Università degli studi di Padova

Department of General Psychology

Ph.D. Course in Psychological Sciences

XXXI Series

The Behavior-Driven Observation

Definition and development of an adaptive observational assessment

Coordinator: *Prof. Giovanni Galfano*

Supervisor: *Ch.mo Prof. Giulio Vidotto*

Co-Supervisor: *Dr. Luisa Sartori*

Ph.D. student: *Umberto Granziol*

Abstract

The observation in psychological assessment provides clinicians with a variety of useful insights about the symptoms of mental disorders. Nonetheless, the application of observational instruments has decreased during the last years, mainly due to their administration complexity and time consumption. A consequence of this general reduction in application is that some innovations fruitfully applied by other psychological assessment instruments, such as the self-reports, are still unexplored. For instance, little focus has been put on the possibility of implementing observational measures with adaptive algorithms. In observational assessment, these algorithms have been applied only by some software developed for observers training; their implementation in observational assessment instruments is still an open challenge. The aim of the present Ph.D. project is to develop an observational adaptive instrument able to help clinicians to generate accurate behavioral response patterns reducing, simultaneously, the time of the observational assessment.

The definition of such an instrument has been a sequential process that started from a deep analysis of the items that should be observed, followed by the consideration of how to observe each of them. These first issues were accounted in Chapter 1, in which an overview of the literature was performed in order to examine all the features necessary to adequately conduct an observational assessment. A specific attention was dedicated on the possible biases that could affect raters, leading to higher probabilities of false positive and negatives on the observed behaviors. Finally, the state of the art relative to the application of adaptive algorithms in observational assessments was introduced and discussed.

The second step toward the definition of the expected instrument consisted in defin-

ing a non adaptive checklist evaluating the behaviors of a mental disorder, possibly based on a formal methodology. In Chapter 2, the Formal Psychological Assessment (FPA) was introduced, describing its deterministic and probabilistic features. FPA is a methodology allowing to define assessment instruments starting from the relation between a set of items and a set of clinical issues of a disorder. In Chapter's end, it was shown how FPA could be extended also to observational assessment composed by multiple measures.

In Chapter 3, the FPA was applied to develop the paper-and-pencil version of the final checklist. The negative symptomatology of schizophrenia was selected as the target mental disorder. A set of 138 items describing nonverbal behaviors was selected from instruments frequently used in the evaluation of schizophrenia. This list was then mapped to a list of 14 negative symptoms, selected in both scientific literature and DSM-5. The application of formal and logical steps provided by FPA led to a final checklist of 22 items, divided into two subscales, exhaustively investigating the 14 negative symptoms. In particular, it emerged how the mapping between items and investigated symptoms defined a deterministic model of assessment in which the clinician could be informed not only of which negative symptoms are evaluated by each item, but also of the relations among items.

This model of assessment was later validated, in order to convert it into a probabilistic model that would have been correctly implemented into an adaptive instrument. In Chapter 4, the validation procedure is described. 172 videos of clinical interviews were observed by two independent raters, who filled the new checklist during one-zero sampling observations and generated modal response patterns for both subscales. Such patterns were used to apply the Basic local Independence Model (BLIM), a probabilistic model allowing to estimate the global fit indexes of the checklist and the false

positive and negative rates for each item. Results showed adequate fit indexes for both subscales of the checklist with acceptable error rates for each item, which were extremely low especially in respect to false positive rates.

The obtained probabilistic model of assessment and its parameters estimates were then used to calibrate an observational adaptive algorithm. In Chapter 5, the first version of the Behavior-Driven Observation (BDO) was introduced, namely the adaptive observational checklist proposed by the present project. After its formulation, the BDO was tested on real data by a simulation study in which both its accuracy and efficiency were examined. Results showed how the BDO algorithm was able to accurately reproduce almost all the non adaptive response patterns, with an average reduction by 38% of suggested items to complete the entire assessment.

Finally, the accuracy and the efficiency of the BDO were tested during real observations, in order to understand if the BDO led to accurately replicate the non adaptive response patterns when used by human raters, with similar savings in terms of efficiency. Two independent trained raters observed twice the videos of twenty patients with a diagnosis of schizophrenia with negative symptoms, filling the two checklist's versions during observations. The observations on the same patient were far one week from each other. A very good intra-rater agreement emerged for each rater, suggesting both a good coherence over time of raters and a good ability of the BDO to replicate the response patterns of its non adaptive counterpart. Likewise, encouraging results were found in regard to BDO's efficiency: The savings in terms of suggested items were the same of the simulation study, for each rater; moreover, such savings corresponded to a reduction of the observational time.

Taken together, the results of this Ph.D. project suggest that is possible to define an adaptive observational checklist able to help clinician to collect information not

otherwise detectable with other assessment modalities. The BDO, in fact, could guide the observation by suggesting which behavior should be observed, taking into account the false positive/negative rates for each behavior. In this way, the accuracy of the final clinical output is increased as well as the efficiency of its generation. Such a clinical output could provide clinicians with a comprehensive set of information, such as the precise response pattern observed during the observation, the most plausible symptoms related to that response pattern and their probability values. All this information, in turn, can be finally integrated with other ones collected from different assessment instruments (e.g., interview, self-report), in order to have a broader frame of patient's condition and, maybe, set an individualized treatment.

Contents

Chapter 1

Observational assessment: State of the art

1.1 Introduction

Clinical observation is not just the act of observing the behavior of a person. It is a systematic data collection procedure, aimed at gathering and further deepening information that are difficult to detect through other assessment methods. The way a person acts or his/her nonverbal behavior are only a subset of information that psychologists can use in order to have a better frame of a person's condition. The set of observational techniques is extremely heterogeneous and covers different psychological areas: For instance, direct observation (both participant and non participant) is frequently used in school and developmental psychology (Bauman, 2015; Hintze, 2005); the Behavioral Observation has been widely applied in clinical psychology and psychotherapy (Hawes, Dadds, & Pasalich, 2013) since the second half of the twentieth century, even as a basis of evidence-based treatments or intervention researches (Snyder et al., 2006);

finally, the use of ethological coding systems as observational instruments is becoming very frequent in psychiatry, especially with psychopathologies like depression and schizophrenia (Troisi, 1999; Troisi, Pompili, Binello, & Sterpone, 2007).

Despite these evidences, the reduction in the use of observation in psychological assessment is not *just a right sensation*. It is a fact: As pointed out by Baumeister, Vohs, and Funder (2007), the trend of studies including behavioral measures is strongly decreased, a phenomenon observed not only for social and personality psychology (Baumeister et al., 2007) but also in clinical psychology (Hawes et al., 2013). The reasons of such a reduction can be found in the observation's nature. The practical and methodological issues related to observational measures make it difficult (and time consuming) both the coding and the analysis of the collected data (Heyman, 2001). These issues can be summarized into two categories: The former includes observer's biases such as halo and anchoring effects. The occurrence of any of these biases have a direct consequence, namely an increased chance of making false positive or false negative errors when deciding on the presence/absence of a behavior. The latter critical issue concerns measurement and reliability. Estimating the reliability of an observation is a complex procedure that needs to consider several factors such as the calculation of the precise length of each sample of time, decisions about the number of observers and, undoubtedly, adequate statistical analyses.

Furthermore, there is another critical point stressed by different authors (Roberts, Chan, & Torous, 2018; Rosenberg, Glueck Jr, & Bennett, 1967; Yanagita, Becirevic, & Reed, 2016): The use of computerized observational instruments. As far as we know, the concept of "computerized" in observational assessment refers to either the electronic version of observational instruments or to software helping the behavioral video analysis, like The Observer software developed by Noldus (Noldus, 1991). All these

applications, although useful in behavioral analysis, are neither able to substantially reduce the administration's time nor to provide clinicians with a comprehensive output describing the behavioral response pattern and the symptoms linked to that pattern. This kind of output, indeed, can be provided by applying in observation the computerized adaptive assessment's logic. Basically, a computerized adaptive instrument suggests the clinician which item should be administered based on the response to the previous items. It mimics the adaptive logic of an interview (Spoto, Serra, Donadello, Granziol, & Vidotto, 2018), where only a subset of questions is asked to a person on the basis of his/her answers. The advantage of using an adaptive observation could be the same found for tests and questionnaires: Ease of administration, reduction of administration's time and an output providing clinicians with (i) the total score, (ii) the person's response pattern and (iii) the related symptoms. The use of adaptive algorithms has a consolidated tradition in psychological testing (Serra, Spoto, Ghisi, & Vidotto, 2017; Wainer, 2000), while its application in observations is still unexplored. The aim of the present doctoral Thesis is to define and develop an observational adaptive instrument able to both minimize the cons and the biases of observational measures and extend the advantages of adaptive questionnaires to observations. In order to reach this goal, an observational checklist evaluating the nonverbal behaviors related to negative symptoms of schizophrenia will be developed by means of the Formal Psychological Assessment methodology (FPA; Spoto, Bottesi, Sanavio, & Vidotto, 2013) and implemented into its adaptive version, called the Behavior-Driven Observation (BDO).

The Thesis is organized as follows: after the above general introduction, Chapter 1 will describe (i) the basic concepts (ii) the methodological issues and (iii) the adaptive assessment's literature related to observational assessment; Chapter 2 will introduce

and describe the FPA; in Chapter 3, the definition and refinement of the non adaptive version of the observational checklist will be described, while in Chapter 4 its validation will be discussed, with a particular focus on the errors parameters' estimation. In Chapter 5, the checklist's implementation into its adaptive version (i.e., BDO) and the results of a simulation study testing its accuracy and efficiency will be presented. Chapter 6 will describe a study in which the BDO is tested by two expert clinicians during real observations. In the last Chapter, all the results, limits and future perspectives will be discussed.

1.2 Observational assessment: features, techniques and instruments

The aims of an observation can be different: The evaluation of interactions between parents and children (Bauman, 2015); the detection of the relationships between a behavior and the environment in which it occurs (Briesch, Volpe, & Floyd, 2018); the indirectly measurement of cognitive processes that are not otherwise measurable by means of tests and questionnaires (Ehrmantrout, Allen, Leve, Davis, & Sheeber, 2011); finally, the observational assessment of a pathological symptoms (Brüne et al., 2008). As several researchers in observation's field suggest, a precise set of rules in order to reliably use an observation should be followed, independently of observation's goals (Groth-Marnat, 2009; Hawes et al., 2013; Haynes & O'Brien, 2000):

1. Definition of the type of behaviors to observe;

2. Definition of a specific behavioral coding system;

3. Operationalization of each behavior to observe. Each behavior must be clearly defined, in order to maximize its probability of being correctly observed. On this purpose, it is straightforward that the text of the item should be as short as possible, avoiding sentences with double negative, redundancies, etc.;

4. Consideration of the users' (i.e., observers) experience and the target population;

5. Decision on the use of audio/video recordings;

6. Check for some observation parameters (e.g., setting, sampling strategy).

In regards to the last point, the parameters that should be taken into account are: Type of data, accessibility and representativeness of the behaviors, setting, type of observer and sampling strategy (Altmann, 1974; Hawes et al., 2013; Haynes & O'Brien, 2000). The first parameter refers to the type of data. Behaviors can be coded based on the extent to which they can capture a fine-grained or a wide information. In fact, a behavior can be molecular or microsocial (Dion et al., 2011; Dishion & Granic, 2004), when it is a discrete and a mutually exclusive unit of behavior. An example of molecular behavior can be "The patient nods", as a measure of social respondence. Molecular behaviors must be salient and detectable with minimal inference. In ethological observations, they are referred as events (Altmann, 1974), since they are instantaneous and evaluated only in terms of occurrence/absence. On the other hand, if the behavior has a longer time-span, it is accounted as a global behavior. The global version of the previously mentioned behavior could be "The patient shows social respondence". In ethological studies they are referred as states (Altmann, 1974). Global behaviors allow the observer to measure not only the frequency, but also the duration of those behaviors. Molecular and global behaviors are not mutually exclusive: As pointed

out by Hops, Davis, and Longoria (1995), in study comprising both of them, the two behavioral categories seems to be highly correlated (Hops et al., 1995).

Once the behaviors have been chosen, it is important to check their accessibility and representativeness (Hawes et al., 2013). A behavior can be more or less accessible, that is prone to be observed, both within and between disorders. This is the case of all the behaviors that occur during a psychotic crisis. The accessibility, defined as the likelihood of the occurrence of a behavior coupled with its difficulty to be observed (Johnston, Pennypacker, & Green, 2010), depends also on the observational setting: For instance, the behaviors observed during an interaction among children at school could not be the same if they were recorded in a laboratory. A set of behaviors must be representative (i.e. typical) of the phenomenon of interest. A researcher should check for both these prerequisites, since representativeness does not imply accessibility. For instance, a tic could be representative during a compulsion but inaccessible without a stimulus or a thought eliciting it.

After the aforementioned parameters' check, the following step is the setting selection (Hartmann & Wood, 1990). In naturalistic observations, the behaviors are observed and recorded in the natural environment of the observed individuals. These observations allow to observe dynamics and behaviors that can be generalized to the real world, even though the control on possible intervening variables is minimal. This aspect lead sometimes researchers to put minimal restrictions during these observations (Dishion & Granic, 2004). When the observation is performed in artificial settings, where the majority of variable are controlled or manipulated, it is called analogue observation (Heyman & Slep, 2004). The pros and cons compared to the naturalistic observation are reversed: Since the possibility of controlling variables and standardizing results is high, a reduced spontaneity can be expected.

Once the setting is selected, the observer needs to clarify his/her role within the observation. On one side of the continuum, there are the participant observations, in which the observer simultaneously interacts with the observed person and evaluates his/her behaviors. On the opposite side, there are the nonparticipant observations, in which the interaction with the observed person is absent. Even in this case, the pros and the cons are reversed: In the participant observation, time and economic costs are reduced, since the observation and the data collection are online and performed by the same rater; this aspect is the disadvantage of nonparticipant observations, where the observation, data collection and analysis are three separate moments, a separation increasing the costs. On the other hand, the absence of interaction with the people observed allows the observer to better focus on target behaviors, leading to more accurate and complex data. In fact, the cognitive load required by participant observations makes necessary the use brief and simple instruments, minimizing the accuracy of the collected data (Haynes & O'Brien, 2000).

After the observation selection, a sampling strategy (Haynes & O'Brien, 2000) is required. The sampling strategies allow the researcher for correctly recording and quantifying the behaviors (Hawes et al., 2013), taking into account the type of behaviors (i.e, molecular or global), the setting (naturalistic, analogue) in which they are expressed and the role of the observer (participant or nonparticipant). All the sampling strategies are described below, including their pros and cons (Altmann, 1974; Hawes et al., 2013; Haynes & O'Brien, 2000; Powell, Martindale, Kulp, Martindale, & Bauman, 1977):

- **Event sampling** consists in recording only the occurrence (i.e., presence/absence) of target behaviors if they occur along an interval of time. It is frequently used in

the observation of behaviors whose instances have clear temporal edges and very high/low frequency rates. This sampling procedure provides data that can be measured in terms of both rates and frequency. Event sampling is easy to design and gives the chance of exhaustively assess all the target behaviors. Nonetheless, it can be applied only with very salient behaviors.

- **Interval sampling** is a method in which the observation is split into equal-length intervals. The interval duration depends on the research design: It has been noted that shorter intervals (e..g., 10-15 seconds) are preferred to detect molecular behaviors, while longer intervals are preferable with global behaviors (Dishion & Snyder, 2004). There are three kinds of interval sampling:

    1. *Partial interval sampling*, in which the behavior is considered present if it occurs at least once during the interval, independently of the moment in which it appears. It is frequently used in studies focused on parent-child interaction or in single case research (Hawes et al., 2013). This method gives the chance of dealing with behaviors without specific temporal edges. On the other hand, it seems that this strategy leads to an overestimation of behaviors frequency (Powell et al., 1977).

    2. *Whole interval sampling*, in which the behavior is considered observed only if it occurs for the entire length of the interval. It has been used in studies focused on attention of children in classrooms (Dion et al., 2011). It has the advantage of providing an accurate measure of behaviors duration, even though it seems to underestimate their frequencies (Powell et al., 1977).

    3. *Event-within interval sampling* is an hybrid sampling method, since it is possible to record the number of times a behavior occurs (as in event sampling)

during a specific time interval (as in interval sampling). It has the advantage of giving a measure of the behaviors' variability over time (Hartmann & Wood, 1990).

- **Real-Time sampling**, in which both the onset and the end of a behavior are collected by means of a clock. This sampling can provide information on frequency and duration of both events and intervals. Nonetheless, its reliability decreases if applied to high-frequency behaviors, unless an electronic device is used (Haynes & O'Brien, 2000).

- **Momentary-time sampling** is a particular case of interval sampling, in which a behavior is systematically recorded during short intervals, during a day. In other words, the observer checks for the behavior occurrence only in a specific moment, during specific intervals of the day. It is used in psychiatric settings in order to understand the onset of specific behaviors during the day (Paul, 1986). Recent studies suggest that this sampling method can be better applied with either high duration behaviors or behaviors lasting about one minute (Sharp, Mudford, & Elliffe, 2015). On the other hand, this sampling method seems to underestimate behaviors' frequencies of behavior having short duration (Powell et al., 1977).

Figure 1.1 summarizes the procedure leading to select an observation's type. Finally, another sampling method has a long tradition in observational assessment, in both psychology and psychiatry. It is called **one-zero sampling** method (Goodenough, 1928) and it will be discussed in details in the following subsection.

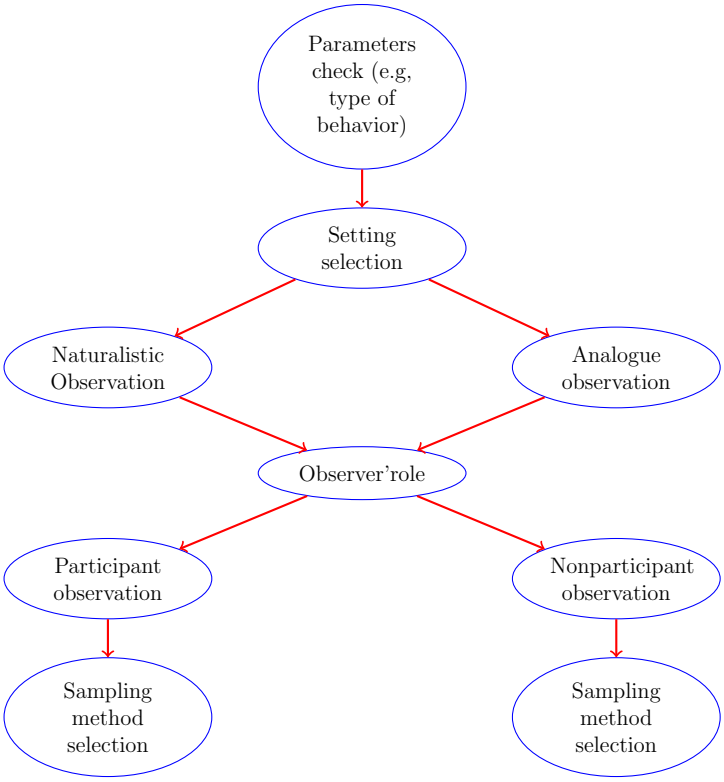


Figure 1.1: Procedure for selecting the type of observation

The One-Zero sampling method

The one-zero sampling is a time sampling method defined by Goodenough in 1928 and applied originally in the observation of animal behavior. It has been applied even in human observation since 1920's, as a method for observing children behavior. During the last decades it has been used in the ethological observations of nonverbal behavior of patients with a diagnosis of depression (Geerts & Brüne, 2009) or schizophrenia (Brüne et al., 2008; Troisi et al., 2007).

It consists in dividing an observation into n equal-length samples (e.g., 15-30 sec) and checking the occurrence/nonoccurrence of a behavior within each interval (Martin, Bateson, & Bateson, 1993). In particular, at the end of each sample (announced

usually by a beeper) the observer scores 1 if he/she observed the target behavior, independently of when it has been observed; if the behavior has not been observed, the observer scores a 0 (i.e., nonoccurrence). The measure provided by the one-zero sampling is a single score for each behavior indicating the proportion of samples in which that behavior occurred, as displayed by Table 1.1.

	$Sample_1$	$Sample_2$	$Sample_3$	$Proportion$
Behavior 1	1	1	0	**0.66**
Behavior 2	0	0	1	**0.33**
Behavior 3	0	1	1	**0.66**

Table 1.1: The score of one-zero sampling method

As pointed out by Martin et al. (1993), this proportion should not to be used as a standard measure of frequency, since a behavior can occur different times during a sample, underestimating the amount of time in which the behavior has been observed. Furthermore, one-zero sampling seems to overestimate the duration of each behavior, since the moment in which it is sampled can vary or last between consecutive samples (Martin et al., 1993). These cons created a great debate in scientific literature (Dunkerton, 1981; Leger, 1977; Powell et al., 1977; Rhine & Linville, 1980; Smith, 1985) and several authors suggested some adjustments in order to both calculate more accurate actual frequency rates and estimate the duration of each behavior. For instance, Suen and Ary (1986) suggested to calculate more correct duration and frequency estimates of a behavior starting from the number of zeroes between two consecutive and not adjacent "1"s (i.e., inter-response time, IRT; see Suen & Ary, 1984, 1986).

Despite the aforementioned critical issues, one zero sampling has undoubted advantages (Smith, 1985): It is an easy sampling method, where the scoring rule is clear and

immediate; it allows researchers to observe several behaviors within the same sample, since their occurrence is not dependent on their onset. Furthermore, it is well-known in literature that one-zero scores can provide good inter and intra-rater agreement indexes (Altmann, 1974; Rhine & Linville, 1980; Troisi et al., 2007), even for multiple behaviors.

The observational techniques described in this section are used as a methodological basis to perform correct and systematic observations. Moreover, they can pave the way to easily use observational instruments as checklists or ethograms. A general description of these tools will be now presented.

1.3 Assessment instrument during observation

During an observation, the set of behaviors observed in each interval or sample can be evaluated in different ways. For instance, a researcher can create a scoring sheet containing a grid in which the samples are in rows and the behaviors in columns. This kind of instrument is recommended when the measures of interest are frequencies or duration rates (Altmann, 1974; Groth-Marnat, 2009; Haynes & O'Brien, 2000). Such grids represent the simplest case of observational instrument.

In general, observational assessment tools can be divided on the basis of the moment in which they are administered. A group of them includes all the instruments used in real time during the observation. In clinical and developmental psychology, the most famous instrument is the Autism Diagnostic Observation Schedule (ADOS; C. Lord, Rutter, DiLavore, & Risi, 1999), a clinical and structured observation able to evaluate communication styles, social interaction skills, stereotyped and creative behaviors of children with a possible diagnosis of autism spectrum disorder (Hus & Lord, 2014). In

its last version, it is organized into five modules (i.e., Toddler, 1, 2, 3 and 4) referred to different age ranges (ADOS-2; C. Lord, Luyster, Gotham, & Guthrie, 2012; C. Lord, Rutter, et al., 2012). For each module, the observer elicits specific behaviors or observes the child's actions during specific tasks, with different scores attributed on the basis of the module. It is available in different versions (Hus & Lord, 2014), some of them presenting a good potential to be implemented into a computerized adaptive assessment (Pino et al., 2018).

Another group includes instruments that are not administered during a direct observation. At one hand, some of them do not require a sampling method. This is the case of the Child Behavior Checklist (CBCL; Achenbach & Ruffle, 2000), an observational instrument used with children and filled by their parents or teachers. Even CBCL has different versions based on chronological age (i.e., CBCL/2-3-4), composed by items referred to specific behaviors that are evaluated on a 3-point scale (i.e., 0 = Not true, 1 = Somewhat or sometimes true, 2 = Very true or often true).

On the other hand, the majority of observational instruments belonging to this second group are used with a videotaped observation, usually sampled with an interval sampling strategy. Good examples of these observational instruments are the ethograms used in clinical psychology and psychiatry. Ethograms are sets of hierarchically ordered behaviors typical of a species (Brüne et al., 2008; Geerts & Brüne, 2009), usually contained in checklists composed by dichotomous items. Ethograms are used to evaluate the occurrence/nonoccurrence of a behavior during a specific time sample. The most used ethogram in psychiatry is the Ethological Coding System for Interviews (ECSI; Troisi, 1999; Troisi et al., 2007). It is a checklist of 37 dichotomous items evaluating molecular nonverbal behaviors related to facial expression, gesture and body movements. The behaviors are clustered into seven high-order factors related to so-

cial skills (i.e., Affiliation, Submission, Prosocial, Flight, Assertion, Displacement and Relaxation). Each behavior is evaluated using a recorded observation, sampled using the one-zero sampling method. It is a gold standard in the evaluation of nonverbal behavior of people with a diagnosis of depression or schizophrenia (Brüne et al., 2008; Geerts & Brüne, 2009; Troisi, 1999; Troisi et al., 2007)

All the observational tools cited in this section are useful and well-established in their psychological areas of application. Beyond their pros and cons, they have to deal with critical issues typical of the observational assessment, a topic described in details in the next section.

1.4 Critical issues in observational assessment

Nowadays psychological assessment is considered as a multi-method process (Meyer et al., 2001) in which all the assessment techniques can contribute to understand the individual. Each of them adds unique information that makes sense only if integrated with those derived from the other assessment techniques. As a result, the clinician has all the elements to set up a personalized treatment (Fischer, 2000; A. J. Fisher & Bosley, 2015). Observational assessment has the advantage of providing information about the behavior of a person less biased by what she/he could say. This is the case of nonverbal behavior, namely a set of related information about facial expressions, gesture, prosody and body movements (Argyle, 2013) determining more than 60% of the personal communication style (Geerts & Brüne, 2009). Another advantage is the possibility of comparing what a person says with the way he/she says it, as in the case of patients with a diagnosis of schizophrenia who exhibit a dissociation between what they report to feel and what they express (Ellgring, 1986; Kring & Caponigro, 2010).

Finally, observational assessment gives the chance of observing how people interact, a typical subject matter of developmental psychology (e.g., the attachment styles of a child; Bretherton, 1992).

Unfortunately, these advantages are only one side of the coin. Beyond all the affective and cognitive biases that can affect who is observed, there are also critical issues related to both the observer and the measures she/he applies within an observation. This section will provide an overview of these critical issues.

1.4.1 Observer's bias

In cognitive psychology it is well known that the more a situation is characterized by uncertainty, the more the chance of using heuristics increases (Tversky & Kahneman, 1974). By definition, the context in which an observation takes place is characterized by a certain degree of uncertainty, since the way a person will react or respond to a stimulus cannot be predicted. Consequently, it is possible that an observer could use oversimplification strategies (Haynes & O'Brien, 2000), which can bias the way of judging the occurrence/nonoccurrence of a behavior, especially in a long and time consuming procedure like observation.

An interesting bias is the so called *anchoring effect*. It consists in giving two different judgments for the same observed case on the basis of the moment and the order in which the information are presented (Friedlander & Stockman, 1983). Once the judgment is defined, it tends to confirm itself, victim of both a confirmatory bias and the tendency of people to maintain a personal internal consistency (Cantor & Mischel, 1979; Friedlander & Phillips, 1984). In terms of observation, such bias could lead to a systematic tendency of the observer to confirm his/her judgment on occurrence/nonoccurrence of a behavior;

this tendency could be different based on the moment the observer perceives a triggering salient information. Such a bias could be stronger if it happens during the first part of the observation, since it could interact with the "five minute impression formation", another bias typical of internals and clinicians (Lee, Barak, & Uhlemann, 1999). In this case, the observer could tend to always judge a behavior as occurred if she/he observed it within the first minutes of the observation.

Another interesting bias is the *halo effect*, a bias in which specific characteristics can influence the judgment on other evaluated dimensions without an empirical justification (Cooper, 1981). Halo effect is very frequent with nonverbal behaviors: As suggested by Mumma (2002), nonverbal behaviors are salient information that can influence the judgment on symptoms' severity during a clinical evaluation of depression (Mumma, 2002). In case of an observation, once the halo effect is active, the occurrence probability of all the behaviors coherent with the halo could increase, leading to overestimate them and, viceversa, underestimate all the incoherent behaviors. It seems that halo effect is caused by a representative heuristic that enhances a selective attention in favor of the formers (Haynes & O'Brien, 2000).

All the aforementioned biases and heuristics have as a direct consequence the increased probability of making two types of error, namely the false positives and false negatives. Assume, for instance, that a behavior has been erroneously judged as observed (i.e., a false positive) at the beginning of the observation. If such a behavior has been perceived as very salient, the anchoring effect could occur and the following judgments on that behavior could be anchored to the first attribution. Consequently, the likelihood of future false positives on that item could be constant or, eventually, increase. Likewise, the influence that the halo effect can create could lead to increase or decrease the gravity of the following behaviors, modifying the threshold of judgment

applied by the observer. Independently of the psychological assessment a clinician decides to use, the false positive/negative issue should always be taken into account, by using assessment methodologies that estimates such errors parameters and consider them during an assessment (Bottesi, Spoto, Freeston, Sanavio, & Vidotto, 2015; Serra, Spoto, Ghisi, & Vidotto, 2015). An insight on the variables that could enhance biases' formation in observers was given by Repp, Nieminen, Olinger, and Brusca (1988), who collected a list of these elements: (a) Especially in direct observation, the observed person can react to the presence of the observer; furthermore, (b) if two observers perform the observation in the same setting, it is possible that they interact in some way, creating a drift effect, that is the change of attribution style during the observation; (c) time intervals could be too short/long or the setting could be inappropriate since eliciting only a specific set of behavior; (d) finally, if the observer starts to think that the behavioral pattern of a person is predictable, he/she will tend to fall into the anchoring effect (Repp et al., 1988).

Even if there is a consistent number of bias and causes, the use of observation should not be discouraged, since it is possible to reduce them by applying some recommendations (for details, see Repp et al., 1988):

- Intensive training for observers (better if balanced on sex), who must not interact neither between each other, nor with the experimenter;
- The experimenter should not interact with observers and should be as unobtrusive as possible. Moreover, the experimenter must not reveal the research hypotheses;
- Use of recording devices, in particular video, in order to make the observation more systematic;

- Use of simple observational codes or instruments, with good psychometric properties, especially in terms of reliability.

In regards to this last point, a particular attention should be paid on the selection of the measure of reliability. This aspect will be the deepened below.

1.4.2 Measurement and reliability

Observational data are usually analyzed according to five units of measurement (Hawes et al., 2013; Hintze, 2005): Frequency, that is the number of times a behavior occurs; Duration, intended as the amount of time in which the behavior is observed; Latency, namely the amount of time before the onset of a behavior; Intensty, namely the strength with which the behavior is showed; finally, there are the permanent products, used as a measure of the effects of a behavior (e.g., environmental objects or scores to a performance). As described in previous sections, all of these units of measurement can be over/underestimated, on the basis of the the selected sampling method (Altmann, 1974; Hawes et al., 2013; Martin et al., 1993). As a consequence, the reliability of the collected data is reduced.

Actually, the over/underestimation issue is not related to the sampling method *per se*, but it is related to the duration and the number of samples typical of the specific sampling. For instance, the first studies using one-zero sampling recommended short samples (e.g., 15 sec) repeated 20 times (Altmann, 1974) or more in studies dealing with human observation (Troisi, 1999). This approach is shared by the interval sampling methods in which the sample duration can be even shorter (5-10 sec) in order to better analyze molecular behaviors (Dion et al., 2011; Dishion & Granic, 2004). There seems to be a general agreement on the samples' duration, stating that it is advisable a big

number of very short intervals to collect reliable data (Altmann, 1974; Hawes et al., 2013; Repp et al., 1988). Unfortunately, this logic cannot be applied to all kind of behaviors, that can have a different duration; moreover, in all interval-like sampling methods, the choice of very short intervals can be ineffective, since the aim is to detect several behaviors. A more moderate position is taken by momentary-time sampling, in which longer time samples are used (e.g., from 30 seconds to 1 minute) during an observation lasting on average 30 min (Brown et al., 2009; Sharp et al., 2015).

Duration and amount of samples are not the only factors affecting the measure of reliability of observational data. In a recent paper, Hallgren (2012) highlighted the factors that are worth of attention when the inter-rater reliability (IRR) of observational data is estimated. In case of multiple raters, it should be *a priori* decided if all the subjects (or only a subset of them) will be used to estimate the IRR. Likewise, it should be decided if all the rates will perform the observation or only a few or them (or just one rater). Once the raters issue is solved, a fundamental check should be done on the assessment tool used and, importantly, an intensive training on it should be attended. One of the best training practice is to fix an inter-rater threshold and train the observers until the threshold is reached. Furthermore, it is important not to select indexes that estimate IRR not taking into account the chance, as for the percentages agreement indexes. In observational studies, the reliability index of choice remains the Cohen's κ, either in its classical (J. Cohen, 1960) or in its modified version that corrects κ for problems related to prevalence and biases in the marginal distributions (Eugenio & Glass, 2004; Hallgren, 2012). The coefficient κ is a standardized measure providing an estimate of the amount of agreement between two (or more) raters, corrected for the agreement that would be reached by chance (J. Cohen, 1960), as displayed by the

equation below:

$$\kappa = \frac{P(o) - P(e)}{1 - P(e)} \tag{1.1}$$

where $P(o)$ is the proportion of observed agreement (calculated by summing all the agreements between the raters) and $P(e)$ is the proportion of agreements expected by chance (calculated by multiplying the marginal frequencies of each rater's rating and dividing this product by the total number of observations). This correction makes κ more precise than other indexes (Hintze, 2005), such as the simple percentage agreement, that focus only on the observed raters' agreement without estimating the weight of the chance. Cohen's κ ranges from -1 (complete disagreement) to +1 (perfect agreement). Several intervals have been proposed in literature in order to correctly interpret the magnitude of the found agreement: For instance, Gelfand and Hartmann (1975) suggest that a κ of .21 to .40 can be considered as sufficient, .41 to .60 as moderate, .60 and above as substantial. In this regard, other authors suggest to use more conservative thresolds (Kirppendorff, 1989; Landis & Koch, 1977). All the issues described could be used by researchers who wants to perform observational studies both maximizing the probability of obtaining valid and reliable data and minimizing the cons of a procedure that remains, indeed, complex and time-consuming. Actually, the matter of time in observation could be conceptualized in a different way by considering the features of the so called adaptive assessment, as it will be described in the next section.

1.5 Perspectives in observation: Adaptive assessment

During a psychological assessment, the number of constructs that clinicians have to reliably evaluate is extremely high (Fliege et al., 2005). This phenomenon is adequately represented by the amount of items that some tests suggest to administer, defining an high-demanding situation in which the cognitive burden of the clinician could be very heavy (Gibbons et al., 2008; Michel et al., 2018). A natural consequence of such a burden is an higher degree of false positives and false negatives on the reported answers, especially during an observation (Yanagita et al., 2016). The use of electronic and automated version of psychological instruments reduce this problem, with the best results obtained by the computerized adaptive assessment ones. The following paragraphs will briefly overview the literature on the applications of these computerized instruments, with a particular focus on observational assessment.

1.5.1 State of the art

A computerized adaptive assessment is an evaluation procedure in which the items to be administered are selected based on the previous responses. This kind of assessment is usually performed by using electronic devices such as PC or tablet, helping clinicians to reach an accurate diagnosis asking less items (Petersen et al., 2006; Spoto et al., 2018). The majority of such adaptive systems are used in the area of testing and are usually called Computerized Adaptive Testing (CAT; Wainer, 2000). A field of application of CATs is the evaluation of students' knowledge. The systems belonging to this category can be clustered on the basis of the formal theory on which they were

developed: At one hand, there are systems based on the Item Response Theory (IRT; F. M. Lord, 1980) such as the Intelligent Evaluation System using Tests for TeleEducation (SIETTE; Conejo et al., 2004), used to help teachers during evaluations in educational environments; likewise, systems that combine IRT with specific statistical methods, like the maximum likelihood, are frequently used in order to cluster people during examination (Eggen & Straetmans, 2000). On the other hand, other systems use different approaches such as Bayesian statistics (EDUFORM; Nokelainen et al., 2001) or mathematical psychology theories to assess students' knowledge. This is the case of the Assessment and LEarning in Knowledge Spaces (ALEKS; Grayce, 2013), a system assessing the amount of information that a student knows on a specific topic (i.e., his/her *knowledge state*) and train him/her to what he/she is ready to learn. Recent studies shown how the application of systems like ALEKS can strengthen the learned information of a student, reducing the memory drop described by the Ebbinghaus forgetting curves (Matayoshi et al., 2018).

Another field of application of CAT is psychological testing. The advantage of using CAT in psychological assessment can be found on their system of functioning: In fact, as pointed out by Spoto et al. (2018), they follow the logic used by expert clinicians during an interview, who adaptively select the questions to ask on the basis of the collected answers, defining a complex inferential system which leads him/her to a diagnostic framework (Spoto et al., 2018). During the last decades, several IRT-based adaptive questionnaires have been developed for the evaluation of several disorders. Michel et al. (2018) have recently developed a multidimensional adaptive questionnaire to assess the quality of life in schizophrenia (Michel et al., 2018). In primary health care settings, Gardner, Kelleher, and Pajer (2002) developed the adaptive version of the Pediatric Symptoms Checklist (PSC; Jellinek et al., 1988) a questionnaire filled by par-

ents assessing their children symptoms and behaviors. Simms et al. (2011) developed the CAT-PD, namely a CAT questionnaire for the evaluation of personality disorders. Gibbons et al. (2008) developed the CAT version of the Mood and Anxiety Spectrum Scales (Dell'Osso et al., 2002). Finally, CAT applications have been implemented to assess depression. Beyond the work of Gibbons, in 2007 Yong, Awang Rambli, and Anh developed a self-help instrument that interacts with patients with a diagnosis of depression and provided them advice about their condition. An adaptive system has been developed also for the Center for Epidemiologic Studies Depression Scale (CES-D; Finkelman, Smits, Kim, & Riley, 2012; Smits, Finkelman, & Kelderman, 2016). Finally, Spoto et al. (2018) implemented an adaptive version of the Qualitative-Quantitative Evaluation for Depressive Symptomatology Questionnaire (QUEDS; Serra et al., 2017), a questionnaire built on theories different from IRT such as the Formal Psychological Assessment (Bottesi et al., 2015; Spoto, Bottesi, et al., 2013).

All the cited instruments represent a step forward in psychological assessment, since help clinicians to easily and efficiently perform an assessment in clinical settings where the resources in terms of both time and personnel are limited. Nonetheless, little is known about the use of computerized adaptive systems in observational assessment, as will be described in the next section.

1.5.2 Application in observation

The need of including electronic systems within observation is not new. Several authors suggested that the standardization and the quantification of observation could lead to a more precise detection of behaviors (Hawes et al., 2013; Kahng & Iwata, 1998; Repp et al., 1988; Rosenberg et al., 1967). Computerized-assisted observations

can bring improvements in terms of setup, duration of data entry, precision in inter-rater reliability calculation, accuracy of the observation itself and costs (Tapp et al., 2006). For these reasons, a lot of interesting computerized software for data collection during observation have been developed (Yanagita et al., 2016). For instance, the *INTerval MANager* (INTMAN; Tapp et al., 2006) is a software which helps observers to collect data during non continuous observations, showing a good potential in reducing time and costs in multiple measurements; moreover it automatically calculates inter-rater agreement indexes. On of the most widely used software is *The Observer* (Noldus, 1991), a computerized event coder available for different platforms. In The Observer, the user is helped in each moment of the observation analysis: In fact she/he can select the type of event recording, set up intervals duration, nest behaviors into mutually exclusive categories and analyze them online, evaluating the occurrence of each behavior separately. The software offers the possibility of calculating frequencies and duration of both single and combined behaviors, with a time sample precision of 0.1 seconds. Moreover, each comment or data is stored as an independent variable to be later analyzed and saved in different formats.

Finally, smartphones or smartwatches are becoming tools able to monitor behavioral patterns of people, since they are able to track movements or actions. Moreover, these devices can be used as collector of biomarkers able to prevent the onset of maniac or psychotic episodes (Roberts et al., 2018).

The step forward done by the development and application of these software and apps is undoubtedly true and can be considered a revolution in observational assessment. Nonetheless, it is self-evident how all of them are used to calculate frequencies, duration or can be used as computerized assistants in raw data collection of behaviors. In clinical settings, indeed, an observational assessment should provide the psy-

chologists/psychiatrists with the same amount of useful information as provided by a questionnaire or an interview. It should be able to collect responses to make possible the formulation of a diagnosis or, at least, a set of information ready to be integrated with others collected with different instruments. It should be adaptive, to provide such information in less time, while preserving accuracy. As far as we know, the only adaptive instrument related to the observation is the *Train-to-Code* software (J. M. Ray & Ray, 2008). It is an adaptive observer system that trains observers not only in coding a behavior (student mode) in terms of occurrence, frequency and duration, but also in teaching to teach (instructor mode); in other words, this software helps the user to set up a coding system that the software itself will be able to teach. Its adaptivity can be found on the methods of teaching that, in line with the operant response-shaping instruction model (R. D. Ray, 1995), gives decreasing feedback during the test phases. Similarly to ALEKS system, the software helps the observer to move from his/her actual level of training to the next one when she/he is ready to step forward; if not, it will continue to give feedback and explanations until the step is ready to be done. Although the Train-to-Code is the first software that use an adaptive system in observational field, it can be applied only to train people to observe; the detection of behaviors for clinical purposes is beyond its aims.

In sum, the present section examined the literature of computerized adaptive systems in psychological assessment. It emerges that the use of these instruments is limited to psychological testing. When moving toward observations, the use of computerized instrument is limited to data collection software or adaptive programs applied in observers training. As a result, the use of adaptive system has not been completely extended to observational instruments, making it impossible to evaluate and monitor patients' behavior (Trull, 2007). At least, until now.

1.6 The point and the aim

The observational assessment is a complex methods of evaluation in which several elements should be taken into account. In fact, the selection of target behaviors and sampling methods is only the top of the iceberg. The researcher should control also that the behavioral coding can be implemented into an observational able to correct reliability data. Likewise, she/he should check for and minimize possible biases belonging to human nature, in order to avoid high error rates and a substantial risk of misinterpretation of the collected data. All combined with the absence of adaptive algorithms that could make the challenge more affordable. Trying to hypothesize an observational assessment instrument able to maximize the pros and minimize the aforementioned critical issues, it should be:

- Built on a clear behavioral code. All the behaviors, independently if molecular or global, must be self-explaining, mutually exclusive and defined by experts in topic of interest.
- Less prone to false positive/negative errors. This goal should be reached both controlling for all the possible observers' biases and estimating the prevalence and error rates by means of mathematical and psychometric techniques.
- Prone to be taught by using a systematic and well-defined training procedure.
- Applied during a sampling method with a precise set of samples.
- Valid, reliable, accurate and efficient.
- Adaptive, or that can be implemented into a computerized adaptive instrument.

- Providing clinician with a score not suggesting biased frequencies or duration, but the trend of the behavioral pattern of a person and the symptoms related to that behavioral pattern.

The present doctoral Thesis is a first attempt to define such a computerized observational instrument. In order to reach this goal, each point of the suggested list has been taken into account and tested applying different techniques and methodologies, that will be explained along the next Chapters. The first step toward the development of the proposed adaptive instrument dealt with the definition of its non adaptive version. This process involved the definition of a behavioral code, namely a list of behaviors necessary and sufficient to evaluate a mental disorder (i.e., the negative symptomatology of schizophrenia in this project) during an observation. As specified by the first points of the aforementioned features, such a list should be composed by precise behaviors, defined by expert clinicians and clear enough to reduce observers' interpretations or biases. All these features, in turn, should be supported by a methodology able to provide a formal/psychometric basis to the final instrument, paving the way for its adaptive implementation. The following Chapter will describe in details the Formal Psychological Assessment, the methodology that made possible the implementation of a behavioral code (defined in Chapter 3) into a non adaptive checklist.

Chapter 2

Formal Psychological Assessment

2.1 Measurement theories of psychological instruments

Measurement in Psychology is one of the most debated topic in modern science (Kline, 2014; Michell, 1997, 2000; Vessonen, 2018). As several authors suggest, attributing a quantitative logic to something that is not necessarily physical, like the temperature measured by a thermometer, requires the adoption of a solid metric system (Kline, 2014; Michell, 1997). Most importantly, such a psychometric system must be able to deal with the unlikelihood for a psychological measure to share the same quantitative properties of physical measurement systems, such as the additivity. This is what happens in psychological assessment, in which tests or interviews scores are used to measure a latent construct assuming a quantitative relation between the obtained score and that latent construct (Michell, 2000). In the history of psychological measurement, several theories gave a solid contribution toward a quantitative measurement ap-

proach: After the first insights emerged by the researches of Weber, the Psychophysics of Fechner (Fechner, 1860) and Spearman's factor analysis (Spearman, 1904), the first fundamental conceptualizations on both the levels and the units of measurement in psychology were provided by Stevens. He defined the four scales of measurement used nowadays in psychology (i.e.,the nominal, ordinal, interval and ratio scale), on the basis of the assumption according to which measuring consists in establishing specific homomorphisms between empirical and numerical relational systems (Stevens, 1946, 1951, 1957). Such an approach was deepened by the Relational Theory of Measurement (RTM; Narens & Luce, 1993; Suppes & Zinnes, 1963; Suppes, Krantz, Luce, & Tversky, 1989), which better analyzed the specific relational axioms which guarantee the relations between empirical and numerical systems. The studies of Stevens marked the beginning of a long tradition of formal and psychometric approaches to psychological measurement, such as the Classical Test Theory (CTT; Gulliksen, 2013; Novick, 1965), the Rasch models (Rasch, 1961) and the Item Response Theory (IRT; F. M. Lord, 1980). Each of these theories represents a gold standard for the definition of psychological assessment instruments, that is the subject matter of this Chapter.

In the CTT models (F. M. Lord, 1959), the relation between the observed score and the target psychological construct can be studied starting from the following equation:

$$X = T + E \tag{2.1}$$

where X is the obtained score, T is the score that would be obtained in absence of measurement error and E is the component error that can reduce to possibility of observing T. The observed score represents, simultaneously, the pro and the cons of CTT. A typical example is provided by psychological self-report measures: At one

hand, the score of a test is used as a good basis to evaluate properties like reliability and validity; on the other hand, a "score-centered" approach could pay less attention to relevant information such as items' difficulty or individuals' abilities (Hambleton, Swaminathan, & Rogers, 1991). A fitting example of this dichotomy can be found in the definition of a short form of an instrument to be applied in screening procedures: The selection of few items could bring the advantage of an high internal consistency between them but, at the same time, such items could be redundant and their range of difficulty could be narrowed on central values. As a result, a floor effect could be expected analyzing the scores of people healthy or whose symptoms are under-threshold, while a ceiling effect could occur with people with over-threshold symptoms (Gardner et al., 2002; McHorney, 1997).

Within IRT models and Rasch models, on the contrary, both the individual ability and the items' difficulty levels are fundamental parameters for the estimation of a latent trait, since they can explain the score obtained by a person. For instance, in the One-Parameter Logistic model, that is equivalent to the Rasch Simple Logistic Model (Bond & Fox, 2013; Rasch, 1960), the score is a function of both the respondent's ability level (θ) and the items' difficulty (β). These parameters determine the probability $P_i(\theta)$ of correctly answering each item i of the test, as depicted by the equation below:

$$P_i(\theta) = \frac{exp(\theta - \beta_i)}{1 + exp(\theta - \beta_i)}.$$

Furthermore, the $\beta \backslash \theta$ parameters are fundamental, since they allow to determine precedence relations among items based on their location on the continuum referred to the latent dimension (Marsman et al., 2018). Consequently, it is possible to locate a person on the latent trait continuum according to his/her responses. The features of Rasch and IRT models make it possible to define not only more precise paper-and-

pencil psychological instruments, but also appreciable adaptive ones (Fliege et al., 2005; Gardner et al., 2002). As pointed out by Spoto et al. (2018), a critical issue of such models concerns the difficulty to implement them when the precedence relations among items are more complex than a linear order. By definition, a linear order is a relation in which, for every given triple of items (x, y, z), four main properties always hold: reflexivity ($x \leq x$), anti-symmetry (for every pair x, y, if $x \leq y$ and $y \leq x$, then $x = y$), transitivity (for every triple x, y, z, if $x \leq y$ and $y \leq z$, then $x \leq z$) and connection (for every pair x, y, either $x \leq y$ or $y \leq x$). Therefore, within a linear order all the items must be comparable (i.e., connected), with respect to this order relation. Nonetheless, in many situations, it could happen some pairs of items could not be compared. In this case, different relation among items should be considered, such as partial order relations. These kind of relations are usually studied by mathematical theories and tools dealing with lattices and posets (Birkhoff, 1937, 1940; Davey & Priestley, 2002). A methodology that takes into account these theories and implements them in order to define (adaptive) psychological instruments, going beyond classical psychometric approaches, is the Formal Psychological Assessment.

2.2 Deterministic features of FPA

The Formal Psychological Assessment (FPA; Spoto, Stefanutti, & Vidotto, 2010; Spoto, Bottesi, et al., 2013) is a methodology aimed at defining assessment instruments able to evaluate specific sets of symptoms of mental disorders during psychological or psychiatric assessments. It has been applied to different mental disorders such as social anxiety disorder (Granziol, Bottesi, Serra, Spoto, & Vidotto, 2017), obsessive compulsive disorder (Bottesi et al., 2015; Spoto et al., 2010) and depression (Serra et al.,

2015, 2017). Although the majority of applications consists in self-reports' development, FPA can be extended to each type of assessment. As introduced in section 1.6, this Thesis is a first attempt to apply FPA in order to create an observational checklist able to evaluate the negative symptoms of schizophrenia. FPA is the formal conjunction and the clinical application of two theories of Mathematical Psychology, namely the Knowledge Space Theory (KST; Doignon & Falmagne, 1999; Falmagne & Doignon, 2011) and the Formal Concept Analysis (FCA; Ganter & Wille, 1999; Wille, 1982).

An instrument defined by FPA allows clinicians to specify and analyze the relations between a nonempty set A of clinical issues (symptoms and diagnostic criteria for a disorder) and a nonempty set Q of items investigating the chosen clinical issues. The collection Q of all the items that can be administered to a person during an assessment of a specific disorder is called *clinical domain*. Practically, it can be created collecting all the items of assessment tools used to evaluate a specific mental disorder. For instance, Table 2.1 shows an hypothetical clinical domain containing items used to evaluate the major depressive disorder from a behavioral point of view:

Item	**Description**
q_1	The person shows sad facial expressions and looks downwards.
q_2	The posture of the person points downwards.
q_3	The person shows sad facial expressions. Moreover, his/her posture points downwards.
q_4	During the conversation, the person often cries.

Table 2.1: An example of clinical domain

The clinical issues investigated by the items of the clinical domain are called *attributes* and can be selected from (a) common clinical practice on the specific disorder, (b) scientific literature, or (c) clinical sources as the DSM-5 (American Psychiatric Association [APA], 2013). Table 2.2 displays an hypothetical set of attributes investigated

by the item of the previous clinical domain:

Attribute	**Description**
a_1	Gaze downwards
a_2	Curve posture
a_3	Crying
a_4	Sad facial expressiveness

Table 2.2: An hyphotetical set of attributes

The relation between the sets Q and A are depicted in the *clinical context*, formally a triple (Q, A, I) where Q is the clinical domain, A is the nonempty set of attributes and I is a binary relation I (i.e., *investigates*) between Q and A. Given an item $q \in Q$ and an attribute $a \in A$, the relation qIa holds if and only if the item q investigates the attribute a. The clinical context is represented by a Boolean matrix containing the items q_i in rows and the attributes a_j in columns; whenever an item q_i investigates an attribute a_j, the cell ij of the context will contain the value 1, otherwise a 0. Table 2.3 shows the clinical context referred to the previous sets of items and attributes.

q/a	a_1	a_2	a_3	a_4
q_1	1	0	0	1
q_2	0	1	0	0
q_3	0	1	0	1
q_4	1	0	1	1

Table 2.3: An example of clinical context

It is easy to check how the item q_2 ("The posture of the person points downwards") investigates the attribute a_2 ("Curve posture"). Beyond the attribute-item relations, the clinical context displays also the relations among items, called *prerequisite relations*. A prerequisite relation $\preceq$ between any two items $x, y \in Q$ is defined by the following rule:

$$A_x \subseteq A_y \iff x \preceq y \tag{2.2}$$

In words, x is a prerequisite of y if and only if the set of attributes investigated by the item x is a subset of attributes investigated by the item y. This means that a scenario in which the item y is endorsed by an individual who does not endorse also item x is not allowed by the model, unless some sort of error in the response behavior occurs. In the clinical context at hand, the item q_2 (i.e., "The posture of the person points downwards") is a prerequisite of the item q_3 (i.e., "The person often shows sad facial expressions. Moreover, his/her posture points downwards."). The prerequisite relations are an essential component of an adaptive assessment: Assume that q_3 is administered by the adaptive algorithm. If q_3 is endorsed by the respondent, according to equation (2.2), its prerequisite q_2 will be considered endorsed too. Consequently, the adaptive algorithm will not suggest the administration of q_2, since its response is inferred. When applied to an instrument composed by several related items, this logic could create a system of inferences that allows for (i) reducing the redundancy caused by the repetition of similar questions and (ii) increasing the assessment efficiency, since the number of asked items to complete the assessment would be substantially reduced. In Chapter 5 the functioning of such a system will be shown in details.

From both the clinical context and the prerequisite relation it is possible to define the *clinical concepts*, namely the pairs (O, S) with $O \subseteq Q$ and $S \subseteq A$ representing the set of all the items endorsed by a person and the necessary and sufficient set of attributes for endorsing such items. Any clinical concept is coherent with the relation I depicted by the clinical context. The collection of all the clinical concepts is called *clinical structure* $\mathcal{C}$. In FPA, a clinical structure is usually represented by a complete lattice containing, in each node, a clinical concept. A clinical structure, given a domain Q, can be obtained in several ways: Through a query to a set of experts (Doignon, 1994; Kambouri, Koppen, Villano, & Falmagne, 1994; Koppen & Doignon, 1990; Koppen,

1993), by defining and implementing skill maps/clinical contexts (Albert & Lukas, 1999; Doignon, 1994; Düntsch & Gediga, 1995; Heller, Augustin, Hockemeyer, Stefanutti, & Albert, 2013; Spoto et al., 2018) or by means of data driven procedures (de Chiusole, Stefanutti, & Spoto, 2017; Robusto & Stefanutti, 2014; Spoto, Stefanutti, & Vidotto, 2016). Within FPA, the clinical structure is delineated by means of the second method, namely through the implementation of a clinical context. In particular, starting from a clinical context and taking a subset of items $O \subseteq Q$ and a subset of attributes $S \subseteq A$, a clinical structure can be delineated starting from two following transformations defining the so called *Galois connection* (Spoto et al., 2010):

$$O' := \{a \in A |\, qIa, \forall q \in O\} \tag{2.3}$$

and

$$S' := \{q \in Q |\, qIa, \forall a \in S\} \tag{2.4}$$

where O' is the collection of all the attributes shared by all the items in O, while S' is the collection of all the items that investigate all the attributes in S. The pair (O, S) is referred to as a *clinical concept* if both conditions $O = S'$ and $S = O'$ are satisfied. The set O is the *extent* of the concept, while the set S is the *intent* of the concept. The bijection between O and S defines the concepts of the clinical structure. In particular, each node corresponds to a clinical concept having as extent the items endorsed by a person and, as intent, the collection of all attributes that all those items share (or investigates). This kind of intent, nonetheless, does not correspond to the set of attributes necessary and sufficient to endorse a set of items. This last is the set of attributes whose identification is the main task of FPA. A method to obtain a clinical

structure whose concepts contain the target sets of attributes has been proposed by Spoto et al. (2010): Given a clinical context (Q, A, I), authors demonstrated that such a structure can be obtained by delineating a clinical context according to the relation I' between items and attributes that is dual to I. In other words, the relation $qI'a$ holds true if and only if the relation I does not hold, that is whenever q does not investigate a:

$$qI'a \iff q\neg Ia \tag{2.5}$$

Table 2.4 displays the conversion of the clinical context (Q, A, I) [3a] into its dual (Q, A, I') [3b].

	a	b	c	d
1	1	0	0	1
2	0	1	0	0
3	0	1	0	1
4	1	0	1	1

3a

	a	b	c	d
1	0	1	1	0
2	1	0	1	1
3	1	0	1	0
4	0	1	0	0

3b

Table 2.4: Example of clinical context (1a) and its dual (1b).

Starting from the concepts of the dual clinical context (Q, A, I') and applying the Galois connections, a clinical structure whose extents are closed under intersection (i.e., $O_1 \cap O_2 \in \mathcal{C}$ for all $O \in \mathcal{C}$) is delineated. The obtained clinical structure is displayed on the left of Figure 2.1[1]. In this structure, the intent $S \subseteq A$ of each concept is the set attributes that are not investigated by any of the items in the extent of the concept. For each concept, the set of attributes necessary and sufficient to endorse all the items in the extent is the dual of S, that is $A \setminus S$. The application of this transformation returns a structure of concepts whose intents are the collections of all the attributes

[1] Both structure have been obtained by means of the software Galicia (Valtchev, Grosser, Roume, & Hacene, 2003); their layout has been adapted for graphical purposes.

necessary and sufficient to endorse all the items in the extent (Figure 2.1, structure on the right. For further details, see Spoto et al., 2010).

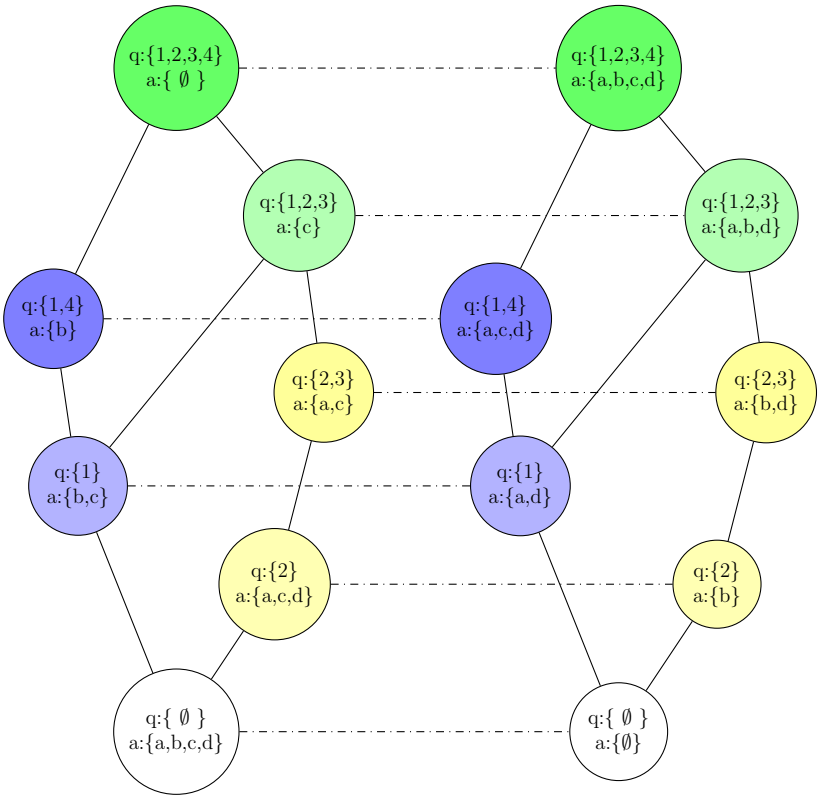


Figure 2.1: Examples of a clinical structure: on the left, each node contains a set of items and the dual set of the their investigated attributes; on the right, the exact set of attributes for each node is displayed. Items are reported in the first lines of the circles, attributes in second lines.

In this example, given the set of attributes $A = \{a, b, c, d\}$ (contained in Table 2.2) and the set of items $Q = \{1, 2, 3, 4\}$ (contained in Table 2.1) it is possible to check all the clinical concepts defined by the context (Q, A, I). Furthermore, the prerequisite relations among items can be verified as well (e.g., the item $\{2\}$ is a prerequisite of the item $\{3\}$). Building a psychological instrument from this set of items, the clinician could know exactly all the admissible clinical outcomes deriving from the administration of such an instrument. Moreover, implementing this hypothetical instrument

into its adaptive version, the clinician could complete an assessment administering less items, due to the prerequisites relations among them. In this way, the accuracy would remain the same with an increased efficiency.

2.3 Probabilistic concepts of FPA

A clinical structure similar to the one represented in Figure 2.1 is an accurate snapshot of the final clinical concepts containing specific sets of attributes investigated by specific sets of items. This structure is, indeed, a deterministic and incomplete basis for an adaptive assessment. Firstly, a precise correspondence between the observed response patterns and the latent clinical concepts cannot be assumed. Secondly, the clinical concepts could occur with different frequencies within the population. Finally, a deterministic clinical structure cannot predict the probability π_C of all the clinical concepts, given the response patterns of patients. This probability is related to both the actual frequencies of the clinical concepts in the population and two parameters, respectively the *false negative* (β) and *false positive* (η). The former refers to the probability that the patient does not endorse an item that she actually presents; in observational terms, the probability of not observing a behavior when it actually occurs. The latter parameter refers to the probability that a patient endorses an item that she does not still present; during an observation, it is the probability of observing a behavior when it has not actually occurred. When all these parameters are present (i.e.,π_C for all $C \in \mathcal{C}$, β and η for each $q \in Q$), it is possible to delineate a *probabilistic clinical structure*, formally a triple $(Q, \mathcal{C}, \pi)$ where $(Q, \mathcal{C})$ is the clinical structure and π is the probability distribution on $\mathcal{C}$. As suggested by Spoto et al. (2010), π can be estimated on a sample of patients. By means of π, each clinical concept $C \in \mathcal{C}$ is

defined by a probability of occurrence in the population. A probability distribution for each response pattern $R \subseteq Q$ is attributed through a response function assigning to R its conditional probability given that a patient is in the concept C (for all concepts $C \in \mathcal{C}$), by applying the unrestricted latent class model displayed by the equation below:

$$P(R) = \sum_{C \in \mathcal{C}} P(R|C)\pi(C). \tag{2.6}$$

The equation (2.6) describes the *Basic Local Independence Model* (BLIM; Doignon & Falmagne, 1999; Falmagne & Doignon, 1988). Within a probabilistic structure defined applying the BLIM, all the responses to the items are assumed to be locally independent. The conditional probability $P(R|C)$ is determined by its probability of a false negative (β_q) and a false positive (η_q) while answering to q, as depicted by Equation (2.7)

$$P(R|C) = \left(\prod_{q \in C \setminus R} \beta_q\right)\left(\prod_{q \in C \cap R} (1-\beta_q)\right)\left(\prod_{q \in R \setminus C} \eta_q\right)\left(\prod_{q \in \overline{R \cup C}} (1-\eta_q)\right), \tag{2.7}$$

were, the lower are β and η, the higher is the probability of observing a specific response pattern R given that a patient is in a specific concept C. More specifically, it is expected that the inequality $\beta + \eta < 1$ holds true for all $q \in Q$. This basic assumption implicitly asserts that the probability of negatively answering an item is higher when the person does not endorse the item than when the student endorses it. Conversely, the probability of positively answering is expected to be higher when the person endorses the item rather than when she/he does not endorse it. As a

consequence, the probability of a endorsing an item q monotonically increases with the probability that belongs to the the clinical concept C, a monotonicity that holds true if and only if the inequality $\beta + \eta < 1$ is respected (Stefanutti, Spoto, & Vidotto, 2018). Such an assumption is fundamental not only for the β and η parameters per se, but also because underlies the reliability of all the model of assessment. In other words, a model containing items with low error rates will be more reliable, since less affected by potential false positive/negative when the responses are collected. A probabilistic model as the proposed one could be eligible of being implemented into an adaptive assessment instrument, since its ability of suggesting to a clinician which are the symptoms corresponding to the obtained clinical concept is coupled with the ability of providing an estimate of the probability related to that clinical concept. In all the applications of FPA (Serra et al., 2015, 2017; Donadello et al., 2017; Pino et al., 2018), the probabilistic structures have been always estimated on data referred to questionnaires responses. In this way, a single response pattern per person was given as an input to the BLIM model, in order to estimates the β and η parameters and the probabilities of the clinical concepts. In other words, one participant, one response patter to insert in the model. In observations, this one-to-one correspondence could not be always applied, since the observation could produce n response patterns relative to n observation samples. A procedure able to estimate a single response pattern from several ones is the first step to expand FPA in observational assessment. The following section will provide a possible solution to this issue.

2.4 FPA and modal scores

Suppose to have two different observational instruments, the former providing only one binary response pattern referring to the entire observation, while the latter producing five binary response patterns, one for each sample of a videotaped observation (e.g., a one-zero sample observation), as showed by Table 2.5:

	$Ov.Re.Pa.$
Item 1	1
Item 2	0
Item 3	0
Item 4	1

1a

	S_1	S_2	S_3	S_4	S_5	$Ov.Re.Pa.$
Item 1	0	1	1	0	1	?
Item 2	1	0	1	1	0	?
Item 3	1	0	1	0	0	?
Item 4	0	1	0	0	1	?

1b

Table 2.5: Response pattern of two different observational instruments. $Ov.Re.Pa.$ stands for "Overall Response pattern"; S stands for "Sample"

The response pattern displayed in the left panel of Table 2.5 is easier to use as a raw datum to define a deterministic/probabilistic clinical structure; it is similar to the response patterns derived from a questionnaire built with FPA. Nonetheless, a single response pattern could not provide enough information, if obtained from a single observation. Moreover, as discussed in the Chapter 1, the amount of possible observer's biases could be higher during a single sample observation. Data provided by the observational instrument displayed in the second panel could solve these critical issues: In fact, the information provided by observing a set of behaviors across multiple observational samples could be more accurate, since less altered by memory interference or observers' biases. Nonetheless, the possibility of collecting several response patterns from an observation could represents a critical aspect, since they are more difficult to manage than a single pattern. It would be more efficient to have a unique response pattern, containing the same amount of accurate information collected from multiple

observations. Moreover, such a pattern would better fit to BLIM models, that usually require only one pattern per person to validate a probabilistic clinical structure. Such a unique response pattern could be composed by items whose value correspond to the proportion of samples in which the behavior occurred: Nonetheless, this measure of frequency is biased and needs relevant attention to be used (see section 1.2). Another possible method to obtain the final response pattern could use the average of samples' patterns, intended as the probability of occurrence of an item, and then find the pattern X that minimizes the distance between the probability profile. In the present project, the solution adopted to extract a unique response patter from multiple ones is represented by the modal response pattern. Such a pattern is composed by items whose values correspond to the their modal occurrence/non occurrence across the n observational samples. A modal response pattern can be calculated from the example at hand, as displayed by Table 2.6:

	S_1	S_2	S_3	S_4	S_5	M
Item 1	0	1	1	0	1	1
Item 2	1	0	1	1	0	1
Item 3	1	0	1	0	0	0
Item 4	0	1	0	0	1	0

Table 2.6: Example of modal Response pattern M. S stands for "Sample"

In particular Item 1 and 2 describe behaviors that have been observed during three samples out of five. Consequently, their modal value will be 1, indicating a modal occurrence across the entire observation. On the contrary, Item 3 and 4 describe behaviors that have been observed during two samples out of five. Therefore, their modal value will be 0, indicating a modal non occurrence across the entire observation. Finally, the modal response pattern will be $M = \{1, 1, 0, 0\}$. This index does not represent only an empirical and reasonable way to extract information from a set

of multiple response patterns. It has a property making it eligible to express the complexity of multiple observation into a unique datum. In fact, a modal response pattern M has the property of minimizing the average symmetric distance Δ between itself and the response patterns $R \in \mathcal{R}$ obtained during n observational samples (Chiu & Douglas, 2013; de Chiusole et al., 2017). The average symmetric distance d is defined as follows:

$$d(M, \mathcal{R}) = \frac{\sum_{i=1}^{|\mathcal{R}|} |(M \,\Delta\, R_i)|}{|\mathcal{R}|} \tag{2.8}$$

where M is the modal response pattern, $\mathcal{R}$ is the set of response patterns collected during each observational sample i, and $|(M \,\Delta\, R_i)|$ is the symmetric distance between the modal response pattern M and the response pattern R_i. This measure of symmetric distance, minimized by the modal response pattern, coincides with the Hamming distance (Hamming, 1950). It is usually as a measure of dissimilarity between two response patterns (Chiu & Douglas, 2013). Formally, it is the distance d between two m-dimensional vectors A and B intended as the number of mismatches between their elements, as expressed by the equation below:

$$d(\mathbf{A},\mathbf{B}) = \sum_{j=1}^{m} \delta(a_j, b_j) \tag{2.9}$$

where

$$\delta(a_j, b_j) = \begin{cases} 1 & \text{if } a_j \neq b_j. \\ 0 & \text{if } a_j = b_j. \end{cases} \tag{2.10}$$

By using modal response patters, it is possible to obtain single data to be used for finding the clinical concepts belonging to the clinical structure. In fact, within an

observational instrument defined by means of FPA, for each modal response pattern M a modal clinical concept C_M should correspond into the clinical structure $\mathcal{C}$, defining a number of concepts more manageable compared to the C_n clinical concepts derived from the n samples of observation. On this regard, an important aspect of this possible $M-C_M$ correspondence must be stressed: If the response pattern R obtained from each sample of observation can correspond to a clinical concept C in the clinical structure $\mathcal{C}$, the same one-to-one correspondence between an M and a modal C cannot be assumed by default, since M is a derived response pattern and could not be equal to any of the originating response patterns. In particular, two situations are likely to occur:

- $M_i = C_i \in \mathcal{C}$: In this case, the modal response pattern corresponds and converges in a clinical concept C_M; consequently, their symmetric distance is equal to zero;

- $M_i \neq C_i \in \mathcal{C}$: In this case, the modal response pattern does not correspond to a clinical concept. Whenever such a scenario occurs, several solutions could be applied in order to obtain useful information, especially in perspective of an adaptive instrument. At one hand, it could be searched the clinical concept $C^* \in \mathcal{C}$ such that the symmetric distance $|(M \,\Delta\, C^*)|$ is minimal. The only disadvantage of this first strategy is that there could be more than one clinical concepts lying at the same symmetric distance from M; consequently, a precise criterion for selecting the final clinical concept should be planned. On the other hand, the clinical concept C^* can be estimated by using a feature of adaptive assessment algorithms: As it will be showed in Chapter 5, these algorithms always give as an output a concept belonging to the structure, estimating the most plausible one given the response pattern used as an input. Therefore, the modal response patterns not directly matching with clinical concepts can be used to

estimate the most plausible concepts given those modal response patterns. A that point, the symmetric distances will be calculated, the items causing such distances will be analyzed and the clinician will be warned about the dissimilarities.

Despite this potential critical issue, modal response patterns could represent a solution to the biased frequencies' topic. M could be used as a measure of behaviors' occurrence. Moreover, modal response patterns can enhance the parameters estimation procedure, that would be otherwise more time consuming if performed for each response pattern obtained by the observation's samples. Finally, an instrument providing modal response patterns has less false positive/negative rates per item, compared to the same instrument administered once at the end of an observation and providing, therefore, a unique response pattern (Chapter 5). In the next Chapter, the non adaptive version of an observational checklist evaluating the nonverbal behavior of schizophrenia and built through FPA will be introduced, called the Nonverbal Assessment of Negative Symptoms (NANS).

Chapter 3

Checklist definition

3.1 The selected disorder

Schizophrenia is a complex mental disorder impairing the way a person lives and perceives the world (Elis, Caponigro, & Kring, 2013). Although its lifetime prevalence seems to be smaller (i.e., 0.3-0.7%) than other disorders (e.g., major depressive disorder; American Psychiatric Association [APA], 2013), the pervasiveness with which schizophrenia affects the cognition, the feelings and the behavior of a person is absolute, representing a stigmatizing burden for this people (Riehle & Lincoln, 2018). According to the last version of the *Diagnostic and Statistical Manual of mental disorders* (DSM-5; APA, 2013), a diagnosis of schizophrenia can be formulated if at least two out of five of the following symptoms cause a functional impairment or persist for six months: Hallucinations, delusions, disorganized speech, disorganized behaviors and negative symptoms. Furthermore, one of the two symptoms must be either delusions, hallucinations (historically called "positive symptoms") or disorganization. The presence of at least one positive symptom to make a diagnosis reflects a dichotomy positive/negative

symptoms that has always characterized the studies of this disorder, usually in favor of positive symptoms (Galderisi, Mucci, Buchanan, & Arango, 2018). The reason of such a greater focus can be explained considering their nature. Positive symptoms, by definition, are exaggerated manifestations of normal behaviors or thoughts, that could lead a person to behave in a dysfunctional way. On the contrary, negative symptoms are usually a reduction in interests, motivation or behaviors. It is more likely that a person seeks for (or is sent to) medical support when an hallucination or a delusion occur. It seems that negative symptoms are more difficult to be detected (Aleman et al., 2017): As pointed out by Selten, Wiersma, and Van Den Bosch (2000), people affected by negative symptoms are usually not aware of them, requesting help for other symptoms. In the last twelve years, indeed, the attention on negative symptoms is increased, starting from their psychological evaluation.

3.1.1 The negative symptoms of schizophrenia

The turning point toward a deepen focus on negative symptoms of schizophrenia happened in 2006, during a consensus conference about a project regarding the Measurement and Treatment Research to Improve Cognition in Schizophrenia (MATRICS), sponsored by the National Institute of Mental Health (NIMH). In this conference a number of researchers on negative symptoms defined a set of guidelines to conceptualize these symptoms, confirming or modifying definitions and approaches to their study. In the same occasion, new challenges in terms of assessment and treatment were proposed (Kirkpatrick, Fenton, Carpenter, & Marder, 2006). The impulse given by MATRICS-NIMH consensus conference lead to a number of studies deepening definitions, prevalence, assessment and treatment of negative symptoms. Following the

new definition, negative symptoms refer to a reduction in (or a lack of) goal-oriented behaviors and activities that are normally performed by most of the people. They are clustered into two factors, respectively diminished emotional expression and apathy-avolition (Blanchard & Cohen, 2006); this bi-factor clustering has been accepted and reported even by DSM-5 (APA, 2013). Diminished emotional expression is composed by *blunted affect* and *alogia*, the former referring to a reduction in facial expression, gesture, body movements and prosody-related aspects (i.e., speed, volume and pitch) that are normally used to enhance social affiliation and interactions; the latter consists in a reduction in both the amount and the fluency of the speech. Apathy-Avolition refers to *asociality*, that is a reduction in desire, interest or motivation in social contacts; *anhedonia*, namely a reduced ability to experience pleasure from activities that people usually judge as enjoyable; finally, *avolition*, that is a lack of the desire to start activities and to later complete them (Elis et al., 2013). These symptoms could be intended as primary or secondary to other symptoms, usually positive ones, or referred to other mental disorders (e.g, mood disorders). It seems, for instance, that depressive symptoms share many negative manifestations with schizophrenia. The definition of a negative symptom as primary or secondary has relevant consequences on their treatment, since the dosages of pharmacological therapy are weighted on the hierarchy of manifested symptoms. In other words, a negative symptom could require different medications or dosage dependently if it is labeled as primary or secondary to other symptoms.

Another classification useful for later setting up a specific treatment concerns the persistency of negative symptoms: As suggested by Buchanan (2007), if a primary or secondary symptom persists for at least six months, after the stabilization of a first episode of psychosis, it could be referred to as a *persistent negative symptom*;

an interesting study suggested, indeed, that persistent negative symptoms are present since the first episode of a psychosis (Hovington, Bodnar, Joober, Malla, & Lepage, 2012). On the other hand, when primary negative symptoms lasts for more than twelve months, the diagnosis should be named as *deficit schizophrenia*.

The persistency issue brought inevitably to questions about the prevalence of negative symptoms. It has been suggested, in fact, that knowing the prevalence and the duration of negative symptoms could make possible to define of a more exhaustive therapeutic plan, reducing the probability of a worse prognosis caused by untreated symptoms (Boonstra et al., 2012). There is a general agreement on the evidence according to which patients with a diagnosis of schizophrenia show at least one negative symptom during prodromal phases (Fusar-Poli et al., 2013) and within early stages of the psychosis (Fulford et al., 2013), maintaining them even in chronic phases (Rabinowitz, Berardo, Bugarski-Kirola, & Marder, 2013). In particular, where the prevalence of negative symptoms during chronic phases can reach the 40%. In a population based-report, the 15.7% of adolescents and young adults showed a lifetime cumulative incidence of negative/disorganized symptoms that predicted the positive ones over time (Dominguez, Saka, Lieb, Wittchen, & Van Os, 2010). Such results were supported by a longitudinal study published by Werbeloff et al. (2015), in which the link between negative symptoms' prevalence and the following percentage of hospitalization was estimated: The 20.2% of the total sample recruited in the screening phase showed at least one negative symptom. Within that group, the 1.6% of patients were hospitalized for a diagnosis of schizophrenia. This study suggested how negative symptoms, when combined with the positive ones, can predict the later occurrence of schizophrenia.

As both the classification and prevalence/persistency issues started to become rel-

evant, the necessity of accurate assessment instruments became pressing. The new instruments should provide integrated information, obtained by clinicians' observation and reports from both patients and their relatives (Azorin, Belzeaux, & Adida, 2014). All the new instruments, such as as the Clinical Assessment Interview for Negative Symptoms (CAINS; Kring, Gur, Blanchard, Horan, & Reise, 2013), the Brief Negative Symptom Scale (BNSS; Kirkpatrick et al., 2011), and the Motor Affective Social Scale (Trémeau et al., 2008), were developed according to the last findings on negative symptoms literature. Each tool presents good psychometric properties, is less unobtrusive for patients and less prone to distraction and fatigue effects (Kring et al., 2013; Millan, Fone, Steckler, & Horan, 2014; Strauss et al., 2012), especially if compared to the "first generation" tools (Garcia-Portilla et al., 2015; Kilian et al., 2015) such as the Scale for the Assessment of Negative Symptoms (SANS; Andreasen, 1982). Furthermore, these instruments seem to adequately implement all the aspects related to the Apathy-Avolition factor. The same result has not been achieved within diminished emotional expression' dimensions. A possible explanation of this poor improvement concern the fact that this factor mostly refers to behavioral dimensions: As pointed out by Galderisi et al. (2018), all the items referred to this factor are, or should be, evaluated through observation. Consequently, all of them could be recoded in an observational fashion, for instance in terms of nonverbal behavior.

3.1.2 Detecting negative symptoms via nonverbal behavior

Nonverbal behavior (NVB) can provide precious elements that can be used during the assessment and for setting up a specific treatment (Ellgring, 1986; Hall, Harrigan, & Rosenthal, 1996; Ramseyer & Tschacher, 2011; Roter, Frankel, Hall, & Sluyter,

2006). In case of schizophrenia, the nonverbal behavior is the target of the Social Skills Training (SST; Bellack, Mueser, Gingerich, & Agresta, 2013; Elis et al., 2013; Turner, van der Gaag, Karyotaki, & Cuijpers, 2014), one of the most effective applied therapies. It has been shown how such a treatment improves the functional outcome of patients with negative symptoms, that is usually extremely poor (Evensen et al., 2012; Harvey & Strassing, 2012).

In schizophrenia, the majority on NVBs can be found in the negative symptoms domain, in both its components (i.e., blunted or flat affect and alogia): Several studies describe blunted affect as a reduction in facial expressions (Troisi et al., 2007), gestures, posture, body movements and prosody (Del-Monte et al., 2014; Messinger et al., 2011; Millan et al., 2014; Trémeau et al., 2008). Prosodic elements are present also when considering alogia, especially in terms of speech fluency (A. S. Cohen, Mitchell, & Elvevåg, 2014; Stassen et al., 1995). All these elements can be found even among the features related to diminished emotional expression factor, which are described among the clinical features of the schizophrenic spectrum of DSM-5 (APA, 2013). Finally, the correlation between NVB and classical measures of negative symptoms has been stressed by several evidences: Patients with a predominance of negative symptoms seem to express an overall reduction of NVB (Brüne et al., 2008; Brüne, Abdel-Hamid, Sonntag, Lehmkämper, & Langdon, 2009; Troisi, Spalletta, & Pasini, 1998); this profile tends to remain constant during clinical interviews (Lavelle, Dimic, Wildgrube, McCabe, & Priebe, 2015). The stability of behaviors related to blunted affect has been observed also in several studies, from the prodromal phase (Malla et al., 2002) to the first onset of the disorder (Shtasel, Gur, Gallacher, Heimberg, & Gur, 1992). In follow-up studies, the blunted affect has emerged as stable for period of one year (Kelley, Haas, & van Kammen, 2008), with a fluctuating course over ten years (Evensen et

al., 2012). The stability over time of blunted affect and its related NVB is strongly related to very poor social and functional outcomes (Evensen et al., 2012; Harvey & Strassing, 2012); these results corroborate evidences in literature according to which such primary negative symptoms are in some way resistant to antipsychotics (Galderisi et al., 2018). It is straightforward the importance to detect as soon as possible such symptoms.

A possible way to assess them is via nonverbal behavior (Granziol, Spoto, & Vidotto, 2018), maybe considering all the items related to diminished emotional expression factor in observational terms, within the frame of NVB. In this regards, Kilian et al. (2015) noted some critical issues concerning how both dimensions of diminished emotional expression factor (i.e., blunted affect and alogia) are investigated within the used instruments. Authors observed how the majority of these instruments considered reduced sets of items to investigate blunted affect (e.g., items referring only to facial and/or vocal expressions). Important items assessing behaviors like eye contact or body movements are less frequently included, especially in the new generation of instruments. This is extremely critical, considering the importance of gaze direction of patients during social interaction (Vail et al., 2018). Likewise, a number of studies suggests how it is possible to infer cognitive impairments from gesture and body movements of a patient (Kupper, Ramseyer, Hoffmann, Kalbermatten, & Tschacher, 2010; Kupper, Ramseyer, Hoffmann, & Tschacher, 2015; Walther et al., 2015). Another example concerns prosodic elements (A. S. Cohen, Mitchell, Docherty, & Horan, 2016) and alogia, whose relationship with the expressive factors is still not well-defined (Alpert, Shaw, Pouget, & Lim, 2002; Kilian et al., 2015). As a result, two key points are still unsolved: At one hand, it remains unclear whether the currently used items can exhaustively detect all of the clinical manifestations of diminished emotional expression,

especially from a nonverbal point of view. On the other hand, the possibility of using such nonverbal behaviors within a specific observational instrument has not been tested.

The aim of the this chapter is to describe the development of the Nonverbal Assessment of Negative Symptoms checklist (NANS), an observational checklist assessing the nonverbal behavior related to negative symptoms of schizophrenia. In order to reach this goal, the guidelines derived from both FPA and the critical issue mentioned in Chapter 1 concerning the definition on an observational instrument will be followed. The first two steps will be focused on the selection of both sets of behaviors and their investigated attributes, in order to define the clinical context depicting their relations.

3.2 Items selection

The procedure to select suitable behaviors was carried out by two independent raters, who had the instruction to search for items describing nonverbal behaviors from other-report instruments applied in the assessment of schizophrenia. The inclusion criteria for instruments were basically two:

- The target instruments should have been used in the assessment of schizophrenia, evaluating negative symptoms, disorganized behavior or both.
- Clinical interviews and observational grids were the target instruments. All self-reports measures had to be discarded.

Once the list of selected instruments was defined, the two raters proceeded to extract the initial list of items, following the criteria described below:

- Only items describing nonverbal behaviors could be selected; items referring to psycho-physiological activation had to be discarded.

- Items had to refer to nonverbal behaviors related to facial expressions, gesture, body movements (i.e., considering the general movement of both the entire body and of its specific parts), gaze, prosody (i.e., speed, volume, pitch of speech) and posture.
- Nonverbal behaviors could have been molecular or global (see section 1.2).
- Whenever items were scored on Likert scale and each point of the scale was specified by a description, all the levels had to be considered separately and the corresponding description analyzed consequently.
- The amount or the content of speech could be considered only if related to verbal-nonverbal synchronicity (Ellgring, 1986; Kring & Caponigro, 2010)

Within the selection of both instruments and items, each element found by the two raters was discussed in order to verify their agreement. Each disagreement was solved by discussion, otherwise a third expert rater was consulted. In general, the mean Cohen's κ was very high ($\kappa = 0.88$). The set of selected items for the first analysis consisted of 138 items, extracted from the following instruments:

i) The Brief Negative Symptom Scale (BNSS; Kirkpatrick et al., 2011). BNSS is an instrument designed to assess the severity of negative symptoms, going beyond clinical trials settings. It is composed by six sub-scales, i.e. Anhedonia, Asociality, Avolition, Distress, Alogia and Blunted Affect, rated from 0 (absent) to 6 (severe). It has the form of a semi-structured interview, so the first four subscales include questions, while the last two include observations. Selected items: 9, 10, 11, 12.

ii) The Inpatient Multidimensional Psychiatric Scale (IMPS; Lorr, 1962). IMPS is used to make a broad-spectrum evaluation of a patient, based on his/her behavior

during a psychiatric interview. It has been selected due to its clarity in describing items referring to NVB; some of them are rated on a 9-level scale (1, 6, 13, 22, 23, 26, 33, 41, 49 in this study) , while others on a 5-level scale (52, 53, 56, 57, 58 in this study).

iii) The Brief Psychiatric Rating Scale 4.0 (BPRS4.0; Overall & Gorham, 1962; Ventura, Green, Shaner, & Liberman, 1993). BPRS 4.0 is used to evaluate the gravity of a psychopathology, especially in case of Depression or Psychotic Disorders. It is formed by 24 items rated from 2 (very mild) to 7 (extremely severe), evaluated during a semi-structured interview. Selected items: 16, 17, 18, 19, 20, 23, 24; all their levels have been analyzed individually.

iv) The Ethological Coding System for Interviews (ECSI; Troisi, 1999). Designed for measuring the nonverbal behavior during interviews, ECSI is 37 dichotomous items checklist, grouped in seven subscales: Eye Contact, Affiliation, Submission, Flight , Assertion, Displacement Activities, Relaxation. All of them were included for the analysis.

v) The Motor Affective Social Scale (MASS; Trémeau et al., 2008). MASS is a 5-minute interview, during which 3 questions are asked by the clinician while she/he evaluates the Number of smiles, Co-verbal gestures and patient's Asked question, as a measure of Alogia. These three aspects are assessed by means of eight items rated on a Likert scale ranging from 1 to 4. Selected items: Number of smiles, Co-verbal gesture.

vi) The Scale of Prodromal Symptoms (SOPS; Miller et al., 1999). SOPS is an instrument investigating prodromal aspects in Psychotic Disorders. It is a part of the

Structured Interview of Prodromal Symptoms (SIPS; Miller et al., 1999). SOPS is composed by 19 items grouped into four sub-scales, named Positive Symptoms, Negative Symptoms, Disorganized Symptoms and General Symptoms, rated from 0 (absent) to 6 (extreme). For this study, only the third item of the Negative Symptoms sub-scale has been selected and decomposed within its three levels. The choice of inserting an instrument used in prodromal phases allowed for taking into account also attenuated symptoms, that are not clearly observable in acute phases.

vii) The Scale for the Assessment of Negative Symptoms (SANS; Andreasen, 1982). SANS is the most used and well-known scale in this field, as introduced above. It can be used during an interview and it is composed by 25 items rated from 0 (none) to 5 (severe) and divided into five sub-scales: Alogia, Blunted Affect, Avolition/Apathy, Asociality/Anhedonia and Attention. Selected items: 1, 2, 3, 4, 5, 6, 7, 9, 12, 14, 16.

Each of the 138 items was treated dichotomously (i.e., in terms of occurrence/nonoccurrence) and inserted in the first version of the clinical context. The next section will deepen the attributes selection procedure.

3.3 Attributes selection

Two sources of data, considered as equally important in defining a set of observable nonverbal behaviors, were used to define the initial set of attributes:

1. The "negative symptoms" criterion for the diagnosis of schizophrenia described by the DSM-5 (APA, 2013). The nonverbal behaviors referred to negative symp-

toms were extracted from both the "diagnostic features" and the "associated features supporting diagnosis" sections located under the table of schizophrenia's diagnostic criteria. All of nonverbal behaviors were considered as attributes.

2. The scientific literature about the nonverbal behaviors of schizophrenia. Search engines as Scopus, PsycINFO, PubMed and Google Scholar were used, by inserting the following terms: "nonverbal behavior" OR "somatic manifestation" OR "facial expression" OR "gesture" OR "motor behavior" OR "body movements" OR "gaze" OR "prosody" OR "voice" OR "body posture" AND "negative symptoms" AND "schizophrenia". The same experts of the previous selection procedure conducted this research independently, and discussed each article. Articles referring to NVBs without any reference to schizophrenia were excluded; likewise, articles describing cognitive or emotional evaluations were discarded. 33 articles emerged from this source: in 9 of them, multiple nonverbal behaviors were investigated; 8 referred only to facial expressions; 7 referred to prosody/voice; 3 referred to behaviors involved in prosocial behaviors; 3 referred to gaze; finally, 3 articles referred specifically to body movements. Whenever at least two articles investigated the same nonverbal behaviors in their results, those behaviors were considered symptoms and were included in the list of attributes.

Table 3.1 displays the list of initial attributes: For most of them, the articles found in scientific literature corroborated their eligibility to attributes. In fact, the attributes obtained from DSM-5 (i.e., from A1 to A12) are the subject matter of several studies. On the other hand, attributes A13, A14, and A15 are not included in the DSM-5 even if they are mentioned in the scientific literature as potential nonverbal behaviors of schizophrenia. Attribute A8 was slightly modified, from "reduction of hands move-

ments" to "reduction of gestures", in order to include more communicative gestures, not only co-verbal ones. Finally, a particular cases is attributes A10: It is included in DSM-5, but among clinical manifestation of "grossly disorganized or catatonic behavior". It was selected since, reading its clinical description, it implies a reduction in behavior (also nonverbal). The set of selected attributes in this phase contained 15 attributes.

ID	Description	Source	References
A1	Reduction of facial expressivity	DSM 5	(Annen, Roser, & Brüne, 2012; Earnst et al., 1996; Ellgring, 1986) (Jones & Pansa, 1979; Lavelle, Healey, & McCabe, 2014) (Mandal, Pandey, & Prasad, 1998; Steimer-Krause, Krause, & Wagner, 1990) (Trémeau et al., 2005; Troisi et al., 2007)
A2	Reduction in head movements	DSM 5	(Annen et al., 2012; Brüne et al., 2008) (Davison, Frith, Harrison-Read, & Johnstone, 1996; Ellgring, 1986)
A3	Alogia	DSM 5	(Stassen et al., 1995)
A4	Reduction in the speed of speech	DSM 5	(A. S. Cohen, Kim, & Najolia, 2013; A. S. Cohen et al., 2016) (Dickey et al., 2012; Püschel, Stassen, Bomben, Scharfetter, & Hell, 1998)
A5	Reduction in the volume of speech	DSM 5	(A. S. Cohen et al., 2016; Dickey et al., 2012) (Leentjens, Wielaert, van Harskamp, & Wilmink, 1998; Püschel et al., 1998)
A6	Reduction in intonation of speech	DSM 5	(Dickey et al., 2012; Leentjens et al., 1998; Murphy & Cutting, 1990) (Stassen et al., 1995)
A7	Reduction of spontaneous movements	DSM 5	(Dimic et al., 2010; Kupper et al., 2010; Morrens, Hulstijn, & Sabbe, 2007)
A8	Reduction of gesture	DSM 5	(Annen et al., 2012; Brüne et al., 2008, 2009; Del-Monte et al., 2014) (Del-Monte et al., 2014; Lavelle, Healey, & McCabe, 2013) (Lavelle et al., 2014; Walther et al., 2015)
A9	Reduction in eye contact	DSM 5	(Brüne et al., 2008; Gaebel, 1989; Troisi et al., 1998; Troisi, 1999) (Annen et al., 2012; Dimic et al., 2010)
A10	Decreased in reactivity to the environment	DSM 5	
A11	Negativism	DSM 5	
A12	Rigid posture	DSM 5	(Hall et al., 1996; Troisi et al., 1998)
A13	Fixed gaze	Literature	(Dowiasch et al., 2016; Gaebel, 1989)
A14	Difficulty in reciprocating social behaviors	Literature	(Kupper et al., 2015; Lavelle et al., 2013)
A15	Dissociation between speech's content and nonverbal behavior	Literature	(Ellgring, 1986; Kring & Caponigro, 2010)

Table 3.1: The set of the nineteen attributes

3.4 Definition of the clinical context

As mentioned in section 2.2, a clinical context is a Boolean matrix containing items in rows and attributes in columns: Whenever an item investigates one or more attributes, the corresponding cells will contain a 1, otherwise a 0 will be present. The first version of the clinical context contained 138 rows and 15 columns. It is important to clarify that the item-attribute assignment for the 138×15 matrix was independently and manually conducted by two experienced clinicians operating in the field of schizophrenia (one male and one female, different from the previous two). The average inter-rater agreement estimated for each cell was very high ($\kappa = 0.91$). As for items and attributes selection, disagreements were solved through direct discussion between the raters or by consulting a third expert. It is straightforward that an observational tool composed by 138 items is not feasible: Beyond their amount, several items could be redundant or could not investigate any of the attributes. A pruning is required to reach an adequate set of specific items. In FPA, the initial clinical context can be pruned by removing:

- Empty rows, containing those items not investigating any attributes (i.e., rows containing only zeros in the Boolean matrix). It is important to stress that an empty row does not necessarily implies the automatic elimination of the corresponding item. It could be the case that such an item describes a relevant aspect of an attribute that has not been considered. Therefore, an attribute *ad hoc* for that item could be defined;

- Empty columns, containing those attributes not investigated by any item. When an empty column is found, a new item investigating it should be created;

- Equivalent rows, containing items that investigate the same sets of attributes

(i.e., rows in which the 1 values are located exactly in the same cells). In this case, a good practice is to insert in the final list of items only the item that is better formulated to investigated the attribute(s);

- Equivalent columns, containing attributes that were investigated by the same sets of items. When two equivalent columns emerge, their corresponding attributes should be better analyzed and, eventually, a new item differentiating them should be defined.

All these configurations convey important and different information regarding the assessment of the nonverbal behaviors in the selected instrument. The next section will show the results concerning the final clinical context and the derived clinical structure.

3.5 Results

3.5.1 Clinical context

The final clinical context of the NANS was composed by 22 items describing nonverbal behaviors investigating 14 attributes. This result was reached by means of a procedure that, as a first step, reduced the number of both items and attributes; then, the nonverbal behavior described within each item was reviewed in order to be more immediate, clear and easy to detect. In regards to the first step, the reduction was obtained by removing the rows and the columns accordingly to the aforementioned configurations. The main findings were:

- Empty columns. The attributes "Fixed gaze" (A13), "Difficulty in reciprocating social behaviors" (A14), and "Dissociation between speech's content and nonverbal behavior" (A15) collected only zeros, meaning that both raters did not

identify any selected item investigating those attributes. At this point, two different solutions were applied: Nonverbal behaviors *ad hoc* for both A13 and A14 were defined (see Table 3.4), since the evidences obtained for both attributes from the scientific literature were corroborated by the clinical common practice of both raters. On the other hand, the difficulty of operationalizing A15, even in terms of global behavior, led to discard this attribute.

- Empty rows. Both raters agreed on discarding 25 items that did not investigate any of the attributes. Such a decision was made on the basis of a deep evaluation of the clinical aspects investigates by those items. For instance, item IMPS58 (i.e., "Glance around at and/or appear to be startled as if hearing voices?"), investigates highly specific nonverbal behaviors related to positive symptoms.

- Equivalent rows. Whenever two or more items investigate exactly the same set of attributes, they define an *equivalence class*. In this clinical context, eight equivalence classes emerged. Each class can be conceived as a tank from which it is possible to randomly select an item to investigate the specific set of attributes. In the present project, each equivalence class is represented in the clinical context by one of its prototypical items. For instance, an equivalence class is displayed in Table 3.2.

Item	*A*7	*A*8	*A*9
SANS2	1	0	0
SANS16	1	0	0

Table 3.2: The equivalence class referring to the attribute A7. This table is an extract of the first clinical context. The item included in the final set of 22 items is "SANS2".

In the example, both items SANS2 ("The patient shows few or no spontaneous movements, does not shift position,moves extremities, etc.") and SANS16 ("The

patient tends to be physically inert. He may sit for hours and not initiate spontaneous activity.") investigate attribute A7 (i.e., "Reduction of spontaneous movements"). Eight equivalence classes were found, containing 90 items altogether.

- Equivalent columns. No equivalent columns were found.

By applying the pruning procedure, a final clinical context of 22 items investigating 14 attributes emerged. This new clinical context contained several rows contained in other rows: As mentioned before, whenever a row is contained in another one, a prerequisite relation occurs. An example of prerequisite relations is displayed by Table 3.3

Item	$A4$	$A5$	$A6$
IMPS1/22	1	0	0
IMPS33	1	1	0
BNSS12	1	1	1

Table 3.3: Example of prerequisite relations. This table is an extract of the final clinical context.

In words, the behavior described by item IMPS1/22 (a behavior created by merging IMPS1 and IMPS22 due to their similar meaning), which investigates the reduction of speech speed (i.e., A4), is a prerequisite of IMPS33, which investigates both the reduction of speech speed AND the reduction of speech volume (i.e., A4, A5). Both of them are prerequisites of BNSS12, which investigates the reduction of speed, volume AND intonation of speech (i.e., A4, A5, A6). In the present context, 40 prerequisite relations were found.

During the definition of the clinical context, all the experts noted how some of the sentences used to describe the nonverbal behaviors of several items were extremely long or less clear than expected. Consequently, the face validity of a checklist defined

by using such nonverbal behaviors could be compromised. For this reason, each item was reviewed and discussed by experts and, if necessary, modified. The modifications were applied in a way that the new instances of the behaviors should preserve both their item-attribute assignment and their main characteristics. The modifications were applied only if all experts agreed on them. The new behavior coding of the NANS, referring to the final clinical context, is displayed in Table 3.4.

FPA	Original Item	Description (The patient...)	Attributes
1*[1]	IMPS6	Exhibits and keeps the same postures (independently if peculiar, unnatural, rigid or bizarre).	17
2*	SANS2	Shows slow, few or no spontaneous movements.	7
3*	BPRS18L6-7[2]	Moves or speaks only if stimulated, otherwise seems blocked, catatonic.	7,10,17
4*	SANS3	Exhibits a reduction or absence of gesture (i.e., movement of hand or other body parts), as an aid in expressing his/her ideas.	8
5*	IMPS16	Manifests slow, few or no movements and gestures. Note: check this item even if movements seem more demanding, labored or delayed.	7,8
6*	BNSS9	Shows a total or nearly total lack of facial expressions.	1
7*	SOPS3a	Presents a flat, constricted, diminished emotional responsiveness, as characterized by: a decrease in facial expressions, reduced gestures and monotone speech. Note: to fill this item please check for ALL the elements of the list.	1,6,8
8*	BNSS11	During the conversation, shows few or no movements of: hands, head and body. Note: to fill this item please check for ALL the elements of the list.	2,7,8
9*	SANS5	Fails to laugh or smile in response to his/her speaker.	1,10
10	New[3]	Presents a lack or reduction of head's movements.	2
11*	BNSS12	Replies with only one or two words, or does not speak at all.	3
12*	IMPS1/22	Manifests a way of speaking that is slowed, characterized by blocking, halting, or irregular interruptions.	4
13*	IMPS33	Speaks in a way slow and difficult to hear. Note: Do not compare the volume of the patient with the speaker's one.	4,5
14	SANS7	Fails to show normal vocal emphasis patterns, is often monotonic.	6
15*	BNSS10	Has a way of speaking characterized by: (i) slowness or irregular interruptions, (ii) reduction in volume and (iii) monotonic speeach (i.e., has a constant tone, independently from what she/he is saying). Note: to fill this item please check for ALL the elements of the list.	4,5,6
16*	BPRS18gen[4]	Beyond slow movements and speech, shows also a monotonic speech.	4,6,7

[1]The notation "*" indicates that the description of the item is a modified version of the original one.

[2]The notation "L6-7", and similar throughout the table, indicates the Likert levels of the item. In the example at hand, the item BPRS18L6-7 indicates that both levels 6 and 7 of item BPRS18 convey the same information, modifying only the strength of the presence of that symptoms

[3]The notation "New" indicates that the item has been defined for the first time.

[4]The notation "gen", and similar throughout the table, indicates the general description of the item was considered relevant to define the behavior.

17	SANS4	Avoids eye contact or "stares through" interview even when speaking.	9
18*	SANS23	The patient appears uninvolved or unengaged. Example: He/she may seem "spacey", distracted.	10
19*	IMPS41	Answers in monosyllables or give only minimal responses or does not speak. Moreover, avoids eye contact.	3,9,10
20*	BPRS20gen	Shows resistance and lack of willingness to cooperate with the interview.	14
21	New	The patient shows a fixed gaze on either the interviewer or another point of the space.	11
22	New	The patient does not respond to prosocial behaviors of the interviewer. Example: she/he does not seem to react to facial expressions, gaze or gesture of the interviewer.	10,12

Table 3.4: The NANS's Clinical Context with the 22 items and their investigated attributes

3.5.2 Clinical structure

The emerged clinical context allowed for delineating a clinical structure, by means of the formal steps described in section 2.2. As said, each node of this lattice represents a clinical concept, containing items describing the nonverbal behaviors observed in the patient and the symptoms related to those behaviors. The connection between the context and the structure, beyond the mathematical aspects, can be found also in the aforementioned configurations leading to the final context, that have their specific counterparts in the structure. In fact, an empty row/column results in the absence of that item/attribute in the structure. Equal rows result in items that are contained exactly in all the same clinical concepts. Finally, even the prerequisite relations are adequately represented: The clinical structure will not include any clinical concept that contains an item but not all of its prerequisites. Within the clinical structure, all the relations among and within items/attributes of the clinical concepts can be graphically

analyzed (Granziol et al., 2018).

The structure obtained from the clinical context contained 9216 concepts, a dramatically smaller number of clinical concepts compared to the cardinality of the power set calculated on the set of items (i.e., $2^{22} = 4,194,304$ with a ratio of $\sim 1/455$). This number of clinical concepts, even if more manageable, was considered still high, since it was defined by the number of items investigating one attribute (i.e., *singletons*). Such a number would have required thousands of patients in order to validate the structure or, at least, to estimate the error parameters, a scenario difficult to achieve considering the observational nature of the study and the prevalence of the negative symptoms. Consequently, a two factor model was defined, splitting the clinical context into two sub-contexts (Table 3.5) and clustering the items and their investigated attributes around two "nonverbal areas of interest".

Item	A1	A2	A7	A8	A10	A17
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

10a

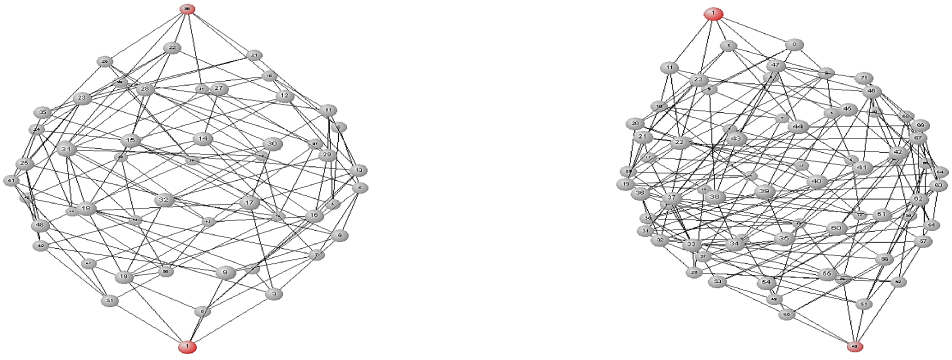
Item	A3	A4	A5	A6	A9	A10	A11	A12	A14
11	1	0	0	0	0	0	0	0	0
12	0	1	0	0	0	0	0	0	0
13	0	1	1	0	0	0	0	0	0
14	0	0	0	1	0	0	0	0	0
15	0	1	1	1	0	0	0	0	0
16	0	1	0	1	0	0	0	0	0
17	0	0	0	0	1	0	0	0	0
18	0	0	0	0	0	1	0	0	0
19	1	0	0	0	1	1	0	0	0
20	0	0	0	0	0	0	0	0	1
21	0	0	0	0	0	0	1	0	0
22	0	0	0	0	0	1	0	1	0

10b

Table 3.5: The sub-contexts related to Movement (1a) and ProsInt (1b) factors.

The former sub-context (Table 10a) contained items describing nonverbal behaviors investigating attributes focused on facial expressions, gesture, head and body movements. It is straightforward how their common denominator is the movement. This sub-context led to a sub-structure composed by 52 concepts, and it was called the *Move-*

ment structure (Figure 3.1a[5]). The latter sub-context (table 10b) contained items describing nonverbal behaviors investigating attributes focused on prosodic features and elements related to social interaction. This second sub-context led to a sub-structure composed by 288 concepts, and it was called the *ProsInt structure*. An extract of the ProsInt structure is displayed in Figure 3.1b.



(a) The Movement structure (b) The ProsInt structure

Figure 3.1: The two substructures. Each node contains the behaviors observed in the patient and the negatie symptoms detected by means of those observed items.

The division into two substructures led to a total number of clinical concepts easier to manage in terms of both model validation and parameters estimation, without losing information or accuracy during an observational assessment. As mentioned in Chapter 2, each node of the two substructures contains both the item describing the nonverbal behaviors observed in a patient and the negative symptoms investigated by those behaviors. The two substructures, consequently, will produce two outputs (i.e., two clinical concepts) containing highly specific information that can be easily combined into a final report or, if preferred, analyzed separately.

[5]As mentioned before, both clincial structure have been obtained by means of the software Galicia (Valtchev et al., 2003); the second one represents an extract, since the real one was extremely large to be represented.

3.5.3 Other results

The lists of items and their investigated attributes represented not only the starting point to define an observational instrument, but gave the chance of making two further considerations. The analysis of the clinical domain (i.e., all the nonverbal behaviors that could be observed to evaluate negative symptoms) confirmed the decrease in nonverbal communication found in scientific literature (Brüne et al., 2009; Lavelle et al., 2013, 2014): In fact, all the items (and their described nonverbal behaviors) of the clinical domain described either a reduction or a lack of a nonverbal dimension. On the other hand, if the assessment instruments from which the items were selected seemed to agree on the reduction of NVB, they showed an interesting variability on which nonverbal behavior should be observed. It emerged how none of the original assessment tools could cover alone the totality of attributes. In particular, items of BNSS and SOPS referred only to the upper part of the face, some gestures and prosody, not considering other nonverbal dimensions such as posture and body movements. ECSI provided a very good example of observational checklist, but was mainly focused on face, head and body movements. Consequently, beyond its high specificity, it completely excluded prosodic elements. The same specificity was found also within MASS, which emerged as very accurate on the few behaviors it examined. Results confirmed how BPRS, SANS and IMPS represented gold standard observational instruments, investigating the majority of attributes, but not their totality. Furthermore, some of their items did not explain the investigated attributes in a direct way, as other instruments did.

3.6 Discussion

The observational assessment of negative symptoms in schizophrenia received great attention during last years, leading to the definition of new instruments assessing several negative dimensions. Nonetheless, assessing those symptoms via nonverbal behavior still remains difficult, due to the costs in terms of training and time to perform such an observational assessment. The assessment tools used until now tried to cope with these issues, usually by observing small sets of nonverbal behavior during or after the interviews. This kind of approach, although representing a good compromise between costs and benefits, has the disadvantage of being extremely reductive and sometimes inaccurate. As far as we know, a list of nonverbal behaviors exhaustively investigating all the clinical manifestations of the negative symptoms (in particular, the diminished emotional expression factor) has not been clearly defined nor implemented into an observational assessment tool. The present Chapter is a first attempt to cope with this challenges, by providing a list of items describing nonverbal behaviors that can be used to efficiently and exhaustively investigate clinical issues related to negative symptoms of schizophrenia. In regards to the former challenge, a set of 22 items investigating 14 symptoms was found, focusing also on nonverbal dimensions partially investigated by other instruments or present only in scientific literature. This is the case of attributes A13 (i.e., Fixed gaze) and A14 (i.e., Difficulty in reciprocating social behaviors), which are clinical manifestations defined under research conditions and, consequently, difficult to observe during a standard consultation. New items were coded in order to investigate them, an extension made it possible by the flexibility of FPA methodology, that allows for the inclusion of attributes or items if they are considered representative of a disorder or typically observed in common clinical practice. This property of FPA

makes it possible another scenario: Considering that mental disorders can change over time, it would be possible to implement the found list with other worthwhile observable behaviors, simply changing the clinical context and re-defining the corresponding clinical structure.

The found list of items paved the way to achieve the other challenge, namely its implementation into an observational instrument. As suggested in Chapter 1, such an instrument should be composed by items describing behaviors easy to be detected and, consequently, less prone to personal interpretation or biases. In this regard, this study reached an interesting result: After a process of behaviors modification to made items clear, immediate and easy to observe, the found list was implemented into a checklist of 22 dichotomous items, namely the NANS. This checklist, indeed, is not only a sum of items, but is a mapping from the described behaviors to the attributes they investigate. In fact, whenever an item has been observed, it helps the clinician to know exactly which symptoms are investigated by that item. More in general, the set of observed behaviors can depict which set of attributes are endorsed by a patient with a specific clinical concept. Moreover, such a mapping provides a model of assessment able to consider and represent, even graphically, all the possible outcomes that could appear once the specific behaviors are observed. This model of assessment is made possible also due to prerequisite relations among items. The advantage of knowing the relationships among items goes beyond the mere formal/mathematical innovation. It is relevant from a clinical point of view: Whenever a specific item is checked by the clinician (i.e., its behavior occurred, it was observed and checked within the checklist), each of its prerequisites is actually endorsed. The time saved by this system of assessment could be used to further investigate other symptoms of a patient, such as her/his personal feelings. These aspects, in turn, allow clinicians to integrate

information coming from different assessment modalities, increasing the quantity and quality of the entire assessment phase.

The study presented in this first Chapter presents some limitations that are are linked to the FPA methodology: For instance, the procedure of item-attribute assignment is not an algorithmic process. Consequently, it is not exempt from inferential errors of each rater, or influence by disagreements among raters. Finally, this procedure is, unfortunately, still time consuming.

Actually, these limitations are counterbalanced by the perspectives introduced by the NANS, namely its implementation into an adaptive observational instrument. Such an instrument will have its foundations in (a) the clinical structure, intended as the deterministic basis of the aforementioned model; (b) the error parameters and the probability values of each clinical concept (elements that will be deepened in the next Chapter) and (c) the clinical context, which can provide outputs based on the attributes endorsed by the response pattern of the patient, rather than on his/her numerical score (Donadello et al., 2017). The adaptive algorithm, by means of the prerequisite relations, would reduce the time spent on checking behaviors that will be surely displayed. All the details of the adaptive checklist will be described in Chapter 5. Finally, the improvements are not only from a quantitative point of view. The qualitative advantages of using such an instrument (independently if standard or adaptive) lies in its discriminative power: Suppose, for instance, that two patients obtain the same score after the observation (i.e., two response pattern with the same number of observed behaviors); the procedure could allow the clinician to delineate their two different and individualized patterns, which could lead to different and personalized treatments.

As mentioned before, the observational checklist created in the present study could be used as a module of a wider assessment procedure, especially if implemented in its

adaptive version. In order to reach this goal, the error parameters estimates and the validation of both structures are necessary. The next Chapter will describe and discuss all these aspects.

Chapter 4

Checklist validation

The definition of an observational checklist is only the first step of a refinement procedure ending with its application on field. Before this endpoint, the model on which the checklist is built needs to be tested. As discussed in previous sections, issues related to the goodness of the model, the reliability of the instrument and the risk of making mistakes (i.e., false positives/negatives) are essential for a psychological instrument, even more for observational tool: It is not new that observational tools are prone to biases and false positives/negative that could undermine the reliability of the collected data (Groth-Marnat, 2009; Repp et al., 1988). In FPA framework, the model testing concerns the validation of the deterministic clinical structure $\mathcal{C}$ by fitting the Basic Local Independence Model (BLIM) to the response patterns obtained from the instrument, in order to obtain the error parameters estimates ($\beta \backslash \eta$) for each item of the checklist and the fit indexes of the structure. Testing a model referring to an observational instrument implies that the response patterns used as input for the BLIM algorithms are more complex than response pattern derived from self-report instruments: As introduced in section 2.4, such patterns are modal response patterns M,

obtained by extracting the modal value of each item across response patterns observed in n samples of observation. Therefore, a procedure to obtain the modal response patterns should be a priori decided. Moreover, these patterns should be reliable, that is each occurrence/nonoccurrence of their nonverbal behaviors requires an high inter-rater agreement. Once all of these aspects are controlled, the model can be tested. The aim of the present Chapter is to present the validation of the two clinical structures belonging to the Nonverbal Assessment of Negative Symptoms (NANS) checklist. In order to achieve this goal, all the steps required to obtain and analyze reliable data, necessary to the model testing, will be described.

4.1 Materials and Methods

4.1.1 Sample

The sample consisted of 172 Italian volunteer participants, including a group of 38 people with a diagnosis of mental disorder (in the sequel, this subsample will be called the clinical group). In particular, the primary diagnoses of people belonging to this group were: Schizophrenia (n = 25, 5 females), Bipolar Disorder with psychotic behavior (BD; n = 6, 0 females), Major Depressive Disorder with psychotic behavior (MDD; n = 3, 3 females) and Obsessive Compulsive Disorder (OCD; n = 4, 4 females). Demographic characteristics of the sample are showed in Table 4.1.

The majority of patients had at least a middle school diploma. Patients with a diagnosis of schizophrenia were treated with antipsychotics of first ($\sim$ 20%) or second generations ($\sim$ 80%), while benzodiazepines, Tricyclic antidepressants (TCAs) and Selective Serotonin Reuptake Inhibitors (SSRIs) were used to treat other disorders

Disorder	Sex	Age (M,SD)	Age range
Schizophrenia	F = 5	42(15)	29-65
	M = 20	47(11)	24-67
BD	F = 0	-	-
	M = 6	32(7)	27-37
MDD	F = 3	51(5)	51-56
	M = 0	51(8)	46-63
OCD	F = 4	40(6)	33-45
	F = 0	-	-

Table 4.1: Demographic characteristics of patients

(i.e., $\sim$ 70% SSRIs). The choice of including people with a different diagnosis was made since other mental disorders share some negative symptoms, as in the case of MDD or depressive phases of BP (Brüne et al., 2009; Geerts & Brüne, 2009; Troisi et al., 1998). Moreover, all the patients without a diagnosis of shizophrenia presented also psychotic behaviors. All the diagnosis were ascertained by expert psychiatrists operating in three psychiatric centers: the Psychiatry Unit of San Salvatore Hospital, L'Aquila, Italy; the Psychiatric Clinic of the Department of Neurosciences, University of Padova, Italy; Department of Clinical Neurosciences, IRCCS San Raffaele Scientific Institute, Milan, Italy. The DSM-IV-TR nosology classification system (APA, 2000) was used to make the diagnoses. Inclusion criteria for the clinical group were: The presence of at least one negative symptom of schizophrenia; being native Italian speakers; finally, being treated with a stable dose of the same pharmacological therapy. Exclusion criteria concerned the presence of: Severe traumatic brain injury or neurological disorders; mental retardation; alcohol or substance abuse in the past six months. The control group was composed by 134 individuals (mostly students) randomly selected from the population and recruited in Padova (100 females). The majority of the control group, whose age ranged from 19 to 67 years ($M = 26, SD = 3.4$), had at least an

high school diploma. The exclusion criteria for non-clinical group were: Absence of one between the aforementioned disorders; mental retardation; alcohol or substance abuse in the past 6 months. All the participants read and filled an informant consent before the interview. The psychiatrists/psychologists explained very carefully that the participation was voluntary, the non-intrusiveness of the study and the possibility to withdraw the interview at any time, without penalization or change in the therapeutic plan, in the case of patients. This study was conducted according the Declaration of Helsinki and was approved by the Ethical Committee of each collaborating center.

4.1.2 Procedure

The experimental procedure was divided into three steps: interview, stimuli definition and scoring.

Interview. All the participants attended a videotaped interview in which the psychiatrist/psychologist asked them a number of question taken from the Positive and Negative Symptoms Scale (PANSS; Kay, Fiszbein, & Opler, 1987), used as a guide to conduct the interview. This interview was performed during the standard assessment phases, in order to be less demanding especially for patients. The two speakers were seated in front of each other: In this way, the video camera located behind the interviewers' right shoulder could record all the body of the participant. The video camera used in the present study was a Sony PJ410, placed on top of a tripod (height: 120 cm); the video camera was remotely controlled and recorded on a 64GB micro memory card. All the operations on the camera (e.g., starting, video extraction, etc.) were performed immediately before or after each interview. All these precautions were taken in order to reduce both interviewer distraction and participants'

sense of being observed. The entire interview lasted from thirty to forty-five minutes. The informed consent was taken before the interview. Once the interview ended, the interviewer explained in details the aim of that interview and answered to possible questions.

Stimuli definition. After each interview, the video camera was transferred into a safe room in which the video was downloaded and linked to a code, unique for each participant. After that, the memory card was formatted. At that point, the samples of video for the scoring phase were extracted. The number of samples was empirically decided and tested. In particular, 10 pilot interviews out of 172 were randomly selected, fixing a priori the sample duration at thirty seconds and one minute (five interviews per sample duration). The number of samples was fixed at fifteen, trying to reach a balance between observation length and amount of information collected. The ten interviews were watched by five psychiatrists who usually work and help people with a diagnosis of schizophrenia: Their task was to fill the NANs after watching each observational sample. After all the observations, a group discussion was conducted with all the psychiatrists in order to gahter their opinions on both samples duration and amount. The one-minute lenght sample duration received the agreement of all the psychiatrists, who suggested that one minute could give the chance of observing the potential occurrence of all the twenty-two nonverbal behaviors of NANS, compared to samples lasting thirty second. Moreover, they agreed on the fact that fifteen samples were enough to collect all the required information. Consequently, each original video was split into fifteen samples. The decision of which sample to select from the original video was carried out by a script coded in Python (Van Rossum et al., 2007) that randomly extracted fifteen time strings from a given interval. Both the first and the last five minutes of interview were excluded, since the former minutes were considered as the habituation time to the

camera, while the latter could be biased by the fatigue caused by the interview. Once the fifteen time strings were defined, the video was edited by means of the Shotcut software (Dennedy, 2011): Before each sample, a countdown sequence and a beeper were added, respectively before and after it; the beeper's adding is a typical procedure used in the one-zero sampling method (Martin et al., 1993). The final set of edited samples was then shuffled and coded as a single video file (i.e, .mp4 extension), ready to be used for the last phase. It is important to stress how both the random selection and the randomization of the fifteen samples were performed to reduce order and sequence effects, minimizing the risks of observer biases described in section 1.4.1. Finally, a note on the data protection protocol deserves mention: All the original videos, or the edited ones, were encrypted and stored in hard disks, located in a locked and safe place whose access was permitted only to the involved researchers.

Scoring phase. The aforementioned procedure was applied to each of the 172 interviews conducted on patients and non patients. Then, each edited video was observed by two independent raters during an observational assessment, that was sampled according to the one-zero sampling method. In particular, the raters observed and rated the fifteen samples of each observation on an iMac 8.1, sitting at 70 cm from a screen of 1680×1050 inches. In order to watch the same screen at the same time, the raters were in the same room but separated by a dividing wall. Since both raters used headphones, they could not have any visual or auditory interference nor contact with each other. Furthermore, they did not interact with the experimenter, who managed the videos remotely. Each rater was trained to observe each item: A detailed description of the NANS was provided to both raters, explaining carefully all the possible manifestations of each item. The scoring rule was explained as well, with a particular attention to

those items containing several nonverbal behaviors. Several simulations of observations were conducted, both with and without the experimenter's help. After each simulation, the response patterns generated by the raters were compared to the ones scored by an expert psychiatrists, who usually works with these patient. The inter-rater reliability was fixed at $\kappa = 0.80$. Each rater continued the training phase until the reliability threshold with the gold standard was reached for three consecutive videos.

During the observational assessment, the rater observed each video sample until the beeper's sound warned her/him that the session was ending. After the beeper, the video was stopped by the experimenter and the rater filled the NANS, signing an item only if the described nonverbal behavior(s) occurred within that sample. A tablet was used to fill the checklist, reducing the compiling time. After the completion of all the fifteen repetitions, the assessment ended and a break of at least twenty minutes was suggested. In order to reduce inaccurate observations due to fatigue, a limit of five observations per day was established. Finally, the fifteen response patterns for each patient were stored to obtain the modal response pattern, for each rater.

4.2 Data analysis

4.2.1 Model fitting and parameters estimation

172 modal response patterns where obtained, for each rater, according to the procedure explained in section 2.4. In order to test the model of assessment provided by the two found clinical structures of the NANS, a combined modal response pattern obtained from the two provided by the raters was defined, collecting their agreements for each item and solving the disagreements by direct discussion. In case of persisting

disagreements, a third rater was involved, usually the psychiatrist/psychologist who interviewed the patients and made his/her diagnosis. If the disagreement persisted, the modal response pattern was discarded from the analysis.

Once obtained, the modal response patterns were used to (i) test the two models, (ii) obtain their fit to data and (iii) the error parameters estimates, by means of an Expectation-Maximization Algorithm (EMA; Dempster, Laird, & Rubin, 1977) implemented in Matlab code (i.e., the CEMBLIM algorithm, see Spoto, 2011). More specifically, data used as input of such an algorithm were the clinical contexts displayed by Table 3.5 and the modal response patterns (i.e, the 172 modal response patterns for both Movement and Prosint subscales). As pointed out by Falmagne and Doignon (2011), an index frequently used to test the goodness of fit of models to data is the Chi-square statistic. The models at hand, in fact, were tested by a Pearsons' χ^2 with the corresponding p-value calculated by means of a parametric bootstrap with 5000 replications. The decision of computing a bootstrapped p-value was made considering the sparseness of the data matrices emerged in this research, for which the asymptotic distribution of the χ^2 is not completely reliable (Reiser & VandenBerg, 1994; Spoto et al., 2010). Beyond the fit indexes, the algorithm was used to estimate all the parameters of the BLIM, namely the probability π_C for each clinical concept $C \in \mathcal{C}$ and the error rates $\beta_q \backslash \eta_q$ for each item q. The $\beta \backslash \eta$ parameters are extremely important for the accuracy of a model of assessment. They can be used, indeed, as fit indexes: As pointed out by Spoto, Stefanutti, and Vidotto (2012), even if the general goodness-of-fit is appropriate, high values of β and η may indicate that the model is misspecified and it should be revised. As mentioned in Section 2.3, the error rates should be low (Stefanutti & Robusto, 2009), in order to provide useful information about the reliability and the validity of both items and the structure of a tool. This

requirement is based on the assumption asserting that the probability to correctly observing an item q monotonically increases with the probability of q belonging to the clinical concept of the person (i.e, $\beta_q + \eta_q < 1$ for all $q \in Q$). Consequently, if the inequality $\eta_q < 1 - \beta_q$ for each item q does not hold, the corresponding model can not be judged as valid, since the probability of observing a false positive on an item q would be greater than the probability of really observing the item q (Spoto et al., 2018; Stefanutti et al., 2018). These assumptions always stay true throughout the Chapter.

4.2.2 Identifiability check

Whenever a probabilistic clinical structure is empirically tested, its model's identifiability should be checked. A probabilistic model can be intended as a triple (Θ, Ω, f) where $\Theta \in \mathbb{R}^n$ is the parameter space of the model (where n is the number model's parameters), $\Omega \in \mathbb{R}^m$ is the outcome space of the model (where m is the number of observable outcomes), and $f : \Theta \to \Omega$ is a mapping, called the prediction function, assigning to each parameter vector $\theta \in \Theta$ a corresponding element $\omega \in \Omega$ of the outcome space. When a probabilistic model like the BLIM is considered:

- A point in the parameter space is a vector θ containing the $\beta_q, \eta_q \in (0, 1)$ values for each item and a probability $\pi_C \in (0, 1)$ for each clinical concept;
- A point in the outcome space Ω is a probability mass distribution on the collection of response patterns.

A probabilistic model will be identifiable whenever f is injective. In other words, there is only one collection of parameters mapping to the same distribution on the

outcome space. On the contrary, if several collection of parameters map to the same point on the outcome space, the model and its parameters will be unidentifiable, since it is not possible to identify which parameters values allow the predictions on the outcome space. Recent evidences showed how the application of the BLIM to structures presenting some particular gradations leads to unidentifiable models (Spoto et al., 2012; Spoto, Stefanutti, & Vidotto, 2013; Stefanutti et al., 2018). This is the case of the *backward-graded* and *forward-graded* structures. A clinical structure $(Q, \mathcal{C})$ is said to be forward-graded in an item q if $C \cup \{q\} \in \mathcal{C}$ for every $C \in \mathcal{C}$. Within a forward-graded structure in an item q, the η_q parameter is unidentifiable. A clinical structure $(Q, \mathcal{C})$ is said to be backward-graded in an item q if $C \setminus \{q\} \in \mathcal{C}$ for every $C \in \mathcal{C}$. Within a backward-graded structure in an item q, the β_q parameter is unidentifiable.

Spoto et al. (2012) defined a way to detect the backward or forward-gradedness of a clinical structure starting from the clinical context: In particular, if a specific attributes is assigned to only one item (i.e., there are not other items investigating it), the clinical structure will be backward-graded in that item. Consequently, the β of that item will by unidentifiable. Moreover, if that item does not investigate other attributes, the clinical structure will be forward-graded in that item. Consequently, the η of that item will by unidentifiable. Once all the potential unidentifiable parameters are found, at least two solutions can be adopted:

1. Reducing to zero the probability of some clinical concepts containing items leading to a backward- or forward-graded structure;

2. Modifying the item-attribute assignment in a way that the backward- or the forward-gradedness is reduced;

3. Fixing the unidentifiable parameters to zero or to constants representing the

maximum possible values for the parameters, avoiding losing them and preserving accuracy (Spoto et al., 2018; Stefanutti et al., 2018).

In the present study, the last approach was selected: Once all the unidenditifable parameters were detected on the basis of the aforementioned rules, a constant value of 0.1 was assigned to the β of items in which the structure could backward-graded; likewise, a constant value of 0.01 was assigned to the η of items in which the structure could be forward-graded. These values were obtained according to the procedure defined by Stefanutti et al. (2018). The decision of setting the β higher than η allowed for maintaining the inequality $\eta_q < 1 - \beta_q$ for each item q and, therefore, the validity of the model.

4.2.3 Accuracy testing

The present Chapter was used to test the accuracy not only of the NANS, but also of the kind of observation in which it should be used. In particular, it was tested the hypothesis according to which multiple observations like the ones sampled according to the one-zero sampling method could provide more accurate data than single observations lasting for long time. As mentioned in section 1.4.1, the latter kind of observation is prone to a series of biases and interference (i.e., anchoring, primacy/recency effects, first minutes impression) increasing the chance of making false positives/negative on each observed behavior. Observations structured to observe the behaviors during multiple and less dependent time intervals could lead to more accurate results. In order to test this hypothesis, an experiment was designed: One month before the observation designed to collect data necessary to validate the NANS, the two raters were asked to watch only the videos of the thirty-eight patients that would have been used as a clinical

group for the later evaluation. In that occasion, the raters' task was to fill the NANS only once after watching the entire video. In order to make the results comparable to the ones obtained from an observation sampled with the one-zero sampling method (i.e., 1-minute $\times$ 15 observational samples), the duration of video was shortened to fifteen minutes; furthermore, the 15-minutes video was extracted from the central part of the clinical interview. In this way, two response patterns were obtained for each patient: A unique pattern obtained from the single observation and the modal response pattern obtained from the one-zero sampling observation. The distances between each pair of items belonging to the single and the modal response patterns were analyzed. In this regard, the choice of not using the symmetric distance allowed for considering also the direction of the distance. All the possible distances between the two patterns are displayed by Table 4.2:

Pattern	I_1	I_2	I_3	I_4
Single	0	1	0	1
Modal	1	1	0	0
Distance	-1	0	0	1

Table 4.2: Example of non symmetric distance between single and modal response patterns. The letter I stands for "Item"

In words, a -1 distance between two responses indicated that an item that was not observed during a single observation was later detected in the multiple one, suggesting a potential underestimation of that item during the single observation. A 0 distance indicated that an item (not) observed during the single observation was (not) observed even during the multiple one, denoting an agreement between the two kind of observations. Finally, a 1 distance suggested that an item observed during a single observation was not detected in the multiple one, suggesting a potential overestimation of that item during the former observation.

4.3 Results

All the 172 modal response patterns were used in order test models fit, since presenting a very high inter-rater agreement as a baseline ($\kappa = 0.94$). Consequently, the remaining disagreements were few enough to be easily solved. Even if the two tested structures were different in terms of the amount of concepts (52 for the Movement structure vs 288 for the ProsInt one), results showed a good fit to the collected data for both models (Movement: $\chi^2_{(952)} = 115.92$, bootstrap-p = 0.14; ProsInt: $\chi^2_{(3784)} = 59.79$, bootstrap-p = 0.11). Interesting results were observed also for the estimates of β and η. As a first step, both clinical context referring to Movement and ProsInt subscales were examined in order to check for items whose error parameters were unidentifiable. In this way, the estimates of such error parameters would have been adjusted. The clinical context of Movement subscale did not show any item causing backward or forward gradedness, while the clinical context of ProsInt susbscale revealed different results: In fact, the ProsInt clinical structure was both backward and forward-graded in items 20 and 21, and only backward-graded in item 22. As mentioned in section 4.2.2, all the unidentifiable parameters were assigned to constant values, namely 0.1 for the β in case of backward-gradedness and 0.01 for η in case of forward-gradedness.

Considering also these adjustments, adequate error parameters were found for all the items of both subscales, as displayed in Tables 4.3:

In regards to the Movement subscale, the estimated η were extremely small. This meant that the probability of committing a false positive for those items was estimated as particularly low. Even among the β parameters several low values were found, with three exceptions: The first one regarded Item 3 who reached a β equal to 0.22, a moderate value but actually explainable. Item 3, in fact, describes a behavior that

Mov			Pro		
Item	β	η	Item	β	η
1	0.0001	0.0004	11	0.0001	0.0001
2	0.0001	0.0001	12	0.0627	0.0004
3	0.2175	0.0001	13	0.0001	0.0001
4	0.0001	0.0001	14	0.0001	0.004
5	0.1304	0.0067	15	0.0001	0.0001
6	0.0528	0.0001	16	0.2534	0.0001
7	0.1694	0.0001	17	0.25	0.0001
8	0.25	0.0001	18	0.3337	0.0003
9	0.4002	0.0001	19	0.0001	0.0001
10	0.0001	0.0001	20	0.1	0.01
			21	0.1	0.01
			22	0.1	0.0001

Table 4.3: Error paramters of both Movement and ProsInt structure's items

needs to be judged not only during the conversation, but also during the pauses between two different questions; this double evaluation could require more attentional efforts, consequently the risk of being misinterpreted could be high. Likewise, Item 8 showed a β equal to 0.25, a value potentially critical attributable to the fact that such a behavior needs the occurrence of a reduction in three body parts (i.e., head, hands and body) to be filled; it could be agreed that, therefore, it is extremely complex to observe. Finally, Item 9 showed a β equal to 0.4: This item requires that the patient reacts to a stimulus provided by the interviewer (i.e., a smile); consequently, its occurrence is more difficult to be checked, since it is linked to an action that involves another person. Similar results emerged with respect to the ProsInt subscale, were almost all the η parameters were very low, despite some backward and forward-gradedness. Finally, even the β parameters of ProsInt were extremely low, except Items 16, 17 ($\beta = 0.25$) and 18 ($\beta = 0.33$), investigating sets of either highly specific (i.e., 16,17) or global (i.e., 18) behaviors that require great focus and expertise to be accurately

observed. These parameters, although some of them could seem still high, could allow a good performance if assigned to an adaptive algorithm, as will be discussed in the next Chapter.

4.3.1 Accuracy testing

Interesting results emerged examining the distances between the response patterns obtained from a single observation compared with the ones obtained by conducting multiple observations. The amount of (dis)agreement between the two types of observations is depicted by Figure 4.1. Descriptive analyses revealed that several responses collected during single observations were not confirmed during the following multiple observations, suggesting an interesting amount of both under- or overestimation of items. The raters overestimated, on average, 2 items per patient: In particular, the occurrence of some behaviors, observed during the single observations, was not confirmed during the later one-zero observation (green dots in Figure 4.1). Interestingly, the majority of the overestimation (80.64%) was found for items which were observed in the single observation but never observed during the one-zero observation, revealing a strong change in judgments that could suggest the occurrence of false positives. The most frequency rates of overestimation were for Item 1 ("The patient exhibits and keeps the same postures (independently if peculiar, unnatural, rigid or bizarre") and 2 ("The patient shows slow, few or no spontaneous movements").

The underestimation was even higher, with an average of 3 items. In fact, some items, whose behaviors were not seen during the single observation, were judged as observed during the one-zero observation. The majority of underestimated items (red dots in Figure 4.1) were observed with high frequency rates during the one-zero ob-

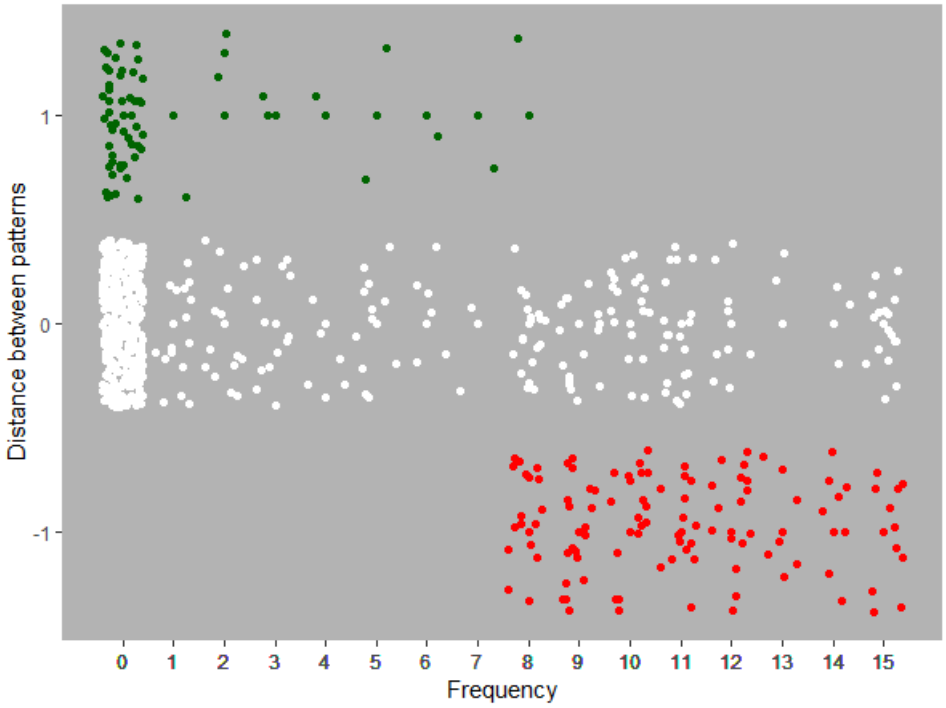


Figure 4.1: Mismatches between patterns derived from the two types of observations. Green dots indicate overestimated items. White dots indicate items that were (not) observed in both observations. Red dots indicate underestimated items.

servation, exceeding the minimal threshold necessary to be 1-scored (i.e.,8 times). For instance, the 61.26% of the underestimation behaviors were observed from 9 to 12 times during the one-zero observation. The most frequent underestimations were found in Item 8 ("During the conversation, the patient shows few or no movements of: hands, head and body. Note: to fill this item please check for ALL the elements of the list") and 16 ("Beyond slow movements and speech, the patient shows also a monotonic speech"), a scenario explainable considering that Items 8 and 16 needs the occurrence of more nonverbal behaviors to be checked, consequently they could be more difficult to observe during a single observation. This considerable amount of underestimation may suggest the occurrence of false negatives. These results suggested that data collected

after a single observation could be less precise compared to the same data collected during a multiple observation.

4.4 Discussion

The validation of a psychological instrument is a delicate procedure, especially considering an observational one. One of the most critical issues concerns the risk of committing false positive and negative errors when using such a tool, especially for mental disorders in which the behavioral component is extremely reduced. In these cases, the procedure of validation should be controlled at each step, from data collection to the model testing, with a particular focus on the estimation of the error parameters. The present Chapter tried to describe a series of controlled steps, in order to validate the NANS by testing the two structures composing it and estimating the error parameters for each item. The aspects controlled to reach these goals were different, starting from the type of observation used to conduct the assessments: A modified version of the one-zero sampling observation was used, composed by observational samples randomly extracted from the original interview and shuffled into a final video. This modification was applied in order to reduce the dependence between the observations. Moreover, the raters who used the NANS were trained according to procedure already applied in the scientific literature concerning the use of ethograms like the NANS. Finally, modal response patterns were used as raw data to insert in the algorithm testing the Movement and ProsInt structures; in order to obtain data as reliable as possible, the final set of 172 modal response patterns (for each subscale) was obtained by merging, for each participant, the modal response patterns provided by both raters and correcting them on the basis of their agreement. Results showed good fit indexes for both models,

with adequate β and η for each item: Only Items 3, 16, 17 and 18 showed higher β or η estimates; this result, although improvable, could be explained by the fact that all of them required either (i) the occurrence of several nonverbal behavior to be checked or (ii) the evaluation of global behaviors that are, by definition, more complex to be observed. Nonetheless, in the next Chapter it will be shown how such slightly higher β and η do not compromise the performance of an adaptive algorithm using them. On the contrary, all the found parameters will be a key factor for the development of the computerized adaptive version of the NANS: It will be described how the efficiency of the new instrument can depend on such parameters, since an item having low error rates can enhance the completion of the adaptive assessment in a more accurate and efficient way.

Another perspective of studying the false positive/negative rates has been deepened in this Chapter, by comparing the response patterns obtained from single and multiple observations (i.e., sampled accordingly to the one-zero sampling method). Results, even if descriptive, showed how a single observation could lead to considerable amounts of both over- or underestimation of some items. In particular, several items were overestimated in the single observation, an interesting result considering that, the majority of times, the judged proportion of occurrence of such items during the one zero sampling was zero. This considerable misinterpretation could be read as a false positive. Likewise, an higher number of underestimated items were found after scoring a single observation; such items were observed several times during one-zero observations, a result that could be read in terms of false negatives' occurrence. An interesting link can be considered between the accuracy estimates and the β/η parameters of some items, especially within the underestimated ones: It seemed that some items presenting larger β (e.g., Items 8 and 16) were also frequently underestimated. In some way, the β

values found for some items in the previous Chapter predicted the mismatch found in these results. This link confirmed the hypothesis concerning the high risk of wrongly observing some behaviors during a single and long observation. The reasons of such under/overestimation could be different. It could be a matter of confidence: The awareness of having only one occasion to judge a behavior as occurred or absent could lead the rater to conservative, choosing to not check an item unless if it clearly occurs; on the contrary, he/she could be extremely confident on his/her skills, overestimating or exaggerating a behavior's occurrence. Another explanation could be more cognitive: A single observation, lasting more time than a one-zero observation, could be more sensitive to memory interference, especially when the behaviors to observe are several; as a consequence, it cannot be excluded that some of them could catch more attention, interfering with the detection and the following memory of the other ones. More studies could test these hypotheses, in order to provide useful insight on this issue.

The present Chapter revealed also some limitations. For instance, the sample used to validate the structures of the NANS can be increased, even if the number of clinical concepts is not extremely high; in case of the ProsInt structure, more than 1000 patient are required to obtain stable results, assuming $\sim$ 4 person as necessary to validate each of concept of that structure. Unfortunately, such a number is not achievable in a few years, since it is difficult to recruit. As explained by Selten et al. (2000), it is unlikely that patients presenting negative symptoms of schizophrenia seek help for those symptoms. Furthermore, fifteen samples could be still a big amount of sample, maybe not affordable in clinical settings. Future studies could focus on how changing the modal threshold could lead to the same result using less observation's samples.

Nonetheless, the present Chapter showed also how a precise control over all the phases of a validation process could provide useful and reliable response patterns lead-

ing to essential information about the error rates of each item, as well as for the instrument's validation. For example, the decision of using the one-zero sampling method for the observation of participant's video allowed for observing the occurrence of multiple behaviors in short intervals of time, avoiding memory load and, consequently, reducing the risk of primacy or recency effects. Likewise, the randomization of observation samples allowed for reducing (i)anchoring, (ii) halo effect and (iii) the probability of forming a general early impression about the nonverbal communication style of the observed person. Moreover, the training of the raters allowed collecting modal response patterns that could be easily overlapped in order to define a final set of patterns to be tested using models like the BLIM. This aspect, in turn made it possible to obtain adequate estimates of the error parameters and probability distributions for clinical concepts. These last elements can be used as components of an algorithms able to implement the proposed observational checklist into a computerized adaptive instrument, able to complete an observational assessment suggesting less items without loss in accuracy. The development of such an adaptive instrument will be the focus of the next Chapter, which will introduce the so called Behavior-Driven Observation, namely the computerized adaptive version of the NANS.

Chapter 5

The Behavior-Driven Observation

Up to this line, the proposed assessment instrument seemed to reach the majority of the requirements listed in section 1.6. It was built with a clear, well-defined behavioral coding strategy, covering different types of behaviors (i.e., molecular and global ones); the new code was nestled in a checklist (i.e., the NANS) by means of FPA, which makes it possible to define a mapping between each item and a set of negative symptoms of schizophrenia. The validation procedure tested the goodness of the instrument and provided false positive/negative estimates for each item. Finally the one-zero sampling method of observation was adapted in order to reliably use this new checklist. Taken together, these results allowed for facing a set of issues related to the use of observational assessment instruments (e.g., behavioral code definition, accuracy of the instrument, focus on error estimates). Nonetheless, a critical issue remained unexplored, namely the efficiency of the NANS, in terms of time saving. The demand of time to perform in details an observational psychological evaluation is high (Yanagita et al., 2016), and it could increase when the time to integrate different information (i.e., deriving from interviews, observations and self-reports) is considered (Gibbons

et al., 2008; Groth-Marnat, 2009; Michel et al., 2018). The NANS, if used in its 22-item version, risked to be as (or less) efficient as the observational instruments it was compared in Chapter 3, since the observation of all the twenty-two items could require more time than expected. In order to reduce such a risk, the NANS was implemented into its computerized adaptive version. As discussed in section 1.5.1, a computerized adaptive assessment instrument is able to suggest the items to administer based on the responses obtained to previous items, reducing the number of asked items to complete the evaluation (Donadello et al., 2017). The use of adaptive assessment algorithms is strongly increased during the last years and different instruments have been developed according to different psychometric approaches (Spoto et al., 2018). For instance, several adaptive assessment instruments have been defined within the theoretical frame or IRT (Fliege et al., 2005; Gibbons et al., 2012). A very explanatory example of their functioning was provided by Fliege et al. (2005): Starting from a set of items, the algorithm selects the one maximizing the information about the score referring to the latent trait of the person; once administered the item, its response is used to update the score by means of methods such as the expected a posteriori estimation (Bock & Mislevy, 1982). The following items are selected among the ones carrying the highest information about the score, that is updated at each step of the assessment. The algorithm stops when a reliability value of $r \geq 0.9$ and a $SE \leq 0.32$ are reached (for further details, see Fliege et al., 2005). Other instrument have been developed, updating different parameters related to the latent trait dimension and covering different mental disorders (for a further description of other adaptive tests developed by means of the IRT, see Fliege et al., 2005; Gardner et al., 2002; Gibbons et al., 2012; Michel et al., 2018). These instrument, although efficient and accurate, present the same disadvantages of instruments build within IRT (see section 2.1).

The assessment instruments developed by means of FPA apply a different kind of adaptive system, called Adaptive Testing System for Psychological Disorders (ATS-PD; Donadello et al., 2017). It is a methodological refinement and an extension to psychological testing of an algorithm designed to evaluate, by means of an adaptive assessment's system, the knowledge of students on a specific topic (Falmagne & Doignon, 2011). So far, the ATS-PD procedure has been tested only on self-report measures investigating obsessive-compulsive disorder (Donadello et al., 2017) and major depressive episode (Spoto et al., 2018). Its extension to observational instruments is still unexplored. The present Chapter is aimed at introducing and further testing the Behavior-Driven Observation (BDO), an adaptive version of the NANS. The Chapter will proceed as follows: After describing in details the functioning of the ATS-PD algorithm, its extension for the BDO will be introduced (section 5.2). Then, section 5.3 will describe a simulation study in which all the parameters and the response patterns of NANS will be passed to the BDO algorithm, in order to test its accuracy and efficiency. Results (section 5.4) and implications in using the BDO will be finally discussed (section 5.5).

5.1 The ATS-PD algorithm

The ATS-PD algorithm was developed in 2016, within the theoretical frame of FPA, with the aim of extending adaptive algorithms used in the assessment of knowledge to psychological instruments (Donadello et al., 2017). The ATS-PD is able to take into account not only the deterministic side of a clinical structure, but also all its probabilistic features, by using all the parameters estimated from the application of the BLIM (i.e., probabilities of the clinical concepts π_C, the false negative β and the false positive η rates of each item). Basically, an ATS-PD algorithm works following

three rules:

1. Questioning rule
2. Updating rule
3. Stopping rule

Each step will be now described, considering also its meaning for a psychological assessment instrument.

5.1.1 The questioning rule

As a first step, the algorithm needs to select an item to ask, from the list of items that are contained in the concepts of the clinical structure. According to the *questioning rule*, the algorithm selects the item that best splits into two equal parts the probability mass of the concepts, namely that item q for which the sum of π_C, for all the clinical concepts containing it, gets closer to 0.50. This item is maximally informative, since it is able to maximize the attainable information irrespective of the received answer. This questioning rule is called also as *split-half* rule (compared to the *informative* rule that select the item reducing as much as possible the entropy of the likelihood on a trial n; Falmagne & Doignon, 2011). Whenever two or more items are eligible to be selected, the algorithm selects one of them at random.

5.1.2 The updating rule

Once the item is selected, the system administers it and collects the answer (i.e., "Yes" or "No"). On the basis of this answer, the algorithm applies the *updating rule*. In particular, the value of the answer is used to update the likelihood $L_n(C)$ of all the

concepts $C \in \mathcal{C}$ at the n-step[1], following this rule: Assuming to assign the vale 1 to a positive response (i.e., $r = 1$) and 0 to the negative one (i.e., $r = 0$), if the algorithm receives a 1 as input from the item q, it will increase the likelihood $L_n(C)$ of all the clinical concepts containing q and it will decrease $L_n(C)$ for all the other concepts. On the contrary, if the algorithm receives a 0 as input from the item q, it will decrease the likelihood $L_n(C)$ of all the clinical concepts containing q and it will increase $L_n(C)$ for all the other concepts. In this way, the likelihood $L_{n+1}(C)$ can be obtained for all the states $C \in \mathcal{C}$ at each step n, as displayed by the equation below:

$$L_{n+1}(C) = \frac{\zeta^C L_n(C)}{\sum_{C' \in \mathcal{C}} \zeta^C L_n(C')} \tag{5.1}$$

where

$$\zeta_{q,r}^C = \begin{cases} \zeta_{q,1} & \text{if } q \in C \ , \ r = 1; \\ 1 & \text{if } q \notin C \ , \ r = 1; \\ 1 & \text{if } q \in C \ , \ r = 0; \\ \zeta_{q,0} & \text{if } q \notin C \ , \ r = 0, \end{cases} \tag{5.2}$$

and ζ is the parameter that directly influences both the updating and the adaptive assessment process' efficiency. Within an adaptive algorithm, the parameter ζ can be fixed to a constant value greater than 1 (Falmagne & Doignon, 2011; Spoto et al., 2018) or it could be estimated by means of β and η for each item q, as displayed by the following two formulas:

$$\zeta_{q,1} = \frac{1 - \beta_q}{\eta_q}; \quad \zeta_{q,0} = \frac{1 - \eta_q}{\beta_q} \tag{5.3}$$

[1]It is important to stress that at the beginning of the assessment, the probability distribution of the concepts is uniform.

The estimation of ζ by means of β and η parameters to update the likelihood of the clinical concepts means that the error parameters can influence the extent to which an item can update the likelihoods. In other words, if a item is highly reliable, its error parameters will be very low, consequently producing a relevant modification on the probability distribution of all the clinical concepts. The more relevant are these modifications, the more efficient will be the algorithm in reaching the final result. Finally, a further refinement can be computed, namely a Bayesian rule able to update the concepts' likelihoods given their observed response patterns:

$$P(C_i|R) = \frac{P(R|C_i)L_n(C_i)}{\sum_{j=1}^{|C|} P(R|C_j)L_n(C_j)} \tag{5.4}$$

where $P(R|C_i)$ is obtained by equation (2.7) and $L_n(C_i)$ is the estimated likelihood of a clinical concept C at the step n of the procedure. This Bayesian refinement could be implemented either at each step n of the adaptive procedure (i.e., online) or when the stopping criterion is reached (i.e., offline), updating only in the end the likelihood of the final clinical concept. A recent study observed how adaptive algorithms implemented with the online Bayesian updating rule can efficiently reproduce a set of non adaptive response patterns, compared to algorithms either not adopting it or using it offline. These results were even more consistent if the Bayesian updating rule adopted a ζ estimated by means of β and η parameters (Spoto et al., 2018).

5.1.3 The stopping rule

The algorithm continues to select questions and update the concepts' probabilities until a stopping criterion is reached. Usually, this *stopping rule* is satisfied when $L_n(C_q)$ exceeds the interval $[0.20, 0.80]$ for each item $q \in Q$, meaning that the item is

splitting into unequal parts the mass probability of the concepts. It has been shown how this stopping rule is equivalent to another one based on the entropy of the system: In particular, when the criterion is reached, the entropy of the system presents an adequately low value, that usually is equal or less than 1 (Donadello et al., 2017). After this last step, the algorithm stops the assessment and the output is generated, containing the response pattern R, the estimated clinical concept C with its related probability value and the amount of time and questions required to end the assessment.

5.2 The Behavior-Driven Observation

The Behavior-Driven Observation is the computerized adaptive version of the NANS checklist developed through FPA. It is aimed at helping psychologists/psychiatrists in efficiently observing the nonverbal behaviors related to the negative symptoms of schizophrenia. As for NANS, it is important to stress that also the BDO has been developed with the idea of being a module of a more comprehensive assessment of such negative symptoms, without substituting any other type of assessment method. The BDO has been coded for the first time on R language (R Core Team, 2018) and later implemented in Shiny R (Chang, Cheng, Allaire, Xie, & McPherson, 2018) for research purposes. An R package containing the entire algorithm is in production phase and it will be usable as soon as possible[2]. The layout is very minimal, as displayed by Figure 5.1.

It can be used for both observations consisting in single trial or offline observations sampled with methods such as the one-zero sampling. In the latter case, the BDO algorithm will perform the assessment for each predetermined observation sample and

[2] Examples of the used functions can be showed on request.

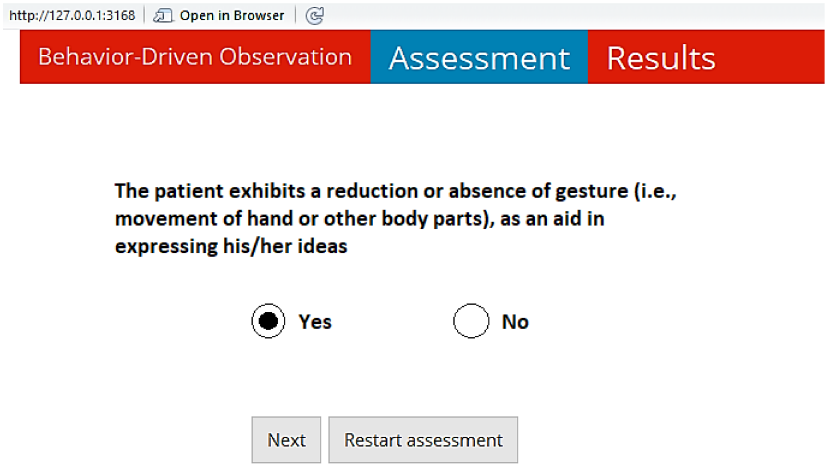


Figure 5.1: Opening page of the Behavior-Driven Observation.

the final output will be the clinical concept referring to the modal response pattern. This clinical concept will be estimated accordingly to each possible scenario discussed in section 2.4.

In general, the BDO algorithm follows the same questioning and stopping rules of the ATS-PD and the Bayesian refinement is implemented. The only modifications were made for the updating rule and the output generation. In regards to the former, the updating of the concepts' likelihood $L_n(C_i)$ at each step n was originally defined according to the following equation:

$$L_{n+1}(C) = \frac{\pi_C^q L_n(C)}{\sum_{C' \in \mathcal{C}} \pi_C^q L_n(C')} \tag{5.5}$$

where the parameter π_C^q represents the item-specific conditional probability of the observed response given the clinical state C. As introduced in section 2.3, this conditional probability is determined by $\beta \backslash \eta$ error parameters related to each item of the

checklist. Consequently, π_C^q can defined as:

$$\pi_C^q = \begin{cases} \beta_q & \text{if } r = 0 \text{ and } q \in C; \\ 1 - \eta_q & \text{if } r = 0 \text{ and } q \notin C; \\ 1 - \beta_q & \text{if } r = 1 \text{ and } q \in C; \\ \eta_q & \text{if } r = 1 \text{ and } q \notin C. \end{cases} \tag{5.6}$$

As mentioned by Heller and Repitsch (2012), starting from π_C^q it is possible to estimate $\zeta_{C,r}^q$ by means of η and β, obtaining the formulas described by equation (5.3). Consequently, even the BDO algorithm was implemented as described by equation (5.1).

In regards to the output, the algorithm is programmed to perform the assessment for fifteen times, referring to the fifteen samples of the one-zero sampling method: In particular, the algorithm calculates the modal response pattern and checks for its corresponding modal clinical concept, if exists; otherwise, it selects the clinical concept of the structure for which the symmetric distance from the original modal pattern is minimal. At that point, the output is generated, containing the modal response pattern M, its estimated clinical state C_M (that usually coincides with M) and the list of the negative symptoms related to C_M, comprehensive of their probability values. The probability of the attributes is estimated by using the bijection between the sets of items and the sets of their investigated attributes. In particular, the probability $P(a)$ of presenting a specific attribute $a \in A$ is estimated by summing the probability values of all the clinical concepts containing items investigating a at each step n of the assessment, as displayed by equation (5.7)

$$P(a) = \sum_{C_a \in \mathcal{C}} L_n(C_a). \tag{5.7}$$

where C_a are the concepts containing items investigating the specific attribute a and $L_n(C_a)$ is their probability value at each step n of the adaptive procedure. An example of BDO's output is in Figure 5.2:

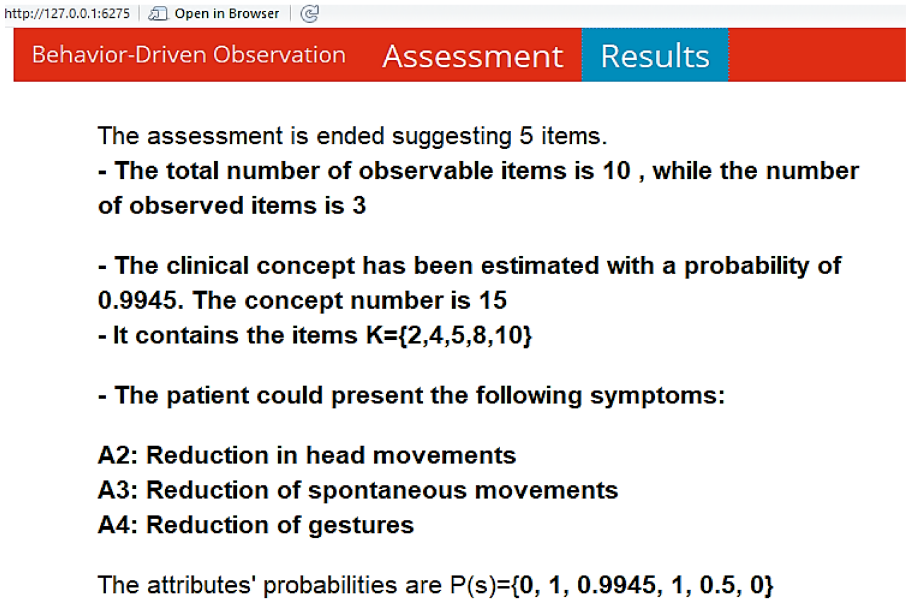


Figure 5.2: Example of output page of the Behavior-Driven Observation, Movement subscale.

In the next section, a simulation study aimed at testing the accuracy and the efficiency of the algorithm underlying the BDO, comparing also the first and the last implemented updating rules in order to test if there are differences between them. The simulation study is also aimed at testing the general efficiency of the BDO.

5.3 Simulation study

A simulation study was designed to test both the accuracy and efficiency of the BDO. The modal response patterns obtained from the NANS and validated in Chapter 4, were simulated by the BDO algorithm, in order to understand (a) if the BDO was able to accurately reproduce the original data, reducing the number of suggested items, and (b) if there were differences in accuracy/efficiency on the basis of the Updating rule implemented.

5.3.1 Methods

As introduced, the manipulated variable was the type of implemented Updating rule: In one version of the algorithm, the Updating rule was implemented on the item-specific conditional probability π_C of the observed response given the clinical concept C, as described by equation (5.5); in the other version, the Updating rule was implemented on the ζ parameter directly estimated from β and η error rates, as described by equation (5.1). The simulated modal response patterns of both algorithm's versions were compared in order to test possible differences in terms of accuracy and/or efficiency. Both versions simulated the 172 non adaptive modal response patterns (for both Movement and Prosint subscales) collected to validate the NANS: Specifically, each response pattern obtained from the fifteen samples of observation was simulated, for each patient; these fifteen response patterns were used to define the simulated modal response patterns. In the end, the simulated and the original patters were compared in order to test BDO's ability of reproducing the original modal response patterns by suggesting less items.

5.3.2 Outcome measures

In general, the average number of suggested items to reach the stopping criterion and to generate the output was used as a measure of efficiency. This index was calculated within each single sample of observation and across all the fifteen samples. The accuracy was tested by calculating the symmetric distance between the modal response patterns obtained from the NANS and the ones simulated by the BDO. Since measures of distance (independently if absolute or symmetric) are used to test dissimilarities, short distances were expected. As pointed out by Spoto et al. (2018), higher distances correspond to a significant inconsistency of the generated information between NANS and BDO output. This possibility could be explained only by some bugs into the algorithm, since the causes related to error parameters were reduced by the validation process. Recalling that:

- M_i is the modal response pattern derived from fifteen samples of observation by using the NANS for a person i;
- C_{M_i} is the clinical concept belonging to $\mathcal{C}$, obtained by the BDO when the input is M_i;
- The response patterns generated by an adaptive instrument as the BDO, will always generate a concept of the clinical structure $C \in \mathcal{C}$, independently if the original response pattern does not belong to $\mathcal{C}$;

it is possible to define the distance $d(C_{M_i}, M_i)$ as the cardinality of $C_{M_i} \Delta M_i$ (Spoto et al., 2018). On the basis of this distance, three scenarios are possible:

1. $M_i = C_{M_i}$: the modal response pattern $M_i \in \mathcal{C}$, consequently $d(M_i, C_{M_i}) = 0$. This means that the output of NANS and BDO exactly converge;

2. $M_i \neq C_{M_i}$: the modal response pattern $M_i \notin \mathcal{C}$, therefore $d(M_i, C_{M_i}) > 0$. This scenario may fall into two categories:

   i) $d(M_i, C_{M_i})$ is minimum, namely there is no $C^* \in \mathcal{C}$ such that $d(C^*, M_i) < d(C_{M_i}, M_i)$;

   ii) $d(M_i, C_{M_i})$ is not minimum, namely exists $C^* \in \mathcal{C}$ such that $d(C^*, M_i) < d(C_{Mi}, M_i)$.

The main hypothesis concerning the accuracy of the BDO was to find the majority of the modal response patterns exactly converging with the simulated ones, meaning that such modal patterns matched with clinical concepts of the clinical structures referring to both Movement and ProsInt subscales.

5.4 Results

Tables 5.1 and 5.2 display all the results in terms of accuracy. As expected, the BDO algorithm was able to reproduce the majority of non adaptive modal response patterns of the NANS, showing a low average symmetric distance between original and simulated patterns for both Movement and ProsInt subscales. Whenever a modal response pattern M_i was also a clinical concept C_{M_i} belonging to the clinical structure $\mathcal{C}$, this modal pattern was perfectly simulated by the BDO algorithm, such that their distance $d(M_i, C_{M_i})$ was zero. This convergence between patterns emerged as the most frequent result. In particular, the 92% of modal response patterns belonging to the Movement subscale and the 96% of modal patterns belonging to the ProsInt subscale exactly converged with the ones simulated by the BDO, independently of the implemented Updating rule.

Accuracy					
Updating	Structure	Δ	Δ_{min}	Δ_{max}	$\Delta \neq 0$
π_C	Movement	0.10	0	2	14
π_C	ProsInt	0.12	0	4	8
(β, η)	Movement	0.10	0	2	14
(β, η)	ProsInt	0.11	0	4	7

Table 5.1: Accuracy of the BDO algorithm, manipulating the updating rule. Δ is the symmetric distance between the original modal response pattern and the simulated one. π_C is the updating rule using the item-specific conditional probability of the observed response given the clinical state C. (β, η) is the updating rule using the β and η parameters.

Whenever the algorithm found that a modal response pattern M did not directly match with a clinical concept C_{M_i} belonging to the clinical structure, it was able to map the simulated modal response pattern into the closest minimal concept $C^*_{M_i}$ such that the distance $d(M_i, C_{M_i})$ was minimal. This was the case of 10 modal response patterns collected from the Movement subscales: For each non adaptive pattern, the algorithm simulated a modal response pattern mapped into a clinical concept $C^*_{M_i}$ such that the symmetric distance $d(M_i, C_{M_i}) = 1$, independently of the applied Updating rule. Likewise, the same scenario was found within the ProsInt subscale: When the BDO algorithm used the Updating rule implemented on π_C (i.e., the item-specific conditional probability of the observed response given the clinical state C), 2 simulated modal response patterns were mapped into clinical concepts $C^*_{M_i}$ such that the symmetric distance $d(M_i, C_{M_i})$ was equal to 1. The same result was reached for 1 modal response pattern simulated by the BDO algorithm whose Updating rule was implemented on the β and η parameters of each item. The algorithm found only a limited number of modal response patterns M such that the distance $d(M_i, C_{M_i})$ was not minimal. For the Movement subscale, 4 modal response patterns presenting a distance $d(C^*_{M_i}, M_i) - d(C_{Mi}, M_i) \leq 2$ were found, meaning that the distance between the concept C_{M_i} and

estimated concept $C^*_{M_i}$ closest to M_i was never greater than 2 items. In regards to the ProsInt subscale, 6 modal response patterns were estimated to a non minimal distance, with a difference of no more than 4 items between C_{Mi} and C^*_{Mi}. Table 5.2 summarizes all the found distances. As reported by Spoto et al. (2018), this rare situation could be caused by the type of Updating rule implemented in the BDO algorithm or just by the sequence of suggested behaviors by the system. The first cause was actually excluded, since the presented results were almost identical for both version of implemented Updating rules.

		$\Delta(C_{Mi}, M_i)$				
Updating	Structure	0	1	2	3	4
π_C	Movement	158	10	4	0	0
π_C	ProsInt	164	2	1	4	1
(β, η)	Movement	158	10	4	0	0
(β, η)	ProsInt	165	1	1	4	1

Table 5.2: Cardinality of the symmetric distances C_{Mi} and M_i between the original modal response patterns of the NANS and their simulated ones by the BDO algorithm. Results are showed for both Movement and ProsInt subscales, across the two implemented Updating rules.

All the results concerning the BDO's accuracy were supported by the efficiency ones. Table 5.3 shows the main findings in terms of efficiency. In general, both versions of the BDO completed the assessment suggesting less items than the NANS, simultaneously maintaining the accuracy introduced above. Infact, the two versions completed the assessment of the Movement subscale by suggesting, on average, 5.5 items (SD= 0.5) per observation sample, out of 10 of suggested by the NANS. This means that the BDO completed the assessment by asking 45% less items across the 15 samples. In regards to the ProsInt sub-scale, the algorithm ended the assessment by asking on average 7.7 items (SD=0.68) per observation sample, out of 12 suggested by the NANS.

Consequently, the average saving across the 15 samples was $\sim 36\%$ of suggested items.

Efficiency			
Updating	Structure	n	N
NANS	Movement	10	150
NANS	ProsInt	12	180
π_C	Movement	5.54	83
π_C	ProsInt	7.75	115.38
(β, η)	Movement	5.52	82
(β, η)	ProsInt	7.68	115.22

Table 5.3: Number of suggested items by the BDO, implemented using two different updating rules (i.e., π_C and (β, η)), starting from the response patterns of the NANS. The n refers to the mean of suggested items per single sample, while N refers to the mean of suggested items across 15 observation samples.

In sum, results suggested that both configurations of the BDO algorithm led to accurately complete the assessment, optimizing both the evaluation and computational time during a real time assessment performed on a machine.

5.5 Discussion

Accuracy and efficiency are fundamental features for a psychological instrument: The former reflects the reliability and the validity of an instrument; the latter can be read as the ability of reaching diagnostic information asking as few questions as possible. The efficiency plays a key role in perspective of an integrated assessment, since it allows to save time while gathering an adequate amount of information to make a diagnosis. So far, the NANS showed good measurement properties, with appreciable goodness of fit in both its subscales and low error parameters obtained from real data. Nonetheless, the procedure on which it was applied and the number of items composing it could reduce its potential in terms of efficiency. The need of producing accurate information

while reducing the time demand to collect them was still unsatisfied. The aim of this Chapter was to implement the NANS into its computerized adaptive observational instrument, the so called Behavior-Driven Observation. It was developed accordingly to the ATS-PD algorithms and extended to observational checklist like the NANS, which could be used during observational procedures such as the one-zero sampling method. In this way, the BDO was able to suggest the items to observe on the basis of the previously observed behaviors, through an algorithm that (a) selected the most informative item, (b) changed the likelihoods of the clinical concepts on the basis of the observer's responses, (c) repeated these two steps until a final clinical concept reach a stopping criterion. Such an algorithm was tested for accuracy and efficiency by means of a simulation study in which all the response patterns used to validate the NANS were simulated by the BDO algorithm. The number of suggested items to end the assessment was used to test the efficiency of the BDO, by implementing the algorithm on two different updating rules in order to find the the most efficient configuration. Results showed how both versions of the BDO completed the assessment asking only 55% of items for the the Movement subscale and the $\sim 64\%$ for the ProsInt subscale. No relevant differences were found in terms of both accuracy between the two implemented updating rules, suggesting that the BDO can replicate an assessment very accurately asking a lower number of items to generate a stable output, in all its configurations. From a general point of view, using 12-13 items instead of 22 to generate the same clinical concept represents an improvement in efficiency that goes beyond expectations: In fact, the algorithm ended the assessment with a reasonable number of suggested items although some of them were characterized by error parameters that could undermine the efficiency of the algorithm itself. It is possible that the noise introduced by those error parameters was counterbalanced by the rest of β and η

that were extremely low. The direct implication of the efficiency showed by the BDO concerns the possibility of reducing the time consumption typical of the observational assessment. Saving 9 or 10 items per assessment implies more time available to deepen delicate personal aspects of the patients.

The efficiency showed by the BDO was supported by substantial results in terms of accuracy. As expected, The BDO algorithm was able to reproduce almost all the original responses, defining simulated modal response patterns that exactly converged with the original ones. A few simulated modal response patterns presented a minimal distance with their non adaptive counterparts; finally, only 10 simulated patterns M (i.e., 4 M for Movement and 6 M for ProsInt subscales) showed a distance not minimal, actually not exceeding the four items. These results are not trivial: Considering that the modal response patterns are not directly observed, the correspondence with a clinical concept cannot be automatically assumed, even if all the response patterns composing it correspond to clinical concepts belonging to the clinical structure. The results at hand showed, indeed, that such a correspondence is not only assumable, but also very frequent. Furthermore, in those few cases in which a direct correspondence did not emerge, the closest clinical concepts were very similar to the original modal patterns. In this way, the clinical output provided by the BDO will be either the exact representation of the observed behavioral pattern of a patient or the most plausible one. Moreover, if the latter case occurs, the BDO is programmed to warn the clinician, by separately highlighting both the symptoms that are directly assumed by the observed items and the most plausible one estimated by the algorithm given such a behavioral pattern.

The study described in this Chapter presents some limitations that are worth of future focus. The efficiency showed by the BDO can solve only partially the problem of

the time consumption; the BDO is not an online evaluation, consequently the clinician should use extra time to use it properly. Even if such an extra time is less than the time spent for an entire clinical interview, it could not be available due to the amount of work required by an hospital's ward. Such an issue, actually, could be solved by making the BDO suitable for online evaluations, maybe performed on a tablet and programmed to automatically create slots of random observations lasting a few minutes, in which the behaviors to observe are suggested. In order to realize such an improvement, a linked methodological limitation should be solved, namely the number of observational samples to generate the modal pattern. Unless the fifteen samples have been obtained a consensus by experts who usually use interviews and observational tools, it could be argued that such a number of samples should be reduced in order to be less time demanding. In other words, the trade-off between a reduced amount of samples and reliable modal response patterns needs more focus.

Despite these limits, the results of the present study seem to suggest interesting future perspectives in the application of computerized adaptive observational assessment. The BDO is a little step forward in this direction, for different reasons. In terms of innovation, it is the first application of an adaptive system to an observational instrument, an issue unexplored and unattempted until now. The algorithm on which the BDO is coded allows it for collecting several accurate information in less time intervals. The only prerequisites to adequately use the BDO are a systematic observational procedure and an adequate training for the rater: Once followed these basic guidelines, the BDO could be applied, independently of the context on which the nonverbal behaviors are observed or the clinical experience of the rater. In the next chapter, this latter issue will be deepened, testing the BDO on field, in order to understand if the accuracy and the saving in terms of evaluation time still hold.

Chapter 6

Application of the BDO

6.1 Introduction

The need of computerized adaptive instruments in psychological assessment derives essentially from practical reasons: In psychological testing, for instance, the administration of high amounts of items to assess a mental disorder could cause patients' fatigue and distress. This is the case of instruments such as the Minnesota Multiphasic Personality Inventory-2 (Ben-Porath, Tellegen, & Graham, 2008; Forbey & Ben-Porath, 2007) that, considering its last reduced version (MMPI-2 RF; Inventory, 2), is composed by a list of 338 items, which a person should fill almost entirely to receive a reliable profile. The burden in terms of administration time for patients (and scoring/interpretation procedure for clinicians) could not justify the accuracy of the provided information, especially if the consequence of such a long assessment is a treatment's delay or elevate costs for the health care system (Kirisci et al., 2012). Observational instruments could be even more burdensome, since both the administration and the scoring/interpretation phases are usually performed by the same person, who

can unintentionally make mistakes in judging the presence (or absence) of a behavior. Computerized adaptive assessment tools seem to reduce these critical aspects, since their underlying algorithms are able to make inferences systematically correct on the basis of the received responses, reducing the administration time and preserving the accuracy of the collected data (Spoto et al., 2018). The accuracy and the efficiency of these adaptive instruments are usually tested by means of two types of study, that pave the way to their application: The first type of studies consists in simulating a set of paper-and-pencil response patterns by means of the adaptive algorithm underlying the instrument, in order to understand if it can reliably converge in the same scores/response patterns by asking less questions and, therefore, saving time. The use of such studies has a long tradition in Computerized Adaptive Testing (Donadello et al., 2017; Forbey & Ben-Porath, 2007; Kirisci et al., 2012; Spoto et al., 2018), since having the advantage of testing the same data twice, without the collateral effects of a double administration (e.g., learning effect or fatigue).

The second group of studies consists in administering both the adaptive and non adaptive versions of the instrument. In these studies, the same group of people is evaluated twice, filling the two versions of the instrument into two assessment phases distant each other one or a few weeks (Zenk et al., 2007). This latter approach received a relevant attention in last decades, especially for those instruments on which a reduction of the number of items could be massive (Simms & Clark, 2005). In both simulation and field studies, the CAT version of all the tested instruments showed a very good ability to convergence with the scores of their non adaptive counterparts; Moreover, those scores were obtained with a greater efficiency and valuable savings in terms of both time and administered items. In regards to observational adaptive instruments, the findings of the previous Chapter show how simulation studies can be applied to

instrument like the Behavior-Driven Observation, obtaining valuable results in terms of reached accuracy and efficiency. On the contrary (and recalling the leitmotiv of this Ph.D. project) an application on field of a computerized adapted observational instrument is still missing.

Different aspects should be taken into account in order to adequately apply an adaptive observational instrument in field, which could be clustered into issues related to inter- and intra-rater agreements. Since the reliability of observational instruments is strongly related to the co-occurring judgments of two raters (at least), it is fundamental that their agreement is kept constant (Castorr et al., 1990); likewise, the internal coherence of each rater across multiple observations should be substantial. In the field of observational assessment, there is a general consensus on the importance of training raters to maximize both the inter-rater and the intra-rater agreements (Castorr et al., 1990; Cusick, Vasquez, Knowles, & Wallen, 2005; Haidet, Tate, Divirgilio-Thomas, Kolanowski, & Happ, 2009; Zenk et al., 2007). The training of a rater can be considered as a sequential procedure allowing her/him to acquire a standardized way of observing that possibly reduces personal interpretations or inferences (Curyto, Van Haitsma, & Vriesman, 2008; Washington & Moss, 1988). Trying to delineate a general outline of a training procedure, it is possible to identify three main phases (Castorr et al., 1990):

1. Training on the use of the observational instrument. This phase consists in introducing the instrument that raters will later use. The trainer should explain each item and provide videotaped examples of the behaviors described by the items, if available (Haidet et al., 2009). The scoring rule should be carefully explained as well, including cases in which the attribution of a score is difficult (Haidet et al., 2009). More time should be spent on those items presenting low

reliability indexes or high rates of false positives and false negatives, since those items are the most prone to be misunderstood. After a deepen discussion on all the possible manifestations of each item and how to consider them, a preliminary testing phase is recommended. All the raters should observe sample of videos and later judge the occurrence/non occurrence of a set of behaviors (Haidet et al., 2009). The obtained response patterns should be compared with a gold standard one collected by an expert user of the instrument. Results should be discussed in group and feedback should be provided to raters (Zenk et al., 2007).

2. Testing the raters. Once completed this preliminary phase, the raters are asked to conduct an observational assessment, individually, on a video containing an interview of a patient. The obtained response pattern is compared with a gold standard one and the inter-rater agreement is calculated, usually by applying the Cohen's κ or Intraclass correlation coefficient (J. Cohen, 1960; R. A. Fisher, 1992). If the overall agreement calculated on each behavior is less than a threshold value (ranging from 0.70 to 0.80, dependently on the research paradigm, Haidet et al., 2009) the assessment is repeated until such a value is reached. Otherwise, the rater could be consider as ready to perform the main experiment (Zenk et al., 2007).

3. Maintaining of the training effects. During the experiment, a number of checks on both inter- and intra-rater agreements should be performed, in order to detect possible changes in the evaluation modality of each rater or consistent disagreement between raters. If one or both scenarios occurred, a retraining phase should be applied (Castorr et al., 1990).

The use of training could minimize the chance of obtaining biased results caused

by poor attention or comprehension of the behaviors to observe, fatigue or changes in the attribution style of raters. These biases, in fact, could mask the goodness of an observational assessment. The benefits of training raters before using an observational assessment, even if adaptive, could overcome the costs in terms of training time, leading to more accurate data and enhancing the application in field research of the proposed observational instrument. The last part of this Ph.D. project is aimed at testing the Behavior-Driven Observation during real observations. In particular, two psychotherapists were trained to use both the NANS and the BDO, which were applied to observe the videos of twenty people with a diagnosis of schizophrenia. Extending the results found in the previous Chapter, the BDO was expected to accurately reach modal response patterns as similar as possible to the ones obtained by using its non adaptive version (i.e., the NANS). In other words, an high intra-rater agreement was expected, for both raters. Furthermore, the target modal response patterns were supposed to be generated by observing less items.

6.2 Material and Methods

Sample. A subsample of twenty-five patients was selected from the clinical group used to validate the NANS. Within this sample, all the twenty-five patients were diagnosed with schizophrenia (n = 25, 5 females; Age (M,SD) = 45.52(11.87); Age range = 24-67). As suggested by the psychiatrists who made the diagnosis, and confirmed by the use of the NANS in previous Chapters, all patients presented at least one negative symptom concerning the reduction of nonverbal behavior. The patients with diagnosis of schizophrenia were treated with anti-psychotics of first ($\sim$ 20%) or second generations ($\sim$ 80%). Both the inclusion and exclusion criteria were those described

in section 4.1.1. All the patients provided oral and written consensus to use their data also for this part of Ph.D. project, which was approved by the Ethical Committee of each center that recruited the patients involved. Even this research was conducted according to the Declaration of Helsinki.

Stimuli. The videos used to validate the NANS represented the main stimuli of this experiment. As described in section 4.1.2, these videos consisted in fifteen sequences of clinical interview, which was extracted from the original interview administered on each patient during ordinary assessment phases. In order to make a video suitable with the one-zero sampling method of observation, each of its sequences was preceded by a countdown and followed by a beeper warning the end of the sequence. The fifteen sequences of each video were defined, shuffled and coded into a single .mp4 file by using the Shotcut software (Dennedy, 2011).

Raters. Two expert female psychotherapists participated in this experiment. The first one is a clinical psychologist trained in CBT therapy with six years of clinical experience; the second rater is a clinical psychologist training in constructivist psychotherapy.

6.2.1 Procedure

Raters' training. Before starting the experimental phase, both raters attended a training. As a first step, the NANS was introduced and each of its items was discussed in details. A considerable amount of time was dedicated to discuss and explain Items 3, 16 and 18 for two main reasons: As discussed in section 4.3, each of them required the joint presence of multiple nonverbal behaviors. Moreover, this set of items is characterized by moderate values of false negative probability. This means that they

require a grater attention in order to not be erroneously observed. Likewise, a relevant part of this first phase was dedicated to the explanation of the difference between microsocial and global behaviors, specifying that the former have a sudden occurrence and last for a few seconds, while the latter are characterized by longer duration. Finally, it was carefully explained that the nonverbal behaviors described by the NANS can be differently manifested by a person. This is the case of Item 14 ("The patient fails to show normal vocal emphasis patterns, is often monotonic") that investigates the reduction of variation in pitch: It is possible that a patient could show two different kinds of monotonically speaking between two observational samples; in that case, the score to Item 14 had to be 1 in both samples. On this regard, the raters were warned about the randomization of the samples, consequently they should considered each sample as independent from the remaining fourteen.

Once all these details were deepened and discussed, the two raters individually evaluated one of the twenty-five patients, by filling the NANS. After the completion of all the fifteen samples of observation, the modal response patterns were calculated for each rater and compared with those used to validate the NANS, intended here as gold standards. The Cohen's κ was calculated for each pair of items contained in the modal response patters provided by raters and the gold standard. Since the average κ for each item was higher than the selected threshold (i.e., 0,70), the training phase was considered concluded and both raters proceeded to the following phase.

Experimental phase. A few days after the training, both raters started the experiment. They observed the remaining twenty-four videos twice, the first time filling the NANS administered on a tablet and the second time filing the BDO, or vice versa. The second assessment for each video was performed exactly one week later, in order to avoid that raters remembered how they previously scored the videos. The sequence

of administration was pseudo-randomly generated, for each rater. Each observational assessment took place in a laboratory of the Department of General Psychology at the University of Padova, previously set in order to let the raters watch the same video at the same time without interacting. In particular, both raters sat at 70 cm from a screen of 1680 $\times$ 1050 inches of an iMac 8.1, divided by a wall designed to both make possible a perfect view of the screen and to hide the view of the other rater. In order to avoid auditory interaction between raters, headphones were provided to raters, who were further asked not to speak neither with the experimenter nor between each other. A scheme of the experimental setting is displayed by Figure 6.1.

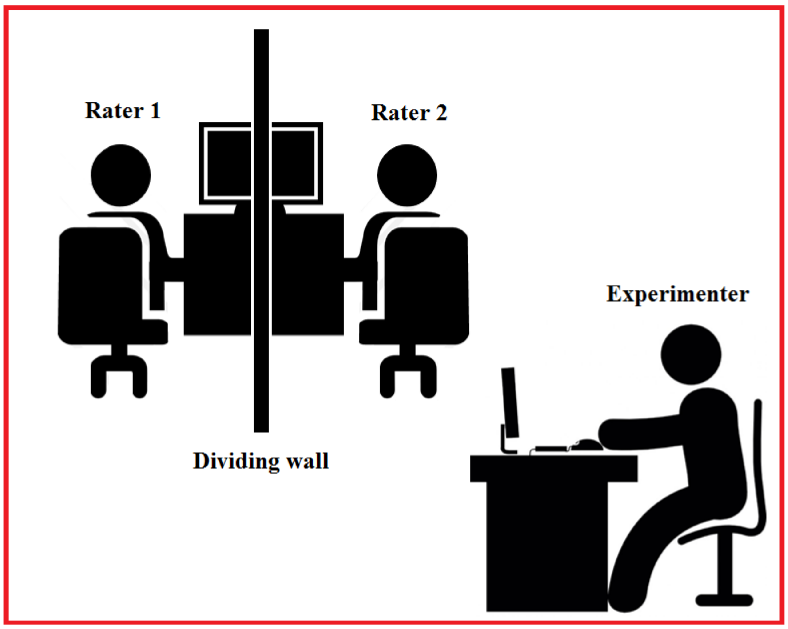


Figure 6.1: The experimental setting.

Each observational assessment was performed according to the one-zero sampling method: After the countdown preceding each observational sample, the raters observed an extract of interview lasting one minute; after the beeper, the video was remotely paused by the experimenter and the raters filled the NANS/BDO. Since the raters

evaluated the videos at the same time and in the same place, even the rater-instrument version assignment was randomly made. In this way, rater performed the assessment using the NANS, while the other rater used the BDO. After the fifteenth sample, the assessment stopped and all the response patterns were collected. In order to avoid biased responses caused by fatigue (Haidet et al., 2009), a break was proposed before a new assessment session and no more than five videos in one day were observed.

6.2.2 Data Analysis

The hypothesis concerning the ability of the BDO of generating modal response patterns converging with the ones collected by using the NANS was tested by estimating the Intraclass Correlation Coefficient (ICC; R. A. Fisher, 1992) and the symmetric distances between modal response patterns obtained by NANS and BDO. The ICC estimates, here used as a measure of intra-rater agreement (Koo & Li, 2016), and their 95% confidence intervals were estimated based on a single rater, absolute agreement, two way mixed effects models; the latter index was used, on the contrary, as a measure of dissimilarity between patterns. Furthermore, the expected efficiency of the BDO was tested by means of a linear mixed effect models, setting the total number of suggested items across each observational sample as a dependent variable, the instrument as predictor (i.e., NANS vs BDO), the intercept for each patient as random factor and the patient as the cluster variable. A significant difference in terms of suggested behaviors was expected, with an average lower number of suggested items for the BDO. Finally, the Cohen's κ was calculated to test the the inter-rater agreement. On this regards, at least a moderate agreement between raters' response was expected, within both NANS and BDO. The disagreements were analyzed by estimating the symmetric

distances between modal response patterns. Both intra- and inter-rater agreements were calculated by means of the *irr* package inside the R statistical software (R Core Team, 2018).

6.3 Results

For the final analyses, the first four videos were discarded, since they were used as a further baseline to make the raters more confident in the use of both NANS and BDO. The expected ability of the BDO to give as output modal response patterns converging with those obtained by the NANS was supported by results concerning the intra-rater agreement, presenting high values of ICC. In particular, Rater 1 obtained an ICC of 0.78, suggesting both a good level of intra-rater reliability and a considerable capacity of the BDO to produce converging modal response patterns. This result was confirmed from the analysis conducted on the amount of disagreement between patterns: The symmetric distances calculated between pairs of responses belonging to the obtained patterns revealed an average mismatch of 1.85 items per patient. Rater 2 obtained slightly higher indexes, showing a stronger intra-rater agreement (ICC = 0.81), with an average symmetric distance between modal response patterns of 1.5 items per patients. Table 6.1 summarizes the intra-rater agreement results.

Rater	ICC	95% CI.LB	95% CI.UB	F Value	df_1	df_2	p value
1	0.78	0.74	0.82	8.26	439	440	$< .001$
2	0.81	0.77	0.84	9.25	439	440	$< .001$

Table 6.1: Intra-rater agreement and symmetric distances between modal response patterns obtained by NANS and BDO. Results are displayed for both raters.95% CI.LB/UP stands for "95% Confidence intervals, Lower and Upper Bound

In regards to the efficiency with which the BDO led to modal response patterns

similar to those obtained by the NANS, the results emerged in the previous Chapter were confirmed. In general, both raters completed the assessment by observing, on average, 13 items suggested by the BDO (Figure 6.2), compared to the 22 of the NANS ($t_{(77)} = 84.3$ $p < .001$). This meant an set of $\sim$ 195 out of 330 suggested items, corresponding to an average items saving of 41%.

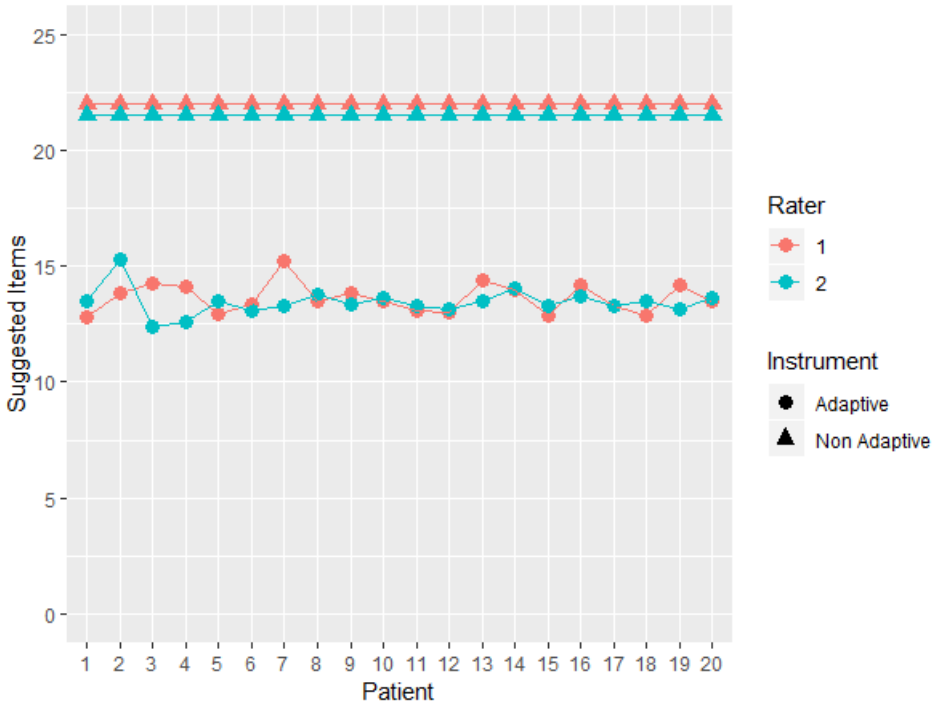


Figure 6.2: Average amount of suggested items per single sample of observational assessment, by using NANS and BDO. Results are displayed for both Raters.

The savings on the total amount of suggested items corresponded to savings in terms of assessment time: excluding the fifteen minutes of observation, the average overall amount of time to fill the BDO for each patient consisted in 23 minutes and 20 seconds (for both raters), compared to a scoring time of $\sim$ 30 minutes and 20 seconds when the NANS was used (Figure 6.3). The average saving for the total observational assessment was around 7 minutes.

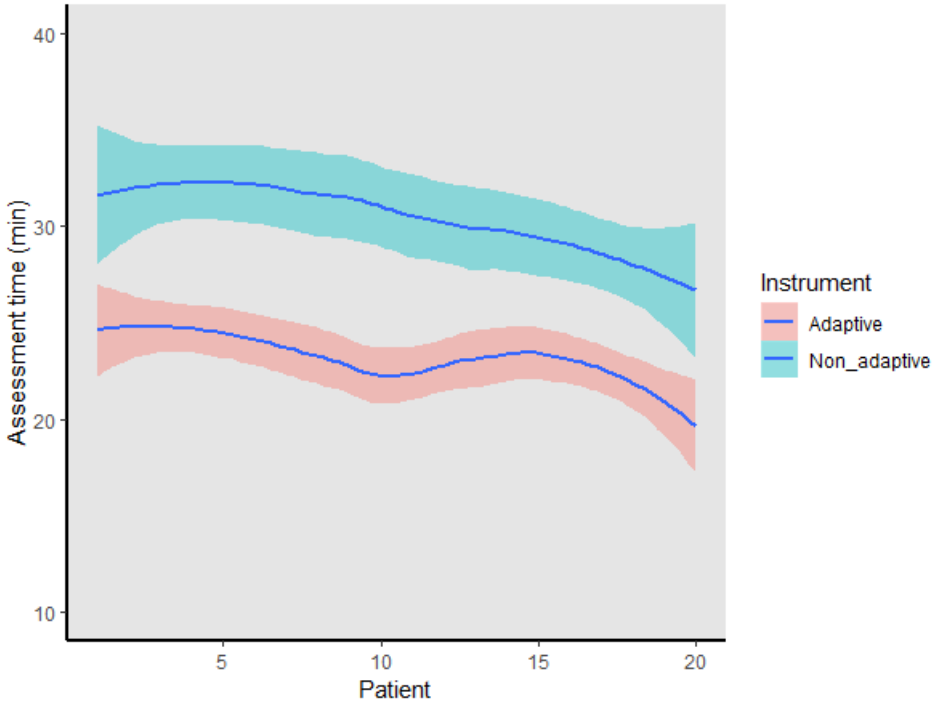
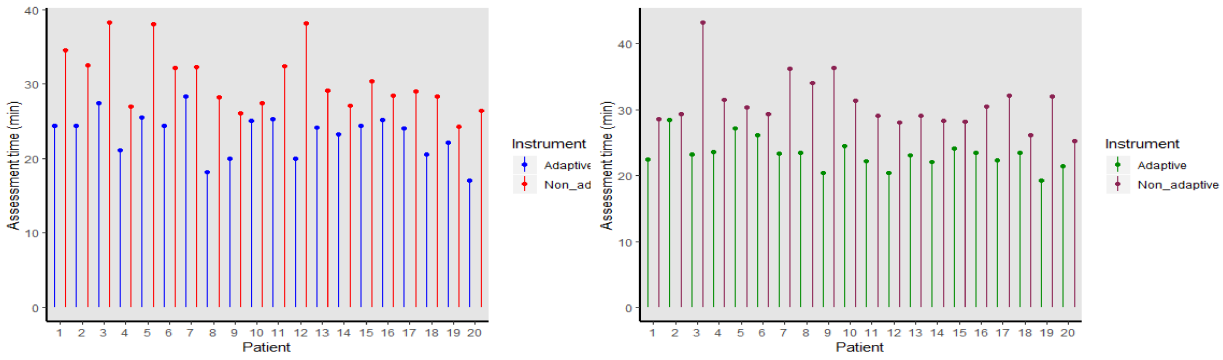


Figure 6.3: Total amount of time to complete the entire observational assessment by using NANS and BDO.

By using the BDO, Rater 1 completed each sample of observational assessment by filling, on average, 13.63 items (SD = 0.64) out of 22 of the NANS; consequently, she completed the entire assessment with 204 suggested items, with an average saving of 38% items to observe. Consequently, the total amount of time to complete the assessment was reduced from 30 minutes and 51 seconds (SD = 4.16 min) to 23 minutes and 20 seconds (Figure 6.4a). Rater 2 reached the target modal response patterns by observing on average 13.43 items (SD =0.50) out of 22, ending the entire assessment by filling 201 items. This saving of observed behaviors led to a reduction in the average assessment time of 8 minutes and 12 second, with a decrease from 31 minutes and 32 seconds (SD = 4.07) to 23 minutes and 20 seconds (Figure 6.4b).

Finally, the analyses conducted to test the inter-rater agreement among the rater's



(a) Rater 1 (b) Rater 2

Figure 6.4: Amount of time to complete the entire assessment of each patient, by using NANS and BDO

modal response patterns led to moderate results. The Cohen's κ calculated between NANS modal response patterns of the two raters as well as between the one defined by the BDO, were moderate, respectively $\kappa = 0.62$ and $\kappa = 0.63$. In the former case, the symmetric distances calculated between modal response patters highlighted an average dissimilarity of 3.2 (non) observed items between raters. Likewise, the modal response patterns obtained by the BDO differed on average of 3.02 items.

6.4 Discussion

As suggested by several authors, the application of simulation studies could not be sufficient to assume that a computerized adaptive instrument can generate response patterns comparable to those obtained by its non adaptive counterpart (Forbey & Ben-Porath, 2007; Kirisci et al., 2012). The efficiency and accuracy of such a tool need to be empirically tested even in live assessment procedure, in order to check if the features characterizing the adaptive algorithm (e.g., item selection, item presentation order or stopping criteria) can adequately fit with real observations. The last part of

this Ph.D. project was dedicated to empirically test the BDO during real observations, extending the results obtained by the simulation study described in Chapter 5. The high values of intra-rater agreement, and the low dissimilarities between the generated modal response patterns suggested how the Behavior-Drive Observation can be successfully applied in field studies. In this regard, results showed how the raters, by using the BDO, were able to generate modal response patterns converging with the ones generated by administering the NANS. In particular, the intra-rater agreement calculated on the two modal response patterns obtanied for each patient was very good, for each rater. Such an agreement between modal response patterns was confirmed by the small number of dissimilarities between their responses: It was found, in fact, that the two modal response patterns delineated for each patient were different, on average, for no more than 1.85 items. These findings suggest interesting information not only on the formal properties of the BDO, but also on the effects of the training procedure. Firstly, it is well-known that the intra-rater agreement is used as an index of reliability for multiple observations conducted by the same rater (Koo & Li, 2016). The high values of the ICC emerged seem to suggest that the BDO could produce reliable data, independently of the rater who applies it. Secondly, the found results suggested that both raters were coherent with themselves across time, confirming their judgment on the occurrence/nonoccurrence of the nonverbal behaviors suggested by the BDO. Such a coherence could be ascribed to a clear definition of the nonverbal behaviors described in the items of the instruments, or to a positive effect of the training phase (or both). Future studies could analyze these possible explanations, both separately and considering their interaction.

Interesting information derived from the results concerning the efficiency of the BDO. The aforementioned convergence between modal response patterns was reached

with a valuable saving in terms of both time and amount of suggested items. The completion of the entire assessment filling only the 61% of items led to a reduction of the assessment time of 26%. Although the time saving was not extremely large, these findings showed how the efficiency of an observational adaptive instrument is not only an hypothesis, but an provable result. Oncoming studies are taking into account the possibility of further refining the adaptive algorithm and its deterministic skeleton, in order improve the efficiency of the instrument maintaining constant its accuracy.

The inter-rater agreement represents a point that deserves future focus: The moderate Cohen's κ obtained for both instruments suggested some differences in the way of observing some behaviors, a result strengthened by the fact that both raters disagreed on 3 items (on average) for each patient. These differences could be explained considering either the training or the raters themselves. At one hand, the training procedure attended by the raters was shorter than expected. It lasted half a day, since both of them resulted extremely responsive and quick in learning the target behaviors. This impression was confirmed during the last part of the training, when they conducted the preliminary observation; in fact, both raters generated modal response patterns showing good inter-rater agreement with the one used as gold standard. It is possible that during time the effect of the training diminished, increasing the chance of judging the occurrence of a behavior on the basis of inferences or personal coding schemes (Curyto et al., 2008; Washington & Moss, 1988). A longer training and multiple checks testing the maintenance of the training effects should be considered in future studies. On the other hand, the "Rater" factor could have equal relevance on this result, independently of the usability of the proposed instrument. Recruiting expert psychotherapists as raters introduces another variable in the model: The experience. Both raters have experience in clinical assessment, consequently they had been already trained on ob-

serving the behavior of a person; this previous knowledge could have interacted with the new training received for this experiment (Haidet et al., 2009). In other words, it is possible that sometimes, and maybe unconsciously, one or both of the raters could have used their personal way of coding and judging a behavior, coherently with their different psychotherapy approach. Oncoming studies are testing the possibility of conducting this experiment by comparing expert psychotherapists with naive observers, possibly students in psychology, accurately trained in the use of NANS and BDO.

Beyond this limit, the results obtained by this last part of Ph.D. project gave interesting insights and future perspectives. The possibility of using the BDO in real observations could be a step forward in the field of adaptive observational assessment. The BDO, in fact, can generate accurate clinical output, highly comparable with those obtained by applying the NANS, by suggesting less behaviors to observed. This aspect means a reduction of the assessment time, independently from whether the observation is a single trial or sampled according to the one-zero sampling method; as a consequence, the time saved could be used to analyze the clinical output and integrate the collected clinical data to other information gathered by means of interview, self-reports or information acquired by direct discussion with people belonging to the social net of the patient. Future studies could use these results to make the BDO more efficient, implemented to be directly used within the clinical interview or during shorter one-zero sampling observations.

Chapter 7

Discussion

The complexity of a psychological assessment is typical of the hierarchical systems: At each level of the hierarchy an error, bias or at least a complication can happen. When the system is aimed at applying a new observational adaptive instrument, it is likely that critical issues will occur starting from the lowest levels of such a hierarchy. The coding of a set of behaviors easy to observe, the minimization of their false positives or negatives rates and the high time demand for evaluating each of them, are perfect examples of critical issue that could make the research on observational assessment extremely challenging. The present Ph.D. project tried to face all these critical issues during the development the Behavior-Driven Observation, a computerized adaptive observational checklist. In order to face the several issues related to the definition of this new observational instrument, different techniques and methodologies have been applied.

The first critical issue concerned the definition of the kind of observational instrument. As discussed in Chapter 1, this issue implies the consideration of several factors: The construction of a behavioral code composed by well-defined and easy-to-observe

behaviors; the implementation of such behavioral code into an instrument that gives the chance of accurately observing multiple behaviors; finally, the possibility of using such an instrument during observations designed according to different sampling methods. Observational instruments able to adequately take into account all these factors are the ethograms. The idea of developing the non adaptive checklist of this Ph.D. project as an ethogram was taken on the basis of some practical advantages characterizing it. It is well-known that ethograms can be composed by sets of behaviors that can be hierarchically grouped in mutually exclusive clusters (Brüne et al., 2008; Geerts & Brüne, 2009). This kind of organization allows observers to evaluate several behaviors without an excessive attentional load: They only need to sequentially concentrate on a cluster at a time and carefully evaluate each behavior contained in it. In this regard, the Nonverbal Assessment of Negative Symptoms checklist (NANS) proved to be a good example of ethogram. It is characterized by two subscales composed of dichotomous items investigating different nonverbal dimensions (i.e., movement and prosody/interaction aspects). The possibility of evaluating several behaviors during an observation could be extremely useful in clinical settings, in which the amount of information to collect from each patient could be considerable and difficult to manage.

Another advantage of using ethograms-like checklists is that they can be applied in observations designed according to different sampling techniques. These instruments, in fact, are composed by specific behaviors frequently rated on dichotomous scales, feature making them easy to administer and extremely versatile. In the present Ph.D. project, the NANS was applied within observations sampled with a modified version of the one-zero sampling method, leading to important insights on a debated issue regarding the use of such method in observational assessment. As pointed out in Chapter 1, when the one-zero sampling method is applied to observations, a video containing the clinical

interview of a patient is split into n equal-length samples: Each sample is watched by one or more raters, who compile an observational instrument or collect data about the occurrence/nonoccurrence of a behavior after each sample. When the last sample ends, the final response pattern is obtained by calculating the proportion of occurrence of each behavior, namely the sum of all the 1-scores for each behavior divided by the total number of samples. Several authors discourage the use of such "score" (Dunkerton, 1981; Leger, 1977; Powell et al., 1977), arguing that it is a biased measure of frequency and duration for the following reason: If a behavior x starts in a sample n_1 and lasts until the following sample n_2, its frequency of occurrence within each sample could be underestimated (Martin et al., 1993). Moreover, the occurrence of the behavior x in the n_2 sample is conditioned on the occurrence of x in the previous sample, defining a condition of dependence between samples that could bias the collected data. In the present project, this limitation was faced from both a methodological and measurement points of view. As a first note, it could be argued that selecting consecutive example is a biasing method per se, since the chance of observing a sequence effect is extremely high. Moreover, if a particular behavior is impaired or expressed in a dysfunctional way, it should be observed for the majority of time, unless it is stimulus-induced. For these reasons, a methodological modification was applied in this project: In particular, the samples were randomly extracted from the original video by a sequence generator programmed in Python language; in this way, the dependence between samples is reduced, as well as the risk of making a biased judgment on the occurrence of each behavior due to sequence or order effects.

Form the measurement point of view, using the proportion of occurrence of a behavior across multiple samples is equivalent to estimate the mean frequency of occurrence of that behavior. Indeed, this mean value could not be informative on the possibility of

judging a behavior as averagely occurring during the observation, since deriving from a series of dichotomous data. This type of data require another estimate of central tendency, that is the mode (Weisberg & Weisberg, 1992; Manikandan, 2011). When the main hypothesis is to understand if a behavior can be considered as occurred, the proposed proportion of occurrence should be substituted with the modal value of that behavior across the observational samples. Consequently, the final measure generated from the fifteen samples of observation used in the present project is a modal response pattern, composed by the modal values obtained for each item of the NANS across the fifteen observational samples. The modal response pattern has the advantage of being easy to understand, since it is composed by a series of one and zeroes indicating if the target behaviors, across all the observational samples, have been occurred or not. This aspect, in turn, could make the formulation of the diagnosis an easier process.

The second step toward the definition of the proposed observational adaptive checklist concerned the possibility to provide such an instrument with psychometric and methodological foundations. The Formal Psychological Assessment, introduced in Chapter 2 and implemented in Chapter 3, made possible to have a precise control over the entire procedure of the Behavior-Driven Observation development. By applying this new methodology, it was possible to define the final version of the NANS, able to exhaustively investigate negative symptoms of schizophrenia starting from a list of twenty-two items describing specific nonverbal behaviors typical of this symptomatology. The bijection emerged between the set of items and the set of negative symptoms allowed to define a model of assessment in which the clinician can study all the relations between items and their investigated attributes. In other words, whenever a clinician observes a particular item, she/he will know exactly the precise set of negative symptoms related to that item, progressively collecting useful information to

delineate the final clinical outcome. Within the NANS, the final clinical outcome is represented by the clinical concept, which contains the nonverbal behaviors observed by the clinician related to the negative symptoms that a patient could present. All these information assume an undoubted clinical relevance, since they can be used as a basis to set a therapeutic plan targeted on a specific set of symptoms. Indeed, the model of assessment provided by the NANS does not provide clinicians with only the target clinical outcome. From a theoretical point of view, the clinician could potentially know all the admissible clinical outputs obtainable from the instrument. In fact, the clinical structures derived from the clinical contexts of the NANS can establish which clinical concepts are admissible out of the total possible response patterns, whose number is the cardinality of the power set on the set of items. The possibility of estimating and taking into account all the admissible clinical outputs given the produced response patterns is an important aspect to consider. This is the case especially for a new observational instruments in which the possible combinations between items could produce a number of clinical outputs difficult to manage. In this regard, the two clinical structures delineated from the Movement and ProsInt clinical context were composed by a manageable number of clinical concepts (i.e., 52 and 288), providing two deterministic models ready to be validated.

The third part of this Ph.D. project concerned exactly the validation of the NANS. The goodness of fit of both subscales showed how the proposed instrument could be adequately used to reach the clinical aims for which it was developed. Moreover, this result suggested how the deterministic model of assessment provided by the NANS could be implemented on a probabilistic model to be used as a basis for the later definition of the final adaptive instrument. In fact, the validation procedure contained in Chapter 4 did not suggest only a good fit of both NANS subscales to data collected

during the observations, but provided further essential information, such as the false positive and negative parameters for each item of the instrument. It is possible to state that the error parameters represent one of the key factors of this project, not only because their values emerged as low or moderate, but for the implementations they allowed. For instance, the knowledge of the false positive/negative estimates made it possible to concentrate more efforts in training all the behaviors that are more prone to be erroneously observed, during the last experiment of Chapter 6. Furthermore, they were estimated also from the amount of under/overestimation made applying the NANS or the BDO in a single observation. In this regard, the comparisons between the modal response patterns generated by means of the NANS showed how a single and long observation could lead to several under/overestimation of observed behaviors, independently of the goodness of the observational assessment applied. All these results suggest how important can be monitoring the possibility of committing a false positive or negative when observing a behavior, since their occurrence could be explained by a variety of causes. Finally, the error parameters' estimates of each item, jointly with the other elements of the validated probabilistic clinical structure (i.e., the clinical concepts and their probability values π_C, the prerequisite relations among items), made it possible to realize the last part of this PhD project, namely the calibration of the adaptive algorithm underlying the Behavior-Driven Observation.

As introduced in Chapter 5, the error parameters were fundamental for both the accuracy and efficiency of the adaptive algorithm, since they can directly influence the updating of the likelihood of the clinical concepts during an adaptive assessment. In particular, items having low error parameters led to maximal increase or decrease of the likelihoods of the clinical concepts, consequently the algorithm had to collect less data to reach a stable results, without a loss in efficiency. This is what exactly

happened within this Thesis. Such a logic of likelihoods' updating, combined with the relations among items define by the prerequisite relations, defined a system of inferences promoting the efficiency of the BDO's adaptive algorithm. An example of the efficiency of this system is provided by considering the rationale of the prerequisite relations: Recalling that an item x is a prerequisite of another item y if and only if its set of investigated attributes is a subset of the attributes' set investigated by y, whenever a rater judged as observed the item y, the item x can be assumed as observed as well. As showed by the result of the last Chapters of this Thesis, when this rationale is implemented on the adaptive algorithm of BDO, a valuable gain in efficiency occurred, since the algorithm did not suggest to observe prerequisites of items considered as observed, avoiding useless redundancy.

A strength of the BDO adaptive algorithm, directly derived from the system explained above, can be found into its ability of mimicking the decisions made by an expert clinician during the assessment procedure. In fact, starting from a situation of high uncertainty, the clinician selects the behavior that could maximize the chance of collecting useful information (i.e., the maximally informative item selected by the questioning rule); after scoring the absence/presence of that behavior, he/she starts to define an hypothesis on the possible diagnosis, giving an higher priority to some behavioral patterns related to such an hypothesis (i.e., the updating rule). In the end, when the clinician believes that the amount of information is enough and other questions/time spent in observing would be redundant, he/she stops the assessment (i.e., stopping rule) and formulates the clinical diagnosis on the basis of the observed behavioral pattern. Results found in Chapter 5 and 6 gave a strong support to the ability of the BDO to mimic this logic of assessment. In the simulation study described in Chapter 5, the BDO algorithm was able to reproduce almost all the modal response

patterns collected by the NANS, completing the simulated assessments suggesting to observe, on average, only the 60% of the NANS' items. Such accuracy and efficiency were replicated in the Chapter 6, in which both NANS BDO was filled by two human expert raters, previously trained. Even in this case, the raters obtained modal response patterns converging almost perfectly with the ones obtained by using the NANS, observing $38-40\%$ items less. Furthermore, the items' saving corresponded to a reduction of the observational time of 26%, a small but consistent result encouraging to continue the research on the implementation of the BDO algorithm, especially considering that its efficiency could have other relevant implications, for both clinicians and patients. If the clinician has to focus on less behaviors, his/her cognitive load is reduced. Consequently, such clinician could feel less fatigue and could use the time saved to more accurately plan the treatment strategy for the patient. This time saving will directly benefit the patient, who could receive a precise feedback of his/her clinical condition without the sensation of being observed for an unnecessary amount of time.

If the BDO's algorithm can be considered as a relevant technical innovation, its clinical output represents an equally important clinical improvement, considering the information that contains. The BDO output is not represented by a numerical score, but includes a series of information such as: All the single response patterns obtained during the observation's samples, the modal response pattern obtained from these single patterns, its corresponding clinical concept (or the most plausible one) and its probability value. The clinical concepts includes, in turn, the set of necessary and sufficient negative symptoms showed in that specific patient, including the probability value of observing those negative symptoms. The information on the symptoms probabilities is extremely useful, since gives the chance of deepening the evaluation of those symptoms that showed a slightly lower probability of occurrence, maybe using

the time saved from the BDO administration. Once the clinician has completed even such a refinement phase, she/he can use the collected clinical output to integrate all the information of the entire assessment and later set an individualized treatment focused only on the symptoms showed by the patient. The benefits of an individualized treatment could involve both the patient, who obviously is the priority, and the entire health care system. At one hand, the patient could be treated with higher accuracy, reducing also the discomfort of staying for long periods in psychiatric wards or mental health services, in favor of more time spent with his/her social net. On the other hand, mental health care services could observe a reduction of treatment costs, which are elevate with patients presenting a diagnosis of schizophrenia, who usually need a pharmacological therapy to enhance the effects of the psychological one. During last years, great efforts have been made on the research focused on personalized treatment of psychosis, especially from a psychodiagnostic point of view. Recent studies report how the definition and the application of instruments able to calculate the risk of developing psychotic symptoms could have massive consequences for the health care system (Cannon et al., 2016; Carrión et al., 2016). An instrument like the BDO, that is designed to detect also under-threshold symptoms, could be advantageously added to these risk calculators: In fact, the BDO could add specific information to these indexes, which could remain general otherwise. Consequently, clinicians could know, quantitatively, how big is the patients' risk of developing psychotic symptoms and, qualitatively, which are these symptoms (or which could occur in case the psychosis get worse).

The clinical output provided by the BDO could be useful for three further clinical purposes: First, the BDO outputs could be used to discriminate patients who obtained similar scores from other instruments, but presenting slightly different behavioral pat-

terns. In these cases, knowing in which items their modal response patterns diverge would allow clinicians to find their different clinical concepts and, in turn, identify which negative symptoms they show differently. A second implementation of BDO clinical output concerns its use within screening phases: It could happen, for instance, that a person may show a small set of negative symptoms with probability values not extremely high. Since these information could not be sufficient to define a diagnosis, the clinician could take into account the possibility that those symptoms could indicate a prodromal phase of a negative symptomatology of schizophrenia. Therefore, she/he could use the clinical output as a starting point to monitor the emerged symptoms over time. The monitoring of the patients' symptomatology over time represents the third application of the BDO's clinical output. In fact, the clinician could administer the BDO over time and compare the evolution or the clinical concept and its related symptoms: If the patient reached a clinical concept lying at a lower level of the structure during later assessments, an improvement of his/her condition could be hypothesized; in case of a suspect worsening of the symptoms, the clinician could recalibrate the treatment plan.

The present Ph.D. project can be considered as a first step toward the use of computerized adaptive instruments in observational assessment. Therefore, it presents aspects that deserve more research. For instance, the methodology that underlies the BDO (and the NANS) requires more efforts in some aspects: At one hand, the response format of instruments built by means of FPA is dichotomous, consequently the meaning of a response concerns either the presence or the absence of a symptom/behavior. It could be useful to collect information also on the gravity of a specific behavior, extending, therefore, the FPA to polytomous data. Such an extension, actually, is one of the recent research lines in field of FPA. As mentioned in Section 3.6, the item-

attribute assignment that brings to the clinical context is a procedure time consuming and prone to inferential errors of the raters who define the context. A data-driven procedure able to delineate the clinical context from existing data would be interesting. From a practical point of view, more data can be collected in order to obtain more robust results, both in terms of reliability and validity of the instrument. Likewise, more observers should be used in order to test the ability of the BDO of adequately reproducing the modal response pattern of its non adaptive counterpart, with high inter-rater reliability rates. On these regards, oncoming studies are taking into account of recruiting naive students to be trained at using the BDO, in order to test if previous clinical experience in observing patients could interfere with the training provided to use to BDO. These limitations are typical of studies trying to introduce a new psychological instrument and, indeed, only one side of the coin. The other side is represented by the future perspectives and implementations planned to make the BDO algorithm more efficient. New researches have been planed to improve it, by allowing the "communication" between the two subscales. Briefly, the main idea is to use the clinical concept estimated from the Movement subscale to update the starting likelihoods of the Prosint clinical concepts. In this way, the results could be the same, but obtained more efficiently. Finally, great efforts are being made to obtain another relevant upgrade: The implementation of the BDO for observations in vivo. This future project could definitely solve the problem of time consumption linked to the use of observational instruments.

To conclude, the present Ph.D. project reached the goal of defining the first computerized adaptive checklist, that can be used even during complex observational assessments. As suggested by its name, the Behavior-Driven Observation could guide clinicians in the observation of the behavioral patterns related the negative symptoms

of schizophrenia, providing them with useful information ready to be integrated with other clinical data. In this way, it will be possible to define a personalized treatment that could help patients during critical moments of their lives.

References

Achenbach, T. M., & Ruffle, T. M. (2000). The child behavior checklist and related forms for assessing behavioral/emotional problems and competencies. *Pediatrics in review*, *21*(8), 265–271.

Albert, D., & Lukas, J. (Eds.). (1999). *Knowledge spaces: Theories, empirical research, and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Aleman, A., Lincoln, T. M., Bruggeman, R., Melle, I., Arends, J., Arango, C., & Knegtering, H. (2017). Treatment of negative symptoms: where do we stand, and where do we go? *Schizophrenia research*, *186*, 55–62.

Alpert, M., Shaw, R. J., Pouget, E. R., & Lim, K. O. (2002). A comparison of clinical ratings with vocal acoustic measures of flat affect and alogia. *Journal of psychiatric research*, *36*(5), 347–353.

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

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

Andreasen, N. C. (1982). Negative symptoms in schizophrenia: definition and reliability. *Archives of General Psychiatry*, *39*(7), 784–788.

Annen, S., Roser, P., & Brüne, M. (2012). Nonverbal behavior during clinical inter-

views: Similarities and dissimilarities among schizophrenia, mania, and depression. *The Journal of nervous and mental disease*, *200*(1), 26–32.

APA. (2000). Diagnostic and statistical manual of mental disorders, text revision. *Washington, DC: American Psychiatric Association*.

Argyle, M. (2013). *Bodily communication*. Routledge.

Azorin, J.-M., Belzeaux, R., & Adida, M. (2014). Negative symptoms in schizophrenia: where we have been and where we are heading. *CNS neuroscience & therapeutics*, *20*(9), 801–808.

Bauman, S. E. (2015). *Utility of the dyadic parent child interaction coding system with children with autism spectrum disorder: An investigation of reliability and validity*. University of South Alabama.

Baumeister, R., Vohs, K., & Funder, D. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*(4), 396-403. doi: 10.1111/j.1745-6916.2007.00051.x

Bellack, A. S., Mueser, K. T., Gingerich, S., & Agresta, J. (2013). *Social skills training for schizophrenia: A step-by-step guide*. Guilford Publications.

Ben-Porath, Y., Tellegen, A., & Graham, J. (2008). *The MMPI-2 symptom validity scale (fbs)*. Minneapolis: University of Minnesota Press.

Birkhoff, G. (1937). Rings of sets. *Duke Mathematical Journal*, *3*(3), 443–454.

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

Blanchard, J. J., & Cohen, A. S. (2006). The structure of negative symptoms within schizophrenia: implications for assessment. *Schizophrenia bulletin*, *32*(2), 238–245.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive eap estimation of ability in a microcomputer environment. *Applied psychological measurement*, *6*(4), 431–444.

Bond, T. G., & Fox, C. M. (2013). *Applying the rasch model: Fundamental measurement in the human sciences*. Psychology Press.

Boonstra, N., Klaassen, R., Sytema, S., Marshall, M., De Haan, L., Wunderink, L., & Wiersma, D. (2012). Duration of untreated psychosis and negative symptomsa systematic review and meta-analysis of individual patient data. *Schizophrenia research*, *142*(1-3), 12–19.

Bottesi, G., Spoto, A., Freeston, M. H., Sanavio, E., & Vidotto, G. (2015). Beyond the score: clinical evaluation through formal psychological assessment. *Journal of personality assessment*, *97*(3), 252–260.

Bretherton, I. (1992). The origins of attachment theory: John bowlby and mary ainsworth. *Developmental Psychology*, *28*(5), 759-775. doi: 10.1037/0012-1649.28.5.759

Briesch, A. M., Volpe, R. J., & Floyd, R. G. (2018). *School-based observation: A practical guide to assessing student behavior*. Guilford Publications.

Brown, W. H., Pfeiffer, K. A., McIver, K. L., Dowda, M., Addy, C. L., & Pate, R. R. (2009). Social and environmental factors associated with preschoolers nonsedentary physical activity. *Child development*, *80*(1), 45–58.

Brüne, M., Abdel-Hamid, M., Sonntag, C., Lehmkämper, C., & Langdon, R. (2009). Linking social cognition with social interaction: non-verbal expressivity, social competence and mentalising in patients with schizophrenia spectrum disorders. *Behavioral and Brain Functions*, *5*(6), 509–527.

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.

Buchanan, R. W. (2007). Persistent negative symptoms in schizophrenia: an overview. *Schizophrenia bulletin*, *33*(4), 1013–1022.

Cannon, T. D., Yu, C., Addington, J., Bearden, C. E., Cadenhead, K. S., Cornblatt, B. A., ... others (2016). An individualized risk calculator for research in prodromal psychosis. *American Journal of Psychiatry*, *173*(10), 980–988.

Cantor, N., & Mischel, W. (1979). Prototypes in person perception1. In *Advances in experimental social psychology* (Vol. 12, pp. 3–52). Elsevier.

Carrión, R. E., Cornblatt, B. A., Burton, C. Z., Tso, I. F., Auther, A. M., Adelsheim, S., ... others (2016). Personalized prediction of psychosis: external validation of the napls-2 psychosis risk calculator with the edippp project. *American Journal of Psychiatry*, *173*(10), 989–996.

Castorr, A., Thompson, K., Ryan, J., Phillips, C., Prescott, P., & Soeken, K. (1990). The process of rater training for observational instruments: Implications for interrater reliability. *Research in Nursing and Health*, *13*(5), 311-318. doi: 10.1002/nur.4770130507

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)

Chiu, C.-Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, *30*(2), 225–250.

Cohen, A. S., Kim, Y., & Najolia, G. M. (2013). Psychiatric symptom versus neurocognitive correlates of diminished expressivity in schizophrenia and mood disorders. *Schizophrenia research*, *146*(1), 249–253.

Cohen, A. S., Mitchell, K., Docherty, N., & Horan, W. (2016). Vocal expression in schizophrenia: Less than meets the ear. *Journal of Abnormal Psychology*, *125*(2),

299-309. doi: 10.1037/abn0000136

Cohen, A. S., Mitchell, K. R., & Elvevåg, B. (2014). What do we really know about blunted vocal affect and alogia? a meta-analysis of objective assessments. *Schizophrenia research*, *159*(2), 533–538.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46. doi: 10.1177/001316446002000104

Conejo, R., Guzmn, E., Milln, E., Trella, M., Luis Prez-De-La-Cruz, J., & Ros, A. (2004). Siette: A web-based tool for adaptive testing. *International Journal of Artificial Intelligence in Education*, *14*(1), 29-61.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological bulletin*, *90*(2), 218-244.

Curyto, K., Van Haitsma, K., & Vriesman, D. (2008). Direct observation of behavior: a review of current measures for use with older adults with dementia. *Research in gerontological nursing*, *1*(1), 52-76. doi: 10.3928/19404921-20080101-02

Cusick, A., Vasquez, M., Knowles, L., & Wallen, M. (2005). Effect of rater training on reliability of melbourne assessment of unilateral upper limb function scores. *Developmental Medicine and Child Neurology*, *47*(1), 39-45. doi: 10.1017/S0012162205000071

Davey, B. A., & Priestley, H. A. (2002). *Introduction to lattices and order*. Cambridge university press.

Davison, P., Frith, C., Harrison-Read, P., & Johnstone, E. (1996). Facial and other non-verbal communicative behaviour in chronic schizophrenia. *Psychological medicine*, *26*(4), 707–713.

de Chiusole, D., Stefanutti, L., & Spoto, A. (2017). A class of k-modes algorithms for extracting knowledge structures from data. *Behavior Research Methods*, *49*(4), 1212-1226. doi: 10.3758/s13428-016-0780-7

Dell'Osso, L., Armani, A., Rucci, P., Frank, E., Fagiolini, A., Corretti, G., ... Cassano, G. B. (2002). Measuring mood spectrum: comparison of interview (SCI-MOODS) and self-report (MOODS-SR) instruments. *Comprehensive psychiatry*, *43*(1), 69–73.

Del-Monte, J., Raffard, S., Capdevielle, D., Salesse, R., Schmidt, R., Varlet, M., ... Marin, L. (2014). Social priming increases nonverbal expressive behaviors in schizophrenia. *PLoS ONE*, *9*(10). doi: 10.1371/journal.pone.0109139

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.

Dennedy, D. (2011). Shotcut [Computer software manual]. Retrieved from `https://shotcut.org/`

Dickey, C. C., Vu, M.-A. T., Voglmaier, M. M., Niznikiewicz, M. A., McCarley, R. W., & Panych, L. P. (2012). Prosodic abnormalities in schizotypal personality disorder. *Schizophrenia research*, *142*(1), 20–30.

Dimic, S., Wildgrube, C., McCabe, R., Hassan, I., Barnes, T. R., & Priebe, S. (2010). Non-verbal behaviour of patients with schizophrenia in medical consultations–a comparison with depressed patients and association with symptom levels. *Psychopathology*, *43*(4), 216–222.

Dion, E., Roux, C., Landry, D., Fuchs, D., Wehby, J., & Dupr, V. (2011). Improving attention and preventing reading difficulties among low-income first-graders: A randomized study. *Prevention Science*, *12*(1), 70-79. doi: 10.1007/s11121-010-0182-5

Dishion, T., & Granic, I. (2004). Naturalistic observation of relationship processes. *Comprehensive Handbook of Psychological Assessment, Volume 3: Behavioral*

Assessment, *3*, 143–161.

Dishion, T., & Snyder, J. (2004). An introduction to the special issue on advances in process and dynamic system analysis of social interaction and the development of antisocial behavior. *Journal of Abnormal Child Psychology*, *32*(6), 575-578. doi: 10.1023/B:JACP.0000047317.96104.ca

Doignon, J.-P. (1994). Knowledge spaces and skill assignments. In G. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics and methodology* (p. 111-121). New York: Springer-Verlag.

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

Dominguez, M.-D.-G., Saka, M., Lieb, R., Wittchen, H.-U., & Van Os, J. (2010). Early expression of negative/disorganized symptoms predicting psychotic experiences and subsequent clinical psychosis: A 10-year study. *American Journal of Psychiatry*, *167*(9), 1075-1082. doi: 10.1176/appi.ajp.2010.09060883

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., Einhuser, 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.

Düntsch, I., & Gediga, G. (1995). Skills and knowledge structures. *British Journal of Mathematical and Statistical Psychology*, *48*, 9-27.

Earnst, K. S., Kring, A. M., Kadar, M. A., Salem, J. E., Shepard, D. A., & Loosen,

P. T. (1996). Facial expression in schizophrenia. *Biological Psychiatry*, *40*(6), 556–558.

Eggen, T., & Straetmans, G. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological measurement*, *60*(5), 713–734.

Ehrmantrout, N., Allen, N., Leve, C., Davis, B., & Sheeber, L. (2011). Adolescent recognition of parental affect: Influence of depressive symptoms. *Journal of Abnormal Psychology*, *120*(3), 628-634. doi: 10.1037/a0022500

Elis, O., Caponigro, J. M., & Kring, A. M. (2013). Psychosocial treatments for negative symptoms in schizophrenia: current practices and future directions. *Clinical psychology review*, *33*(8), 914–928.

Ellgring, H. (1986). Nonverbal expression of psychological states in psychiatric patients. *European archives of psychiatry and neurological sciences*, *236*(1), 31–34.

Eugenio, B. D., & Glass, M. (2004). The kappa statistic: A second look. *Computational linguistics*, *30*(1), 95–101.

Evensen, J., Røssberg, J. I., Barder, H., Haahr, U., ten Velden Hegelstad, W., Joa, I., ... others (2012). Flat affect and social functioning: a 10 year follow-up study of first episode psychosis patients. *Schizophrenia research*, *139*(1-3), 99–104.

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.

Fechner, G. T. (1860). *Elemente der psychophysik (2 vols)*. Breitkopf und Härtel.

Finkelman, M. D., Smits, N., Kim, W., & Riley, B. (2012). Curtailment and stochastic curtailment to shorten the CES-D. *Applied Psychological Measurement*, *36*(8),

632–658.

Fischer, C. T. (2000). Collaborative, individualized assessment. *Journal of Personality Assessment*, *74*(1), 2–14.

Fisher, A. J., & Bosley, H. G. (2015). Personalized assessment and treatment of depression. *Current Opinion in Psychology*, *4*, 67–74.

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

Forbey, J., & Ben-Porath, Y. (2007). Computerized adaptive personality testing: A review and illustration with the MMPI-2 computerized adaptive version. *Psychological Assessment*, *19*(1), 14-24. doi: 10.1037/1040-3590.19.1.14

Friedlander, M. L., & Phillips, S. D. (1984). Preventing anchoring errors in clinical judgment. *Journal of consulting and clinical psychology*, *52*(3), 366-371.

Friedlander, M. L., & Stockman, S. J. (1983). Anchoring and publicity effects in clinical judgment. *Journal of clinical psychology*, *39*(4), 637–644.

Fulford, D., Niendam, T. A., Floyd, E. G., Carter, C. S., Mathalon, D. H., Vinogradov, S., ... Loewy, R. L. (2013). Symptom dimensions and functional impairment in early psychosis: more to the story than just negative symptoms. *Schizophrenia research*, *147*(1), 125–131.

Fusar-Poli, P., Borgwardt, S., Bechdolf, A., Addington, J., Riecher-Rössler, A., Schultze-Lutter, F., ... others (2013). The psychosis high-risk state: a comprehensive state-of-the-art review. *JAMA psychiatry*, *70*(1), 107–120.

Gaebel, W. (1989). Visuomotor behavior in schizophrenia. *Pharmacopsychiatry*,

22(suppl 1), 29–34.

Galderisi, S., Mucci, A., Buchanan, R., & Arango, C. (2018). Negative symptoms of schizophrenia: new developments and unanswered research questions. *The Lancet Psychiatry*. doi: 10.1016/S2215-0366(18)30050-6

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

Garcia-Portilla, M., Garcia-Alvarez, L., Saiz, P., Al-Halabi, S., Bobes-Bascaran, M., Bascaran, M., ... Bobes, J. (2015). Psychometric evaluation of the negative syndrome of schizophrenia. *European Archives of Psychiatry and Clinical Neuroscience*, *265*(7), 559-566. doi: 10.1007/s00406-015-0595-z

Gardner, W., Kelleher, K., & Pajer, K. (2002). Multidimensional adaptive testing for mental health problems in primary care. *Medical Care*, *40*(9), 812-823. doi: 10.1097/00005650-200209000-00010

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.

Gelfand, D. M., & Hartmann, D. P. (1975). *Child behavior analysis and therapy*. New York: Pergamon.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., ... Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services*, *59*(4), 361–368.

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.

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

Grayce, C. J. (2013). A commercial implementation of knowledge space theory in college general chemistry. In J.-C. Falmagne, D. Albert, C. Doble, D. Eppstein, & X. Hu (Eds.), *Knowledge spaces: Applications in education* (pp. 93–113). New York: Springer.

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

Gulliksen, H. (2013). *Theory of mental tests*. Routledge.

Haidet, K., Tate, J., Divirgilio-Thomas, D., Kolanowski, A., & Happ, M. (2009). Methods to improve reliability of video-recorded behavioral data. *Research in Nursing and Health*, *32*(4), 465-474. doi: 10.1002/nur.20334

Hall, J. A., Harrigan, J. A., & Rosenthal, R. (1996). Nonverbal behavior in clinician-patient interaction. *Applied and Preventive Psychology*, *4*(1), 21–37.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, *8*(1), 23–34.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System technical journal*, *29*(2), 147–160.

Hartmann, D. P., & Wood, D. D. (1990). Observational methods. In *International handbook of behavior modification and therapy* (pp. 107–138). Springer.

Harvey, P. D., & Strassing, M. (2012). Predicting the severity of everyday functional disability in people with schizophrenia: cognitive deficits, functional capacity, symptoms, and health status. *World Psychiatry*, *11*(2), 73–79.

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.

Heller, J., Augustin, T., Hockemeyer, C., Stefanutti, L., & Albert, D. (2013). Recent developments in competence-based knowledge space theory. In *Knowledge spaces* (pp. 243–286). Springer.

Heller, J., & Repitsch, C. (2012). Exploiting prior information in stochastic knowledge assessment. *Methodology*, *8*(1), 12-22. doi: 10.1027/1614-2241/a000035

Heyman, R. E. (2001). Observation of couple conflicts: Clinical assessment applications, stubborn truths, and shaky foundations. *Psychological assessment*, *13*(1), 5.

Heyman, R. E., & Slep, A. M. S. (2004). Analogue behavioral observation. In *Comprehensive handbook of psychological assessment, vol. 3. behavioral assessment* (pp. 162–180). Hoboken, NJ, US: John Wiley & Sons Inc.

Hintze, J. (2005). Psychometrics of direct observation. *School Psychology Review*, *34*(4), 507-519.

Hops, H., Davis, B., & Longoria, N. (1995). Methodological issues in direct observation: Illustrations with the living in familial environments (life) coding system. *Journal*

of Clinical Child Psychology, *24*(2), 193–203.

Hovington, C., Bodnar, M., Joober, R., Malla, A., & Lepage, M. (2012). Identifying persistent negative symptoms in first episode psychosis. *BMC Psychiatry*, *12*. doi: 10.1186/1471-244X-12-224

Hus, V., & Lord, C. (2014). The autism diagnostic observation schedule, module 4: revised algorithm and standardized severity scores. *Journal of autism and developmental disorders*, *44*(8), 1996–2012.

Inventory, P. (2). *Restructured form): Manual for administration, scoring, and interpretation.* Minneapolis, MN: University of Minnesota Press.

Jellinek, M. S., Murphy, J. M., Robinson, J., Feins, A., Lamb, S., & Fenton, T. (1988). Pediatric symptom checklist: screening school-age children for psychosocial dysfunction. *The Journal of pediatrics*, *112*(2), 201–209.

Johnston, J. M., Pennypacker, H. S., & Green, G. (2010). *Strategies and tactics of behavioral research.* Routledge.

Jones, I. H., & Pansa, M. (1979). Some nonverbal aspects of depression and schizophrenia occurring during the interview. *Journal of Nervous and Mental Disease*, *167*(7), 402–409.

Kahng, S., & Iwata, B. A. (1998). Computerized systems for collecting real-time observational data. *Journal of Applied Behavior Analysis*, *31*(2), 253–261.

Kambouri, M., Koppen, M., Villano, M., & Falmagne, J.-C. (1994). Knowledge assessment: Tapping human expertise by the query routine. *International Journal of Human-Computer Studies*, *40*(1), 119–151.

Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The positive and negative syndrome scale (panss) for schizophrenia. *Schizophrenia bulletin*, *13*(2), 261–276.

Kelley, M. E., Haas, G. L., & van Kammen, D. P. (2008). Longitudinal progression of

negative symptoms in schizophrenia: a new look at an old problem. *Schizophrenia research*, *105*(1-3), 188–196.

Kilian, S., Asmal, L., Goosen, A., Chiliza, B., Phahladira, L., & Emsley, R. (2015). Instruments measuring blunted affect in schizophrenia: A systematic review. *PloS one*, *10*(6).

Kirisci, L., Tarter, R., Reynolds, M., Ridenour, T., Stone, C., & Vanyukov, M. (2012). Computer adaptive testing of liability to addiction: Identifying individuals at risk. *Drug and Alcohol Dependence*, *123*(SUPPL.1), S79-S86. doi: 10.1016/j.drugalcdep.2012.01.016

Kirkpatrick, B., Fenton, W. S., Carpenter, W. T., & Marder, S. R. (2006). The nimh-matrics consensus statement on negative symptoms. *Schizophrenia bulletin*, *32*(2), 214–219.

Kirkpatrick, B., Strauss, G. P., Nguyen, L., Fischer, B. A., Daniel, D. G., Cienfuegos, A., & Marder, S. R. (2011). The brief negative symptom scale: psychometric properties. *Schizophrenia bulletin*, *37*(2), 300–305.

Kirppendorff, K. (1989). Content analysis: An introduction to its methodology. *Beverley Hills: Sage*.

Kline, P. (2014). *The new psychometrics: science, psychology and measurement*. Routledge.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155–163.

Koppen, M. (1993). Extracting human expertise for constructing knowledge spaces: An algorithm. *Journal of mathematical psychology*, *37*(1), 1–20.

Koppen, M., & Doignon, J.-P. (1990). How to build a knowledge space by querying

an expert. *Journal of Mathematical Psychology*, *34*(3), 311–331.

Kring, A. M., & Caponigro, J. M. (2010). Emotion in schizophrenia: where feeling meets thinking. *Current directions in psychological science*, *19*(4), 255–259.

Kring, A. M., Gur, R. E., Blanchard, J. J., Horan, W. P., & Reise, S. P. (2013). The clinical assessment interview for negative symptoms (cains): final development and validation. *American Journal of Psychiatry*, *170*(2), 165-172.

Kupper, Z., Ramseyer, F., Hoffmann, H., Kalbermatten, S., & Tschacher, W. (2010). Video-based quantification of body movement during social interaction indicates the severity of negative symptoms in patients with schizophrenia. *Schizophrenia research*, *121*(1-3), 90–100.

Kupper, Z., Ramseyer, F., Hoffmann, H., & Tschacher, W. (2015). Nonverbal synchrony in social interactions of patients with schizophrenia indicates socio-communicative deficits. *PLoS One*, *10*(12), e0145882.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174. doi: 10.2307/2529310

Lavelle, M., Dimic, S., Wildgrube, C., McCabe, R., & Priebe, S. (2015). Non-verbal communication in meetings of psychiatrists and patients with schizophrenia. *Acta Psychiatrica Scandinavica*, *131*(3), 197–205.

Lavelle, M., Healey, P. G., & McCabe, R. (2013). Is nonverbal communication disrupted in interactions involving patients with schizophrenia? *Schizophrenia bulletin*, *39*(5), 1150–1158.

Lavelle, M., Healey, P. G., & McCabe, R. (2014). Nonverbal behavior during face-to-face social interaction in schizophrenia: A review. *The Journal of nervous and mental disease*, *202*(1), 47–54.

Lee, D., Barak, A., & Uhlemann, M. (1999). Forming clinical impressions during the

first five minutes of the counseling interview. *Psychological Reports*, *85*(3 PART 1), 835-844.

Leentjens, A. F., Wielaert, S. M., van Harskamp, F., & Wilmink, F. W. (1998). Disturbances of affective prosody in patients with schizophrenia; a cross sectional study. *Journal of Neurology, Neurosurgery & Psychiatry*, *64*(3), 375–378.

Leger, D. W. (1977). An empirical evaluation of instantaneous and one-zero sampling of chimpanzee behavior. *Primates*, *18*(2), 387–393.

Lord, C., Luyster, R., Gotham, K., & Guthrie, W. (2012). Autism diagnostic observation schedule second edition (ados-2) manual (part ii): Toddler module. *Torrance, CA: Western Psychological Services*.

Lord, C., Rutter, M., DiLavore, P., & Risi, S. (1999). *Autism diagnostic observation schedule: Ados*. Western Psychological Services Los Angeles, CA.

Lord, C., Rutter, M., DiLavore, P., Risi, S., Gotham, K., & Bishop, S. (2012). Autism diagnostic observation schedule second edition (ados-2) manual (part 1): Modules 1–4. *Torrance, CA: Western Psychological Services*.

Lord, F. M. (1959). Statistical inferences about true scores. *Psychometrika*, *24*(1), 1–17.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Lorr, M. (1962). *Inpatient multidimensional psychiatric scale (imps)*. Consulting Psychologists Press.

Malla, A. K., Takhar, J. J., Norman, R. M., Manchanda, R., Cortese, L., Haricharan, R., ... Ahmed, R. (2002). Negative symptoms in first episode non-affective psychosis. *Acta Psychiatrica Scandinavica*, *105*(6), 431–439.

Mandal, M. K., Pandey, R., & Prasad, A. B. (1998). Facial expressions of emotions

and schizophrenia: A review. *Schizophrenia bulletin*, *24*(3), 399–412.

Manikandan, S. (2011). Measures of central tendency: Median and mode. *Journal of Pharmacology and Pharmacotherapeutics*, *2*(3), 214-215. doi: 10.4103/0976-500X.83300

Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L., ... Maris, G. (2018). An introduction to network psychometrics: Relating ising network models to item response theory models. *Multivariate Behavioral Research*, *53*(1), 15-35.

Martin, P., Bateson, P. P. G., & Bateson, P. (1993). Recording methods. In *Measuring behaviour: an introductory guide* (pp. 84–100). Cambridge University Pres.

Matayoshi, J., Granziol, Doble, C., Uzun, H., , & Cosyn, E. (2018). Forgetting curves and testing effects in an adaptive learning and assessment system. In *11th international conference on educational data mining (edm 2018), Buffalo, NY, USA.*

McHorney, C. (1997). Generic health measurement: Past accomplishments and a measurement paradigm for the 21st century. *Annals of Internal Medicine*, *127*(8 II SUPPL.), 743-750.

Messinger, J., Trmeau, F., Antonius, D., Mendelsohn, E., Prudent, V., Stanford, A., & Malaspina, D. (2011). Avolition and expressive deficits capture negative symptom phenomenology: Implications for dsm-5 and schizophrenia research. *Clinical Psychology Review*, *31*(1), 161-168. doi: 10.1016/j.cpr.2010.09.002

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., ... Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American psychologist*, *56*(2), 128–165.

Michel, P., Baumstarck, K., Lancon, C., Ghattas, B., Loundou, A., Auquier, P., &

Boyer, L. (2018). Modernizing quality of life assessment: development of a multidimensional computerized adaptive questionnaire for patients with schizophrenia. *Quality of Life Research*, *27*(4), 1041-1054. doi: 10.1007/s11136-017-1553-1

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*(3), 355–383.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology*, *10*(5), 639–667.

Millan, M. J., Fone, K., Steckler, T., & Horan, W. P. (2014). Negative symptoms of schizophrenia: clinical characteristics, pathophysiological substrates, experimental models and prospects for improved treatment. *European Neuropsychopharmacology*, *24*(5), 645–692.

Miller, T. J., McGlashan, T. H., Woods, S. W., Stein, K., Driesen, N., Corcoran, C. M., ... Davidson, L. (1999). Symptom assessment in schizophrenic prodromal states. *Psychiatric Quarterly*, *70*(4), 273–287.

Morrens, M., Hulstijn, W., & Sabbe, B. (2007). Psychomotor slowing in schizophrenia. *Schizophrenia bulletin*, *33*(4), 1038–1053.

Mumma, G. (2002). Effects of three types of potentially biasing information on symptom severity judgments for major depressive episode. *Journal of Clinical Psychology*, *58*(10), 1327-1345. doi: 10.1002/jclp.10046

Murphy, D., & Cutting, J. (1990). Prosodic comprehension and expression in schizophrenia. *Journal of Neurology, Neurosurgery & Psychiatry*, *53*(9), 727–730.

Narens, L., & Luce, R. D. (1993). Further comments on the nonrevolution arising from axiomatic measurement theory. *Psychological Science*, *4*(2), 127–130.

Nokelainen, P., Niemivirta, M., Kurhila, J., Miettinen, M., Silander, T., & Tirri, H.

(2001). Implementation of an adaptive questionnaire. In *Proceedings of the ed-media conference* (pp. 1412–1413).

Noldus, L. P. (1991). The observer: a software system for collection and analysis of observational data. *Behavior Research Methods, Instruments, & Computers*, *23*(3), 415–429.

Novick, M. R. (1965). The axioms and principal results of classical test theory. *ETS Research Report Series*, *1965*(1), 1–18.

Overall, J. E., & Gorham, D. R. (1962). The brief psychiatric rating scale. *Psychological reports*, *10*(3), 799–812.

Paul, G. L. (1986). *The time sample behavioral checklist: Observational assessment instrumentation for service and research.* Champaign 1L: Research Press.

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. *Quality of Life Research*, *15*(3), 315–329.

Pino, M. C., Spoto, A., Mariano, M., Granziol, U., Peretti, S., Masedu, F., ... Vidotto, G. (2018). Formal psychological assessment for autism spectrum disorder diagnosis: A new methodology to build an adaptive testing system. *The Open Psychology Journal*, *11*(1), 112–122.

Powell, J., Martindale, B., Kulp, S., Martindale, A., & Bauman, R. (1977). Taking a closer look: time sampling and measurement error. *Journal of Applied Behavior Analysis*, *10*(2), 325-332. doi: 10.1901/jaba.1977.10-325

Püschel, J., Stassen, H., Bomben, G., Scharfetter, C., & Hell, D. (1998). Speaking behavior and speech sound characteristics in acute schizophrenia. *Journal of psychiatric research*, *32*(2), 89–97.

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rabinowitz, J., Berardo, C. G., Bugarski-Kirola, D., & Marder, S. (2013). Association of prominent positive and prominent negative symptoms and functional health, well-being, healthcare-related quality of life and family burden: a catie analysis. *Schizophrenia research*, *150*(2-3), 339–342.

Ramseyer, F., & Tschacher, W. (2011). Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. *Journal of consulting and clinical psychology*, *79*(3), 284–295.

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

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the fourth berkeley symposium on mathematical statistics and probability* (Vol. 4, pp. 321–333).

Ray, J. M., & Ray, R. D. (2008). Train-to-code: An adaptive expert system for training systematic observation and coding skills. *Behavior research methods*, *40*(3), 673–693.

Ray, R. D. (1995). A behavioral systems approach to adaptive computerized instructional design. *Behavior Research Methods, Instruments, & Computers*, *27*(2), 293–296.

Reiser, M., & VandenBerg, M. (1994). Validity of the chi-square test in dichotomous variable factor analysis when expected frequencies are small. *British Journal of Mathematical and Statistical Psychology*, *47*(1), 85-107. doi: 10.1111/j.2044-8317.1994.tb01026.x

Repp, A., Nieminen, G., Olinger, E., & Brusca, R. (1988). Direct observation: Factors affecting the accuracy of observers. *Exceptional Children*, *55*(1), 29-36. doi: 10.1177/001440298805500103

Rhine, R. J., & Linville, A. K. (1980). Properties of one-zero scores in observational studies of primate social behavior: The effect of assumptions on empirical analyses. *Primates*, *21*(1), 111–122.

Riehle, M., & Lincoln, T. (2018). Investigating the social costs of schizophrenia: Facial expressions in dyadic interactions of people with and without schizophrenia. *Journal of Abnormal Psychology*, *127*(2), 202-215. doi: 10.1037/abn0000319

Roberts, L. W., Chan, S., & Torous, J. (2018). New tests, new tools: mobile and connected technologies in advancing psychiatric diagnosis. *npj Digital Medicine*, *1*(1), 6.

Robusto, E., & Stefanutti, L. (2014). Extracting a knowledge structure from the data by a maximum residuals method. *TPM: Testing, Psychometrics, Methodology in Applied Psychology*, *21*(4), 421-433. doi: 10.4473/TPM21.4.4

Rosenberg, M., Glueck Jr, B. C., & Bennett, W. L. (1967). Automation of behavioral observations on hospitalized psychiatric patients. *American Journal of Psychiatry*, *123*(8), 926–929.

Roter, D. L., Frankel, R. M., Hall, J. A., & Sluyter, D. (2006). The expression of emotion through nonverbal behavior in medical visits. *Journal of general internal medicine*, *21*(S1), S28–S34.

Selten, J., Wiersma, D., & Van Den Bosch, R. (2000). Discrepancy between subjective and objective ratings for negative symptoms. *Journal of Psychiatric Research*, *34*(1), 11-13. doi: 10.1016/S0022-3956(99)00027-8

Serra, F., Spoto, A., Ghisi, M., & Vidotto, G. (2015). Formal psychological assessment

in evaluating depression: a new methodology to build exhaustive and irredundant adaptive questionnaires. *PloS one*, *10*(4), e0122131.

Serra, F., Spoto, A., Ghisi, M., & Vidotto, G. (2017). Improving major depressive episode assessment: A new tool developed by formal psychological assessment. *Frontiers in psychology*, *8*, 214.

Sharp, R. A., Mudford, O. C., & Elliffe, D. (2015). A data-based method for selecting parameters of momentary time sampling to provide representative data. *European Journal of Behavior Analysis*, *16*(2), 279–294.

Shtasel, D. L., Gur, R. E., Gallacher, F., Heimberg, C., & Gur, R. C. (1992). Gender differences in the clinical expression of schizophrenia. *Schizophrenia Research*, *7*(3), 225–231.

Simms, L. J., & Clark, L. (2005). Validation of a computerized adaptive version of the schedule for nonadaptive and adaptive personality (snap). *Psychological Assessment*, *17*(1), 28-43. doi: 10.1037/1040-3590.17.1.28

Simms, L. J., Goldberg, L. R., Roberts, J. E., Watson, D., Welte, J., & Rotterman, J. H. (2011). Computerized adaptive assessment of personality disorder: Introducing the cat–pd project. *Journal of personality assessment*, *93*(4), 380–389.

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.

Smits, N., Finkelman, M. D., & Kelderman, H. (2016). Stochastic curtailment of questionnaires for three-level classification: Shortening the ces-d for assessing low, moderate, and high risk of depression. *Applied Psychological Measurement*, *40*(1), 22–36.

Snyder, J., Reid, J., Stoolmiller, M., Howe, G., Brown, H., Dagne, G., & Cross, W.

(2006). The role of behavior observation in measurement systems for randomized prevention trials. *Prevention Science*, *7*(1), 43-56. doi: 10.1007/s11121-005-0020-3

Spearman, C. (1904). "General Intelligence", objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–292.

Spoto, A. (2011). Formal psychological assessment theoretical and mathematical foundations.

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 of Mathematical 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

Stassen, H., Albers, M., Püschel, J., Scharfetter, C., Tewesmeier, M., & Woggon, B. (1995). Speaking behavior and voice sound characteristics associated with negative schizophrenia. *Journal of psychiatric research*, *29*(4), 277–296.

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 blims unidentifiability: Forward and backward parameter transformation groups. *Journal of Mathematical Psychology*, *82*, 38–51.

Steimer-Krause, E., Krause, R., & Wagner, G. (1990). Interaction regulations used by schizophrenic and psychosomatic patients: studies on facial behavior in dyadic interactions. *Psychiatry*, *53*(3), 209–228.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*.

Stevens, S. S. (1951). *Mathematics, measurement, and psychophysics.* Oxford, England: Wiley.

Stevens, S. S. (1957). On the psychophysical law. *Psychological review*, *64*(3), 153.

Strauss, G. P., Keller, W. R., Buchanan, R. W., Gold, J. M., Fischer, B. A., McMahon, R. P., ... Kirkpatrick, B. (2012). Next-generation negative symptom assessment for clinical trials: validation of the brief negative symptom scale. *Schizophrenia research*, *142*(1), 88–92.

Suen, H. K., & Ary, D. (1984). Variables influencing 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.

Suppes, P., Krantz, D. H., Luce, R. D., & Tversky, A. (1989). *Foundations of mea-*

surement: Geometrical, threshold, and probabilistic representations (Vol. 2). Academic Pr.

Suppes, P., & Zinnes, J. (1963). Basic measurement theory. In R. D. Luce & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 1–76). Oxford: Wiley.

Tapp, J., Ticha, R., Kryzer, E., Gustafson, M., Gunnar, M. R., & Symons, F. J. (2006). Comparing observational software with paper and pencil for time-sampled data: A field test of interval manager (intman). *Behavior research methods*, *38*(1), 165–169.

Trémeau, F., Goggin, M., Antonius, D., Czobor, P., Hill, V., & Citrome, L. (2008). A new rating scale for negative symptoms: the motor-affective-social scale. *Psychiatry research*, *160*(3), 346–355.

Trémeau, F., Malaspina, D., Duval, F., Corrêa, H., Hager-Budny, M., Coin-Bariou, L., ... Gorman, J. M. (2005). Facial expressiveness in patients with schizophrenia compared to depressed patients and nonpatient comparison subjects. *American Journal of Psychiatry*, *162*(1), 92–101.

Troisi, A. (1999). Ethological research in clinical psychiatry: The study of nonverbal behavior during interviews. *Neuroscience and Biobehavioral Reviews*, *23*(7), 905-913. doi: 10.1016/S0149-7634(99)00024-X

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), 475-481. doi: 10.1016/j.pnpbp.2006.11.016

Troisi, A., Spalletta, G., & Pasini, A. (1998). Non-verbal behaviour deficits in schizophrenia: an ethological study of drug-free patients. *Acta Psychiatrica Scandinavica*, *97*(2), 109–115.

Trull, T. (2007). Expanding the aperture of psychological assessment: Introduction to the special section on innovative clinical assessment technologies and methods. *Psychological Assessment*, *19*(1), 1-3. doi: 10.1037/1040-3590.19.1.1

Turner, D. T., van der Gaag, M., Karyotaki, E., & Cuijpers, P. (2014). Psychological interventions for psychosis: a meta-analysis of comparative outcome studies. *American Journal of Psychiatry*, *171*(5), 523–538.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Vail, A., Baltruaitis, T., Pennant, L., Liebson, E., Baker, J., & Morency, L.-P. (2018). Visual attention in schizophrenia: Eye contact and gaze aversion during clinical interactions. In (Vol. 2018-January, p. 490-497). doi: 10.1109/ACII.2017.8273644

Valtchev, P., Grosser, D., Roume, C., & Hacene, M. R. (2003). Galicia: an open platform for lattices. In *Using conceptual structures: Contributions to the 11th intl. conference on conceptual structures (iccs'03)* (pp. 241–254).

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

Ventura, J., Green, M. F., Shaner, A., & Liberman, R. P. (1993). Training and quality assurance with the brief psychiatric rating scale:" the drift busters". *International Journal of Methods in Psychiatric Research*, *3*(4), 221-244.

Vessonen, E. (2018). The complementarity of psychometrics and the representational theory of measurement. *The British Journal for the Philosophy of Science*, axy032.

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

Walther, S., Stegmayer, K., Sulzbacher, J., Vanbellingen, T., Mri, R., Strik,

W., & Bohlhalter, S. (2015). Nonverbal social communication and gesture control in schizophrenia. *Schizophrenia Bulletin*, *41*(2), 338-345. doi: 10.1093/schbul/sbu222

Washington, C., & Moss, M. (1988). Pragmatic aspects of establishing interrater reliability in research. *Nursing Research*, *37*(3), 190-191.

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

Werbeloff, N., Dohrenwend, B., Yoffe, R., Van Os, J., Davidson, M., & Weiser, M. (2015). The association between negative symptoms, psychotic experiences and later schizophrenia: A population-based longitudinal study. *PLoS ONE*, *10*(3). doi: 10.1371/journal.pone.0119852

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

Yanagita, B. T., Becirevic, A., & Reed, D. D. (2016). Computer-assisted technologies for collecting and summarizing behavioral data. In *Computer-assisted and web-based innovations in psychology, special education, and health* (pp. 95–116). Elsevier.

Yong, S., Awang Rambli, D. R., & Anh, N. (2007). Depression consultant expert system.

Zenk, S., Schulz, A., Mentz, G., House, J., Gravlee, C., Miranda, P., ... Kannan, S. (2007). Inter-rater and test-retest reliability: Methods and results for the neighborhood observational checklist. *Health and Place*, *13*(2), 452-465. doi: 10.1016/j.healthplace.2006.05.003