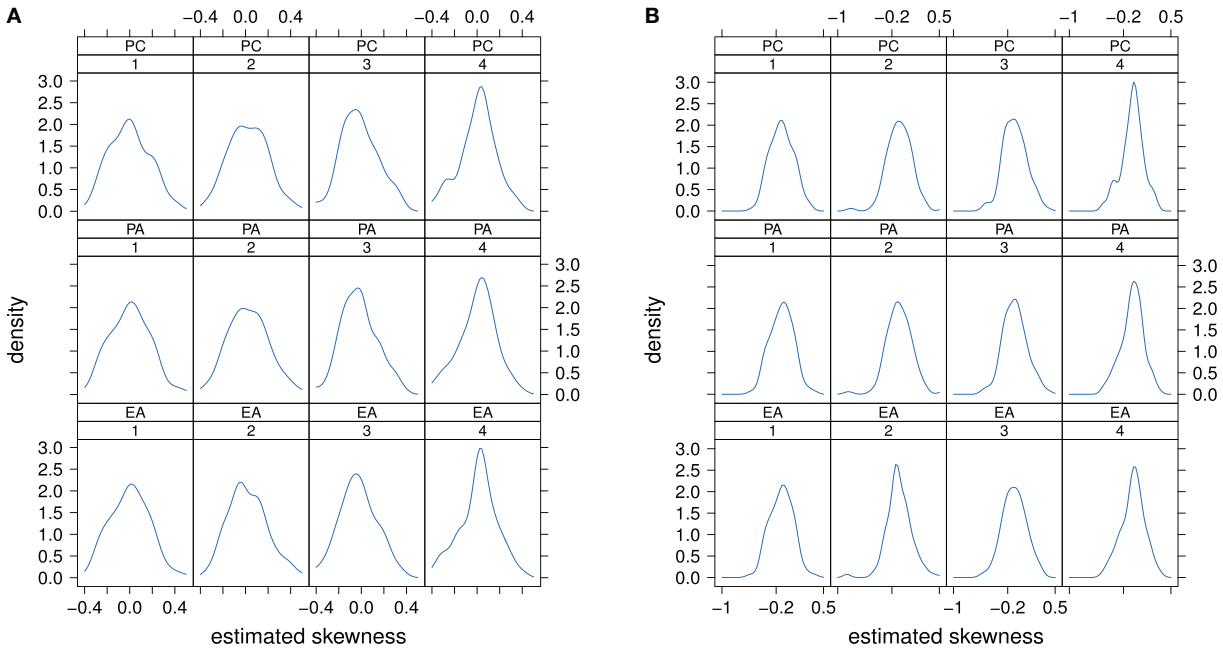
FIGURE 5 | Relative frequencies of the numbers of extracted factors, for $n = 600$ and $k = 8$. Factor models are principal component analysis (PCA, or PC), exploratory factor analysis (EFA, or EA), and principal axis analysis (PAA, or PA). Kaiser-Guttman criterion (KG), scree test (ST), and parallel analysis (PL) serve as factor extraction criteria.

considered ("orthogonal" case), or two-dimensional if any of the four combinations of correlated factors {A, L} × {I, R} is postulated (oblique case). In the latter case, "Overall Reading" may be assumed a higher order dimension common to the four factors. Booklet number 4 covers all these four sorts of PIRLS items.

A total of $n = 526$ students worked on booklet number 4. We investigated these data using principal component analysis, exploratory factor analysis, and principal axis analysis. For determining the number of underlying dimensions, the Kaiser-Guttman criterion, the scree test, and the method of parallel analysis were used. The results of the analyses can be found in **Table 5**.

The situation at this point is comparable to what we have reported in simulation in **Figure 3**. The scree test unveils unidimensionality of the test data independent of factor model. The numbers of factors extracted by the parallel analysis method depend on the factor model that was used. For PCA, again as for the scree test, unidimensionality is detected, however for the error component models EFA and PAA, four dimensions are uncovered (see also below). It seems that these "inferential" or "distributional" factor models, to some degree, are sensitive to dependencies among factors. According to the Kaiser-Guttman criterion, which performs worst, there are six dimensions underlying the data for any of the three factor models.

The varimax rotated loading matrices for the exploratory factor analysis and principal axis analysis models with four factors are reported in **Tables 6** and **7**. Once again, the situation is comparable to what we have obtained in simulation in **Table 4** or **Figure 9**. The estimated loading matrices under EFA and PAA are very similar. Highlighted factor loadings $\hat{l}_{xy} \geqslant 0.30$, for instance, are identically located in the matrices. As can be seen from **Tables 6** and **7**, substantially different items in regard to their PIRLS contents load on the same factors, and moreover, there are items of same PIRLS contents that show substantial loadings on different factors. We suppose that this may be a consequence of the factors, in this example, most likely being correlated with a postulated common single dimension underlying the factors.



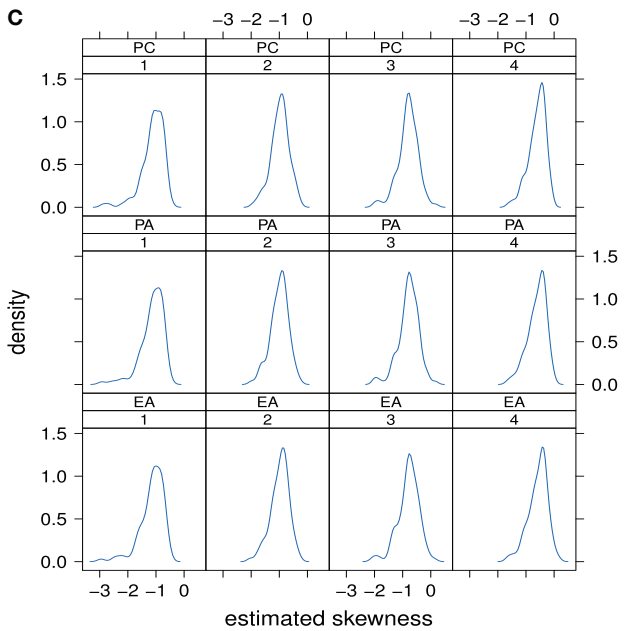


FIGURE 6 | Distributions of the estimated factor score skewness values as a "function" of factor model and factor position. Factor models are principal component analysis (PCA, or PC), exploratory factor analysis (EFA, or EA), and principal axis analysis (PAA, or PA). Numbers 1, 2, 3, and 4 stand for 1st, 2nd, 3rd, and 4th factors, respectively. The normal, slightly skewed, and strongly skewed distribution conditions are depicted in the panels **A**, **B**, and **C**, respectively.

7. CONCLUSION

7.1. SUMMARY

Assessing construct validity of a test in the sense of its factorial structure is important. For example, we have addressed possible implications for the analysis of criterion-referenced tests or for such large scale assessment studies as the PISA or PIRLS. There are a number of latent variable models that may be used to analyze the factorial structure of a test. This paper has focused

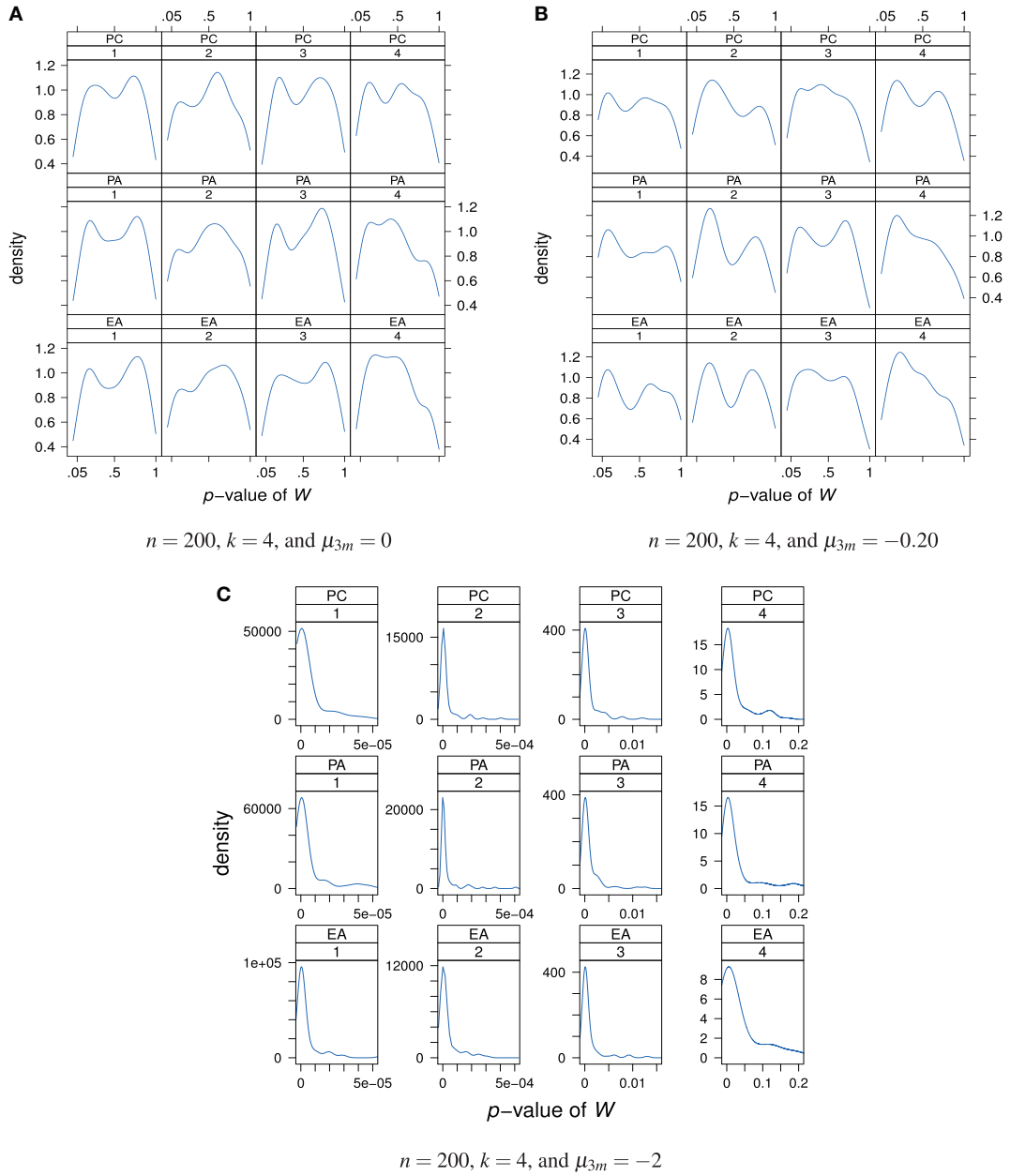


FIGURE 7 | Distributions of the *p*-values of the Shapiro-Wilk test statistic *W* as a "function" of factor model and factor position. Factor models are principal component analysis (PCA, or PC), exploratory factor analysis (EFA, or EA), and principal axis analysis (PAA, or PA). Numbers 1, 2, 3, and 4 stand for 1st, 2nd, 3rd, and 4th factors, respectively. The normal, slightly skewed, and strongly skewed distribution conditions are depicted in the panels **A**, **B**, and **C**, respectively.

on the following classical factor analytic approaches: principal component analysis, exploratory factor analysis, and principal axis analysis. We have investigated how accurately the factorial structure of test data can be estimated with these approaches, when assumptions associated with the procedures are not satisfied. We have examined the scope of those methods for estimating properties of the population latent ability distribution, especially when that distribution is slightly or strongly skewed (and not normal).

The estimation accuracy of the classical factor analytic approaches has been investigated in a simulation study. The study has in particular shown that the estimation of the true number of factors and of the underlying factor loadings seems to be relatively robust against a skewed population ability or factor score distribution (see Sections 5.1 and 5.3, respectively). Skewness and distribution of the estimated factor scores, on the other hand, have been seen to be sensitive concerning the properties of the true ability distribution (see Section 5.2). Therefore, the classical

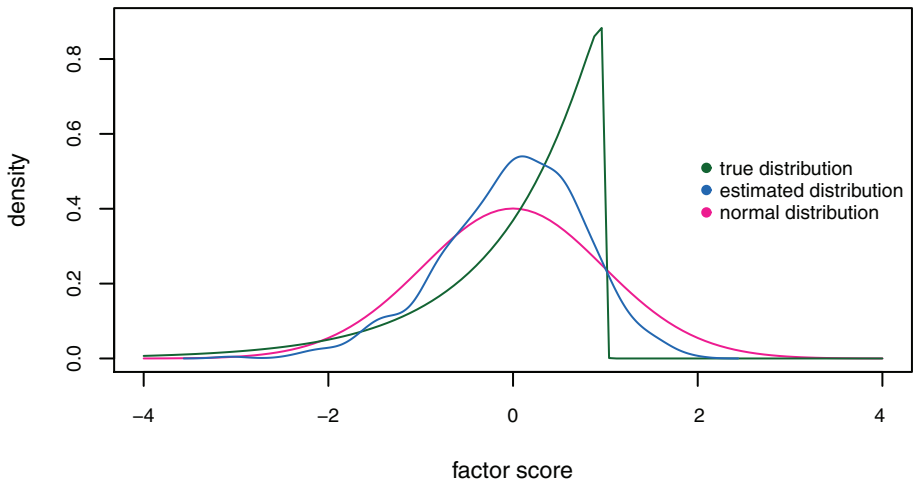


FIGURE 8 | Distributions of the estimated (blue curve) and true (green curve) factor scores on the fourth factor under the exploratory factor analysis model for sample size $n = 1{,}000$, factor space dimension $k = 4$, and true skewness $\mu_3 = -2$. The unit normal distribution is plotted as a reference (red curve).

Table 4 | Discrepancy averages and standard deviations $\bar{D}$ and *s*, respectively.

n	*k*	Model	Latent variable distribution					
			Normal		Slightly skewed		Strongly skewed	
			$\bar{D}$	*s*	$\bar{D}$	*s*	$\bar{D}$	*s*
200	4	PCA[a]	0.143	0.156	0.143	0.156	0.158	0.173
		EFA[b]	0.129	0.142	0.124	0.136	0.141	0.154
		PAA[c]	0.128	0.139	0.124	0.136	0.137	0.150
600	4	PCA	0.076	0.087	0.075	0.086	0.091	0.106
		EFA	0.063	0.072	0.062	0.072	0.080	0.095
		PAA	0.066	0.075	0.064	0.074	0.082	0.096
200	8	PCA	0.165	0.169	0.162	0.166	0.172	0.176
		EFA	0.154	0.157	0.152	0.155	0.156	0.159
		PAA	0.135	0.138	0.134	0.138	0.143	0.146
600	8	PCA	0.119	0.123	0.118	0.123	0.125	0.130
		EFA	0.106	0.112	0.107	0.112	0.115	0.120
		PAA	0.097	0.101	0.095	0.099	0.102	0.105

[a] *PCA, principal component analysis;* [b] *EFA, exploratory factor analysis;* [c] *PAA, principal axis analysis.*

factor analytic procedures, even though they are performed with metric scale indicator variables, seem not to be appropriate for estimating properties of ability in the "non-normal case." Significance of this result on sensitivity of factor score estimation to the nature of the latent distribution has been discussed for the PISA study, which is an international survey with impact on education policy making and the education system in Germany (see Sections 1 and 3.1). In addition to that discussion, the classical factor analytic approaches have been examined in more detail on PIRLS large scale assessment data, corroborating the results that we have obtained from the simulation study (see Section 6).

A primary aim of our work is to develop some basic understanding for how and to what extent the results of classical factor analyses (in the present paper, PCA, EFA, and PAA) may be affected by a non-normal latent factor score distribution. This has to be distinguished from non-normality in the manifest variables, which has been largely studied in the literature on the factor analysis of items (cf. Section 3.2). In this respect, regarding the investigation of non-normal factors, the present paper is novel. However, this is important, since it is not difficult to conceive of the possibility that latent variables may be skewed. Interestingly, moreover we have seen that a purely computational dimensionality reduction method can perform surprisingly well, as compared to the

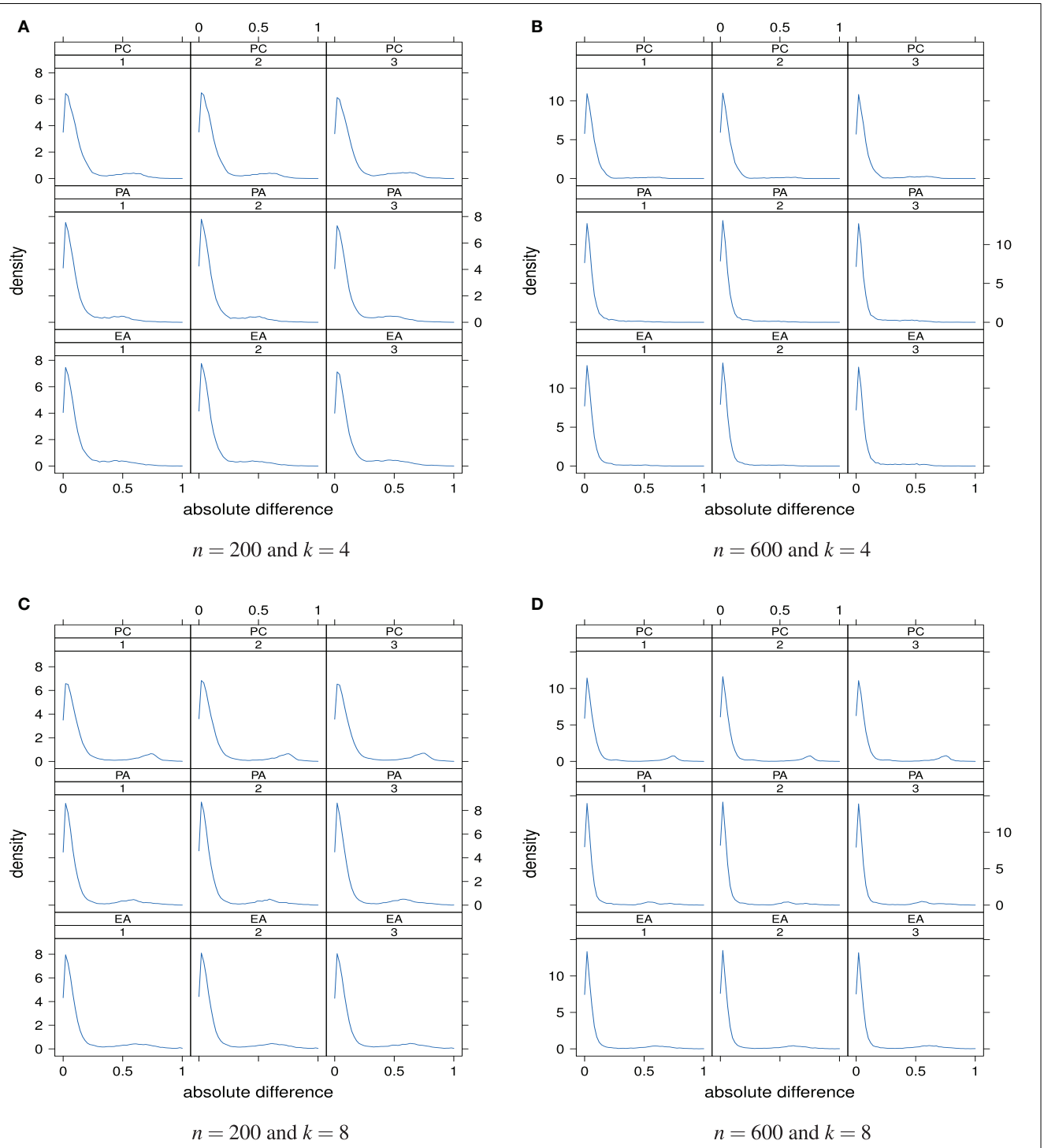


FIGURE 9 | Distributions of the absolute differences $\left|\hat{l}_{i;xy} - l_{xy}\right|$ as a "function" of factor model and skewness of the latent ability distribution. Factor models are principal component analysis (PCA, or PC), exploratory factor analysis (EFA, or EA), and principal axis analysis (PAA, or PA). Numbers 1, 2, and 3 stand for normal, slightly skewed, and strongly skewed population latent ability values, respectively. The panels are for the different sample sizes and numbers of underlying factors.

results obtained based on latent variable models. This observation may possibly be coined a general research program: whether genuine statistical approaches (originally based on variables without a measurement error) can work well, perhaps under specific restrictions to be explored, when latent variables are basically postulated, seemingly more closely matching the purpose of analysis.

Table 5 | Number of extracted dimensions for the PIRLS 2006 test booklet number 4, German sample.

Extraction method	Factor model		
	PCA[a]	EFA[b]	PAA[c]
Kaiser–Guttman criterion	6	6	6
Scree test	1	1	1
Parallel analysis method	1	4	4

[a]*PCA, principal component analysis;* [b]*EFA, exploratory factor analysis;* [c]*PAA, principal axis analysis.*

Table 6 | Loading matrix for four factors exploratory factor analysis of the PIRLS 2006 data for test booklet number 4, German sample.

Item	Factor			
	1	2	3	4
R011A01C^{A,R}	0.15	0.26	**0.39**	−0.05
R011A02M^{A,R}	0.14	0.28	**0.34**	0.19
R011A03C^{A,R}	0.16	0.24	0.09	0.03
R011A04C^{A,I}	**0.39**	0.19	0.10	0.06
R011A05M^{A,R}	0.22	0.08	0.19	0.21
R011A06M^{A,R}	0.20	0.03	0.14	0.06
R011A07C^{A,R}	**0.50**	0.20	0.22	0.15
R011A08C^{A,R}	**0.35**	0.04	**0.38**	−0.09
R011A09C^{A,I}	**0.55**	0.18	0.11	0.00
R011A10M^{A,I}	0.28	0.27	0.22	0.11
R011A11C^{A,I}	**0.37**	0.06	0.14	0.02
R021E01M^{L,R}	0.08	**0.45**	0.19	−0.06
R021E02M^{L,R}	0.02	**0.49**	0.09	0.24
R021E03M^{L,R}	0.14	**0.34**	−0.02	0.02
R021E04M^{L,R}	0.17	0.28	0.15	0.02
R021E05C^{L,R}	0.22	0.23	**0.32**	0.12
R021E06M^{L,R}	0.17	**0.44**	0.09	0.28
R021E07C^{L,I}	0.13	0.06	**0.48**	0.22
R021E08M^{L,I}	**0.32**	0.23	0.04	**0.48**
R021E09C^{L,I}	**0.45**	0.24	0.02	0.20
R021E10C^{L,I}	0.27	0.23	0.17	0.07
R021E11M^{L,I}	0.00	0.01	0.06	**0.40**
R021E12C^{L,I}	**0.38**	0.17	**0.31**	0.22

Factor loadings greater or equal 0.30 are highlighted.

A, Acquire and Use Information; L, Literary Experience; R, Retrieving and Straightforward Inferencing; I, Interpreting, Integrating, and Evaluating.

7.2. OUTLOOK

We have discussed possible implications of the findings for criterion-referenced tests and large scale educational assessment. The assumptions of the classical factor models have been seen to be crucial in these application fields. We suggest, for instance, that the presented classical procedures should not be used, unless with special caution if at all, to examine the factorial structure of dichotomously scored criterion-referenced tests. Instead, if model violations of the "sensitive" type are present, better suited or more sophisticated latent variable models can be used (see Skrondal and Rabe-Hesketh, 2004). Examples are item response theory parametric or non-parametric models for categorical response data (e.g., van der Linden and Hambleton, 1997). Furthermore, we would like to mention item response based factor analysis approaches by Bock and Lieberman (1970) or Christoffersson (1975, 1977). We may also pay attention to tetrachoric or polychoric based structural equation models by Muthén (1978, 1983, 1984) and Muthén and Christoffersson (1981).

Table 7 | Loading matrix for four factors principal axis analysis of the PIRLS 2006 data for test booklet number 4, German sample.

Item	Factor			
	1	2	3	4
R011A01C^{A,R}	0.15	0.26	**0.40**	−0.06
R011A02M^{A,R}	0.14	0.29	**0.33**	0.18
R011A03C^{A,R}	0.16	0.24	0.09	0.02
R011A04C^{A,I}	**0.38**	0.20	0.10	0.06
R011A05M^{A,R}	0.22	0.07	0.19	0.24
R011A06M^{A,R}	0.19	0.02	0.14	0.07
R011A07C^{A,R}	**0.50**	0.20	0.22	0.16
R011A08C^{A,R}	**0.36**	0.03	**0.38**	−0.08
R011A09C^{A,I}	**0.54**	0.19	0.12	0.00
R011A10M^{A,I}	0.28	0.27	0.22	0.11
R011A11C^{A,I}	**0.38**	0.07	0.13	0.02
R021E01M^{L,R}	0.07	**0.45**	0.19	−0.06
R021E02M^{L,R}	0.03	**0.49**	0.09	0.24
R021E03M^{L,R}	0.14	**0.33**	−0.02	0.02
R021E04M^{L,R}	0.17	0.26	0.16	0.04
R021E05C^{L,R}	0.21	0.23	**0.32**	0.12
R021E06M^{L,R}	0.17	**0.44**	0.08	0.27
R021E07C^{L,I}	0.13	0.06	**0.47**	0.23
R021E08M^{L,I}	**0.32**	0.24	0.05	**0.46**
R021E09C^{L,I}	**0.45**	0.24	0.02	0.19
R021E10C^{L,I}	0.27	0.24	0.17	0.06
R021E11M^{L,I}	0.00	0.02	0.05	**0.40**
R021E12C^{L,I}	**0.38**	0.17	**0.30**	0.22

Factor loadings greater or equal 0.30 are highlighted.

A, Acquire and Use Information; L, Literary Experience; R, Retrieving and Straightforward Inferencing; I: Interpreting, Integrating, and Evaluating.

As with factor analysis a general problem (e.g., Maraun, 1996), we had to deal with the issue of rotational indeterminacy and of selecting a specific rotation. We have decided to use varimax rotation, due to the fact that this rotation is most frequently used in empirical educational studies (for better interpretability of the factors). Future research may cover other rotations (e.g., quartimax or equimax) or the evaluation of parameter estimation by examining the communality estimates for each item (which are not dependent on rotation, but are a function of the factor loadings). Moreover, the orthogonal factor model may not be realistic, as factors are correlated in general. However, in the current study, it may be unlikely that having non-zero population factor loadings for correlated dimensions would substantially affect the findings. In further research, we will have to study the case of the oblique (non-orthogonal) factor model.

The results of this paper provide implications for popular research practices in the empirical educational research field. The methods that we have utilized are traditional and often applied in practice (e.g., by educational scientists), for instance to determine the factorial validity of criterion-referenced tests or to study large scale assessment measurement instruments. In addition, to consider other, more sophisticated fit statistics can be interesting and valuable. For example, such model fit statistics as the root mean square residual, comparative fit index, or the root mean squared error of approximation may be investigated. Albeit these fit statistics are well-known and applied in the confirmatory factor analysis (CFA) context, they could be produced for exploratory factor analysis (given that CFA and EFA are based on the same common factor model).

We conclude with important research questions related to the PISA study. In the context of PISA, principal component analysis is used, in the purely computational sense. Other distributional, inferential, or confirmatory factor models, especially those for the verification of the factorial validity of the PISA context questionnaires, have not been considered. Interesting questions arise: are there other approaches to dimensionality reduction that can perform at least as well as the principal component analysis method in PISA data (e.g., multidimensional scaling; Borg and Groenen, 2005)? Is the 95% extraction rule in principal component analysis of PISA data an "optimal" criterion? How sensitive are PISA results if, for example, the parallel analysis method is used as the extraction criterion? Answering these and other related questions is out of the scope of the present paper and can be pursued in more in-depth future analyses. Nonetheless, the important role of these problems in the PISA context is worth mentioning. The PISA procedure uses not only manifest background information but also principal component scores on complex constructs in order to assign literacy or plausible values to students. Future research is necessary to investigate the effects and possible implications of potentially biased estimates of latent or complex background information on students' assigned literacy values, and especially, their competence levels, based on which the PISA rankings are reported.

ACKNOWLEDGMENTS

The authors wish to thank Sabine Felbinger for her critical reading and helpful comments. In particular, we are deeply indebted to Jason W. Osborne, Chief Editor, and four reviewers. Their critical and valuable comments and suggestions have improved the manuscript greatly.

REFERENCES

Adams, R., Wilson, M., and Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Appl. Psychol. Meas.* 21, 1–23.

Bartholomew, D., Knott, M., and Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach.* Chichester: John Wiley & Sons.

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, D., and Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika* 35, 179–197.

Bolt, D. M. (2005). "Limited- and full-information estimation of item response theory models," in *Contemporary Psychometrics*, eds A. Maydeu-Olivares and J. J. McArdle (Mahwah, NJ: Lawrence Erlbaum Associates), 27–72.

Borg, I., and Groenen, P. J. F. (2005). *Modern Multidimensional Scaling: Theory and Applications.* Berlin: Springer.

Browne, M. (1974). Generalized least squares estimators in the analysis of covariance structures. *South Afr. Stat. J.* 8, 1–24.

Burt, C. (1909). Experimental tests of general intelligence. *Br. J. Psychol.* 3, 94–177.

Carroll, J. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika* 10, 1–19.

Cattell, R. (1966). The scree test for the number of factors. *Multivariate Behav. Res.* 1, 245–276.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika* 40, 5–32.

Christoffersson, A. (1977). Two-step weighted least squares factor analysis of dichotomized variables. *Psychometrika* 42, 433–438.

Collins, L., Cliff, N., McCormick, D., and Zatkin, J. (1986). Factor recovery in binary data sets: a simulation. *Multivariate Behav. Res.* 21, 377–391.

Cronbach, L., and Meehl, P. (1955). Construct validity in psychological tests. *Psychol. Bull.* 52, 281–302.

Cudeck, R., and MacCallum, R. (eds). (2007). *Factor Analysis at 100: Historical Developments and Future Directions.* Mahwah, NJ: Lawrence Erlbaum Associates.

Dolan, C. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: a comparison of categorical variable estimators using simulated data. *Br. J. Math. Stat. Psychol.* 47, 309–326.

Ferguson, E., and Cox, T. (1993). Exploratory factor analysis: a users' guide. *Int. J. Sel. Assess.* 1, 84–94.

Ferguson, G. (1941). The factorial interpretation of test difficulty. *Psychometrika* 6, 323–329.

Flora, D., LaBrish, C., and Palmers, R. (2012). Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. *Front. Psychol.* 3:55. doi:10.3389/fpsyg.2012.00055

Gorsuch, R. (1997). Exploratory factor analysis: its role in item analysis. *J. Pers. Assess.* 68, 532–560.

Green, S. (1983). Identifiability of spurious factors using linear factor analysis with binary items. *Appl. Psychol. Meas.* 7, 139–147.

Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika* 19, 149–161.

Harman, H. (1976). *Modern Factor Analysis.* Chicago, IL: University of Chicago Press.

Horn, J. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika* 30, 179–185.

Hotelling, H. (1933a). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* 24, 417–441.

Hotelling, H. (1933b). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* 24, 498–520.

IEA. (2007). *PIRLS 2006 Assessment.* Boston, MA: TIMSS & PIRLS International Study Center.

Jöreskog, K. G. (1966). Testing a simple structure hypothesis in factor analysis. *Psychometrika* 31, 165–178.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika* 32, 443–482.

Kaiser, H. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 23, 187–200.

Kaiser, H., and Dickman, K. (1959). Analytic determination of common factors. *Am. Psychol.* 14, 425.

Kasper, D. (2012). *Klassische und Dichotome Faktorenanalyse auf dem Prüfstand: Zum Einfluss der Fähigkeits- und Aufgabenparameter auf die Schätzgenauigkeit verschiedener Faktorenmodelle [Classical and Dichotomous Factor Analysis put to Test: On the Impact of the Ability and Item Parameters on the Estimation Precision of Various Factor Models].* Münster: Waxmann.

Kelley, T. L. (1935). *Essential Traits of Mental Life.* Cambridge, MA: Harvard University Press.

Klauer, K. (1987). *Kriteriumsorientierte Tests: Lehrbuch der Theorie und Praxis lehrzielorientierten Messens [Criterion-Referenced Tests: Textbook on Theory and Practice of Teaching Goal Oriented Measuring].* Göttingen: Hogrefe.

Lienert, G., and Raatz, U. (1998). *Testaufbau und Testanalyse [Test Construction and Test Analysis].* Weinheim: Psychologie Verlags Union.

MacCallum, R. (2009). "Factor analysis," in *The SAGE Handbook of Quantitative Methods in Psychology*, eds R. E. Millsap and A. Maydeu-Olivares (London: Sage Publications), 123–147.

Maraun, M. D. (1996). Metaphor taken as math: indeterminacy in the factor analysis model. *Multivariate Behav. Res.* 31, 517–538.

Martin, M., Mullis, I., and Kennedy, A. (2007). *PIRLS 2006 Technical Report.* Boston, MA: TIMSS & PIRLS International Study Center.

Mattson, S. (1997). How to generate non-normal data for simulation of structural equation models. *Multivariate Behav. Res.* 32, 355–373.

Maydeu-Olivares, A. (2005). "Linear item response theory, nonlinear item response theory and factor analysis: a unified framework," in *Contemporary Psychometrics*, eds A. Maydeu-Olivares and J. J. McArdle (Mahwah, NJ: Lawrence Erlbaum Associates), 73–102.

McDonald, R. (1985). *Factor Analysis and Related Methods.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Mulaik, S. (2009). *Foundations of Factor Analysis.* Boca Raton, FL: Chapman & Hall/CRC.

Mullis, I. V., Kennedy, A. M., Martin, M. O., and Sainsbury, M. (2006). *PIRLS 2006 Assessment Framework and Specifications.* Boston, MA: TIMSS & PIRLS International Study Center.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika* 43, 551–560.

Muthén, B. (1983). Latent variable structural equation modeling with categorical data. *J. Econom.* 22, 43–65.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* 49, 115–132.

Muthén, B. (1989). Dichotomous factor analysis of symptom data. *Sociol. Methods Res.* 18, 19–65.

Muthén, B., and Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika* 46, 407–419.

Muthén, B., and Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: a note on the size of the model. *Br. J. Math. Stat. Psychol.* 45, 19–30.

OECD. (2005). *PISA 2003 Technical Report.* Paris: OECD Publishing.

OECD. (2012). *PISA 2009 Technical Report.* Paris: OECD Publishing.

Pearson, K. (1901). On lines and planes of closest fit to a system of points in space. *Philos. Mag.* 2, 559–572.

R Development Core Team. (2011). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

Ramberg, J. S., Tadikamalla, P. R., Dudewicz, E. J., and Mykytka, E. F. (1979). A probability distribution and its uses in fitting data. *Technometrics* 21, 201–214.

Reinartz, W. J., Echambadi, R., and Chin, W. W. (2002). Generating non-normal data for simulation of structural equation models using Mattson's method. *Multivariate Behav. Res.* 37, 227–244.

Roznowski, M., Tucker, L., and Humphreys, L. (1991). Three approaches to determining the dimensionality of binary items. *Appl. Psychol. Meas.* 15, 109–127.

Seier, E. (2002). Comparison of tests for univariate normality. Retrieved from: http://interstat.statjournals.net/YEAR/2002/articles/0201001.pdf [March 4, 2013].

Shapiro, S., and Wilk, M. (1965). An analysis of variance test for normality (complete samples). *Biometrika* 52, 591–611.

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

Spearman, C. (1904). "General intelligence," objectively determined and measured. *Am. J. Psychol.* 15, 201–292.

Sturzbecher, D., Kasper, D., Bönninger, J., and Rüdel, M. (2008). *Evaluation der theoretischen Fahrerlaubnisprüfung: Methodische Konzeption und Ergebnisse des Revisionsprojekts [Evaluation of the Theoretical Driving License Test: Methodological Concepts and Results of the Revision Project].* Dresden: TÜV|DEKRA arge tp 21.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychol. Rev.* 38, 406–427.

Thurstone, L. L. (1965). *Multiple Factor Analysis.* Chicago, IL: University of Chicago Press.

van der Linden, W., and Hambleton, R. (eds). (1997). *Handbook of Modern Item Response Theory.* New York: Springer.

Velicer, W., and Jackson, D. (1990). Component analysis versus common factor analysis: some issues in selecting an appropriate procedure. *Multivariate Behav. Res.* 25, 1–28.

Wang, C. (2001). *Effect of Number of Response Categories and Scale Distribution on Factor Analysis.* Master's thesis, National Taiwan University, Taipei.

Weng, L., and Cheng, C. (2005). Parallel analysis with unidimensional binary data. *Educ. Psychol. Meas.* 65, 697–716.

Widaman, K. F. (2007). "Common factors versus components: principals and principles, errors and misconceptions," in *Factor Analysis at 100: Historical Developments and Future Directions*, eds R. Cudeck and R. C. MacCallum (Mahwah, NJ: Lawrence Erlbaum Associates), 177–204.

Wirth, R., and Edwards, M. (2007). Item factor analysis: current approaches and future directions. *Psychol. Methods* 12, 58–79.

Conflict of Interest Statement: 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.

Received: 10 June 2012; accepted: 17 February 2013; published online: 27 March 2013.

Citation: Kasper D and Ünlü A (2013) On the relevance of assumptions associated with classical factor analytic approaches. Front. Psychol. ***4****:109. doi:10.3389/fpsyg.2013.00109*

This article was submitted to Frontiers in Quantitative Psychology and Measurement, a specialty of Frontiers in Psychology.