On the implementation of computerized adaptive observations for psychological assessment

Abstract

The use of observational tools in psychological assessment has decreased in recent years, mainly due to its personnel and time costs, and researchers have not explored methodological innovations like adaptive algorithms in observational assessment. In the present study, we introduce the behavior-driven observation procedure to develop, test, and implement observational adaptive instruments. In Study 1, we use a preexisting observational checklist to evaluate nonverbal behaviors related to psychotic symptoms and to specify the adaptive algorithm's model. We fit the model to observational data collected from 114 participants. The results support the model's goodness of fit. In Study 2, we use the estimated model parameters to calibrate the adaptive procedure and test the algorithm for accuracy and efficiency in adaptively reconstructing 58 non-adaptively collected response patterns. The results show the algorithm's good accuracy and efficiency, with a 40% average reduction in the number of administered items. In Study 3, we used real raters to test the adaptive checklist built with behavior-driven observation. The results indicate adequate intra-rater agreement and good consistency of the observed response patterns. In conclusion, the results support the possibility of using behavior-driven observation to create accurate and affordable (in terms of resources) observational assessment tools.

Keywords: adaptive psychological assessment; behavioral observation; behavior-driven observation; one-zero sampling; modal response patterns; schizophrenia, cross-validation.

On the implementation of computerized adaptive observations for psychological assessment

In assessing mental disorders, the need for much high-quality information often collides with the time required to collect it. The time required to entirely administer diagnostic tools, such as observational checklists or semi-structured interviews (e.g., the Positive and Negative Symptoms Scale; Kay, et al., 1987), sometimes could not fit the clinical routine in health-care settings (Kølbæk et al., 2018), even if the tools provide extremely accurate information. On the other hand, diagnostic tools such as questionnaires offer quick administration but sometimes redundant and inexhaustive information, from a qualitative point of view (Serra, et al., 2015). Researchers often balance the quality and quantity of information with administration time by defining short forms of assessment instruments. In recent years, another increasingly frequent solution in clinical research involves computerized adaptive assessment. This evaluation procedure—usually performed using electronic devices such as a PC or a tablet—consists of adaptively administering items based on people's previous responses (Petersen et al., 2006; Spoto et al., 2018).

Clinicians have used adaptive assessments previously. In fact, this method already occurs in semi-structured interviews, such as in the Structured Clinical Interview for the DSM (SCID; First, 2014), where the clinician moves through the questions based on the collected responses. Clinicians also widely apply adaptive assessment in psychological testing via the so-called computerized adaptive testing (CAT; Wainer, 2000). Several articles in the last two decades showed that CAT allows clinicians to (a) reduce the length of the questionnaires (Petersen et al., 2006), (b) collect the same amount of information as tests' extended (traditional) versions (Spoto et al., 2018), (c) avoid accuracy loss, and (d) increase the assessment's efficiency (Donadello et al., 2017). While adaptive questionnaires are becoming frequently adopted by researchers (Gibbons et al., 2012; Michel et al., 2018; Serra, et al., 2017; Spoto et al., 2018), the same cannot be said for adaptive observational

assessments. Until now, the logic of adaptivity in observations is mainly related to training behavior detection (e.g., the Train-to-Code software; Ray & Ray, 2008), and the detection of behaviors for clinical purposes lacks sufficient research. In this sense, Pino et al. (2018) first showed how some modules of the latest Autism Diagnostic Observation Schedule version (ADOS-2; C. Lord, Luyster, et al., 2012; C. Lord, Rutter, et al., 2012) have good implementation potential in a computerized adaptive assessment system. Unfortunately, no concrete adaptive observational tool currently exists.

Defining an adaptive observation involves the same critical issues encountered in constructing traditional observational tools:

- the definition of a deterministic assessment model comprising a list of clear and easy-to-operationalize behaviors (Hawes, et al., 2013; Haynes & O'Brien, 2000);
- the possibility of using multiple observations per person to obtain more precise response patterns (observational measures require a remarkable amount of resources in terms of personnel and time, so observational assessments often consist of a single scoring of an instrument during [i.e., online] or after [i.e., offline] the observation, which could lack accuracy because the memory interference could introduce bias); and
- the implementation and testing of the found model according to a probabilistic framework, which should account for common sources of error during the observation, such as the halo effect (Mumma, 2002), confirmatory bias (e.g., Cantor & Mischel, 1979), five-minute impression formation (Lee, Barak, & Uhlemann, 1999), and primacy and recency effects (Groth-Marnat, 2009).

After accomplishing these steps, observers should implement the adaptive observational tool into a computerized algorithm to guide clinical observations. Finally, observers must have a report containing information such as the score, the person's behavioral pattern, and the set of symptoms endorsed by that pattern.

This methodological study aims at introducing the behavior-driven observation (BDO), a method used to build computerized adaptive observational instruments that could help psychologists conduct exhaustive, effective, and efficient observations. This study particularly aims at describing, step-by-step, a procedure that people could use to build and test an adaptive observation. To reach this goal, we applied the BDO method as a practical example to upgrade a preexisting observational instrument into an adaptive tool.

We describe the BDO procedure in this order: after describing the methodology used to define an instrument (General Methods section), we present a first study to describe how to define the deterministic basis of an instrument built within the BDO and how to test it using a probabilistic model of assessment (Study 1). We then use Study 1's results in Study 2 to implement an adaptive algorithm of the BDO. Study 2 aims also at testing, through a simulation, the adaptive algorithm's accuracy and efficiency. Finally, in a pilot study, we test the adaptive instrument with real raters in Study 3. Finally, we discuss all the results, limitations, and future perspectives.

General Methods

In the next subsections, we describe the theoretical and methodological components of the BDO and introduce a practical example to clarify the main technical issues.

The Formal Psychological Assessment

Researchers can use the Formal Psychological Assessment (FPA; Spoto et al., 2013) methodology to build questionnaires that can exhaustively investigate the symptoms of a given psychological disorder (Granziol et al., 2017). The FPA formally connects two theories of mathematical psychology with clinical psychology, namely the Knowledge Space Theory (KST; Doignon & Falmagne, 1999; Falmagne & Doignon, 2011) and the Formal Concept Analysis (FCA; Ganter & Wille, 1999; Wille, 1982) to build assessment instruments.

The FPA's core aspect concerns the possibility of delineating and analyzing relationships

between the nonempty set *A* of clinical issues (i.e., diagnostic criteria or symptoms linked to a mental disorder; *attributes* in the FPA) and the nonempty set *Q* of items investigating those same clinical issues. The term *clinical domain* refers to the *Q* collection of all the items investigating a disorder.

When considering the evaluation of a major depressive disorder from a behavioral point of view, for example, the clinical domain comprises a set of all the possibly explored items, such as "The posture of the person points downward". Moreover, the collection of clinical issues from common clinical practices, scientific literature, or clinical sources (e.g., the DSM-5; American Psychiatric Association [APA], 2013), referred to as the diagnosis of major depressive episode, makes up set *A* of attributes to investigate. The symptom "Curved posture" exemplifies a possible element in this set. Each item in the domain may investigate one or more symptoms. For instance, the previous item "The posture of the person points downward" investigates the attribute "Curved posture."

The *clinical context* displays all the links between items and attributes in the form of a Boolean matrix (i.e., a binary table of 0 and 1) containing the items in rows and the attributes in columns; in a table with as many rows as the number of items and as many columns as the number of attributes, each cell contains 1 whenever item *q* in the row investigates the attribute *a* in the column (it contains 0 otherwise). For instance, the item "The posture of the person points downward" (q_1) in Table 1 investigates the attribute "Curved posture" (a_1); the item "Both posture and gaze of the person point downward" (q_2) investigates a_1 and the attribute "Gaze downward" (a_2). In fact, Table 1 contains a 1 in the cell corresponding to q_1 and a_1, and both cells corresponding to a_1 and a_2 contain 1 for item q_2.

[INSERT TABLE 1 HERE]

The clinical context allows an analysis of the *prerequisite relations* among items. An item, *p*, constitutes a prerequisite for another item, *q*, only if the set of attributes investigated by item *p* comprises a subset of the attributes investigated by item *q*. Assuming no error, we cannot observe a positive answer to item *q* without observing an affirmative answer to item *p*. We can easily check in

Table 1 if item q_1 constitutes a prerequisite of item q_2. In fact, the symptoms needed to positively answer item q_2 also affirmatively answer item q_1. Importantly, we can use the prerequisite relation in adaptive assessment to infer the presence of certain symptoms without directly investigating them.

The clinical context and the prerequisite relation define all the admissible clinical outputs of an assessment instrument built with FPA, namely the *clinical concepts*. These nonnumerical outputs collect all the items endorsed by a person and the set of attributes investigated by those items. The clinical context defines every clinical concept, so it follows that we can know all the clinical outputs a priori from the clinical context (see Spoto et al., 2010). We call the collection of all the clinical concepts the *clinical structure* ($\mathcal{C}$). Spoto et al. (2010, 2016) details the construction of a clinical structure from a clinical context. We do not focus on these issues in the present research, so we will not explain them further.

Figure 1 displays the clinical structure defined from Table 1's clinical context. Each concept of this structure, each node on the graph**,** contains the set of attributes necessary and sufficient to observe all the items in the concept.

[INSERT FIGURE 1 HERE]

Given the set of attributes $A= \{a_1, a_2\}$ and the set of items $Q= \{q_1, q_2\}$ (both shown in Table 1), this example allows us to check all the clinical concepts defined by the clinical context. Respectively, they comprise the clinical concept in which the patient does not present any symptoms (i.e., the empty clinical concept, $\varnothing$); the clinical concept in which the patient presents only the symptom "Curved posture" (i.e., the clinical concept $\{\{q_1\};\{a_1\}\}$); and the clinical concept in which the patient presents all the symptoms (i.e., the clinical concept $\{\{q_1, q_2\};\{a_1, a_2\}\}$). Moreover, we can verify the prerequisite relation among items (e.g., item q_1 constitutes a prerequisite of q_2).

A clinical structure like Figure 1 represents a deterministic and incomplete basis for an adaptive assessment, which also requires a probabilistic framework for three main reasons. First, the observed

response patterns could imperfectly represent an individual's clinical concept because of noise in the assessment, represented by the *false negative* (β_q) and *false positive* (η_q) rates for each item. From an observational perspective, the false negative rate shows the probability of not observing a behavior that actually occurs, and the false positive rate shows the probability of reporting a behavior that has not actually occurred. Second, the clinical concepts could occur with different frequencies within the population; that is, each concept has its own probability, π_C, of occurrence in the population. A deterministic approach implicitly assumes equiprobability for each concept. Third, a deterministic clinical structure cannot predict the probability of all the clinical concepts, given people's response patterns (i.e., the observed pattern necessarily coincides with the concept). The mentioned parameters (i.e., π_C for every clinical concept in the clinical structure and $\mathcal{C}$, β_q and η_q for every item in the clinical domain) allow researchers to define a *probabilistic clinical structure*, namely an assessment model that can assign a probability of occurrence in the population to each clinical concept of the structure. The probabilistic model applied to clinical structures comprises the Basic Local Independence Model (BLIM; Doignon & Falmagne, 1999; Falmagne & Doignon, 1988).

Within the BLIM, we calculate each response pattern's occurrence probability by multiplying the conditional probability of the pattern—given that a patient exists in a given clinical concept—by the probability of the clinical concept. We determine the conditional probability of the response pattern given the state via the false negative and false positive rates of each item, so such error parameters support the entire assessment model (formal details on the BLIM appear in Appendix S1). In fact, higher error parameters yield lower reliability for the collected answers. In all the previous applications of FPA (e.g., Donadello et al., 2017; Pino et al., 2018; Serra et al., 2015, 2017), a single response pattern per person was sufficient to fit the BLIM and, consequently, to validate the instrument from a probabilistic point of view because the instruments used comprised self-report questionnaires. As mentioned before, using a single scoring in observation could reduce response patterns' accuracy

because several distortions and false positives or negatives can occur. We would have to obtain a single pattern derived by multiple scores. The following section describes a procedure for estimating a single response pattern out of several.

One-Zero Sampling and Modal Response Patterns

In observational studies, people frequently use multiple observations to correctly detect or quantify a behavior's frequency or duration. In particular, researchers use specific sampling strategies to accurately organize these multiple data collections (Altmann, 1974; Hawes et al., 2013; Powell et al., 1977). The one-zero sampling method has a long tradition in observational assessment (Goodenough, 1928). Originally developed for observing animal behavior, it also appears in human observation, such as in the assessment of nonverbal behavior in patients diagnosed with schizophrenia (Brüne et al., 2008; Troisi et al., 2007) or with depression (Geerts & Brüne, 2009).

The one-zero sampling method consists of two main phases: splitting the observation into *n* equal-length samples (e.g., 15–30 s) and then checking a behavior's occurrence within each sample (Martin et al., 1993). Specifically, at the end of each sample (usually announced by a beeper), observers should score 1 if they observe the target behavior at any time during the sample. Thus, a single score exists for each behavior via the proportion of samples in which it appears. Authors debate the adequacy of such a score in the literature (e.g., Rhine & Linville, 1980; Smith, 1985). Some authors discouraged its usage, arguing that it represents a biased measure of frequency or duration (Altmann, 1974; Dunkerton, 1981). Other authors proposed some adjustments to its computation that, while maintaining the procedure's advantages, allow researchers to obtain accurate frequency and duration rates based on the response patterns (Smith, 1985; Suen & Ary, 1984, 1986). In general, researchers consider one-zero sampling an easy-to-apply method that allows them to observe several behaviors within the same sample. Finally, one-zero scores can enhance inter- and intra-rater agreement (Altmann, 1974; Rhine & Linville, 1980; Troisi et al., 2007). In this study, we propose a solution to use

the information of multiple one-zero samples and to organize them into a single measure: the *modal response pattern*. We can obtain this pattern by establishing for each behavior whether occurrence or non-occurrence is modal across *n* observational samples. Table 2 displays an example of a modal response pattern:

[INSERT TABLE 2 HERE]

In this case, Items 1, 4, and 5 describe patterns observed in three out of five samples. Therefore, their modal value will be 1; that is, "occurrence" takes place throughout the observation. Items 2 and 3 describe patterns observed during two samples; therefore, their modal value will be 0; that is, "non-occurrence" takes place throughout the observation. This example has a modal response pattern of $Mo = \{1,0,0,1,1\}$.

The modal response patterns approach conveys accurate information out of a set of response patterns because the mode acts as the most appropriate central tendency index for dichotomous data, such as those of one-zero sampling. Other centrality indexes could inappropriately represent dichotomous data: for instance, the median and the mode in this case coincide, but the mean represents the proportion of ones rather than representing a centrality indicator (Manikandan, 2011; Weisberg & Weisberg, 1992). Moreover, the present study's main objective of observation concerns establishing a behavior's occurrence or absence across *n* observational samples and counting its occurrences in various samples does not convey this information. In fact, several authors (e.g., Dunkerton, 1981; Leger, 1977; Martin et al., 1993; Powell et al., 1977) argue that using counts as a score can result in a biased measure of frequency.

Furthermore, a modal response pattern appears closest to all other response patterns collected during the observation in terms of similar responses. This property makes modal response patterns good candidates for summarizing information derived from multiple observations into a single measure (a detailed description of this property and the corresponding example appears in Appendix S2). After

reconstructing univocal information out of the patterns observed in each sample, we can implement a probabilistic framework for the deterministic model of the observational assessment to define its adaptive counterpart.

In the next sections, we will describe how to define and implement a nonadaptive deterministic model of observational assessments within a probabilistic structure according to the FPA. In particular, we report how we determined our sample size, all data exclusions, all manipulations, and all measures to first define a nonadaptive checklist (a demonstration of the BDO method) and then test its model fit and estimate its error parameters (Study 1). We then use Study 1's results to calibrate the adaptive algorithm and test its accuracy and efficiency in Study 2. Finally, we test the algorithm with real raters in Study 3.

Study 1

To develop the proposed observation method, we built a nonadaptive observational checklist via the FPA. Importantly, the definition and the testing of the nonadaptive checklist should only act as a demonstration of the BDO's method. In particular, experts refined preexisting lists of items and attributes used to evaluate the nonverbal behavior of schizophrenia. We used both lists to define the deterministic side of the assessment model. We later tested the model fit and parameter estimates of the assessment model using real data. We describe all the steps below.

Methods

Definition of the Nonadaptive Tool

We defined the nonadaptive observational instrument starting with the list of nonverbal behaviors proposed by Granziol et al. (2018) and constructed it according to the FPA. The list consisted of a set of 23 items containing nonverbal behaviors related to negative symptoms frequently observed in schizophrenia. We rephrased the items (e.g., shortening some items) somewhat to provide more precise behavioral coding. Furthermore, we added some attributes to the original set, which

investigated relevant symptoms such as gaze fixation (Dowiasch et al., 2016; Gaebel, 1989). Two experts in the field of schizophrenia conducted both operations starting with the original lists of items and attributes.

The experts then defined two clinical contexts independently. The experts' agreement on this task was quite high (κ = .91). After discussing the few remaining disagreements, the experts defined the final clinical context, consisting of 20 items that investigated all the attributes. Tables S1 and S2 in the supplemental material contain the final lists of attributes and items, respectively.

Starting from the context, we then constructed the corresponding clinical structure, which contained 6,336 clinical concepts. To have a more convenient clinical structure size, we split the clinical context into a number of subcontexts (Table 3) by clustering the items and their investigated attributes around nonverbal areas of interest. In doing so, we aimed to have theoretically consistent subcontexts and obtain substructures that shared the minimum possible number of investigated symptoms. The most convenient solution was creating two subcontexts. One contained items that investigated head and body Movements, as well as gesture and facial expressivity (Table 3a); the other collected items related to prosodic features and prosocial interactive manifestations (Table 3b). The former cluster, which we named *Movement*, generated a structure containing 40 concepts. The latter context, which we named *ProsInt*, generated a structure containing 180 clinical concepts. These two clusters shared only one attribute, "Decreased reactivity to the environment," enhancing between-cluster independence.

[INSERT TABLE 3 HERE]

This division into two substructures led to a total number of clinical concepts that was easier to manage in terms of model testing and parameter estimation without losing information or accuracy during observational assessments.

Testing the Nonadaptive Tool

The testing of the deterministic structures focused on various issues: selecting samples, planning a systematic observation to minimize observer bias, defining reliable modal response patterns, using the modal patterns as BLIM input for testing structure fit, and estimating item error parameters. Such analyses represent only one step of the BDO procedure. As specified above, a comprehensive validation of the nonadaptive checklist is beyond the aim of the present study. In the next section, we describe the sampling strategy for the clinical and nonclinical subsamples.

Sample

A total of 172 Italian participants enrolled in this study on a voluntary basis. We established the sample size based on previous similar studies aimed at conducting similar estimations (Donadello et al., 2017; Spoto et al., 2010). The focus of sample size definition is the number of error parameters per item to estimate; in this case, we have a false positive and a false negative parameter for each item. Therefore, a minimum of two participants per item is necessary. In the case at hand, we had to estimate 22 parameters for the ProsInt substructure (i.e., 11 items × 2 parameters per item) and 18 parameters for the Movement substructure (9 items × 2 parameters per item). This sample size allowed us to count approximately nine persons per item. We recruited a clinical subsample of 38 residential inpatients with the following primary diagnoses: schizophrenia, major depressive disorder with psychotic behavior and bipolar disorder with psychotic behavior. We recruited all patients diagnosed with bipolar disorder during the depressive phase. All patients without a diagnosis of schizophrenia presented at least one negative symptom. Patients were enrolled in three psychiatric centers in Italy: The Psychiatry Unit of San Salvatore Hospital in L'Aquila; the Padova University Hospital; and the Department of Clinical Neurosciences, IRCCS San Raffaele Scientific Institute in Milan.

Psychiatrists who were experts in the schizophrenia field ascertained diagnoses using the DSM-IV-TR as nosology classification system (American Psychiatric Association [APA], 2000). Patients with a diagnosis of schizophrenia received first- (~20%) or second-generation (~80%)

antipsychotics to treat their disorder, whereas patients with other disorders received benzodiazepines, tricyclic antidepressants, and selective serotonin reuptake inhibitors (SSRIs; ~70%). Coadministration of benzodiazepines was possible. Inclusion criteria for the clinical group were as follows: the presence of at least one psychotic symptom; ongoing treatment with a stable dose of pharmacological therapy; being a native speaker of Italian. Severe traumatic brain injury, neurological disorders, intellectual disability, and alcohol or substance abuse in the past six months were the exclusion criteria. The nonclinical sample comprised 134 people, mostly students, selected from the population and recruited in Padova. The exclusion criteria for nonclinical sample were self-reported presence of at least one of the disorders mentioned above, intellectual disability, and self-reported alcohol or substance abuse in the past six months. We used a snowball sampling strategy (Goodman, 1961). The initial recruitment used online advertisements. Each time participants answered the advertisement and agreed to participate, we encouraged them to ask other people to participate. In this way, we increased the randomness of the recruitment. Table 4 shows the demographics of the entire sample. Table S3 contains levels of education of all the participants, divided by group.

[INSERT TABLE 4 HERE]

Although the established sample is adequate to verify model fit (e.g., Spoto et al., 2010), the proportion of patients and nonclinical individuals may be not completely adequate to represent the real population. Given the prevalence of psychotic spectrum disorder (0.3–0.7% lifetime; APA, 2013), the proportion of patients with a psychotic spectrum disorder diagnosis in our sample was larger than in in the actual population. Nonetheless, given that the main aim of the present research was to introduce a method for building, testing, and implementing an adaptive observational tool, this imbalance did not represent a serious problem.

All participants read and filled out an informed consent form before the experimental phase. The psychiatrists and psychologists explained the voluntary basis of participation and the nonintrusive

nature of the study very carefully. Because the study consisted of a videotaped interview, we stated explicitly to the participants that they could withdraw from the interview at any time without penalty or a change in the therapeutic plan. We conducted this study according to the Declaration of Helsinki and obtained approval from each collaborating center's ethical committee.

Stimuli and Procedure

We videotaped participants during the administration of the Positive and Negative Symptoms Scale (PANSS; Kay et al., 1987), a semi structured interview frequently used in psychiatric settings by psychiatrists at the collaborating centers. A psychologist administered the PANSS to the nonclinical group. For the patients, the psychiatrists administered the PANSS during their standard assessment phases, which was less demanding on patients. In all cases, the speakers sat in front of each other with a remotely controlled Sony PJ410 video camera placed on top of a 120-cm tripod and positioned behind the interviewer's right shoulder. This position allowed the interviewer to record the participant's whole body without the interference of the expert. The interviews lasted from 30 to 45 minutes. Interviewers performed all camera operations (e.g., video extraction, charging) immediately before or after each interview. This allowed the interviewers to be less distracted and reduced the participants' feelings of being observed. After the interview, the experts provided the participants with a detailed explanation of the interview's aim and answered any questions the participants raised.

Videos were then downloaded and linked to a code that was unique for each participant. We extracted the video samples for the scoring phase to build the stimuli according to the one–zero sampling method, and we determined the samples' number and duration empirically decided based on 10 pilot interviews. To provide a good tradeoff between the length of the observations and the accuracy of the collected information, we selected 15 one-minute samples.

We created the stimuli as follows. First, we split each original video into 15 samples. A Python script then randomly determined which sample to extract from the original video (Van Rossum et al.,

2007). The script excluded the first and last five minutes. We then edited the 15 samples, adding a countdown sequence and a beeper before and after each sample, respectively. Finally, we shuffled and coded this set of edited samples as a single MP4 video file[1] for use during the testing phase. We used the random selection and the subsequent randomization of the 15 samples to reduce order and sequence effects and minimize the risk of biases such as the first-minute impression.

Independent raters (one male and one female) observed and coded the 172 final videos. The raters were psychiatrists selected based on their considerable expertise in mental disorders belonging to the psychotic spectrum, based on the number of years they have worked with these disorders and related scientific publications. In particular, both raters watched and rated the 15 samples of each video on a personal computer while seated 70 cm from the screen. They received instructions not to talk or have any contact with each other or with the experimenter. During the pre-experimental phase, the raters received a detailed description of items along with some examples of their manifestations. We also explained the scoring rule and paid particular attention to those items investigating several nonverbal behaviors.

During the observational assessment, raters observed each video sample until the beeper's sound warned them of the session's end. At that point, the video stopped, and the raters filled out the nonadaptive checklist (provided in paper-and-pencil form), checking an item only if the described nonverbal behaviors occurred within that sample. When the raters observed several cases per day, we suggested they take a break every two observed patients and limit themselves to four observations per day.

Model Fitting and Parameters Estimate

The one–zero sampling and modal response pattern procedures led to 344 modal response

[1] We stored all original and edited videos on hard disks located in a locked, safe place that only the project's researchers could access.

patterns (i.e., two patterns per patient, or, one for each rater). We obtained the final set of 172 modal response patterns by combining each pair of modal patterns into a single pattern by solving the few disagreements between the raters using direct discussion. Finally, we split this final set into two subsets to perform a cross-validation. Specifically, we used 114 modal response patterns (i.e., 89 from the nonclinical group and 25 from the clinical group) to test the structures' fit and estimate the item error parameters β and η. We used the remaining 58 modal response patterns (i.e., 45 from the nonclinical group and 13 from the clinical group) to perform Study 2.

Concerning model fitting and the parameters' estimate analyses, we used an expectation-maximization algorithm (EMA; Dempster et al., 1977) implemented in MATLAB. The Pearson's chi-square statistic (Falmagne & Doignon, 2011) evaluated the two structures' goodness of fit to the data, and we calculated the corresponding *p* value using a parametric bootstrap with 5000 replications. We computed the *p* value by bootstrap due to the sparseness of the data matrices, for which the asymptotic distribution of the χ^2 was not completely reliable (Reiser & VandenBerg, 1994; Spoto et al., 2010). Moreover, we used the algorithm to estimate all the BLIM parameters, namely the π_c for each clinical concept of the structures and the error rates (i.e., β_q and η_q) for each item on both subscales. High error rates for an item indicated low reliability of the responses collected through that item, meaning a possible misfit of the model.

In recent studies (Spoto, Stefanutti et al., 2012; 2013; Spoto, Serra, et al., 2018; Stefanutti & Robusto, 2009; Stefanutti et al., 2018), error rates between 0 and .1 were low, values between .11 and .2 were moderate, and values between .21 and the upper boundary of .5 were high. Appendix S3 contains our detailed methodological rationale for the proposed thresholding.

Results

In general, the 172 pairs of modal response patterns presented a very high inter-rater agreement (κ = .94), and the few disagreements were easily solved. The results of the first participants' subset

(n = 114) showed a good fit of both structures to the data with adequate error parameters, as shown in Table 5.

[INSERT TABLE 5 HERE]

Among the β parameters of the Movement structure, Item 5 showed a moderate value (β = .13; range: .11–.2); Item 8 (β = .36; range: .21–.5) showed a high value (Table 6).

[INSERT TABLE 6 HERE]

However, the estimated η parameters were extremely small for all items, indicating the probability of a false positive was particularly low. Even among items belonging to the ProsInt structure, all η parameters were very low. Finally, the only item that obtained high β estimates was Item 16 (β = .27), which investigated a set of highly specific behaviors. In other words, with few exceptions, all items presented low false positive and negative rates.

The results of Study 1 indicate the possibility of defining a procedure to address some critical issues related to the construction of an observational instrument as first step of the BDO method. In particular, a rater-friendly operationalization of the behaviors and their integration according to FPA methodology resulted in the construction of a nonadaptive observational checklist. The testing procedure revealed that the constructed model fit real data adequately. The missing steps are the application of the model to conduct an adaptive procedure for observation and its subsequent testing in the field. In Study 2, we implemented the checklist built and tested in Study 1 in an adaptive fashion by defining an adaptive observation algorithm. Finally, we tested the new computerized tool's accuracy and efficiency with real raters (Study 3).

Study 2

In the last two decades, several algorithms have implemented adaptive assessment instruments starting from psychometric or mathematical frameworks. For instance, various algorithms were coded according to the item response theory and mainly implemented with self-report measures (Fliege et al.,

2005; Gibbons et al., 2012; Michel et al., 2018). The adaptive system usually applied to FPA instruments is the so-called Adaptive Testing System for Psychological Disorders (ATS-PD; Donadello et al., 2017). The ATS-PD was coded within the FPA theoretical framework; therefore, it can account for the deterministic and probabilistic features of a clinical structure. In other words, ATS-PD uses all the parameters estimated from the application of the BLIM (i.e., probabilities of the clinical concepts π_c, the false negative β_q, and the false positive η_q rates of each item) to create an adaptive instrument. In this regard, the ATS-PD procedure has been tested only on self-report measures investigating obsessive-compulsive disorder (Donadello et al., 2017) and major depressive episodes (Spoto et al., 2018). In this study, we applied it within the BDO to implement the adaptive version of the previously refined checklist.

Methods

Behavior Driven Observation Algorithm

An adaptive algorithm implemented using ATS-PD and here applied within the BDO is based on three main rules: the questioning rule, the updating rule and the stopping rule. Appendix S4 describes the formal aspects of the algorithm's functioning.

The BDO algorithm selects items to observe according to the questioning rule. In particular, the algorithm selects the most informative item. In other words, it selects the item that provides the largest amount of information regardless of the collected response. Whenever two or more items are eligible or observation, the algorithm selects one randomly.

The algorithm collects the answer to the proposed question (i.e., "yes" or "no") and then applies the updating rule. This rule states that if an item receives a positive answer from the rater, the probability that the final output (i.e., the clinical concept) contains that item increases, and the probability decreases for those outputs that do not contain this item. This symmetric reasoning is valid for a collected negative response. It is important to note that, even for the algorithm, error parameters

play a key role because they directly influence the extent to which an item can update the aforementioned probabilities. If an item has low error parameters, it will produce a substantial modification of the probability distribution of all clinical concepts (see Appendix S4). In other words, the lower the error parameters for each item, the higher the reliability of the collected responses, and, therefore, the higher the accuracy and efficiency of the algorithm in delineating the final output.

The algorithm continues asking questions and updating concepts' probabilities until it reaches a stopping criterion**.** In particular, this stopping rule is satisfied when all items are either very unlikely to be included in the final output or are almost certainly included in the final output. Therefore, none of the items contributes substantially to the definition of the final output. Once the algorithm reaches this criterion, it stops the assessment, stores the response pattern and its corresponding clinical concept, and is ready to start a new assessment. An instrument built using the BDO method can administer multiple observations. In this study, the instrument performed 15 observations in this study.[2] After the last observational sample, the algorithm ends the entire assessment, calculates the modal response pattern, and defines the corresponding modal clinical concept. In the end, the BDO's algorithm generates the output, which consists of the modal response pattern M, its estimated clinical concept C_M (which usually coincides with M), the list of the symptoms related to C_M, their probability estimates, and the number of questions required to end the assessment. Figure 2 shows an example output provided by the BDO algorithm for a single session of a simulated assessment of the Movement subscale. The first section of the output contains the number of items selected by the algorithm to complete the assessment out of the total number of items in the scale. In this example, the algorithm asked five of the nine items in the Movement subscale (see tables S1 and S2 for details); furthermore, in this section of the output is reported the number of items that received an affirmative response; in the case at hand, three items out of the five asked received a positive answer. The second section of the output reports information about

[2] The BDO algorithm is suitable also for a single observation, if necessary.

the items included in the clinical concept and its associated probability value. In the case at hand, the clinical concept is $C = \{2,4,5,9\}$ with a probability value of .98. Notice that, even if the algorithm collected only three affirmative responses, the final clinical concept C contains four items. This is due to the prerequisite relations (see General methods section) among those four items. The last section of the output reports, first, the attributes that the patient may present with their associated probability; second, the attributes that the patient did not prove to present, with their associated probability. In the case at hand, the subject displayed three out of six attributes investigated by the Movement subscale: a reduction in head movements, spontaneous movements and gesture, with high probability of occurrence for each of them (i.e., 1, .99 and 1). On the other hand, the procedure suggests that the attributes "reduction in facial expressivity", "decreased reactivity to the environment", and "rigid posture" (with probability 0, .5 and 0, respectively) should not characterized the patient.

[INSERT FIGURE 2 HERE]

We coded the BDO algorithm in R (R Core Team, 2018) and implemented it in Shiny R (Chang et al., 2018) for research purposes.[3] In addition, we tested the algorithm's accuracy and the efficiency, as described below.

Simulation Design

We designed a simulation study to test the BDO algorithm's accuracy and efficiency when reproducing the original data. To accomplish these goals, we coded the algorithm to run again the observational assessments used to test the nonadaptive checklist. In particular, we adaptively simulated the remaining subset of 58 response patterns (i.e., obtained from their 15 observational samples) and used these response patterns to define the simulated modal response patterns for each patient. Finally, we compared the simulated and original modal patterns to have a measure of accuracy, and we used the

[3] An R package containing the entire algorithm is in production phase. Examples of the used functions are available upon request.

number of suggested items as a measure of efficiency.

Outcome Measures

To measure efficiency, we calculated the average number of suggested items within each sample of observation and across all 15 samples. We expected an average lower number of suggested items for the adaptive instrument compared to its nonadaptive counterpart's 20 items (i.e., 9 items for the Movement structure and 11 for the ProsInt structure).

We tested accuracy by analyzing two symmetric distances. The first was the distance between the modal response patterns we obtained using the nonadaptive and adaptive versions of the tool. Such distances corresponded to the number of discrepant answers between the two modal response patterns and represented the first index of accuracy. The second symmetric distance was the number **of** discrepant answers between the clinical concepts (i.e., the final outputs) derived from both versions of the instrument. This represented the second index of accuracy. Appendix S5 contains a formal description of such indexes. We expected small distances, which indicated more similarities. Higher distances corresponded to a relevant inconsistency in the generated information between the two outputs (Spoto et al., 2018). We expected most of the simulated modal response patterns to be equal to the original ones for both Movement and ProsInt structures.

Results

Table 7 displays the main results concerning the algorithm's efficiency. The algorithm completed the entire assessment, suggesting checking fewer items compared to the nonadaptive version of the instrument. In particular, the algorithm simulated and completed the Movement structure assessment by suggesting an average of 5.1 items out of 9 (*SD* = 0.3) per observation sample. In terms of overall observation, the assessment suggested only 57% of items across the 15 observational samples. Concerning the ProsInt subscale, the algorithm asked an average of 7.08 items out of 11 (SD = 0.4) per observation sample. Consequently, observation across the 15 samples was completed by

suggesting only 64% of items.

[INSERT TABLE 7 HERE]

Table 8 displays the distances between the modal response patterns obtained from both versions of the observational tool. The BDO algorithm perfectly simulated most of the original modal response patterns. In particular, for 95% of the modal response patterns simulated by the BDO, the symmetric distance between the original and the simulated response patterns was zero.

[INSERT TABLE 8 HERE]

The algorithm found a limited number of modal response patterns whose symmetric distance was not zero. This occurred for only three modal response patterns in both structures: for each of them, the algorithm simulated a modal response pattern whose symmetric distance was equal to one item.

For the second accuracy index (related to clinical concepts), the symmetric distance was zero for all comparisons between the concepts of the Movement structure, meaning that the modal patterns calculated from both versions of the instruments converged into the same clinical concept. The same result was found within the ProsInt structure, where only one comparison led to a symmetric distance different from zero, namely equal to two items. In sum, the results suggested that the BDO algorithm completed the assessment accurately and optimized it in real time.

Study 3

As a final step, we conducted a pilot study to test the tool implementation according to the BDO procedure. We expected that real raters using the adaptive checklist could reach modal response patterns **consistent** with those obtained through the nonadaptive checklist and save time.

Methods

Raters

Four female raters, different from those of Study 1, were recruited for this last study. Their ages ranged between 22 and 31 years old (M = 26, SD = 3.79). Two raters (R1 and R2) were experienced

psychotherapists: R1 was trained in cognitive behavioral therapy, while R2 was trained in constructivist psychotherapy. The other two raters (R3 and R4) were psychology students working on their master's theses.

Sample

A subsample of 10 patients was randomly selected from the clinical group of Study 1. All patients in this sample were diagnosed with schizophrenia. This sample was equally distributed for sex, and participants' ages ranged between 24 and 67 years old (M = 45.52, SD = 11.87). All patients presented at least one negative symptom and were treated with first- (20%) or second-generation (80%) antipsychotics. All patients gave consent for this study.

Stimuli

The stimuli for this study were the videos of the 10 patients included in Study 3 that were collected for Study 1.

Training

Each rater attended a brief training session before the experimental phase. The experimenter explained and discussed each item belonging to the nonadaptive checklist with the raters. Great attention was devoted to items describing multiple behaviors, and items whose estimated error parameters in Study 1 were moderate or high. Then, the four raters were made aware that (a) each video sample was independent from the others, and (b) no time sequence was followed in the samples' presentation. Finally, raters watched pilot videos (of both patients and controls) and scored the nonadaptive checklist for them. Their modal response patterns were matched with those obtained from Study 1 and Cohen's κ were calculated. Once an average κ of .8 was reached, the training was stopped, and the experimental phase began. Four videos per rater were necessary to reach this threshold.

Experimental Phase

We asked each rater to watch each of the 10 videos twice, then complete the checklist once in

its adaptive form and once in its nonadaptive form. The order of adaptive and nonadaptive versions was randomized among raters. One week was set as the inter-observational interval between the adaptive and nonadaptive checklist completions. Each rater watched the video 70 cm away from a 21-inch screen of an iMac 8.1 with standard resolution of 1680 × 1050. During the observation, participants wore headphones to avoid distraction. The experimenter instructed each rater to watch each video sample until its conclusion and, only at that point, to complete the checklist and press a button to watch the next sample. Raters were also asked not to interact with the experimenter. Raters were blinded to the particular form of the checklist they used (i.e., adaptive or nonadaptive). Thus, to make the two administrations more similar to one another, the items of the nonadaptive version of the checklist were administered in a random order, which varied in every sample. To reduce the risk of biased responses caused by fatigue (Haidet et al., 2009), a break was planned after the evaluation of each patient. Moreover, no more than two patients per day were observed by raters.

Data analysis

The intraclass correlation coefficient (ICC; Fisher, 1992) was used as a measure of intra-rater agreement (Koo & Li, 2016) to test whether the adaptively collected modal response patterns converged with those defined via the nonadaptive checklist. The *irr* package (Gamer et al., 2012) in the R statistical software was used to compute ICC and its 95% confidence interval. The time-saving efficiency of the adaptive checklist was tested using a generalized linear mixed-effects model (GLMM). The mean time needed to administer the checklist was used as the response variable, the instrument version (adaptive vs. nonadaptive) was set as its predictor (i.e., GLMM's fixed effect), and the intercept for each patient was a random factor. The *lme4* package (Bates et al., 2015) was used to calculate the GLMM, while the *car* package (Fox & Weisberg, 2011) was used to obtain the *p* values. The decision process leading to this analytic plan is described in Appendix S6.

Results

Table 9 contains the results of the efficacy and efficiency tests for each rater. In general, each rater showed high intra-rater agreement, ranging from .75 to .87 (*M* ICC = .82; *SD* = .05; 95% CI = [.53, .92]). Expert raters were more **consistent** and took less time to fill out the checklists.

[INSERT TABLE 9 HERE]

Regarding general efficiency, a significant effect of the instrument version emerged ($\chi^2(1) = 141.18$, $p < .001$). We observed that administering a nonadaptive checklist required more time on average than its adaptive counterpart across the 15 samples (time difference: 6 min 28 s, $t(5) = 11.88$, $p < .001$, Cohen's $d = 1.97$). This result was similar for all raters (see Table 9). These results suggest that using an adaptive checklist built through a procedure such as BDO can lead to accurate data across multiple observations with significant savings in time.

General Discussion

Observational assessment can provide experts with information that is not detectable by other assessment modalities (e.g., nonverbal behavior, dynamics of interaction). In psychology and psychiatry, this information is extremely useful for completing exhaustive descriptions of patients' symptomatology. Nonetheless, in recent years, the rate of research and new developments in observational tools for psychological assessment has decreased, probably due to difficulties related to observation such as time consumption and inter-rater agreement.

Our methodological study attempted to introduce an adaptive observational method called BDO, which succeeded in conducting a more efficient and equally accurate observational assessment, compared to traditional observation methods. BDO implements the FPA methodology in the observational assessment framework by using the one-zero sampling procedure for multiple observations. For this paper, BDO was applied to the evaluation of nonverbal behaviors related to some negative symptoms frequently observed in psychotic-spectrum disorders, mainly schizophrenia. As with typical FPA applications, the BDO algorithm (i.e., the clinical concept) had appreciable clinical

relevance because it can be used to set a therapeutic plan targeted at a patient's specific set of symptoms.

In its first part, the application of BDO overlaps that of FPA and consists of construction of clinical context, definition of the clinical structure, and probabilistic testing based on administration of an observational task in a nonadaptive fashion. In the application described in this article, two contexts were built: one for prosody and prosocial behaviors, and another for Movement behaviors. These contexts produced two clinical structures counting 40 Movement and 180 ProsInt clinical concepts. Test results for these structures showed good fit indexes of the models to the data as well as adequate values for both the β and η parameter estimates for each item. These results are encouraging because they suggest that obtaining valid assessment models is possible even for observation, with low probabilities of false positive or negative assessments when observing multiple behaviors. One of the main advantages of defining and testing an instrument using BDO is the flexibility of this method. BDO's flexibility results in a very efficient way to cope with both inter-rater disagreement on item-attribute assignments and changes in the attributes selected for investigation (these could correspond to either possible theoretical modifications in the definition of the disorder, or to different theoretical perspectives of the experts.). Whenever disagreements among experts on item-attribute assignments occur, two possibilities are provided by BDO. The first possibility refers to the solution of disagreements by discussion between the experts; this is the solution adopted in our study. The second possibility is the selection of the item-attribute assignment that fits the data best. On the other hand, changes in the detection of the relevant attributes for a disorder can easily be implemented by slightly modifying the clinical context (the Boolean matrix) and verifying the effects of the modifications on model fit.

Furthermore, the different techniques and methods applied by the BDO algorithm yielded some interesting outcomes from a clinical point of view: (a) The application of the one-zero sampling for

multiple observations allowed us to monitor for the occurrence of several behaviors in short intervals of time, reducing the risk of primacy or recency effects; (b) the randomization of observational samples could have reduced effects, such as anchoring, halo, and early impression; and (c) the collection of modal response patterns from each rater gave us the chance to use a single and reliable datum for each person to test the structures, overcoming the debated criticisms of one-zero patterns (Dunkerton, 1981; Rhine & Linville, 1980; Smith, 1985).

All these advantages sum to the intrinsic advantage of the adaptive administration of the tool, which solves one of the most crucial drawbacks of observation: time consumption. The last two parts of BDO application consist of combining the two structures and the parameter estimates into two modules (one for the ProsInt structure and one for the Movement structure) of adaptive assessment tools. This is aimed at helping to accurately complete an observation and reduce its time, even with real raters. In general, the time reduction is positively affected by two main factors: the size of the structure and the item error parameter values. The larger the structure, the weaker the relations among the items; therefore, the adaptive procedure would be less efficient. On the other hand, the lower the error parameter values, the more reliable the information collected through each item administration, and, therefore, the more efficient the adaptive assessment. In the applications described in this paper, the parameters were low enough that in the adaptive administration, a strong reduction of the requested items was observed, and a reliable reconstruction of the modal patterns was obtained. The responses' convergence was corroborated even in the last pilot study, where high intra-rater agreement was observed across four different raters regardless of their clinical experience. Finally, a modest reduction in terms of evaluation time was observed when the adaptive checklist was used. This result is reasonable considering that it is the first trial of a new procedure. Indeed, the final aim of the BDO is the construction of an online adaptive observation tool. The online adaptive modality would require fewer multiple observation samples, collapse the observation and evaluation phases, and take complete

advantage of BDO's potential. From this perspective, our results are encouraging.

The studies included in this paper present some limitations: For instance, the response scale of the proposed checklist is binary (yes/no). The implementation of a rating scale checklist could be useful for collecting information about the gravity of a specific behavior, and this is a topic of research within FPA methodology. Another partial limitation is the number of observations conducted to test the structures, even if the collected data are adequate for reliable estimation of the error parameters. It is common practice, in any case, to assume uniform distribution of the concepts as the starting point for an adaptive procedure. Because the collection of a sufficient amount of data to reliably estimate concept probabilities is impossible (it would be necessary to collect more than 1,000 observations conducted by at least two different raters), this solution makes application of the adaptive assessment feasible, reliable, and efficient.

A final limitation is the number of samples per observation. While the 15 samples can solve the problem related to the reliability of the response patterns, it can be argued that such a number of samples is still time-demanding and should be reduced**.** In fact, Study 3 showed how a statistically significant reduction in time occurred but did not reflect the relevant savings in terms of observable items shown in the Study 2. As a first step, this work focused on the intended efficiency as a reduction of the number of asked items. More effort must be made to achieve clinical time saving. Current studies are focused on solving these issues and finding the minimal number of samples for obtaining accurate and reliable outputs, with the goal of creating multiple adaptive online observations that could almost completely solve the time consumption issue.

Beyond these limitations, it is important to stress some implications of our study. The BDO algorithm was allowed to merge with and take advantage of several techniques and methods: the formal and metric characteristics of the FPA, the use of multiple observations from the one-zero sampling method, the possibility of applying all of them through modal response patterns, and the calibration of

adaptive systems such as the ATS-PD. As the results show, all these factors made it possible to develop a procedure that could define easy-to-use adaptive observational instruments and provide clinicians with help not only in conducting observations, but also with gathering sensitive information. In particular, data contained within the clinical concepts are potentially the most important clinical elements of an instrument defined by BDO. As shown in Figure 2, the clinical output of this checklist consists of a list of several elements: the number of observed items during the entire assessment out of the total number of items in the checklist, the corresponding concept within the structure and its probability, and the precise list of the observed behaviors related to specific clinical symptoms, and their probability of occurrence. Each of these elements can be used by clinicians during the assessment phase, providing them with more exhaustive data to formulate a case. For instance, the number of items observed can be used as a patient's raw score; the list of the specific symptoms can discriminate between patients with the same score but different symptomatology. In this way, a specific and detailed therapeutic plan can be set. Moreover, the list of symptoms can be used to focus on the core aspects to be treated. Finally, the clinical output of observational instruments built with BDO can be used for two other important applications. On the one hand, BDO instruments can be used to train raters because their output is immediate and can be examined and discussed among trainers and trainees. On the other hand, the clinical concepts can be used as an index of treatment efficacy, as the increase or decrease of symptoms' number across multiple assessments can give clinicians an idea about what (and how much) is changing thanks to treatment.

To conclude, the BDO can be used to help experts during complex observational assessments by defining instruments that provide them with a clinical output that immediately conveys the specific symptoms presented by patients. This information can be used to precisely formulate cases and define personalized treatment strategies to help patients with specific critical conditions. Furthermore, our results pave the way for other attempts at implementing observations with computerized adaptive

686 algorithms.

687

References

Altmann, J. (1974). Observational study of behavior: Sampling methods. *Behaviour , 49*(3-4), 227-266. doi: 10.1163/156853974X00534

American Psychiatric Association [APA]. (2000). *Diagnostic and statistical manual of mental disorders IV -Text Revision*. Washington, DC: American Psychiatric Association.

American Psychiatric Association [APA]. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5*. Washington, DC: American Psychiatric Association.

Bates, D., Maechler, M., Bolker, B. and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67(1),* 1-48. doi:10.18637/jss.v067.i01.

Brüne, M., Sonntag, C., Abdel-Hamid, M., Lehmkämper, C., Juckel, G., & Troisi, A. (2008). Nonverbal behavior during standardized interviews in patients with schizophrenia spectrum disorders. *The Journal of nervous and mental disease*, *196*(4), 282-288.

Cantor, N., & Mischel, W. (1979). Prototypes in person perception. In *Advances in experimental social psychology* (Vol. 12, pp. 3-52). Elsevier.

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2018). shiny: Web application framework for r [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=shiny (R package version 1.1.0)

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1-38.

Doignon, J.-P., & Falmagne, J.-C. (1999). *Knowledge spaces*. New York: Springer.

Donadello, I., Spoto, A., Sambo, F., Badaloni, S., Granziol, U., & Vidotto, G. (2017). Ats-pd: An adaptive testing system for psychological disorders. *Educational and Psychological*

Measurement, *77*(5), 792-815. doi:10.1177/0013164416652188

Dowiasch, S., Backasch, B., Einhäuser, W., Leube, D., Kircher, T., & Bremmer, F. (2016). Eye movements of patients with schizophrenia in a natural environment. *European Archives of Psychiatry and Clinical Neuroscience*, *266*(1), 43-54. doi: 10.1007/s00406-014-0567-8

Dunkerton, J. (1981). Should classroom observation be quantitative? *Educational Research*, *23*(2), 144-151.

Falmagne, J.-C., & Doignon, J.-P. (1988). A class of stochastic procedures for the assessment of knowledge. *British Journal of Mathematical and Statistical Psychology*, *41*(1), 1-23.

Falmagne, J.-C., & Doignon, J.-P. (2011). *Learning spaces*. New York: Springer.

First, M. B. (2014). Structured clinical interview for the dsm (scid). The encyclopedia of clinical psychology, 1-6.

Fisher, R. A. (1992). Statistical methods for research workers. In *Breakthroughs in statistics* (pp. 66-70). Springer.

Fliege, H., Becker, J., Walter, O., Bjorner, J., Klapp, B., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research*, *14*(10), 2277-2291. doi: 10.1007/s11136-005-6651-9

Fox, J. and Weisberg, S. (2011). *An {R} Companion to Applied Regression, Second Edition.* Thousand Oaks CA: Sage. URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Gaebel, W. (1989). Visuomotor behavior in schizophrenia. *Pharmacopsychiatry*, *22*(suppl 1), 29-34.

Gamer, M., Lemon, J., & Fellows, I. (2012). *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84. https://CRAN.R-project.org/package=irr

Gamer, M., Lemon, J., & Fellows, I. (2012). *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84. https://CRAN.R-project.org/package=irr

Ganter, B., &Wille, R. (1999). *Formal concept analysis: mathematical foundations*. Berlin-Heidelberg:

Springer Verlag.

Geerts, E., & Brüne, M. (2009). Ethological approaches to psychiatric disorders: focus on depression and schizophrenia. *Australian & New Zealand Journal of Psychiatry*, *43*(11), 1007-1015.

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2012). Development of a computerized adaptive test for depression. *Archives of general psychiatry*, *69*(11), 1104-1112.

Goodenough, F. L. (1928). Measuring behavior traits by means of repeated short samples. *Journal of Juvenile Research*, *12*(230), 35.

Goodman, L. A. (1961). Snowball sampling. *The annals of mathematical statistics*, 148-170.

Granziol, U., Bottesi, G., Serra, F., Spoto, A., & Vidotto, G. (2017). New perspectives on the assessment of the social anxiety disorder: The formal psychological assessment. *Journal of Evidence-Based Psychotherapies*, *17*(2), 53-68.

Granziol, U., Spoto, A., & Vidotto, G. (2018). The assessment of nonverbal behavior in schizophrenia through the formal psychological assessment. *International Journal of Methods in Psychiatric Research*, *27*(1). doi: 10.1002/mpr.1595

Groth-Marnat, G. (2009). *Handbook of psychological assessment*. John Wiley & Sons.

Haidet, K. K., Tate, J., Divirgilio-Thomas, D., Kolanowski, A., & Happ, M. B. (2009). Methods to improve reliability of video-recorded behavioral data. *Research in nursing & health*, *32*(4), 465-474.

Hawes, D. J., Dadds, M. R., & Pasalich, D. (2013). Observational coding strategies. *The oxford handbook of research strategies for clinical psychology*, 120-141.

Haynes, S. N., & O'Brien, W. H. (2000). Principles and strategies of behavioral observation. In *Principles and practice of behavioral assessment* (pp. 225-263). Springer.

Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The positive and negative syndromescale (panss) for

759 schizophrenia. *Schizophrenia bulletin*, *13*(2), 261-276.
760 Kølbæk, P., Blicher, A. B., Buus, C. W., Feller, S. G., Holm, T., Dines, D., . . .Østergaard, S. D.
761 (2018). Inter-rater reliability of ratings on the six-item positive and negative syndrome scale
762 (panss-6) obtained using the simplified negative and positive symptoms interview (snapsi).
763 *Nordic Journal of Psychiatry*, *72*(6), 431-436. (PMID: 30037286) doi:
764 10.1080/08039488.2018.1492014
765 Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation
766 coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155-163.
767 Lee, D., Barak, A., & Uhlemann, M. (1999). Forming clinical impressions during the first five minutes
768 of the counseling interview. *Psychological Reports*, *85*(3 PART 1), 835-844.
769 Lord, C., Luyster, R., Gotham, K., & Guthrie, W. (2012). Autism diagnostic observation schedule
770 second edition (ados-2) manual (part ii): Toddler module. *Torrance, CA: Western Psychological*
771 *Services.*
772 Lord, C., Rutter, M., DiLavore, P., Risi, S., Gotham, K., & Bishop, S. (2012). Autism diagnostic
773 observation schedule second edition (ados-2) manual (part 1): Modules 1-4. *Torrance, CA:*
774 *Western Psychological Services.*
775 Martin, P., Bateson, P. P. G., & Bateson, P. (1993). Recording methods. In *Measuring behaviour: an*
776 *introductory guide* (pp. 84-100). Cambridge University Pres.
777 Michel, P., Baumstarck, K., Lancon, C., Ghattas, B., Loundou, A., Auquier, P., & Boyer, L. (2018).
778 Modernizing quality of life assessment: development of a multidimensional computerized
779 adaptive questionnaire for patients with schizophrenia. *Quality of Life Research*, *27*(4),
780 1041-1054. doi: 10.1007/s11136-017-1553-1
781 Mumma, G. (2002). Effects of three types of potentially biasing information on symptom severity
782 judgments for major depressive episode. *Journal of Clinical Psychology*, *58*(10), 1327-1345.

783 doi: 10.1002/jclp.10046
784 Petersen, M. A., Groenvold, M., Aaronson, N., Fayers, P., Sprangers, M., Bjorner, J. B., et al. (2006).
785 Multidimensional computerized adaptive testing of the EORTC QLQ-C30: Basic developments
786 and evaluations. *Quality of Life Research*, 15(3), 315-329.
787 Pino, M. C., Spoto, A., Mariano, M., Granziol, U., Peretti, S., Masedu, F., . . .Vidotto, G. (2018).
788 Formal psychological assessment for autism spectrum disorder diagnosis: A new methodology
789 to build an adaptive testing system. *The Open Psychology Journal* , *11*(1), 112-122.
790 Powell, J., Martindale, B., Kulp, S., Martindale, A., & Bauman, R. (1977). Taking a closer look: time
791 sampling and measurement error. *Journal of Applied Behavior Analysis*, *10*(2), 325-332. doi:
792 10.1901/jaba.1977.10-325
793 R Core Team. (2018). R: A language and environment for statistical computing [Computer software
794 manual]. Vienna, Austria. Retrieved from https://www.R-project.org/
795 Ray, J. M., & Ray, R. D. (2008). Train-to-code: An adaptive expert system for training systematic
796 observation and coding skills. *Behavior research methods*, *40*(3), 673-693.
797 Reiser, M., & VandenBerg, M. (1994). Validity of the chi-square test in dichotomous variable factor
798 analysis when expected frequencies are small. *BritishJournal of Mathematical and Statistical*
799 *Psychology*, *47*(1), 85-107. doi: 10.1111/j.2044-8317.1994.tb01026.x
800 Rhine, R. J., & Linville, A. K. (1980). Properties of one-zero scores in observational studies of primate
801 social behavior: The effect of assumptions on empirical analyses. *Primates*, *21*(1), 111-122.
802 Serra, F., Spoto, A., Ghisi, M., & Vidotto, G. (2015). Formal psychological assessment in evaluating
803 depression: a new methodology to build exhaustive and irredundant adaptive questionnaires.
804 *PloS one*, *10*(4), e0122131.
805 Serra, F., Spoto, A., Ghisi, M., & Vidotto, G. (2017). Improving major depressive episode assessment:
806 A new tool developed by formal psychological assessment. *Frontiers in psychology*, *8*, 214.

Smith, P. K. (1985). The reliability and validity of one-zero sampling: misconceived criticisms and unacknowledged assumptions. *British Educational Research Journal*, *11*(3), 215-220.

Spoto, A., Bottesi, G., Sanavio, E., & Vidotto, G. (2013). Theoretical foundations and clinical implications of formal psychological assessment. *Psychotherapy and psychosomatics, 82*(3), 197-199.

Spoto, A., Serra, F., Donadello, I., Granziol, U., & Vidotto, G. (2018). New perspectives in the adaptive assessment of depression: The ATS-PD version of the QuEDS. *Frontiers in Psychology*, *9*(JUL). doi: 10.3389/fpsyg.2018.01101

Spoto, A., Stefanutti, L., & Vidotto, G. (2010). Knowledge space theory, formal concept analysis, and computerized psychological assessment. *Behavior Research Methods*, *42*(1), 342-350.

Spoto, A., Stefanutti, L., & Vidotto, G. (2012). On the unidentifiability of a certain class of skill multi map based probabilistic knowledge structures. *Journal of Mathematical Psychology*, *56*(4), 248-255.

Spoto, A., Stefanutti, L., & Vidotto, G. (2013). Considerations about the identification of forward-and backward-graded knowledge structures. *Journal ofMathematical Psychology*, *57*(5), 249-254.

Spoto, A., Stefanutti, L., & Vidotto, G. (2016). An iterative procedure for extracting skill maps from data. *Behavior Research Methods*, *48*(2), 729-741. doi: 10.3758/s13428-015-0609-9

Stefanutti, L., & Robusto, E. (2009). Recovering a probabilistic knowledge structure by constraining its parameter space. *Psychometrika*, *74*(1), 83-96.

Stefanutti, L., Spoto, A., & Vidotto, G. (2018). Detecting and explaining blim's unidentifiability: Forward and backward parameter transformation groups. *Journal of Mathematical Psychology*, *82*, 38-51.

Suen, H. K., & Ary, D. (1984). Variables inuencing one-zero and instantaneous time sampling outcomes. *Primates*, *25*(1), 89-94.

Suen, H. K., & Ary, D. (1986). A post hoc correction procedure for systematic errors in time-sampling duration estimates. *Journal of psychopathology and behavioral assessment*, *8*(1), 31-38.

Troisi, A., Pompili, E., Binello, L., & Sterpone, A. (2007). Facial expressivity during the clinical interview as a predictor functional disability in schizophrenia. a pilot study. Progress in Neuro-*Psychopharmacology and Biological Psychiatry*, *31*(2),n475-481. doi: 10.1016/j.pnpbp.2006.11.016

Van Rossum, G., et al. (2007). Python programming language. In *Usenix annual technical conference* (Vol. 41, p. 36).

Wainer, H. (2000). CATs: Whither and whence. *ETS Research Report Series(2)*.

Weisberg, H., & Weisberg, H. F. (1992). *Central tendency and variability* (No. 83). Sage.

Wille, R. (1982). Restructuring lattice theory: An approach based on hierarchies of concepts. In I. Rival (Ed.), *Ordered Sets* (pp. 445-470). Dordrecht: Reidel.

Tables

Table 1

Example of Clinical Context

q/a	a_1	a_2
q_1	1	0
q_2	1	1

Note. q_1 = "the person's posture points downward;" q_2 = "both posture and gaze of the person point downward;" a_1 = curved posture; a_2 = downward gaze.

Table 2

Example Modal Response Patterns

	S_1	S_2	S_3	S_4	S_5	Mo
Item 1	0	1	1	0	1	1
Item 2	1	0	1	0	0	0
Item 3	0	1	0	0	0	0
Item 4	1	0	1	1	1	1
Item 5	1	0	1	1	1	1

Note. *Mo* = modal response pattern. S_1–S_5 denote observational samples.

Table 3

Clinical Subcontexts for Movement and ProsInt Scales

Movement

Item	A_1	A_2	A_7	A_8	A_{10}	A_{12}
1	0	0	0	0	0	1
2	0	0	1	0	0	0
3	0	0	1	0	1	1
4	0	0	0	1	0	0
5	0	0	1	1	0	0
6	1	0	0	0	0	0
7	1	0	0	1	0	0
8	0	1	1	1	0	0
9	1	0	0	0	1	0
10	0	1	0	0	0	0

3a

ProsInt

Item	A_3	A_4	A_5	A_6	A_9	A_{10}	A_{11}	A_{13}	A_{14}
11	1	0	0	0	0	0	0	0	0
12	0	1	0	0	0	0	0	0	0
13	0	1	1	0	0	0	0	0	0
14	0	0	0	1	0	0	0	0	0
15	0	1	1	1	0	0	0	0	0
16	0	1	0	1	0	0	0	0	0
17	0	0	0	0	1	1	0	0	0
18	1	0	0	0	1	1	0	0	0
19	0	0	0	0	0	0	1	0	0
20	0	0	0	0	0	1	0	1	0
21	0	0	0	0	0	1	0	0	1

3b

Table 4

Demographic Characteristics

Group	*n*	Age Range		Sex	Occupation	Marital Status
		M	± 1 *SD*			
Nonclinical	134	26	3.40	19-67	F: 100 M: 34	Student: 125 Working student: 7 Employed: 1 Retired: 1	Single: 130 Married: 3 Divorced: 1
Clinical	38	42	7.35	24-67	F: 10 M: 28	Student: 1 Unemployed: 30 Employed: 6 Retired: 1	Single: 24 Married: 14
Diagnosis: SCH	25	44.5	13	24-67	F: 5 M: 20	Student: 1 Unemployed: 22 Employed: 1 Retired: 1	Single: 18 Married: 7
Diagnosis: MDD	5	51	6.5	46-63	F: 3 M: 2	Unemployed: 2 Employed: 3	Single: 1 Married: 4
Diagnosis: BD	8	36	6.5	27-45	F: 2 M: 6	Unemployed: 6 Employed: 2	Single: 5 Married: 3

Note. SCH = schizophrenia; MDD = major depressive disorder; BD = bipolar disorder. Descriptors of the clinical group are further split across diagnoses.

Table 5

Model Fit Indexes for Movement and ProsInt Structures

Structure	χ^2	*df*	Bootstrap *p*
Movement	43.44	454	.06
ProsInt	23.10	1846	.11

Table 6

Error Parameter Estimates for Movement and ProsInt Structure Items

Movement			ProsInt		
Item	**β**	**η**	**Item**	**β**	**η**
1	.1	.01	**10**	<.001	<.001
2	<.001	<.001	**11**	<.001	<.001
3	.1	.01	**12**	<.001	<.001
4	<.001	<.001	**13**	<.001	.052
5	.133	<.001	**14**	<.001	<.001
6	<.001	.068	**15**	.273	<.001
7	.076	<.001	**16**	<.001	<.001
8	.364	<.001	**17**	<.001	<.001
9	<.001	<.001	**18**	.1	.01
			19	.1	.01
			20	.1	.01

Note. β = false negative; η = false positive.

Table 7

Number of Suggested Items in Nonadaptive and Adaptive Versions of Observational Tool

		Efficiency	
Structure	**Instrument**	***n***	***N***
Movement	Nonadaptive	9	135
Movement	Adaptive	5.10	76.50
ProsInt	Nonadaptive	11	165
ProsInt	Adaptive	7.08	106.20

Note. n = mean number of suggested items per sample; N = mean number of suggested items across 15 observation samples.

Table 8

Accuracy of BDO Algorithm

		Accuracy	
Structure	**Δ**	**d = 0***	**d = 1****
Movement	0.05	55	3
ProsInt	0.05	55	3

Note. Δ = average symmetric distance between original and simulated modal response patterns.
*Number of modal simulated patterns at distance 0 from original.
**Number of simulated modal patterns at distance 1 from original.

Table 9

Accuracy and Efficiency of BDO Adaptive Tool, Real Raters

Rater	*Mean ICC*	95% *CI*	Nonadaptive Minutes*	Adaptive Minutes	Difference
R1	.87	.60, .92	17.30 (2.37)	12.55 (3.31)	4.55 (3.28)
R2	.83	.60, .92	17.09 (3.04)	10.46 (3.04)	6.55 (2.32)
R3	.85	.49, .89	32.36 (4.18)	24.05 (3.16)	8.07 (3.01)
R4	.75	.44, .88	32.28 (5.16)	24.20 (2.36)	7.43 (6.33)

Note. *M*s and *CI*s of ICCs across 10 patients for each rater.
*In the last three columns, the average (± *SD*) amount of time required for the completion of the nonadaptive and adaptive forms are reported, together with their average (± *SD*) differences among the 10 patients.