New Perspectives in the Adaptive Assessment of Depression: The ATS-PD Version of the QuEDS

Andrea Spoto[1]*, Francesca Serra[1], Ivan Donadello[2], Umberto Granziol[1] and Giulio Vidotto[1]

[1] Quantitative Psychology Laboratory, Department of General Psychology, University of Padua, Padua, Italy, [2] Fondazione Bruno Kessler, Trento, Italy

Measurement is a crucial issue in psychological assessment. In this paper a contribution to this task is provided by means of the implementation of an adaptive algorithm for the assessment of depression. More specifically, the Adaptive Testing System for Psychological Disorders (ATS-PD) version of the Qualitative-Quantitative Evaluation of Depressive Symptomatology questionnaire (QuEDS) is introduced. Such implementation refers to the theoretical background of Formal Psychological Assessment (FPA) with respect to both its deterministic and probabilistic issues. Three models (one for each sub-scale of the QuEDS) are fitted on a sample of 383 individuals. The obtained estimates are then used to calibrate the adaptive procedure whose performance is tested in terms of both efficiency and accuracy by means of a simulation study. Results indicate that the ATS-PD version of the QuEDS allows for both obtaining an accurate description of the patient in terms of symptomatology, and reducing the number of items asked by 40%. Further developments of the adaptive procedure are then discussed.

Keywords: adaptive psychological assessment, formal psychological assessment, depression, qualitative and quantitative assessment, item response theory (IRT)

1. INTRODUCTION

Measurement in Psychology is a challenging issue that rose since the very beginning of the history of Psychology as a science. The first formalization of measurement in Psychology is due to the empirical research of Weber and to the Psychophysics of Fechner (1860), while Spearman (1904) paved the way for the measurement of theoretical constructs through methodologies such as factor analysis. The lack of a consistent formulation of the measurement problem in psychology was first addressed by Stevens by means of direct methods for psychological measurement (Stevens, 1946, 1951, 1957). An axiomatic definition of the measurement scales appeared within the theoretical framework of the Relational Theory of Measurement (RTM; Suppes and Zinnes, 1963; Suppes et al., 1989; Narens and Luce, 1993).

Currently, in psychological measurement the classical test theory (CTT; Spearman, 1904; Novick, 1965; Gulliksen, 2013) and the item response theory (IRT; Rasch, 1960; Lord, 1980) are the formal and methodological frameworks for the construction of measurement tools. The classical test theory relies on the evaluation of the reliability, validity and factorial structure of a defined psychological measure. An important limit lies in the impossibility to distinguish and compare the parameters related to the individuals (abilities) and those relative to the items (difficulties). On the other hand, the item response theory and the Rasch model (Rasch, 1960) explain the test

performance of individuals by referring to the presence of latent traits. A relation between the latent traits and the observed scores is postulated, so that information on the first ones are inferred starting from the observable performance of the individual in answering the set of items. The relationship between test score and latent trait is expressed by a mathematical model, defined *a priori*. IRT models are a family of mathematical models which describe a wide number of contexts starting from the simple logistic model. The interest in applying these models is growing thanks to the numerous advantages compared to the classical test methods. IRT models represent a fundamental formal tool for applying adaptive measurement in psychology since they allow the definition of precedence relations among items according to their location on the latent trait dimension (Marsman et al., 2018).

In recent years another approach to measurement has been adopted in psychological assessment. This approach allows for expanding the measurement properties obtained through IRT by considering complex precedence relations among items (i.e., beyond the linear order). It refers to an axiomatic formulation of the relations among sets of items and sets of attributes investigated by them. It refers to the possibility of depicting the relations among items according to well established mathematical tools such as *lattices* and *posets* (Birkhoff, 1937, 1940; Davey and Priestley, 2002). It is the Formal Psychological Assessment (FPA; Spoto, 2011; Spoto et al., 2013a; Bottesi et al., 2015; Serra et al., 2015, 2017). It was developed with the aim of providing detailed information about the clinical features endorsed by a patient who answered a specific set of items of a questionnaire.

In the present article, this last methodology is employed to implement an adaptive algorithm for the assessment of depression. More specifically, the main aim of this article is to present the adaptive form of the Qualitative-Quantitative Evaluation of Depressive Symptomatology questionnaire (QuEDS; Serra et al., 2017).

The paper is structured as follows: In section 2, the general concepts concerning recent developments in psychological assessment are addressed; In section 3 a brief outline of the main deterministic and probabilistic issues related to the FPA is presented; section 4 introduces the main concepts of the adaptive algorithm implemented in this study; In sections 5 and 6 a simulation study aimed at testing the accuracy and the efficiency of the adaptive procedure is presented; Final remarks, limitations and future perspectives are explored in section 7.

2. PSYCHOLOGICAL ASSESSMENT: STATE OF THE ART

Measurement is a crucial issue in many fields of psychology. One of them is psychological assessment. The main tools adopted for carrying out measurement in psychological assessment are self-report questionnaires, observation and interviews. Clinical interviews and observation have the capacity for gathering and deepening several information such as nonverbal aspects, which are essential to make a diagnosis (Annen et al., 2012; Fiquer

et al., 2013; Girard et al., 2014). This last is the main aim of clinical assessment, and therefore, the main goal of measurement in clinical psychology. For example, negative emotions and social behaviors are indicators of the severity of depression and relevant predictors of its clinical remission (Philippot et al., 2003; Uhlmann et al., 2012) that are beyond the control and awareness of the patient (Andersen, 1999; Geerts and Brüne, 2009). Nonetheless, both observation and clinical interview are time consuming and prone to inferential errors by clinicians (Strull et al., 1984; Nordgaard et al., 2013).

The self-report questionnaires provide scores that are supposed to indicate the severity of the symptomatology and the impairment level (Groth-Marnat, 2009). The score of a questionnaire is helpful in distinguishing individuals with critical clinical features, but it is not sufficient, in the form so far provided by both CTT and IRT, to differentiate patients with different symptom configurations who obtained similar scores (in the limit, the same score) to the test (Spoto et al., 2013a; Bottesi et al., 2015; Serra et al., 2015, 2017). Moreover, not all the items have the same "weight" from the clinical point of view, since they reflect different symptoms that may be more or less severe (Gibbons et al., 1985; Serra et al., 2017).

Therefore, measurement in clinical psychology should attempt to evaluate individual data in a broad perspective and it should account for individual specific features (Meyer et al., 2001). For example, the construct of depression, as represented by the score, can sometimes be misleading. Indeed, depression can manifest with a variety of different symptoms that may be due to a different culture or a different etiology (Benazzi et al., 2002; Goodwin and Jamison, 2007). Only through personalized assessment it is possible to clearly distinguish among such different manifestations of the disorder (Groth-Marnat, 2009; Serra et al., 2017). For these reasons, an increasing number of self-report assessment tools are validated according to the IRT framework (Gibbons et al., 2008; Embretson and Reise, 2013). In this way the assessment can be much more focused on the objective measure of the uniqueness of a particular clinical configuration. For instance, Balsamo et al. (2014) applied Rasch analysis to the item selection for the Teate Depression Inventory, a self-report depression tool; it has been highlighted by a number of papers that this tool, built according to the IRT methodology, performs better than tools developed within the CTT with respect to many different measurement properties such as, for instance, convergent-divergent validity (e.g., Innamorati et al., 2013; Balsamo et al., 2015a,b).

As mentioned above, IRT is also a crucial stepping stone for implementing adaptive testing, which in turn is an important way to implement the administration of a questionnaire in a personalized fashion. Each individual is administered with different scale items on the basis of the specific answers he/she provided to the previous ones (Wainer, 2000; Fliege et al., 2005). Within this field, Computerized Adaptive Testing (CAT; Wainer, 2000) is an approach which presents multiple advantages. Different studies showed that questionnaires could be shortened without loss of information by means of CAT, achieving a more efficient and equally accurate assessment (Petersen et al., 2006). CAT's procedure mimics the semi-structured interview

(i.e., clinical interview where only some items, out a list, are posed according to specific adaptive selection criteria), letting the algorithm to carry out inferences by accounting for all the information collected and following a logically correct process (Spoto, 2011). It should be quite easy to understand why the assessment of knowledge is one of the core areas in which such procedures have been developed. For instance, Eggen and Straetmans (2000) combined IRT with statistical procedures, like sequential probability ratio test and weighted maximum likelihood, for classifying people under exam. Other systems use Bayesian statistical techniques instead of IRT in the evaluation of students' knowledge. Examples are EDUFORM (Nokelainen et al., 2001), and PARES (Marinagi et al., 2007). In the field of knowledge assessment the ALEKS (Assessment and LEarning in Knowledge Spaces) system implements the theoretical framework of Knowledge Space Theory (KST; Doignon and Falmagne, 1985, 1999; Falmagne and Doignon, 2011) for the adaptive assessment of the so called *knowledge state* of a student, i.e., the set of items that he/she is able to solve about a specific topic (Grayce, 2013).

The formulation of an adaptive algorithm is clearly more difficult in the clinical setting. In fact, the objectivity of the questions and therefore of the answers given by the subject is much more questionable, and the probability of misinterpretations in the answers is increased. Despite this, research has demonstrated that both IRT and CAT (Baek, 1997) can be applied to the measurement of attitudes and personality variables (Reise and Waller, 1990). In the clinical framework, Spiegel and Nenh (2004) developed an expert system, which calculates possible symptom combinations given the answers of a patient and returns all possible diagnoses coherent with such combination. Yong et al. (2007) developed an interactive self-help system for depression diagnosis that provides advice about patients levels of impairment. Simms et al. (2011) developed the CAT for Personality Disorders (CAT-PD) aimed at realizing a computerized adaptive assessment system. CAT has been applied also in developing adaptive classification tests by means of stochastic curtailment using CES-D for depression (Finkelman et al., 2012; Smits et al., 2016). Gibbons et al. (2008) used the combination of item response theory and CAT in mood and anxiety disorder assessment. In particular, they applied a bi-factor structure consisting of a primary dimension and four sub-factors (mood, panic-agoraphobia, obsessive-compulsive, and social phobia) to build the CAT version of the Mood and Anxiety Spectrum Scales (MASS; Dell'Osso et al., 2002). Results of this study showed that the adaptive tool allowed to both administer a small set of the items (the most relevant for a given individual) with no loss of information compared to the classical form of the MASS, and strongly reduce time consumption as well as patient and clinician burden. In six patients with mood disorders (three major depressive disorder and three bipolar disorder) who were interviewed by the psychiatrist, many of the CAT items investigating important information, such as a history of manic symptoms, potentially risky behaviors, etc., were both endorsed and not documented in the psychiatric evaluation through SCID-I (First et al., 1996). Gibbons' study is an important example of how adaptive testing can be effective and efficient.

Although there have been several attempts to apply adaptive clinical assessment, as far as we know, no system is able to combine adaptivity, quantitative and qualitative information, and punctual estimates of error parameters. Within clinical psychology, the Formal Psychological Assessment (Spoto et al., 2010, 2013a; Serra et al., 2015; Granziol et al., 2018) represents an important contribution in the improvement of adaptive psychological assessment, allowing to overcome the obstacles encountered up to now in this field. The main deterministic and probabilistic concepts of this methodology are presented in the next section.

3. THE FORMAL PSYCHOLOGICAL ASSESSMENT

An adaptive testing, most of the time, relies on a formal substratum composed by several relations among items (Donadello et al., 2017). The FPA is a methodology which allows defining assessment tools able to detect specific symptoms in several mental disorders, independently of the kind of assessment used, such as self-report (Serra et al., 2015, 2017) or behavioral observations (Granziol et al., 2018). FPA makes it possible applying two theories of Mathematical Psychology in psychological assessment: The Knowledge Space Theory (KST; Doignon and Falmagne, 1999; Falmagne and Doignon, 2011) and the Formal Concept Analysis (FCA; Wille, 1982; Ganter and Wille, 1999). The core characteristic of FPA is the definition of a relation between a set of items and a set of clinical criteria. In the next two subsections are reported separately the deterministic and probabilistic main concepts of FPA.

3.1. FPA Deterministic Concepts

In FPA methodology, a very basic concept is the *clinical domain*, intended as a nonempty set Q of questions that can be asked to a patient for investigating a certain psychopathology. Each item is referred as an *object*. The complete list of the items included in the QuEDS, namely the objects in the present article, are listed in **Table 1**.

For instance, the item QuEDS34 "I feel sad" is an object. For the sake of simplicity, the term item will be preferred to object in the sequel. The subset $K \subseteq Q$ of all the items that are endorsed by a patient is called the *clinical state* of that patient. Each item investigates one or more *attributes*, intended as a diagnostic criteria of a psychopathology selected from either clinical sources like the DSM-5 (American Psychiatric Association, 2013) or the scientific literature, or both. The complete list of the attributes investigated by the QuEDS is displayed in **Table 2**.

For instance, the diagnostic criterion "Depressed mood" of the DSM-5 is investigated by the aforementioned item. The collections of both items and attributes make it possible to build the so called *clinical context*, formally a triple (Q, M, I), where Q is the set of items, M is a set of attributes and I is a binary relation between the sets Q and M which assigns to each item $q \in Q$ the attributes $m \in M$ it investigates. The clinical context can be represented as a Boolean Matrix, having the items in the rows and the attributes in the columns: whenever an item q investigates an

TABLE 1 | The list of items of the QuEDS grouped by sub-scale.

Item	Description
	Cognitive scale
QuEDS5	I am stressed by feelings of guilt
QuEDS6	I think the world is cruel and unhappy
QuEDS9	I think my life is hell and I only deserve to feel bad
QuEDS10	I feel incapable of facing life's events
QuEDS14	I feel incapable and totally useless
QuEDS19	I have thought about killing myself
QuEDS20	Sometimes I think it would be better if I were dead
QuEDS21	I am really worried about my health
QuEDS24	I am afraid about everything that will happen to me because I am incapable
QuEDS25	I feel like I don't have any more power over my empty and sad life
QuEDS27	Making choices is hard for me
QuEDS30	I feel I am a burden other people and it would be better if I killed myself
QuEDS32	I am disappointed in myself and the choices I have made
QuEDS33	I have problems making decisions
QuEDS41	I often feel like a loser
	Somatic scale
QuEDS1	I feel like I don't have the same energy to have sex
QuEDS2	I often wake up in the middle of the night and I can't go back to sleep
QuEDS3	I feel like that my thinking is slowing down
QuEDS4	I have sleep problems
QuEDS11	I suffer from somatic disorders (e.g., headache, stomach ache)
QuEDS13	I am less interested in sex
QuEDS16	My desire to eat is not the same
QuEDS22	My weight has had significant changes
QuEDS23	I've visibly lost (or gained) weight
QuEDS26	My appetite has changed
QuEDS28	I feel I'm slowing down in my daily routines
QuEDS31	I don't have much energy and I feel tired
QuEDS35	My ability to think and memorize has been reduced
QuEDS39	I feel so tired and without any energy that I need help to look after myself
	Affective scale
QuEDS7	I keep crying very easily
QuEDS8	I get irritated very easily
QuEDS12	I have lost interest in the future which doesn't have anything good for me
QuEDS15	I see the same unhappiness I have now in the future
QuEDS17	I often feel like crying, but I cannot do it
QuEDS18	I can't have interest and pleasure in people and things as before
QuEDS29	I feel helpless and inhibited facing my incapacity to concentrate
QuEDS34	I feel sad
QuEDS36	I don't have any interest and desire to do anything
QuEDS37	I am agitated about the idea that this sadness won't ever leave me
QuEDS38	I feel agitated
QuEDS40	I am better in the evening more than in the morning

TABLE 2 | The list of the attributes investigated by the items of the QuEDS.

Attribute	Description
A1	Depressed mood
A2	Diminished interest and pleasure
A3	Decreased interest in sex
A4	Increase or loss of weight
A5	Gain or loss of appetite
A6	Insomnia or hypersomnia
A7	Agitation
A8	Psycho-motor retardation
A9	Fatigue or energy loss
A10	Feelings of worthlessness (or Beck's negative view of self)
A11	Feelings of guilt
A12	Diminished ability to think and concentrate
A13	Indecision
A14	Recurrent thoughts of death
A15	Suicidal ideation or attempted suicide
A16	Beck's negative view of the world
A17	Beck's negative expectation of the future
A18	Seligman's learned helplessness
A19	Irritability
A20	Apathy
A21	Health concern
A22	Somatic disorders
A23	More positive mood in the evening

attribute m (i.e., whenever the relation qIm holds true), the qm cell contains the value 1, otherwise 0.

Starting from the clinical context, the *clinical structure* $\mathcal{K}$ can be delineated (Spoto et al., 2010, 2016). A clinical structure is a collection of clinical states, containing at least the empty set (∅) and the whole clinical domain (Q) and it represents the implications among the items of the domain. Whenever a clinical structure is closed under set union (i.e., $K_1 \cup K_2 \in \mathcal{K}$ for all $K_1, K_2 \in \mathcal{K}$) it is a *clinical space*. On the other hand, a clinical structure closed under intersection (i.e., $K_1 \cap K_2 \in \mathcal{K}$ for all $K_1, K_2 \in \mathcal{K}$), it is a *clinical closure space*. In order to obtain a structure where all the states K are in a one to one connection with the set of attributes endorsed by all the items in K, it is necessary to modify the clinical context by defining a relation R between items and attributes, which is dual to I:

$$qRm \Longleftrightarrow q\neg Im.$$

According to this relation, a clinical closure space is obtained, where each state is in a one to one correspondence with the set of attributes endorsed by the items in each state (for details refer to Spoto et al., 2010). In other words the relation R allows for representing in the structure a principle often used in clinical and medical practices: if a patient endorses an item, he should present all the attributes investigated by that item. From a practical point of view, a clinical structure is useful since it includes only the clinical states, that is all and only the admissible response patterns given the clinical context. Any state $K \in \mathcal{K}$

is coherent with the theoretical framework and, therefore, it does not violate specific order relation among items. In FPA, the relation required is based on the attributes investigated by each item and it is called *prerequisite relation*, stating that whenever an item q investigates a subset of attributes of another item r, q is a prerequisite for r. For instance, taken the subset of QuEDS's selected attributes composed by {A1, A17} and two items of the QuEDS, namely QuEDS34 which investigates only A1, and QuEDS15, investigating both A1 and A17. The two rows of the clinical context representing QuEDS15 and QuEDS34 will be in an inclusion relation with respect to the attributes investigated (**Table 3**).

Among the following response patterns:

(a) {∅};
(b) {*QuEDS*34};
(c) {*QuEDS*15};
(d) {*QuEDS*34, *QuEDS*15},

only the patterns a, b, and d are clinical states. In fact, it is not possible (excluding errors) that a patient who endorses item QuEDS15 does not endorse item QuEDS34 (i.e., the pattern c). A clinical structure can be represented as a complete lattice displaying the partial order among the items of a domain, where each node contains a subset of items (investigating a specific subset of attributes; Granziol et al., 2018). In the case at hand, the clinical structure is a clinical closure space where each node contains the items endorsed by the patient and the uniquely determined set of attributes (symptoms) corresponding to that set of questions.

By delineating a clinical structure, it is possible to obtain the deterministic skeleton for defining a computerized adaptive assessment, which needs to be completed also from a probabilistic point of view. The next section will deepen the probabilistic features needed to implement the algorithm.

3.2. FPA Probabilistic Concepts

A deterministic clinical structure provides a fundamental starting point for the procedure aimed at creating an adaptive assessment. Nonetheless, such a structure is incomplete from both a theoretical and practical point of view. In fact, each state could be present with different frequencies in the population; moreover, the observed response pattern of a subject could not represent his/her real state. The probability of observing each clinical state π_K is then related to both its actual frequency in the population, and to two further parameters, namely the *false negative* (β) and the *false positive* (η). The false negative refers to the probability that the patient does not endorse an item that he/she actually presents. The false positive parameter, on the contrary, refers

to the probability that a patient endorses an item that he/she does not present. By means of all the aforementioned parameters (i.e., π_K for each $K \in \mathcal{K}$; β_q and η_q for each item $q \in Q$) a *probabilistic clinical structure* (Donadello et al., 2017) can be obtained. Formally, it is a triple $(Q, \mathcal{K}, \pi)$ where $(Q, \mathcal{K})$ is the clinical structure and π is the probability distribution on $\mathcal{K}$ estimated through a sample of patients (Spoto et al., 2010). The probability distribution for each response pattern $R \subseteq Q$ is obtained by means of a response function assigning to R its conditional probability given a state K (for all states $K \in \mathcal{K}$), as displayed by the unrestricted latent class model represented by Equation 1:

$$P(R) = \sum_{K \in \mathcal{K}} P(R|K)\pi(K). \tag{1}$$

This model is the so called *basic local independence model* (BLIM; Falmagne and Doignon, 1988; Doignon and Falmagne, 1999). Within the probabilistic clinical structure, the responses to the items are assumed to be locally independent. The conditional probability $P(R|K)$ is determined by the probability of false negative (β_q) and false positive (η_q) while answering to q, as displayed by Equation (2):

$$P(R|K) = \left(\prod_{q \in K \setminus R} \beta_q \right) \left(\prod_{q \in K \cap R} (1 - \beta_q) \right) \left(\prod_{q \in R \setminus K} \eta_q \right) \left(\prod_{q \in \overline{R \cup K}} (1 - \eta_q) \right). \tag{2}$$

In the present study, the expectation-maximization algorithm (Dempster et al., 1977) has been used in order to estimate both the β and η parameters and the probability distribution for $\mathcal{K}$. These estimates have been carried out on a sample of 383 individuals, according to the same procedure employed in, e.g., Spoto et al. (2013a), Bottesi et al. (2015), and Donadello et al. (2017). Such estimates were then used to implement the adaptive algorithm whose general functioning is detailed in the next section.

4. THE ADAPTIVE TESTING SYSTEM FOR PSYCHOLOGICAL DISORDERS ALGORITHM

In this section we aim at introducing the Adaptive Testing System for Psychological Disorders Donadello et al. (ATS-PD; 2017) developed starting from the clinical structure and the parameters' estimate via the BLIM.

Within this framework, the clinical structure is the deterministic skeleton defining the starting point for an adaptive assessment which, if no error is assumed, could reasonably proceed as follows:

i) Select the item that is closest to be present in 50% of the clinical states of the structure;
ii) Ask this item to the patient;
iii) Register the dichotomous answer;

TABLE 3 | The precedence relation between items QuEDS15 and QuEDS34 as depicted in the clinical context.

Items	A1	⋯	A17	⋯	A23
QuEDS15	1	⋯	1	⋯	0
QuEDS34	1	⋯	0	⋯	0

iv) Exclude all the states not containing the investigated item if the answer is "yes," or vice versa, all the states that contain the item if the answer is "no."

These steps are repeated on the remaining states until only one state remains. The output is the clinical state with all the attributes (diagnostic criteria) satisfied by all the items of the state. This procedure applied in a real context would almost surely fail due to the absence of a probabilistic model defining both the probabilities of the different states, and the error probabilities for the answers. As it has been shown in the previous section, the probabilistic model which accounts for both these issues could be the BLIM. Therefore, an adaptive procedure should make an appropriate use of the probabilities of the states (π_K) and of the false positive (η) and false negative (β) rates for each item.

Thus, the above outlined procedure can be therefore modified as follows:

i) Detect and administer the item q which best splits into two equal parts the probability mass of the states (*questioning rule*), and register the dichotomous answer;

ii) Update the probability π_K of all the states $K \in \mathcal{K}$ according to the following *updating rule*:

- if an affirmative answer to q is observed: increase π_K for all $K \in \mathcal{K}$ which contain q, and decrease π_K for the remaining states;
- if a negative answer to q is observed: decrease π_K for all $K \in \mathcal{K}$ which contain q, and increase π_K for the remaining states.

iii) Repeat the previous steps until a given condition is satisfied (*stopping rule*).

The algorithm used in this research implements these three main steps. The questioning rule selects the item to ask, i.e., the item $q \in Q$ that is "maximally informative." This characteristic is satisfied by the item(s) for which the sum of π_K for all the states containing q best approaches 0.50. In other words, this item maximizes the obtainable information irrespectively of its observed answer (i.e., either "yes" or "no"). If many items are equally informative, one of them is chosen at random. We call $L_n(K)$ the probability of the state K at the step n. At each step of the procedure, the subject's response is collected by the system. Then, the updating rule is applied to obtain the likelihood $L_{n+1}(K)$ for all the states $K \in \mathcal{K}$. More precisely, let us denote an affirmative response with $r = 1$ and a negative one with $r = 0$. It is then possible to formalize the updating rule of the probability $L(K)$ for each $K \in \mathcal{K}$ as follows:

$$L_{n+1}(K) = \frac{\zeta^K L_n(K)}{\sum_{K' \in \mathcal{K}} \zeta^K L_n(K')} \quad (3)$$

where

$$\zeta_{q,r}^K = \begin{cases} \zeta_{q,1} & \text{if } q \in K, r = 1; \\ 1 & \text{if } q \notin K, r = 1; \\ 1 & \text{if } q \in K, r = 0; \\ \zeta_{q,0} & \text{if } q \notin K, r = 0. \end{cases} \quad (4)$$

In this formulation ζ is a parameter always greater than 1 that increases the likelihood and influences the efficiency of the adaptive assessment process. The higher ζ, the more reliable are considered the answers provided by the subject, and therefore, the more efficient (but potentially less accurate) the adaptive procedure. It has been observed by Falmagne and Doignon (2011) that ζ values less than 2 make the assessment redundant. On the contrary, fixing the ζ value to an excessively high number could affect algorithm accuracy. It has been proven that an adequate value of ζ could be 21 (Falmagne and Doignon, 2011). This value allows for an accurate and efficient detection of the state of individuals in several applications, e.g., ALEKS (Falmagne et al., 2013). An alternative way to estimate ζ is based on the η_q and β_q parameters of each item (see Falmagne and Doignon, 2011, p. 265). The estimate is carried out according to the following formulas:

$$\zeta_{q,1} = \frac{1 - \beta_q}{\eta_q} \qquad\qquad \zeta_{q,0} = \frac{1 - \eta_q}{\beta_q}$$

This rule is local since it takes into account both the η and β of the last item asked in order to update the probability of the states. According to this method the "weight" of each item in updating the probabilities is a function of its error rates. Namely, an item whose error rates are low (i.e., whose answer is more reliable) will produce a significant modification on the probability distribution of the states, while a less reliable item will have a weaker effect in updating of the probabilities of the states.

In order to further refine the updating of the states probabilities given the pattern observed at a specific step n of the adaptive assessment, a Bayesian rule can be introduced according to what described by Donadello et al. (2017):

$$P(K_i|R) = \frac{P(R|K_i)L_n(K_i)}{\sum_{j=1}^{|K|} P(R|K_j)L_n(K_j)} \quad (5)$$

Where $P(R|K_i)$ is obtained by Equation (2), and $L_n(K)$ is the estimated probability of the state K at the step n.

All these steps are replicated until a given stopping criterion is reached. In the present article, the stopping rule is satisfied whenever $L_n(K_q)$ is outside the interval $[0.20, 0.80]$ for all $q \in Q$. In this way the algorithm stops as soon as any possible item to be asked splits the probability mass in very unequal parts, indicating that it is almost surely either inside or outside the state. This choice is coherent with previous literature (Falmagne and Doignon, 2011, p. 362). When the stopping criterion is matched, the algorithm concludes the assessment and provides as output the response pattern R, the estimated state K (with its estimated probability) and the amount of time needed to complete the assessment.

In the next section, we will present a simulation study aimed at testing the algorithm under different conditions in order to identify the best performing configuration of the procedure. Before conducting such a simulation we estimated the parameters of the BLIM in order to provide the deterministic skeleton with probabilistic weights.

5. A SIMULATION STUDY

Testing the adaptive procedure with respect to both its accuracy and efficiency is a necessary operation in order to guarantee that (i) the information collected through the adaptive form of the questionnaire is reliable, and (ii) that the administration time of the questionnaire is actually reduced. In order to reach these goals, the traditional form of the QuEDS containing 41 dichotomous items ("Yes"/"No") grouped into three sub-scales (namely: Cognitive, Somatic and Affective) was administered to a sample of 383 individuals. Using the collected data, the parameters of the BLIM were estimated. Then, such values were passed to the adaptive procedure. Finally, the adaptive algorithm, under different conditions, simulated the administration of the test starting from the available 383 response patterns, and the results were analyzed. The details of all these passages are provided below.

5.1. Sample

The sample of 383 Italian individuals included a clinical group consisting of 38 subjects with Major Depressive Episode (who were diagnosed with either major depressive disorder or bipolar disorder). These patients were recruited by the Neurosciences Mental Health and Sensory Organ (NESMOS) Department of La Sapienza University, Rome. The psychiatrists of NESMOS Department evaluated the presence of Major Depressive Episode in participants of the clinical group by means of the clinical interview and the SCID-I. The diagnosis was then formulated according to the DSM-IV-TR nosology classification system. The exclusion criteria were mental retardation and psychotic traits in order to guarantee a correct interpretation of the meaning of QuEDS items. The 47% of participants were males and the remaining 53% were females. The majority of the participants had a high school diploma, and their age ranged between 21 and 69 years old ($M = 33.5$; $SD = 4.8$). The remaining 345 individuals were randomly selected from the general population and recruited in Padova (68% were females). The majority of participants had a high school diploma, and their age ranged between 19 and 58 years old ($M = 27.5$; $SD = 6.4$). Participants of the non-clinical group did not undergo a psychological assessment. Before the beginning of the administration of the test they were asked to indicate whether they were currently under pharmacological or psychotherapeutic treatment for MDE. The exclusion criterion in the non-clinical group was the presence of MDE (i.e., individuals under pharmacological or psychotherapeutic treatment for depression).

The study was conducted in accordance with the Declaration of Helsinki and the research protocol was approved by the Psychology Ethical Committee of the University of Padua. All participants entered the study of their own free will and provided their written informed consent before taking part. They were informed in detail about the aims of the study, the voluntary nature of their participation, and their right to withdraw from the study at any time and without being penalized in any way. Furthermore, participants were allowed for asking the restitution about their own score, providing authors with their own auto generated code, used during the administration phase.

5.2. Procedure

All participants completed informed consent and sociodemographic forms before answering the questionnaire items. All participants completed the written form of QuEDS according to the following instructions: "Please answer "Yes" or "No" to the following statements on the basis of how you felt in the last 15 days." No time limit was imposed. Clinical participants provided written, informed consent for potential research analysis and anonymous reporting of clinical findings in aggregate form, at clinical intake.

5.3. Parameters Estimate

As mentioned in the previous sections, the estimate of BLIM's parameters (i.e., π_K for each state, β_q and η_q for each item), as well as of the fit of the model, were performed with a specific version of the Expectation-Maximization Algorithm (Dempster et al., 1977). For the details of the algorithm, refer to Spoto (2011). The tested models were obtained from the formal contexts displayed in **Tables 4–6** according to the methodology described in Spoto et al. (2010).

The fit of each of the three models has been tested by Pearson's Chi-square. It is well established that for large data matrices (as those used in the present study) the asymptotic distribution of χ^2 is not reliable. Therefore, a p-value for the obtained χ^2 was calculated by parametric bootstrap with 5,000 replications. An important fit index is provided by the estimates of the error rates. In general, they are expected to be low, but it is crucial that for each item the following inequality holds: $\eta_q < 1 - \beta_q$. If this condition is not satisfied, then the assessment loses its meaning, since the probability of observing a false positive (η) on an item q would be greater than the probability of observing an actual affirmative answer to q. Spoto et al. (2012) established a specific connection between the characteristics of the context and the identifiability of the error parameters of the items. In this respect the value of the unidentifiable parameters (Spoto et al., 2013b; Stefanutti et al., 2018) was fixed to a constant corresponding to the maximum possible value of the parameter (Stefanutti et al., 2018) in order to both preserve the computability of the adaptive procedure and adopt the maximally conservative approach from a diagnostic point of view, preferring accuracy to efficiency.

This first step of the study provided the needed parameters for calibrating the adaptive procedure.

5.4. Simulation Design

Six different conditions for testing the adaptive algorithm were generated by manipulating the following two variables:

1. Estimate of ζ parameter (2 levels): (i) 21 in one case; (ii) estimated via the values of η and β in the second;
2. Implementation of Bayesian updating rule (3 levels): (i) on-line (at each step n of the adaptive procedure), (ii) off-line (when the stopping criterion is reached), (iii) absent (no Bayesian updating is applied).

The adaptive procedure was then run according to each of the six conditions described above in order to simulate the 383 response patterns collected in the previous part of the study. The task of

TABLE 4 | The clinical context for the Cognitive sub-scale of the QuEDS.

Items	A1	A10	A11	A12	A13	A14	A15	A16	A17	A18	A21
QuEDS5	0	0	1	0	0	0	0	0	0	0	0
QuEDS6	0	0	0	0	0	0	0	1	0	0	0
QuEDS9	1	1	1	0	0	0	0	1	0	0	0
QuEDS10	0	0	0	0	0	0	0	0	0	1	0
QuEDS14	0	1	1	0	0	0	0	0	0	1	0
QuEDS19	0	0	0	0	0	1	1	0	0	0	0
QuEDS20	0	0	0	0	0	1	0	0	0	0	0
QuEDS21	0	0	0	0	0	0	0	0	0	0	1
QuEDS24	0	0	0	0	0	0	0	0	1	1	0
QuEDS25	1	0	0	0	0	0	0	1	0	1	0
QuEDS27	0	0	0	1	1	0	0	0	0	0	0
QuEDS30	1	1	1	0	0	1	0	1	0	0	0
QuEDS32	0	1	1	0	0	0	0	0	0	0	0
QuEDS33	0	0	0	0	1	0	0	0	0	0	0
QuEDS41	0	1	0	0	0	0	0	0	0	0	0

TABLE 5 | The clinical context for the Somatic sub-scale of the QuEDS.

Items	A3	A4	A5	A6	A7	A8	A9	A12	A18	A20	A22
QuEDS1	1	0	0	0	0	0	1	0	0	0	0
QuEDS2	0	0	0	1	1	0	0	0	0	0	0
QuEDS3	0	0	0	0	0	1	0	1	0	0	0
QuEDS4	0	0	0	1	0	0	0	0	0	0	0
QuEDS11	0	0	0	0	0	0	0	0	0	0	1
QuEDS13	1	0	0	0	0	0	0	0	0	1	0
QuEDS16	0	0	1	0	0	0	0	0	0	0	0
QuEDS22	0	1	0	0	0	0	0	0	0	0	0
QuEDS23	0	1	0	0	0	0	0	0	0	0	0
QuEDS26	0	0	1	0	0	0	0	0	0	0	0
QuEDS28	0	0	0	0	0	1	1	0	0	0	0
QuEDS31	0	0	0	0	0	0	1	0	0	0	0
QuEDS35	0	0	0	0	0	1	1	1	0	0	0
QuEDS39	0	0	0	0	0	1	1	0	1	1	0

TABLE 6 | The clinical context for the Affective sub-scale of the QuEDS.

Items	A1	A2	A7	A12	A17	A19	A20	A23
QuEDS7	1	0	0	0	0	1	0	0
QuEDS8	0	0	0	0	0	1	0	0
QuEDS12	0	1	0	0	1	0	0	0
QuEDS15	1	0	0	0	1	0	0	0
QuEDS17	1	0	0	0	0	0	1	0
QuEDS18	0	1	0	0	0	0	1	0
QuEDS29	0	0	0	1	0	0	1	0
QuEDS34	1	0	0	0	0	0	0	0
QuEDS36	0	1	0	0	0	0	1	0
QuEDS37	1	0	1	0	0	0	0	0
QuEDS38	0	0	1	0	0	0	0	0
QuEDS40	0	0	0	0	0	0	0	1

the algorithm was to administer the QuEDS in an adaptive form and provide the clinical state as output.

In order to check for the accuracy and the efficiency of the procedure, a number of indexes were used. First, the average number of items asked to converge (i.e., to match the stopping criterion) was used to test the efficiency of the ATS-PD version of QuEDS in terms of reduction of the number of items administered to a patient. Second, the distance between the reconstructed state and the paper and pencil pattern observed for each specific patient was used to evaluate procedure's accuracy. In fact, the higher the distance between the reconstructed state and the response pattern, the greater is the amount of information that is inconsistent between the two modalities of administration of the test. This, in turn, may be due to the error parameters of the items, to a misspecification of the model, or to problems with the algorithm. Since the first two options are excluded given the good fit and the acceptable error estimates (described in the previous section), a strong divergence between the observed pattern and the reconstructed state could be due to some errors in the algorithm; therefore, the measured distance is expected to be as low as possible if the algorithm is accurate. In this respect some further concepts need to be introduced.

A response pattern R_i is the list of the observed answers provided by a subject i to the written version of QuEDS. K_i is the state in $\mathcal{K}$ that is the output of the adaptive assessment when the input is the response pattern R_i. It is important to emphasize that the adaptive procedure always produces a state $K \in \mathcal{K}$ as output even if the observed response pattern $R_i \notin \mathcal{K}$. Thus, we define the distance $d(K_i, R_i)$ as the cardinality of $K_i \Delta R_i$. As a consequence, the results of the simulation for each subject can fall into one of the following mutually exclusive categories:

1. $K_i = R_i$: this happens when the response pattern $R_i \in \mathcal{K}$. In this case $d(K_i, R_i) = 0$. In the specific condition, the output of the adaptive assessment is exactly the same of the output obtained with the written version of the sub-scale;
2. $K_i \neq R_i$: This happens whenever $R_i \notin \mathcal{K}$. In this case of course, $d(K_i, R_i) > 0$. In this situation two alternatives may occur:
   i) $d(K_i, R_i)$ is minimum, that is, there is no $K^* \in \mathcal{K}$ such that $d(K^*, R_i) < d(K_i, R_i)$;
   ii) $d(K_i, R_i)$ is not minimum, that is, there exists $K^* \in \mathcal{K}$ such that $d(K^*, R_i) < d(K_i, R_i)$.

Of course, the occurrence of this last situation should be as rare as possible if the adaptive algorithm is accurate.

6. RESULTS

Results are presented separately for the model fitting analysis, and for the algorithm efficacy and effectiveness test.

6.1. Model Fitting

It is important to stress that the size of the three structures was relatively small counting 124 states for the Cognitive scale, 163 for the Somatic scale, and 142 for the Affective scale. The results of the model fitting for the three structures demonstrated an adequate fit of the models to the set of collected data [Cognitive:

$\chi^2_{(32,144)} = 23,348$, bootstrap-$p = 0.07$; Somatic: $\chi^2_{(15,972)} = 7,237$, bootstrap-$p = 0.16$; Affective: $\chi^2_{(3,928)} = 8,696$, bootstrap-$p = 0.06$]. Therefore, a general adequate fit of the structures was observed.

Another fundamental information obtained by the model fitting was the estimate of both the η and β parameters for each item of the scale. With respect to the η parameters, the estimated values are in general adequately small for almost all items, ranging between 0.01 and 0.18. Only few items presented relatively high values of the estimated β. For the Cognitive scale such items are QuEDS9 with $\beta = 0.50$, and QuEDS30 with $\beta = 0.44$. This criticism may be explained by the phrasing of the items which both include two coordinate sentences. In the Affective scale of QuEDS two items reported high β estimates: namely item QuEDS7 with $\beta = 0.44$, and item QuEDS17 with $\beta = 0.45$. Interestingly both these items are related to crying, suggesting that either the subjects could intentionally fake the specific answer, or that subjects' answer could be affected by a poor introspection about "crying." Although not fully satisfying, it will be shown in the next section that these values did not affect the performance of the adaptive procedure.

Given the number of participants and the number of states, the estimate of the π_K for some states $K \in \mathcal{K}$ was 0. Therefore, we did not use such information in the following part of the simulation and we fixed the starting values of all π_K in the adaptive procedure according to the uniform distribution. This specific implementation is quite common in several applications.

In general, both the fit indexes and the parameters' estimates were satisfactory and were, then, implemented in the adaptive procedure whose performance is analyzed in the next subsection.

6.2. Clinical State Reconstruction

The results of both accuracy and effectiveness tests supported the goodness of the adaptive algorithm. First of all (and actually, as expected) the system, in all the tested versions, was able to correctly reproduce the patient's pattern whenever the pattern R_i was a state in $\mathcal{K}$. Moreover, for the great majority of the patterns, whenever $R_i \notin \mathcal{K}$, the algorithm mapped the pattern into the closest state in the structure, thus, $d(K_i, R_i) = min$ in most cases. Only in a limited number of cases happened that the algorithm mapped a response pattern $R_i \notin \mathcal{K}$ to a state that was not at the minimum distance, thus, for some cases happened that there was a state $K_i^* \in \mathcal{K}$ such that $d(K_i^*, R_i) < d(K_i, R_i)$. This could depend on the sequence of questions asked by the system, which in turn is affected by the error parameters, and by the type of update used by the system. However, this situation has rarely occurred in the simulations.

Table 7 summarizes the results with respect to both accuracy and efficiency of the algorithm.

The table displays that the best performing configuration of the ATS-PD version of the QuEDS is the one with on-line Bayesian update and the parameter ζ computed as a function of η_q and β_q. In the cognitive scale, which has 15 items in total, with this configuration we had a maximum of 11 questions asked and a minimum of 7 to reach the stopping criterion; the average is 8.83 items asked ($SD = 0.47$). It means that the saving in terms

TABLE 7 | Main results of the adaptive algorithm testing.

ζ	Bayesian update	Cognitive		Somatic		Affective	
		Max items	Not Min	Max items	Not Min	Max items	Not Min
21	off-line	17	12	9	8	8	40
21	on-line	11	15	9	6	8	35
21	absent	17	11	9	8	8	39
(η_q, β_q)	off-line	20	7	11	8	23	39
(η_q, β_q)	on-line	**11**	**10**	**9**	**5**	**8**	**29**
(η_q, β_q)	absent	20	7	11	8	29	34

The first column contains the levels of the variable ζ; The second column refers to the implementation of the Bayesian update. The remaining columns contain, for each sub-scale, the maximum number of questions asked to reach the stopping criterion, and the number of response patterns mapped to a state K_i such that there exists $d(K^*, R_i) < d(K_i, R_i)$ for the three sub-scales. In bold the best performing configuration.

of question posed is between 31 and 53%. We found 10 response patterns R in which the distance $d(K_i, R_i)$ was not minimal. In the specific case, $d(K_i, R_i) - d(K_i^*, R_i) \leq 2$. It means that the distance between the output state K_i and the state K_i^* that was the closest one to R_i was never greater than 2 (that is, no more than two answers in K_i were different from K_i^*, whose distance was minimum from R_i).

The somatic scale has 14 items in total. In the best performing configuration we observed a maximum of 9 items asked to reach the stopping criterion and a minimum of 8 item asked to achieve the output of the assessment; the average is 8.42 items asked ($SD = 0.82$). The saving in terms of questions posed is between 36 and 50%. Out of the 173 observed different response patterns, five were mapped to a state whose distance from the pattern was not minimal. Also in this case $d(K_i, R_i) - d(K_i^*, R_i) \leq 2$.

The affective scale counted a total 12 items in the written version. In the best performing configuration we had a maximum of 8 items asked and a minimum of 7; the average is 7.66 items asked ($SD = 0.47$). The saving in terms of questions posed is between 33 and 42%. In this scale 29 response patterns were mapped to a state at a non minimal distance. In the specific case $d(K_i, R_i) - d(K_i^*, R_i) \leq 3$. This last scale seemed to perform in a less accurate, although still adequate and effective, way.

It is important to stress how the procedure is carried out on-line: this means that the questioning rule, the updating rule (together with the Bayesian correction) and the stopping rule are applied in real time even on a standard machine. This indicates an adequate optimization of the computational costs of the procedure.

7. DISCUSSION

This paper aimed at presenting the adaptive version of the three sub-scales (namely, cognitive, somatic, affective) of the QuEDS questionnaire. The computerized algorithm was implemented for the new questionnaire based on an extension of an already existing algorithm for the assessment of knowledge (Doignon and Falmagne, 1999; Falmagne and Doignon, 2011). The parameters of the probabilistic model (i.e., π_K for each clinical state $K \in \mathcal{K}$, η_q and β_q for every $q \in Q$) were estimated through an iterative procedure based on maximum likelihood (Dempster et al., 1977)

on data from the 383 participants. The estimated parameters were then used to calibrate the adaptive algorithm. The simulation study was carried out to test the efficiency and accuracy of the implemented adaptive procedure. Results supported that the adaptive version of the QuEDS provides clinicians with accurate information collected in an efficient way. Moreover, the information collected by means of the adaptive version of QuEDS allows the differentiation of individuals with the same score but with different symptoms (i.e., with different clinical states) and, possibly, different severity of the episode. These properties represent a relevant improvement in the amount, and quality of the collected diagnostic information, as well as in the amount of time needed for case formulation.

The parameter estimates provided the starting point for the implementation of the questioning rule and of the updating rule. This last was tested under different conditions with respect to the computation of the multiplicative parameter ζ. Finally, the opportunity to apply a Bayesian update was tested. Results showed that the most efficient and accurate implementation of the algorithm included the estimate of ζ via the η and β parameters, and the application of an on-line Bayesian updating.

It is important to highlight how the adaptive version of the questionnaire allows for a consistent reduction of the number of questions asked. In the classical written form of the QuEDS, each participant had to answer all 41 items, 15 for the cognitive sub-scale, 14 for the somatic sub-scale, 12 for the affective sub-scale. In the adaptive form of QuEDS only a percentage ranging between 50 and 70% of the items is asked.

The present study and the adaptive version of the QuEDS presents some limitations which should be addressed in future research. Although the sample size is adequate to obtain reliable estimates of the error parameters of the items and of the model fit, it is too small to achieve reliable estimates of the clinical states' probabilities. In fact, given the size of the three structures (respectively 124, 163, and 142 states) and the obtained error parameters of the items, a reliable estimate of the π_K parameters would need a sample of approximately 1,000 individuals. Notice that this limitation is not crucial since, in general, with large structures the *a priori* probability of each state is very low, thus in the adaptive form, the possibility of starting from a uniform distribution on the states is not that strong and generally

accepted. The second limitation of the present study is the relatively low number of patients in the clinical sub-sample. This limitation, which could appear critical in the perspective of classical methodologies in psychological measurement, within the framework of FPA is not that crucial, since the estimate of the parameters and the fit involve the sample in the whole rather than taking into account different sub-samples. Nonetheless, the recruitment of a greater number of patients to refine the estimates and the efficiency-efficacy of the adaptive version of the QuEDS will be the subject matter of future research. One final limitation deserves mention: The present version of the QuEDS does not contain any control scale for social desirability. The inclusion of social desirability scale into self-report tools for the assessment of depression is an important and debated issue (e.g., Langevin and Stancer, 1979; Pichot, 1986; Tanaka-Matsumi and Kameoka, 1986; Cappeliez, 1990; Balsamo and Saggino, 2007) and it will receive further attention in future research in order to provide users of the QuEDS with complete information about this specific issue.

To conclude, this new form of the QuEDS allows a clinician to differentiate the individual's depressive symptoms beyond the score and to administer only the items related to its symptomatology following the logical flow of question-answer. Thus, two patients who obtain the same score to the test can be treated differently according to their symptoms, since answering the same number of items does not mean having the same symptoms configuration.

The future directions of the development of the adaptive version of the QuEDS questionnaire are twofold: on the one hand it is necessary to improve the user interface in order to achieve a simple graphical output able to provide the clinician with a helpful and accessible way to interact with the system. On the other hand, the formal definition of the suggestions for further investigation on the patient have to be formalized and implemented. This last issue is in continuity with the operational approach adopted by the Cognitive Behavioral Assessment 2.0 (CBA 2.0; Bertolotti et al., 1990; Sanavio et al., 2008), and represents the fundamental philosophical approach implemented by FPA methodology. Furthermore, several refinements of ATS-PD system can be implemented, for example the possibility of simplifying the updating rule for real-time application of QuEDS as implemented by Augustin et al. (2013). Another important future direction will be the extension of this approach to the case of polytomous items. The implementation of these extensions would allow FPA to be used with Likert scales, promoting its wider application in both psychological measurement and clinical practice.

AUTHOR CONTRIBUTIONS

AS defined the framework of the paper, conducted the parameter estimates analysis and prepared the final version of the manuscript. FS created the items of the questionnaire, supervised the definition of the formal contexts and contributed to the introduction and discussion of the paper. ID implemented the adaptive algorithm and tested the accuracy and efficiency of the procedure. UG prepared the description of the main deterministic and probabilistic issues of Formal Psychological Assessment, and prepared part of the introduction. GV supervised the whole research project.

REFERENCES

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders: DSM-5, 5th Edn.* Washington, DC: American Psychiatric Association.

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

Annen, S., Roser, P., and Brüne, M. (2012). Nonverbal behavior during clinical interviews: similarities and dissimilarities among schizophrenia, mania, and depression. *J. Nerv. Ment. Dis.* 200, 26–32. doi: 10.1097/NMD.0b013e31823e653b

Augustin, T., Hockemeyer, C., Kickmeier-Rust, M. D., Podbregar, P., Suck, R., and Albert, D. (2013). The simplified updating rule in the formalization of digital educational games. *J. Comput. Sci.* 4, 293–303. doi: 10.1016/j.jocs.2012.08.020

Baek, S.-G. (1997). Computerized adaptive testing using the partial credit model for attitude measurement. *Object. Meas.* 4:37.

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

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

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

Balsamo, M., and Saggino, A. (2007). Test per l'assessment della depressione nel contesto italiano: un'analisi critica [tests for depression assessment in italian context: a critical review]. *Psicoter. Cogn. Comport.* 13:167.

Benazzi, F., Helmi, S., and Bland, L. (2002). Agitated depression: unipolar? bipolar? or both? *Ann. Clin. Psychiatry* 14, 97–104. doi: 10.3109/10401230209149096

Bertolotti, G., Zotti, A. M., Michielin, P., Vidotto, G., and Sanavio, E. (1990). A computerized approach to cognitive behavioural assessment: an introduction to cba-2.0 primary scales. *J. Behav. Ther. Exp. Psychiatry* 21, 21–27. doi: 10.1016/0005-7916(90)90045-M

Birkhoff, G. (1937). Rings of sets. *Duke Math. J.* 3, 443–454. doi: 10.1215/S0012-7094-37-00334-X

Birkhoff, G. (1940). *Lattice Theory.* Providence, RI: American Mathematical Society.

Bottesi, G., Spoto, A., Freeston, M. H., Sanavio, E., and Vidotto, G. (2015). Beyond the score: clinical evaluation through formal psychological assessment. *J. Pers. Assess.* 97, 252–260. doi: 10.1080/00223891.2014.958846

Cappeliez, P. (1990). Social desirability response set and self-report depression inventories in the elderly. *Clin. Gerontol.* 9, 45–52. doi: 10.1300/J018v09n02_06

Davey, B. A., and Priestley, H. A. (2002). *Introduction to Lattices and Order.* Cambridge: Cambridge University Press.

Dell'Osso, L., Armani, A., Rucci, P., Frank, E., Fagiolini, A., Corretti, G., et al. (2002). Measuring mood spectrum: comparison of interview (SCI-MOODS) and self-report (MOODS-SR) instruments. *Comprehens. Psychiatry* 43, 69–73. doi: 10.1053/comp.2002.29852

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Methodol.).* 39, 1–38.

Doignon, J.-P., and Falmagne, J.-C. (1985). Spaces for the assessment of knowledge. *Int. J. Man Mach. Stud.* 23, 175–196. doi: 10.1016/S0020-7373(85)80031-6

Doignon, J.-P., and Falmagne, J.-C. (1999). *Knowledge Spaces.* New York, NY: Springer.

Donadello, I., Spoto, A., Sambo, F., Badaloni, S., Granziol, U., and Vidotto, G. (2017). Ats-pd: an adaptive testing system for psychological disorders. *Educ. Psychol. Meas.* 77, 792–815. doi: 10.1177/0013164416652188

Eggen, T., and Straetmans, G. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educ. Psychol. Meas.* 60, 713–734. doi: 10.1177/00131640021970862

Embretson, S. E., and Reise, S. P. (2013). *Item Response Theory*. Psychology Press.

Falmagne, J.-C., Albert, D., Doble, C., Eppstein, D., and Hu, X. (2013). *Knowledge Spaces: Applications in Education*. Berlin: Springer.

Falmagne, J.-C., and Doignon, J.-P. (1988). A class of stochastic procedures for the assessment of knowledge. *Brit. J. Math. Stat. Psychol.* 41, 1–23. doi: 10.1111/j.2044-8317.1988.tb00884.x

Falmagne, J.-C., and Doignon, J.-P. (2011). *Learning Spaces*. New York, NY: Springer.

Fechner, G. T. (1860). *Elemente der Psychophysik, Vol. 2*. Leipzig: Breitkopf und Härtel.

Finkelman, M. D., Smits, N., Kim, W., and Riley, B. (2012). Curtailment and stochastic curtailment to shorten the CES-D. *Appl. Psychol. Meas.* 36, 632–658. doi: 10.1177/0146621612451647

Fiquer, J. T., Boggio, P. S., and Gorenstein, C. (2013). Talking bodies: nonverbal behavior in the assessment of depression severity. *J. Affect. Disord.* 150, 1114–1119. doi: 10.1016/j.jad.2013.05.002

First, M., Gibbon, M., Spitzer, R. L., and Williams, J. (1996). *User's Guide for the Structured Clinical Interview for DSM-IV Axis i Disorders' Research Version*. New York, NY: Biometrics Research Department, New York State Psychiatric Institute.

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., and Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Qual. Life Res.* 14:2277. doi: 10.1007/s11136-005-6651-9

Ganter, B., and Wille, R. (1999). *Formal Concept Analysis: Mathematical Foundations*. Berlin; Heidelberg: Springer Verlag.

Geerts, E., and Brüne, M. (2009). Ethological approaches to psychiatric disorders: focus on depression and schizophrenia. *Aust. N.Z. J. Psychiatry* 43, 1007–1015. doi: 10.1080/00048670903270498

Gibbons, R. D., Clark, D. C., Cavanaugh, S. V., and Davis, J. M. (1985). Application of modern psychometric theory in psychiatric research. *J. Psychiatr. Res.* 19, 43–55. doi: 10.1016/0022-3956(85)90067-6

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., et al. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatr. Serv.* 59, 361–368. doi: 10.1176/ps.2008.59.4.361

Girard, J. M., Cohn, J. F., Mahoor, M. H., Mavadati, S. M., Hammal, Z., and Rosenwald, D. P. (2014). Nonverbal social withdrawal in depression: evidence from manual and automatic analyses. *Image Vis. Comput.* 32, 641–647. doi: 10.1016/j.imavis.2013.12.007

Goodwin, F. K., and Jamison, K. R. (2007). *Manic-Depressive Illness: Bipolar Disorders and Recurrent Depression, Vol. 1*. Oxford: Oxford University Press.

Granziol, U., Spoto, A., and Vidotto, G. (2018). The assessment of nonverbal behavior in schizophrenia through the formal psychological assessment. *Int. J. Methods Psychiatr. Res.* 27:e1595. doi: 10.1002/mpr.1595

Grayce, C. J. (2013). "A commercial implementation of knowledge space theory in college general chemistry," in *Knowledge Spaces: Applications in Education*, eds J.-C. Falmagne, D. Albert, C. Doble, D. Eppstein, and X. Hu (New York, NY: Springer), 93–113.

Groth-Marnat, G. (2009). *Handbook of Psychological Assessment*. New York, NY: John Wiley & Sons.

Gulliksen, H. (2013). *Theory of Mental Tests*. New York, NY: Routledge.

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

Langevin, R., and Stancer, H. (1979). Evidence that depression rating scales primarily measure a social undesirability response set. *Acta Psychiatr. Scand.* 59, 70–79. doi: 10.1111/j.1600-0447.1979.tb06948.x

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. New York, NY: Routledge.

Marinagi, C. C., Kaburlasos, V. G., and Tsoukalas, V. T. (2007). "An architecture for an adaptive assessment tool," in *Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual*, T3D–11. (Milwaukee, WI: IEEE).

Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L., et al. (2018). An introduction to network psychometrics: relating ising network models to item response theory models. *Multivar. Behav. Res.* 53, 15–35. doi: 10.1080/00273171.2017.1379379

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment: a review of evidence and issues. *Amer. Psychol.* 56:128. doi: 10.1037/0003-066X.56.2.128

Narens, L., and Luce, R. D. (1993). Further comments on the nonrevolution arising from axiomatic measurement theory. *Psychol. Sci.* 4, 127–130. doi: 10.1111/j.1467-9280.1993.tb00475.x

Nokelainen, P., Silander, T., Tirri, H., Nevgi, A., and Tirri, K. (2001). "Modeling students' views on the advantages of web-based learning with bayesian networks," in *Proceedings of the Tenth International PEG Conference: Intelligent Computer and Communications Technology–Learning in On-Line Communities* (Tampere), 101–108.

Nordgaard, J., Sass, L. A., and Parnas, J. (2013). The psychiatric interview: validity, structure, and subjectivity . *Eur. Arch. Psychiatry Clin. Neurosci.* 263, 353–364. doi: 10.1007/s00406-012-0366-z

Novick, M. R. (1965). The axioms and principal results of classical test theory. *ETS Res. Rep. Ser.* 1965, 1–18. doi: 10.1002/j.2333-8504.1965.tb00132.x

Petersen, M. A., Groenvold, M., Aaronson, N., Fayers, P., Sprangers, M., Bjorner, J. B., et al. (2006). Multidimensional computerized adaptive testing of the EORTC QLQ-C30: basic developments and evaluations. *Qual. Life Res.* 15, 315–329. doi: 10.1007/s11136-005-3214-z

Philippot, P., Schaefer, A., and Herbette, G. (2003). Consequences of specific processing of emotional information: impact of general versus specific autobiographical memory priming on emotion elicitation. *Emotion* 3:270. doi: 10.1037/1528-3542.3.3.270

Pichot, P. (1986). "A self-report inventory on depressive symptomatology (qd2) and its abridged form (qd2a)," in *Assessment of Depression*, eds N. Sartorius and T. A. Ban (Berlin; Heidelberg: Springer), 108–122.

Rasch, G. (1960). *Studies in Mathematical Psychology: I. Probabilistic Models for Some Intelligence and Attainment Tests*. Oxford: Nielsen & Lydiche.

Reise, S. P., and Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Appl. Psychol. Meas.* 14, 45–58. doi: 10.1177/014662169001400105

Sanavio, E., Bertolotti, G., Michielin, P., Vidotto, G., and Zotti, A. (2008). *Cba-2.0 Scale Primarie: Manuale. Una Batteria ad Ampio Spettro per l'Assessment Psicologico*. Firenze: Organizzazioni Speciali.

Serra, F., Spoto, A., Ghisi, M., and Vidotto, G. (2015). Formal psychological assessment in evaluating depression: a new methodology to build exhaustive and irredundant adaptive questionnaires. *PLoS ONE* 10:e0122131. doi: 10.1371/journal.pone.0122131

Serra, F., Spoto, A., Ghisi, M., and Vidotto, G. (2017). Improving major depressive episode assessment: a new tool developed by formal psychological assessment. *Front. Psychol.* 8:214. doi: 10.3389/fpsyg.2017.00214

Simms, L. J., Goldberg, L. R., Roberts, J. E., Watson, D., Welte, J., and Rotterman, J. H. (2011). Computerized adaptive assessment of personality disorder: introducing the cat–pd project. *J. Pers. Assess.* 93, 380–389. doi: 10.1080/00223891.2011.577475

Smits, N., Finkelman, M. D., and Kelderman, H. (2016). Stochastic curtailment of questionnaires for three-level classification: shortening the ces-d for assessing low, moderate, and high risk of depression. *Appl. Psychol. Meas.* 40, 22–36. doi: 10.1177/0146621615592294

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

Spiegel, R., and Nenh, Y. (2004). An expert system supporting diagnosis in clinical psychology. *WIT Trans. Inform. Commun. Technol.* 31, 145–154. doi: 10.2495/CI040151

Spoto, A. (2011). *Formal Psychological Assessment Theoretical and Mathematical Foundations*. Doctoral Dissertation, University of Padua. Available online at: http://paduaresearch.cab.unipd.it/3477

Spoto, A., Bottesi, G., Sanavio, E., and Vidotto, G. (2013a). Theoretical foundations and clinical implications of formal psychological assessment. *Psychother. Psychosomat.* 82, 197–199. doi: 10.1159/000345317

Spoto, A., Stefanutti, L., and Vidotto, G. (2010). Knowledge space theory, formal concept analysis, and computerized psychological assessment. *Behav. Res. Methods* 42, 342–350. doi: 10.3758/BRM.42.1.342

Spoto, A., Stefanutti, L., and Vidotto, G. (2012). On the unidentifiability of a certain class of skill multi map based probabilistic knowledge structures. *J. Math. Psychol.* 56, 248–255. doi: 10.1016/j.jmp.2012.05.001

Spoto, A., Stefanutti, L., and Vidotto, G. (2013b). Considerations about the identification of forward-and backward-graded knowledge structures. *J. Math. Psychol.* 57, 249–254. doi: 10.1016/j.jmp.2013.09.002

Spoto, A., Stefanutti, L., and Vidotto, G. (2016). An iterative procedure for extracting skill maps from data. *Behav. Res. Methods* 48, 729–741. doi: 10.3758/s13428-015-0609-9

Stefanutti, L., Spoto, A., and Vidotto, G. (2018). Detecting and explaining blims unidentifiability: forward and backward parameter transformation groups. *J. Math. Psychol.* 82, 38–51. doi: 10.1016/j.jmp.2017.11.001

Stevens, S. S. (1946). On the theory of scales of measurement. *Science* 103, 677–680. doi: 10.1126/science.103.2684.677

Stevens, S. S. (1951). *Mathematics, Measurement, and Psychophysics.* Oxford: Wiley.

Stevens, S. S. (1957). On the psychophysical law. *Psychol. Rev.* 64:153. doi: 10.1037/h0046162

Strull, W. M., Lo, B., and Charles, G. (1984). Do patients want to participate in medical decision making? *J. Am. Med. Assoc.* 252, 2990–2994. doi: 10.1001/jama.1984.03350210038026

Suppes, P., Krantz, D. H., Luce, R. D., and Tversky, A. (1989). *Foundations of Measurement: Geometrical, Threshold, and Probabilistic Representations, Vol. 2.* New York, NY: Academic Press.

Suppes, P., and Zinnes, J. (1963). "Basic measurement theory," In *Handbook of Mathematical Psychology*, eds R. D. Luce and E. Galanter (Wiley, Oxford), 1–76.

Tanaka-Matsumi, J., and Kameoka, V. A. (1986). Reliabilities and concurrent validities of popular self-report measures of depression, anxiety, and social desirability. *J. Consult. Clin. Psychol.* 54:328. doi: 10.1037/0022-006X.54.3.328

Uhlmann, E. L., Leavitt, K., Menges, J. I., Koopman, J., Howe, M., and Johnson, R. E. (2012). Getting explicit about the implicit: a taxonomy of implicit measures and guide for their use in organizational research. *Organ. Res. Methods* 15, 553–601. doi: 10.1177/1094428112442750

Wainer, H. (2000). *CATs: Whither and Whence.* Princeton, NJ: ETS Research Report Series.

Wille, R. (1982). "Restructuring lattice theory: an approach based on hierarchies of concepts," in *Ordered Sets*, ed I. Rival (Dordrecht: Reidel), 445–470.

Yong, S., Awang Rambli, D. R., and Anh, N. (2007). "Depression consultant expert system," in *6th Annual Seminar on Science and Technology* (Tawau).

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.