Further data on the reliability of the mentalization imbalances scale and of the modes of mentalization scale

Giulia Gagliardini, Laura Gatti, Antonello Colli

Department of Humanites, "Carlo Bo" Univeristy of Urbino, Italy

ABSTRACT

The aim of this study was to provide data on the Inter-Rater Reliability (IRR) and the test-retest reliability of the Mentalization Imbalances Scale (MIS) and the Modes of Mentalization Scale (MMS) in two different studies. Three junior raters and two senior raters assessed blindly 15 session transcripts of psychotherapy of five patients, using both the MIS and the MMS. The same 15 sessions were rated after the junior raters completed a training at the use of the scales and after on month from the end of the training to assess test-retest reliability. Four therapists used the MIS and the MMS to provide different ratings of 22 patients undergoing a psychotherapy in different settings. Intraclass Correlation Coefficient (ICC) values ranged from sufficient to good and increased after the training. Test re-test reliability was sufficient for both scales (Study 1). ICC values ranged from sufficient to good, and were globally higher than the ones found in the first study sample (Study 2). Our results provide support to the inter-rater reliability of the MIS and the MMS.

Key words: Mentalization; reflective function; assessment; inter rater reliability.

Introduction

Since its introduction in psychological science, the construct of mentalization has been a subject of growing interest among various authors, especially in recent years (Bateman & Fonagy, 2016; Katznelson, 2014). Mentalization represents "the mental process by which an individual implicitly and explicitly interprets the actions of himself and others as meaningful on the basis of intentional mental states such as personal desires, needs, feelings and reasons" (Bateman & Fonagy, 2004, p.XXI). Moreover, problematics in mentalization have been observed in different clinical populations, such as personality disorders (Bateman, Bolton, & Fonagy, 2013; Gagliardini et al., 2018), anxiety disorders (Rudden, Milrod, Target, Ackerman, & Graf, 2006), depression (Taubner, Kessler, Buchheim, Kächele, & Staun, 2011), and eating disorders (Skårderud, 2007). In recent years, several measures of mentalization have been developed. At the present time, these measures can be divided into four main categories (Bateman & Fonagy, 2016): i) interviews/narrative coding systems, ii) questionnaires, iii) experimental observational tasks, and iv) projective measures. We will not consider here the projective measures, which are represented specifically only by the Projective Imagination Test (Blackshaw, Kinderman, Hare, & Hatton, 2001).

One of the most used and validated narrative-based measures of mentalization is the Reflective Functioning Scale (RFS; Fonagy, Target, Steele, & Steele, 1998), which is rated on the basis of the Adult Attachment Interviews (AAI) and shows good psychometric properties. The RFS allows for the rater to assess patients mentalization on each item (*i.e.* patients' answers to interviewers' questions) of the interview and to provide a global score of the whole transcript. According to RFS scoring, AAI's questions are divided into "permit" questions (questions which allow for the patient to answer in mentalizing terms but that do not require it explicitly) and "demand" questions (questions which explicitly require for the patient to adopt mentalization in order to answer properly). Demand

Correspondence: Antonello Colli, Department of Humanites, "Carlo Bo" Univeristy of Urbino, Italy.
E-mail: antonello.colli@uniurb.it

Citation: Gagliardini, G., Gatti, L., & Colli, A. (2020). Further data on the reliability of the mentalization imbalances scale and of the modes of mentalization scale. *Research in Psychotherapy: Psychopathology, Process and Outcome, 23*(1), 88-98. doi: 10.4081/ripppo.2020.450

Conflict of Interest: The authors declare no conflict of interest.

Acknowledgments: The authors would like to thank the clinicians and raters who participated to this study by providing their evaluations.

Received for publication: 22 January 2020.
Accepted for publication: 8 April 2020.

Research in Psychotherapy:
Psychopathology, Process and Outcome 2020; 23:88-98
doi:10.4081/ripppo.2020.450

questions are always rated with the RFS while permit questions may not be rated if they do not provide information on patient's mentalization. Taubner et al. (2013) have tested RFS reliability on a sample of 74 verbatim transcripts of AAIs from different patients and using Intraclass Correlation Coefficient (ICC) found that the inter-rater reliability of "demand" items ranged from 0.27 to 0.45, while the reliability of the global score was 0.71. These results are in line with the study carried out by Fonagy et al. in 1996, in which the authors have assessed the Inter-Rater Reliability (IRR) between two senior raters who did a double training, the first in 1987 and the second after seven years. The RFS showed good inter-rater reliability index after a three-day training at the use of the measure, with a level of agreement of 0.91 between two different senior raters.

The RFS represents an expert rating of mentalization useful for empirical purposes and can provide a global score on a Likert scale ranging from - 1 (negative reflective functioning) to + 9 (marked reflective functioning). The Reflective Function Rating Scale (RFRS; Meehan, Levy, Reynoso, Hill, & Clarkin, 2009) represents a multi-item rating scale for assessing mentalization that can be applied to a range of data sources (*e.g.*, interviews) by informants such as therapists or observers rating interactions on the basis of three dimensions: i) defensive/distorted, ii) awareness of mental states, and iii) developmental.

Mentalization can also be measured through questionnaires, such as the Reflective Functioning Questionnaire (RFQ; Fonagy et al., 2016), the Mentalization Questionnaire (MZQ; Hausberg et al., 2012), and the Mentalization Scale (MentS; Dimitrijevíc, Hanak, Altaras Dimitrijevíc, & Marjanovíc, 2018), which can be self-reported by patients without being time consuming. The RFQ has shown good internal consistency and can discriminate between clinical samples and normal controls (Fonagy et al., 2016). The questionnaire is composed by two factors named Uncertainty about Mental States and Certainty about Mental States. The MZQ has shown good internal consistency; the scale is composed by four factors: refusing self-reflection, emotional awareness, psychic equivalence mode, and regulation of affect (Hausberg et al., 2012). Finally, the MentS is composed of three factors: other-related mentalization, self-related mentalization, and motivation to mentalize (Dimitrijevíc et al., 2018). All these measures have shown a good validity and can discriminate between clinical populations and healthy control samples.

Experimental-observational tasks, such as the Reading the Mind in the Eyes Test (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001), are based on the recognition of mental states through the observation of facial emotions presented to patients and have been used in a number of studies. Some authors, however, have criticized the assumption that the identification of mental states through the observation of facial expressions can be considered on the whole as an indicator of mentalization or theory of mind (Oakley, Brewer, Bird, & Catmur, 2016). Moreover, experimental-observational tasks share a criticism that is common to all the aforementioned methodologies of assessing mentalization: They are mostly focused on the explicit aspects of mentalizing and do not assess the automatic and implicit facets of the construct.

All the measures described above have shown good psychometric properties, however there are some criticisms which must be noted also in relation to self-report measures and interview-based methodologies. Self-report measures can be considered helpful and not time-consuming assessment tools; at the same time, in some cases the assessment may be biased by the fact that patients with personality disorders could not be reliable when filling out mentalization measures, because they have problems with self-awareness (Davidson, Obonsawin, Seils, & Patience, 2003; Huprich, Bornstein, & Schmitt, 2011). Moreover, patients with borderline features manifest limitations of their insight into the relative disadvantages in the capacity for cooperative relationships and a limited ability to approach life in a non-impulsive manner, which may limit their capacity to complete self-report measures (Morey, 2014). The RFS and the other interview-based measures are highly reliable; however, they are time consuming because they require therapy session transcripts or interviews for the assessment and long trainings to be applied reliably. This restricts their application in large-scale studies and limits their use in clinical contexts. Moreover, all these measures are not explicitly focused on the imbalances on the dimensions of mentalization, nor on the pre-mentalizing modalities of thought, which are pivotal facets of the theory of mentalization. Some authors have therefore enlightened the importance of developing specific measures for the assessment of mentalizing dimensions and modalities (Luyten, Fonagy, Lowyck, & Vermote, 2012).

In light of these considerations, two clinician-reports for the assessment of mentalization have been developed, the Mentalization Imbalances Scale (MIS; Gagliardini et al., 2018) and the Modes of Mentalization Scale (MMS; Gagliardini & Colli, 2019). By introducing these measures the authors have tried to overcome the limitations of the previously published mentioned measures enlightened above, namely an excessive focus on the explicit dimensions of mentalization; the impossibility to use them to rate the majority of the dimensions or modalities of mentalization described in the theoretical literature on the topic; a large amount of time required for the evaluations; and the problematics that self-reports arise in the evaluations of ego-syntonic traits. These measures have been preliminarily validated in previous studies, in which both the scales have shown good psychometric properties. In relation the MIS, six different factors related to pathological imbalances in the dimensions of mentalization were found from a confirmative factor analysis, namely: imbalance on the self, imbalance on the others, affective im-

balance, cognitive imbalance, automatic imbalance, and external imbalance (Gagliardini et al., 2018). In the previously published study, the reliability of the MIS was tested and all scales had good Cronbach's alphas, with values ranging from 0.70 (automatic imbalance) to 0.89 (cognitive imbalance). These factors were coherently related to personality disorders and different clinical features. In relation to the MMS, five different factors related to the quality of mentalization have emerged from an explorative factor analysis, namely: excessive certainty, concrete thought, good mentalization, teleological thought, and intrusive pseudomentalization (Gagliardini & Colli, 2019). In the previously published study, the reliability of the MMS was tested and all scales had good coefficient alphas, with values ranging from 0.67 (intrusive pseudomentalization) to 0.91 (excessive certainty). These factors were coherently related to specific attachment styles in a clinical population of patients with personality disorders. Previous studies were focused on the validity of both the MIS and the MMS, and did not investigate the reliability of the scales.

A crucial point in the validation of clinician report measures is represented by their IRR. Although several researches have shown that clinicians tend to make highly reliable evaluations if their observations and inferences are quantified using psychometrically sophisticated instruments (Blagov, Bi, Shedler, & Westen, 2012) this remains an important issue to be addressed. In relation to this problematic, it is important to note that it might be quite difficult to have data on the IRR of clinician reports, since it would imply that different clinicians are able to rate the same patients, a condition which is not always possible to fulfill for practical reasons. In order to overcome this limit it is possible to test the reliability of clinician report measures by using them to rate psychotherapy session transcripts, which have the major limitation of not providing information on the implicit and procedural facets of mentalization, but may allow for different raters to assess the same clinical material, a condition which is not easily obtainable in everyday clinical practice.

A further important topic is related to how the expertise of the raters may influence their reliability and if, eventually, a training for the reliable use of these measures is required. These two concerns are crucial for the assessment of the reliability of clinician report measures in order to understand who and how could use them reliably, especially in everyday clinical practice.

Aims and Hypothesis

In light of the aforementioned considerations, the aim of this work is to test the IRR of the MIS and the MMS in different conditions. Study 1 addressed the agreement between junior raters and gold standard ratings and the impact that a training for the use of the scales could have on their reliability and test-retest reliability on a sample of psychotherapy session transcripts (N = 15).

Study 2 addressed the agreement on the rating of the same patients (N = 22) by four different raters working in a community service.

In doing so, we made some a-priori predictions:

– that the reliability between junior raters and gold standard evaluations would be sufficient pre-training;
– that the reliability between junior raters and gold standard evaluations would increase after a specific training on the use of the scales;
– that the agreement between clinicians working in different settings with the same patients would be sufficient to good;
– that the agreement between clinicians working with real patients would be higher than the agreement of raters working on session transcripts.

Methods

Study 1

Patients

From a database of 400 verbatim transcripts of psychotherapy sessions of 50 different cases, we selected randomly 5 adult Caucasian patients (4 women, 1 man) and for each case we randomly picked three sessions related to different phases of the treatment (beginning, middle, end). The final sample consisted of 15 session transcripts. The mean age of the patients was 25.4 years (min. = 19, max. = 39). Before entering psychotherapy, all patients received a DSM–5 diagnosis (APA, 2013). Two patients had a personality disorder diagnosis (borderline and dependent personality disorders) and three patients had anxiety disorders and depressive disorders.

Therapists

Four therapists (mean age = 41; min = 37; max = 50) were psychodynamic-oriented psychologists and psychotherapists, while one therapist reported a cognitive-behavioral approach. Three therapists were seeing the selected patient in a private setting and two in mental health institutions.

Raters and Procedure

The rating group was composed by three junior raters (*i.e.*, graduate students without any clinical, empirical or assessment experience) that evaluated each session independently. The authors ***** evaluated blindly the 15 sessions, the ratings then were jointly revised and finally classified as gold standard. During the first phase the three junior raters have independently assessed 15 sessions: Raters were asked to read one session at a time. They were asked to read the whole session and afterwards to use the MIS and MMS to provide a rating of the mentalizing capacities of the patients on the basis of patients' communications during the session, by using the scales' items.

Raters were asked to work independently and blindly. The specific instruction for each rater were: "Use the following list of items to assess patient's mental functioning. In doing so, please consider both the explicit content and narratives the patient has provided and the way he or she interacts with the therapist during the course of the session. Rate each item on a scale from 0 (absolutely not descriptive) to 5 (absolutely descriptive)." The raters did not have any additional information in relation to patients' characteristics nor in relation to the measures. After this first phase the three junior raters participated to a 12-hours training for the use of the MMS and to a 12-hours training for the use of the MIS. Each training was organized in four sessions of three hours. The training was provided by one of the authors of the scales. During the training each item of the scales was discussed and explained by providing clinical examples. Raters were also asked to complete the ratings of five session transcripts at home; the ratings were then discussed with the coordinator of the training. During the training the sessions already evaluated in the first phase were not discussed and the session transcripts rated during the training have not been included for reliability evaluation in the present study. On the whole, the number of hours required for the training for both the scales was 34, including the homework (10 hours) and the training (24 hours).

To evaluate the effect of training on reliability, after the training the three raters evaluated independently the same 15 sessions (raters had a deadline to complete evaluations of four weeks). We decided to use the same 15 sessions in order to avoid the possibility that the differences between the IRR pre- and post-training might be related to specific characteristics of the sessions or of the patients. The decision to use sessions related to different phases of the therapy was motivated by the necessity to maximize the heterogeneity of the sessions, in order to have the possibility to rate different modalities and different imbalances in the dimensions of mentalization, which can change throughout the course of the therapy. For example, it is probable that in the first phases of a psychotherapy patients show more pre-mentalizing modalities of thought and more imbalances in the dimensions of mentalization, while in the final phases we may expect to see a more compelling presence of good mentalization. In order to assess the quality of the sessions in terms of the mentalizing material provided by the patients, in line with the rating of the RFS (Fonagy et al., 1998), we calculated the number of demand questions (*i.e.* questions provided by the therapist/interviewer which explicitly ask for the patient to mentalize, *e.g.* "Why do you think your parents behaved as they did?") for each session. Overall, the mean number of demand questions was 4.67 (min. 2; max. 12). This number is in line with the number of demand questions normally rated with the RFS applied to the AAI, suggesting a good quality of the selected sessions in terms of the material available for the evaluation of patients' mentalization. To evaluate test re test reliability after one month the raters evaluated again the same sessions.

Measures

Mentalization Imbalances Scale (MIS) (Gagliardini et al., 2018)

The MIS represents a clinician report assessment measure of mentalizing imbalances in adult patients. It is composed by 22 items rated on a Likert scale from 0 ("absolutely not descriptive") to 5 ("absolutely descriptive") and represents an assessment measure of mentalizing imbalances on the basis of six subscales: (1) imbalance towards self (4 items) indicating an excessive focus on patient's own mind which prevents from the possibility to connect with others' thoughts and feelings and perspectives (*e.g.*, "P. doesn't seem capable of assuming other people's points of view when interpreting other people's behavior"); imbalance towards others (3 items) indicating an excessive focus on other peoples' mental states rather than patient's own (*e.g.*, "P. can easily be influenced by other people's emotions"); affective imbalance (4 items) indicating an hyper-activation of affects and emotions not adequately balanced by cognition (*e.g.*, "When experiencing an intense emotion, P. can think clearly"); cognitive imbalance (5 items) indicating an excessive focus on the cognitive facets of mentalization (which can lead to intellectualizing) that is not balanced by the affective facets of experience (*e.g.*, "Even when talking about painful and/or emotionally intense themes, P. seems to be detached"); automatic imbalance (3 items), indicating the ability to automatically and implicitly recognize mental states, which, however, is not paired by the capacity to explicitly and declaratively reflect on them, even when actively solicited by others (*e.g.*, a therapist) (*e.g.*, "P. fails to reflect on the first impression he or she has of a person or a situation"); external imbalance (3 items), indicating those cases in which a person excessively relies on the external cues of mental states (*i.e.*, facial expressions, body postures, etc.) without reflecting on inner mental states (*e.g.*, beliefs, desires, thoughts, emotions) (*e.g.* "P. seems to have a "sixth sense" about other people's (including the therapist) mental states"). Reliability analysis showed that the Cronbach's alpha values were: 0.89 (cognitive imbalance), 0.83 (affective imbalance), 0.81 (imbalance toward others), 0.78 (imbalance toward self and external imbalance), and 0.70 (automatic imbalance).

Modes of Mentalization Scale (MMS) (Gagliardini & Colli, 2019)

The MMS is a clinician-report assessment measure of the modes of mentalization on five different sub-scales: (1) excessive certainty (6 items), indicating an over-activation of mentalization, in which patients show an excessive cer-

tainty about mental states and think that they can provide all of the answers regarding other people's inner worlds; concrete thinking (6 items), indicating the tendency to interpret reality on the basis of heuristics and prejudices and/or on the basis of physical or invariant constraints, to use common-sense explanations or clichés to explain emotions, and to adopt bizzarre explanations of behaviors; (3) good mentalization (5 items), indicating a good capacity to recognize and coherently describe mental states, united with a curious stance toward the same and an awareness that people can experience contrasting feelings and desires; (4) teleological thought (3 items), indicating a tendency to rely more on the physical manifestations of mental states (*i.e.*, actions) rather than interpreting the world in terms of beliefs, desires, or thoughts, to focus more on what people do (and not on what they think or feel), and to be more focused on the physical, practical, resolution of a problem rather than on the meanings related to the situation; (5), intrusive pseudomentalization (4 items), related to a more malign form of hyper- or pseudo-mentalization, indicating a tendency to intrude on and manipulate other people's life, in which the reflections of one's inner world don't seem to be genuine. The factor structure of the scale was explored in a previous study that enlightened good psychometric properties, with alphas of 0.91 (excessive certainty), 0.83 (good mentalization), 0.79 (concrete thought), 0.77 (teleological thought), and 0.67 (intrusive pseudomentalization).

Statistical analysis

All analyses were conducted using SPSS Statistics 21 for Windows (IBM, Armonk, NY). In order to assess the inter-rater reliability between the junior raters and the gold standard evaluation (pre- and post-training), the ICC was calculated using Two-Way mixed effects model, single measures absolute agreement (Shrout & Fleiss, 1979). ICC scores ≤ 0.40 indicate an insufficient level of agreement; scores of ≤ 0.40 and ≤ 0.60 indicate a sufficient level of agreement; scores of ≤ 0.60 and ≤ 0.80 indicate a good level of agreement and > 0.80 indicate an excellent level of agreement (Shrout & Fleiss, 1979). To evaluate the ICC pre- and post- training we evaluated the agreement of each junior rater with the gold standard evaluation and then calculated the mean ICC value. Pearson correlation was used to analyze the retest reliability.

Results

Three junior raters used the MIS and the MMS to rate patients' mentalizing capacities on the basis of 15 session transcripts of psychotherapy sessions. The authors ***** evaluated blinded the 15 sessions, the ratings then were jointly revised and finally classified as gold standard. The same sessions were assessed by the three junior raters after a training at the use of the scales provided by the authors. Intraclass Correlation Coefficient was used to assess the inter-rater reliability of the three junior raters with the gold standard, pre- and post-training, for each item of the MIS (Table 1) and of the MMS (Table 2). Regarding the MIS, pre-training values ranged from 0.27 to 0.86 and the mean value is 0.60, while post-training values ranged from 0.43 to 0.89 and the mean value is 0.68. Regarding the MMS, pre-training values ranged from .06 to 0.84 and the mean value is 0.58, while post-training values ranged from 0.35 to 0.89 and the mean value is 0.69.

Pearson correlation was used to evaluate test retest reliability and produced sufficient values both for MIS and for MMS with mean values of r = 0.742 $p \leq .001$ for the MIS and r = 0.735 $p \leq .001$ for the MMS.

Study 2

Therapists

The sample is composed by four Caucasian therapists, two of which were males (age 42 and 45) and two females (age 30 and 58). Therapists were working in two different residential therapeutic communities for the treatment of patients with personality disorders and substance use disorder (two therapists were working in a community and two in the other). One therapist had an eclectic/integrated theoretical approach while two had a psychodynamic theoretical approach and one a systemic approach. Therapists' mean clinical experience was 9 years (*SD* = 7.9; min. = 1; max. =. 20). In each community, one therapist was seeing the selected patients in a group therapy setting, while the other was working with the same patients in an individual setting. The present study was approved by the IRB of the authors.

Patients

The sample is composed by 22 Caucasian patients with substance use disorder, treated in psychotherapy. This overall sample is composed by two different samples of 10 and 12 patients treated in the two therapeutic community centers. Twelve patients were male and 10 female; their mean age was 23.45 years (*SD* = 3.65; min. = 18; max. = 34). Patients were diagnosed with different primary addictions, more specifically: heroin (*n* = 17), cocaine (*n* = 2), drugs (*n* = 1); cannabis (*n* = 1) alcool (*n* = 1). The average age of the first episode of substance abuse was 14.8 years old (*SD* = 2.07; min. = 12; max. = 19). In eighteen patients, substances induced mental disorders, more specifically depressive disorder (*n* = 10), anxiety disorder (*n* = 12), sleep disorder (*n* = 4), bipolar disorder and related disorders (*n* = 4) and sexual disorders (*n* = 2). All patients were in a controlled environment. The average length of treatment at the moment of the evaluation was 12.3 months (*SD* = 7.05; min. = 2; max. = 24). Sixteen patients had also a diagnosis of personality disorder according to the DSM-5, more specifically: avoidant personality disorder (*n* = 9), histrionic personality disorder (*n* = 2), obsessive-compulsive dis-

order (n = 2), schizoid personality disorder (n = 1), dependent personality disorder (n = 1), and antisocial personality disorder (n = 1).

Six patients had at least one previous hospitalization and two patients reported self-harming behaviors. Eight patients were, by the time of the assessment, undergoing a pharmacotherapy and one patient had previously attempted suicide.

Measures

See the Measures Section of Study 1.

Statistical analysis

All analyses were conducted using SPSS Statistics 21 for Windows (IBM, Armonk, NY). In order to assess the inter-rater reliability between the junior raters and the gold standard evaluation (pre- and post-training), the ICC was calculated using Two-Way mixed effects model, single measures absolute agreement (Shrout & Fleiss, 1979). ICC scores ≤ 0.40 indicate an insufficient level of agreement; scores of ≤ 0.40 and ≤ .60 indicate a sufficient level of agreement; scores of ≤ .60 and ≤ 0.80 indicate a good level of agreement and > 0.80 indicate an excellent level of agreement (Shrout & Fleiss, 1979). To evaluate the IRR

Table 1. Inter Rater Reliability of MIS Items and Subscales (N = 15) Pre and Post Training.

MIS Item Text	Pre-Training ICC	Post Training ICC
P. is excessively focused on the facial expressions and/or nonverbal cues when communicating with others(including the therapist).	0.61	0.69
P. seems to inhibit the expression of emotions.	0.58	0.61
P. often seems to lack words to describe his/her ownfeelings and emotions.	0.71	0.71
P. can't assume other people's perspective whenreflecting on behaviors.	0.85	0.83
P. can easily be influenced by other people's emotions.	0.53	0.59
P. feels that his/her emotions are out of his/her control.	0.76	0.80
P. understands people more on a cognitive level thanon an affective one.	0.35	0.48
P. can't consider points of view that are different fromhis/her own.	0.82	0.82
P. tends to (consciously and/or unconsciously) imitateother people.	0.54	0.59
P. is impulsive.	0.86	0.89
When speaking about emotions, P. seems to be caughtin an intellectual game and/or use abstract terms.	0.32	0.68
P. can misunderstand other people's behavior.	0.67	0.75
P. seems to have a "sixth sense" about other people's(including the therapist) mental states.	0.56	0.69
P.'s emotions can change rapidly.	0.71	0.76
P. seems to be unconsciously attuned to other people'semotions.	0.37	0.49
P.'s emotions overcome his/her capacity to think.	0.66	0.75
P. seems to be detached from emotions.	0.73	0.77
When solicited (*e.g.*, with questions or confrontations),P. fails to reflect on his/her own behaviors.	0.75	0.80
P. fails to reflect on the first impression he or she hasof a person or a situation.	0.27	0.43
Even when discussing painful feelings, P. seems to bedetached.	0.80	0.86
P. provides some bizarre and/or unlikely explanationsof other people's behavior or reactions.	0.32	0.44
P. seems to preverbally intuit people's feelings orthoughts.	0.36	0.43
MIS Subscales	**Pre-Training ICC**	**Post Training ICC**
Cognitive Imbalance	0.56	0.70
External Imbalance	0.51	0.60
Affective Imbalance	0.75	0.80
Imbalance Toward Others	0.50	0.56
Imbalance Toward Self	0.67	0.71
Automatic Imbalance	0.58	0.65

ICC: Average Intraclass Correlation Coefficient: ≤ 0.40 poor reliability; ≤ 0.40 and ≤ 0.60 sufficient reliability; ≤ 0.60 and ≤ 0.80 good reliability; >0.80 excellent reliability (Shrout & Fleiss, 1979). MIS: Mentalization Imbalances Scale.

first we calculated the ICC for each couple of clinicians (working in the same therapeutic community) and then we calculated the mean global ICC for the four clinicians.

Results

Four clinicians used the MIS and the MMS to assess 22 patients with substance use disorder seen in two different therapeutic communities in different settings: For each community one clinician was working with the patients individually and one in a group setting. Intraclass Correlation Coefficient was used to assess the inter-rater reliability of the four clinicians for each item of the MIS (Table 3) and of the MMS (Table 4). Regarding the MIS, ICC values for each item ranged from 0.48 to 0.94 and the mean value was 0.81. In relation to the subscales, ICC values for the MIS ranged from .63 (external imbalance) to 0.92 (automatic imbalance). As for the MMS, ICC values for each item ranged from 0.59 to 0.94 and the mean

Table 2. Inter Rater Reliability of MMS Items and Subscales (N =15) Pre- and Post- Training.

MMS Item Text	Pre-Training ICC	Post Training ICC
P. seems to use his/her mental capacities to manipulateother people.	0.21	0.35
P. adopts unlikely explanations of behaviors.	0.83	0.89
P. understands that people can experience contrastingfeelings or desires.	0.24	0.41
P. seems to recognize the interest of significant othersonly if it is supported by concrete actions.	0.62	0.71
P. tends to adopt prejudice or generalization to explainhis/her own or others behavior.	0.53	0.60
P. seems to be intrusive towards other people.	0.80	0.85
P. interprets his/her own or other people's behavior interms of situational or physical constraints.	0.59	0.70
P. seems to treat therapy as an intellectual game.	0.19	0.49
P. tends to express an excessive certaintyupon other people's thoughts or feelings.	0.72	0.81
P. seems to excessively rely on the fact that externalchanges can change his/her moods.	0.84	0.84
P. tends to rely in an excessive way to his/her intuitivecapacity.	0.06	0.36
P. can describe coherently mental states.	0.75	0.83
P. can't consider point of view that differs from his/her own.	0.81	0.84
P. seems to focus more on what people do ratherthan on what they think or feel.	0.78	0.83
P.'s reflections on his/her inner world seemto be not genuine.	0.49	0.69
P. is excessively sure of the motivations and/orthoughts and/or emotions of others.	0.84	0.85
When solicited with specific questions, P. interpretsbehaviors in term of mental states.	0.68	0.76
P. seems to have all the answers regarding his/herown and/or other people's behavior.	0.72	0.79
P. tends to interpret behaviors in term of physicalcauses (*e.g.*, illness) and/or stable characteristics (*e.g.*,race, cultural background, or intelligence) and/or in termsof social external factors.	0.60	0.66
P. seems to be more focused on the practical resolutionof a problem rather than on the underpinning meanings.	0.76	0.73
P. believes he/she often knows what someone else isthinking or feeling.	0.82	0.82
P. is curious about the comprehension of his/her ownor other people's functioning.	0.13	0.55
P. uses common-sense explanations or cliché to explainaffects or feelings.		0.50 0.62
P. spontaneously interprets behaviors in term of mentalstates.	0.49	0.54
MMS Subscales	**Pre-Training ICC**	**Post Training ICC**
Excessive Certainty	0.66	0.75
Concrete Thinking	0.65	0.72
Good Mentalization	0.46	0.62
Teleological Thought	0.72	0.76
Intrusive Pseudomentalization	0.42	0.60

ICC: Average Intraclass Correlation Coefficient: ≤ 0.40 poor reliability; ≤ 0.40 and ≤ 0.60 sufficient reliability; ≤ 0.60 and ≤ 0.80 good reliability; >0.80 excellent reliability (Shrout & Fleiss, 1979). MMS: Modes of Mentalization Scale.

value is 0.84. In relation to the subscales, ICC values for the MMS ranged from 0.80 (concrete thinking and good mentalization) to 0.90 (teleological thought).

Discussion and Conclusions

The present studies represent an attempt to assess the inter-rater reliability and test-retest reliability of two clinician report measures for the assessment of mentalizing dimensions (MIS) and mentalizing failures (MMS). The first aim of this work was to assess the reliability of the measures with raters who did not have any clinical, empirical or assessment experience, and consequently to assess the effect of a training at the use of the measures (Study 1). By doing so we wanted to test the reliability of raters without any clinical experience pre and post training, in order to see if it can be considered as adequate. For this purpose, three junior raters used the MIS and the MMS to rate patients' mentalizing capacities on the basis of 15 session transcripts of psychotherapy sessions. The authors evaluated blinded the 15 sessions, the ratings then were jointly revised and finally classified as gold standard. The same sessions were assessed by the three junior raters after a training at the use of the scales provided by the authors. The pre-training IRR was overall sufficient, with mean values of .60 for the MIS and of 0.58 for the

Table 3. Inter Rater Reliability of MIS Items and Subscales (N = 22)

MIS Item text	ICC
P. is excessively focused on the facial expressions and/or nonverbal cues when communicating with others (including the therapist).	0.73
P. seems to inhibit the expression of emotions.	0.92
P. often seems to lack words to describe his/her own feelings and emotions.	0.94
P. can't assume other people's perspective when reflecting on behaviors.	0.81
P. can easily be influenced by other people's emotions.	0.94
P. feels that his/her emotions are out of his/her control.	0.85
P. understands people more on a cognitive level than on an affective one.	0.78
P. can't consider points of view that are different from his/her own.	0.80
P. tends to (consciously and/or unconsciously) imitate other people.	0.83
P. is impulsive.	0.84
When speaking about emotions, P. seems to be caught in an intellectual game and/or use abstract terms.	0.72
P. can misunderstand other people's behavior.	0.89
P. seems to have a "sixth sense" about other people's (including the therapist) mental states.	0.48
P.'s emotions can change rapidly.	0.63
P. seems to be unconsciously attuned to other people's emotions.	0.77
P.'s emotions overcome his/her capacity to think.	0.86
P. seems to be detached from emotions.	0.78
When solicited (*e.g.*, with questions or confrontations), P. fails to reflect on his/her own behaviors.	0.88
P. fails to reflect on the first impression he or she has of a person or a situation.	0.94
Even when discussing painful feelings, P. seems to be detached.	0.82
P. provides some bizarre and/or unlikely explanations of other people's behavior or reactions.	0.94
P. seems to preverbally intuit people's feelings or thoughts.	0.69
MIS Subscales	**ICC**
Cognitive Imbalance	0.80
External Imbalance	0.63
Affective Imbalance	0.80
Imbalance Toward Others	0.85
Imbalance Toward Self	0.90
Automatic Imbalance	0.92

ICC: Average Intraclass Correlation Coefficient: ≤ 0.40 poor reliability; ≤ 0.40 and ≤ 0.60 sufficient reliability; ≤ 0.60 and ≤ 0.80 good reliability; >0.80 excellent reliability (Shrout & Fleiss, 1979). MIS: Mentalization Imbalances Scale.

MMS. The evaluation of the pre-training IRR has enlightened some difficulties of the raters in relation to specific items. The most problematic items, both in the MIS and in the MMS, seem to be related to pseudomentalization (*e.g.*, "When speaking about emotions, P. seems to be caught in an intellectual game and/or use abstract terms."). It is possibile that this result is related to the difficulty to discriminate between good mentalization and a patient who "pretends to mentalize" but who is not in contact with the genuine, authentic facets of experience (Bateman & Fonagy, 2016). Moreover, some of the items of the MMS did not overcome 0.40 and in three cases remained close to that threshold after training. This result may be related to the lack of clinical experience of the raters, since the results of the raters working with patients have enlightened higher scores on all the items of the MMS. Another hypothesis is that this result is related to the procedure of using clinician report measures to assess psychotherapy sessions. In line with this consideration we must note that higher scores were found in relation to the same items when the measures were used to rate real patients (Study 2). For this reason, we are now working on

Table 4. Inter Rater Reliability of MMS Items and Subscales (N =22).

MMS Item text	ICC
P. seems to use his/her mental capacities to manipulate other people.	0.88
P. adopts unlikely explanations of behaviors.	0.74
P. understands that people can experience contrasting feelings or desires.	0.73
P. seems to recognize the interest of significant others only if it is supported by concrete actions.	0.88
P. tends to adopt prejudice or generalization to explain his/her own or others behavior.	0.94
P. seems to be intrusive towards other people.	0.84
P. interprets his/her own or other people's behavior in terms of situational or physical constraints.	0.59
P. seems to treat therapy as an intellectual game.	0.81
P. tends to express an excessive certaintyupon other people's thoughts or feelings.	0.92
P. seems to excessively rely on the fact that external changes can change his/her moods.	0.76
P. tends to rely in an excessive way to his/her intuitive capacity.	0.85
P. can describe coherently mental states.	0.82
P. can't consider point of view that differs from his/her own.	0.87
P. seems to focus more on what people do rather than on what they think or feel.	0.94
P.'s reflections on his/her inner world seemto be not genuine.	0.90
P. is excessively sure of the motivations and/or thoughts and/or emotions of others.	0.93
When solicited with specific questions, P. interprets behaviors in term of mental states.	0.89
P. seems to have all the answers regarding his/her own and/or other people's behavior.	0.88
P. tends to interpret behaviors in term of physical causes (*e.g.*, illness) and/or stable characteristics (*e.g.*, race, cultural background, or intelligence) and/or in terms of social external factors.	0.84
P. seems to be more focused on the practical resolution of a problem rather than on the underpinning meanings.	0.86
P. believes he/she often knows what someone else is thinking or feeling.	0.74
P. is curious about the comprehension of his/her own or other people's functioning.	0.81
P. uses common-sense explanations or cliché to explain affects or feelings.	0.90
P. spontaneously interprets behaviors in term of mental states.	0.72
MMS Subscales	**ICC**
Excessive Certainty	0.87
Concrete Thinking	0.80
Good Mentalization	0.80
Teleological Thought	0.90
Intrusive Pseudomentalization	0.86

ICC: Average Intraclass Correlation Coefficient: ≤ 0.40 poor reliability; ≤ 0.40 and ≤ 0.60 sufficient reliability; ≤ 0.60 and ≤ 0.80 good reliability; >0.80 excellent reliability (Shrout & Fleiss, 1979). MMS: Modes of Mentalization Scale.

an observer version of the measures, which should be more easily used on psychotherapy session transcripts.

The post-training IRR was higher than pre-training IRR, with mean values of 0.68 for the MIS and of 0.69 for the MMS with an increment of the reliability especially for the items which were not sufficient: 100% of the MIS items and 60% the MMS items which had an insufficient value of ICC pre-training (ICC < 0.40) were sufficiently reliable post-training (ICC > 0.40). Finally, test-retest reliability was overall fair, and suggested a sufficient stability throughout time of the evaluations. The pre-training IRR of junior raters without training was on the whole sufficient, however since the ICC increased after a training at the use of the scale, it is advisable to participate to a specific training in order to use the scales reliably to evaluate session transcripts for junior raters without any clinical or empirical experience.

In Study 2, four clinicians used the MIS and the MMS to assess 22 patients with substance use disorder seen in two different therapeutic communities in different settings: For each community one clinician was working with the patients individually and one in a group setting. Intraclass Correlation Coefficient was used to assess the inter-rater reliability of the four clinicians for each item of the MIS (Table 3) and of the MMS (Table 4). Regarding the MIS, ICC mean value for each item was 0.81 and in relation to the subscales, ICC values for the MIS ranged from 0.63 (external imbalance) to 0.92 (automatic imbalance). As for the MMS, ICC mean value for each item was 0.84. In relation to the subscales, ICC values for the MMS ranged from 0.80 (concrete thinking and good mentalization) to 0.90 (teleological thought). It is important to note that, even if the IRR in both studies ranged from sufficient to good, Study 2, in which the clinician reports were used in their natural context, *i.e.* everyday clinical practice, was related to higher score than Study 1, with good values of IRR for every item of the MMS and a sufficient value only for one item of the MIS and moderate to good values for all the other items. In relation to the reliability of the subscales of the measures, all the subscales were characterized by excellent values (*i.e.* $0.80 \leq ICC \leq 0.92$) with the exception of external imbalance that presented a good value.

It is possible that working on "real" patients (rather than on session transcripts) may be associated with better ratings for different reasons: For example, since mentalization is mostly implicit (Bateman & Fonagy, 2016), it is possible that working on a transcript, and not on the here-and-now of the therapeutic relationship, may block some crucial information, and this overall lowers the quality of the ratings of the scales. Further data should be necessary in relation to different sources, for example audio or video recordings of sessions, which could provide more information on the procedural and implicit facets of the interaction between the therapist and the patient (*e.g.* the voice tone, the prosody, the body posture, etc.). Moreover, Study 2 clinicians have clinical experience (while Study 1 raters were junior raters without any clinical experience) and this may help them at having a more nuanced understanding of their patients. Further studies are required in order to address the issue of the necessity of comparing ratings (pre- and post- training) of raters with different levels of clinical experience (*e.g.* undergraduate *vs* trainees of psychotherapy programs).

It is also important to note that the sample of the first study was composed by a limited number of transcripts of therapy sessions related to different patients, therefore it seems reasonable to assume that the information that Study 1 raters have on the patients are lower than the ones that a therapist seeing the patents for at least two months (either in a group or in an individual setting) could have. A further consideration on to the differences in the IRR between transcripts-related ratings and evaluations made on patients treated in psychotherapy by the clinician/rater is related to the fact that both the measures (MIS and MMS) were not originally meant for transcript-based evaluations. They differ from observer-rated measures in various ways and, most importantly, they do not provide a manual for the ratings. Moreover, the mean time for the assessment with the MIS and the MMS on session transcripts may be quite long (approximately an hour) when compared to the mean rating time of real patients (15 minutes). In light of these considerations and of the data provided by the present work, we are now conducting a study on the reliability and validity of an observer rated measure of the MIS and the MMS provided with a coding manual. However, the decision to use psychotherapy session transcripts with a clinician report measure represents a forced method and a compromise between the necessity to have also data on the IRR for clinician reports, and the impossibility to always have ratings from different therapists on the same patients. We can't exclude the possibility that working with this material may cause the loss of pivotal material for the ratings, since the raters do not have all the information that a clinician working vis-a-vis with the patient would have.

Both studies have different limitations. For example, as noted before, to use a clinician-report measure on psychotherapy session transcripts may be considered as a forced and limited methodology. However, this solution seems to be one of the only few accessible at the present time in order to have data on the IRR of clinician report measure. The MIS and the MMS are clinician report measures, but we have seen that they can reliably be applied to psychotherapy session transcripts. At the present time we are working on the observer-rater version of the measures, which will necessarily include also a rating of the quality of the sessions, in terms of the material available for the assessment, *e.g.* in terms of the number of demand questions provided by the therapist. Our data seem to suggest that the selected sessions are adapt for the rating of this capacity in a naturalistic setting (*i.e.* psy-

chotherapy sessions) however, future studies with an adequate sample should also consider this variable as a moderator for a subsequent sensitivity analysis.

Related to this issue, we must note that Study 2 is based on a quite limited sample, since having data on the same patients by different clinicians is possible only in specific and limited settings such as, for example, residential structures. Moreover, Study 2 sample was composed predominantly by patients with heroin abuse and we can't exclude that this is a source of bias in our results.

Finally, our results suggest that both the clinician reports can be reliably used, especially for the assessment of patients in everyday clinical practice, even without a specific training at the use of the scales. However, the reliability of the measures when applied to different coding material (*e.g.* psychotherapy session transcripts) may be sufficient, especially after a specific training at the use of the scales, but requires further investigation.

References

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The 'Reading the Mind in the Eyes' test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *J Child Psychol Psychiatry 42*(2):241–51.

Bateman, A.W., Bolton, R., & Fonagy, P. (2013). Antisocial personality disorder: A mentalizing framework. *J Lifelong Learning Psychoanal XI* (2):178–86. doi: 10.1176/appi.focus.11.2.178

Bateman, A.W., & Fonagy, P. (2004). *Psychotherapy for borderline personality disorder. Mentalization based treatment.* Oxford: Oxford University Press.

Bateman A.W., & Fonagy, P. (2016). *Mentalization Based Treatment for Personality Disorders: A Practical Guide*. Oxford: Oxford University Press.

Blackshaw, A.J., Kinderman, Hare, D. J., & Hatton, C. (2001). Theory of mind, causal attribution and paranoia in Asperger syndrome. *Autism* 5,:147–63. doi:10.1177/1362361301005002005

Blagov, B., Bi, W., Shedler, J., & Westen, D. (2012). The Shedler-Westen Assessment Procedure (SWAP): Evaluating psychometric questions about its reliability, validity, and fixed score distribution. *Assessment 19*(3):370–82. doi: 10.1177/1073191112436667.

Davidson, K. M., Obonsawin, M. C., Seils, M., & Patience, L. (2003). Patient and clinician agreement on personality using the SWAP-200. *J Personality Disord* 17(3):208-18. doi:10.1521/pedi.17.3.208.22148

Fonagy, P., Leigh, T., Steele, M., Steele, H., Kennedy, R., Mattoon, G., (...) Gerber, A. (1996). The relation of attachment status, psychiatric classification, and response to psychotherapy. *J Consulting Clin Psychol* 64(1):22-31. doi:10.1037/0022-006X .64.1.22

Fonagy. P., Luyten, P., Moulton-Perkins, A., Lee, Y.W., Warren, F., Howard, S. … Lowyck, B. (2016). Development and Validation of a Self-Report Measure of Mentalizing: The Reflective Functioning Questionnaire. *PloS One* 11(7). doi:10.1371/journal.pone.0158678

Fonagy, P., Target, M., Steele, H., & Steele, M. (1998). *Reflective- Functioning manual: Version 5 for application to adult attachment interview.* Unpublished manual. London: University College.

Gagliardini, G., & Colli, A. (2019). Assessing mentalization: Development and preliminary validation of the Modes of Mentalization Scale. *Psychoanal Psychol 36*(3):249-58. doi: 10.1037/pap0000222

Gagliardini, G., Gullo, S., Caverzasi, E., Boldrini, A., Blasi, S., & Colli, A. (2018). Assessing mentalization in psychotherapy: First validation of the Mentalization Imbalances Scale. *Research in Psychotherapy: Psychopathology, Process and Outcome 21*(3):164-77. doi: 10.4081/ripppo.2018.339

Hausberg, M.C., Schulz, H., Piegler, T., Happach, C.G., Klöpper, M., Brütt, A.L. (…) Andreas, S. (2012). Is a self-rated instrument appropriate to assess mentalization in patients with mental disorders? Development and first validation of the mentalization questionnaire (MZQ). *Psychother Res* 22(6):699-709. doi: 10.1080/10503307.2012.709325

Huprich, S.K., Bornstein, R.F. & Schmitt, T.A. (2011). Self-report methodology is insufficient for improving the assessment and classification of Axis II personality disorders. *J Personality Disord* 25(5):557-70. doi: 10.1521/pedi.2011.25.5.557.

Katznelson, H. (2014). Reflective function: A review. *Clin Psychol Rev 34*(2): 107–17. doi: 10.1016/j.cpr.2013.12.00

Luyten, P., Fonagy, P., Lowyck, B., & Vermote, R. (2012). Assessment of mentalization. In A.W. Bateman & P. Fonagy (Eds.), *Handbook of mentalizing in mental health practice* (pp. 43–66). Arlington, VA: American Psychiatric Publishing.

Meehan, K.B., Levy, K.N., Reynoso, J.S., Hill, L.L., & Clarkin, J.F. (2009). Measuring reflective function with a multidimensional rating scale: comparison with scoring reflective function on the AAI. *J Am Psychoanal Assoc* 57(1):208-13. doi: 10.1177/0003065109057001008

Morey, L.C. (2014). Borderline features are associated with inaccurate trait self-estimations. *Borderline Personality Disord Emotion Dysregulation* 1(1):4. doi: 10.1186/2051-6673-1-4

Oakley, B.F.M., Brewer, R., Bird, G., & Catmur, C. (2016). Theory of mind is not theory of emotion: A cautionary note on the Reading the Mind in the Eyes Test. *J Abnorm Psychol* 125(6):818-23. doi:10.1037/abn0000182

Rudden, M., Milrod, B., Target, M., Ackerman, S., & Graf, E. (2006). Reflective functioning in panic disorder patients: A pilot study.*J Am Psychoanal Assoc 54*(4):1339-43. doi.org/10.1177/00030651060540040109

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bull* 86(2):420–28. doi:10.1037/0033-2909.86.2.420

Skårderud, F. (2007). Eating one's words, part II: The embodied mind and reflective function in anorexia nervosa – theory. *European Eating Disord Rev* 15(4):243–52. doi:10.1002/erv.778

Taubner, S., Hörz, S., Fischer-Kern. M., Doering, S., Buchheim, A., & Zimmermann, J. (2013). Internal structure of the Reflective Functioning Scale. *Psychological Assessement* 25(1):127-35. doi: 10.1037/a0029138

Taubner, S., Kessler, H., Buchheim, A., Kächele, H., & Staun, L. (2011). The role of mentalization in the psychoanalytic treatment of chronic depression. *Psychiatry 74*(1):49–57. doi: 10.1521/psyc.2011.74.1.49