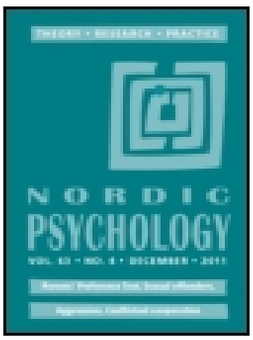

Nordic Psychology

ISSN: 1901-2276 (Print) 1904-0016 (Online) Journal homepage: http://www.tandfonline.com/loi/rnpy20

The mentalization-based therapy adherence and quality scale (MBT-AQS): Reliability in a clinical setting

Sebastian Simonsen, Sophie Juul, Mickey Kongerslev, Sune Bo, Espen Folmo & Sigmund Karterud

To cite this article: Sebastian Simonsen, Sophie Juul, Mickey Kongerslev, Sune Bo, Espen Folmo & Sigmund Karterud (2018): The mentalization-based therapy adherence and quality scale (MBT-AQS): Reliability in a clinical setting, Nordic Psychology, DOI: 10.1080/19012276.2018.1480406

To link to this article: https://doi.org/10.1080/19012276.2018.1480406

Published online: 04 Jun 2018.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=rnpy20

Nordic Psychology, 2018
https://doi.org/10.1080/19012276.2018.1480406

ARTICLE

Check for updates

The mentalization-based therapy adherence and quality scale (MBT-AQS): Reliability in a clinical setting

SEBASTIAN SIMONSEN[1], SOPHIE JUUL[1], MICKEY KONGERSLEV[2,3], SUNE BO[3,4], ESPEN FOLMO[5] & SIGMUND KARTERUD[5,6]

Correspondence address: Sebastian Simonsen, Stolpegaard Psychotherapy Centre, Stolpegardsvej 20, 2820, Gentofte, Denmark. Email: sebastian.simonsen@regionh.dk

Abstract

The *Mentalization-Based Treatment Adherence and Quality Scale* (MBT-AQS) is a 17-item measure of treatment adherence and quality of individual mentalization-based therapy (MBT). Until now, reliability research on the scale has primarily been conducted by highly experienced raters from the Norwegian MBT Quality Lab who were part of its development. Hence, it can be questioned whether only experts in research settings can achieve satisfying levels of reliability on the scale. In this study, we investigated whether a satisfying level of reliability on the MBT-AQS could be obtained by experienced MBT therapists in a clinical setting following a brief one-day training course. The overall reliabilities for six raters were good for adherence (.67) and for quality (.62). Thus, the MBT-AQS was found to be an appropriate MBT adherence rating instrument with clinical and educational utility outside of the Norwegian MBT Quality Lab. However, ambiguity of some constructs, low frequency of certain item ratings and low levels of MBT quality challenge reliability. This is discussed in the context of utilizing the scale for clinical and supervising purposes

Keywords: treatment integrity, adherence, quality, assessment, mentalization-based treatment, MBT

Introduction

In psychotherapy research, treatment integrity "refers to the extent to which the intervention was implemented as intended" (Perepletchikova, Treat, & Kazdin, 2007, p. 829). Treatment integrity is assessed by rating therapy interventions in terms of both theoretically prescribed and proscribed actions and often in terms of both the therapist's *adherence* to the treatment protocol, and the therapist's *competence* (or quality) in delivering therapy in a skillful manner. Skill is often referred to as the therapist's ability to respond to the therapeutic context appropriately. Thus, competence presupposes adherence, but adherence does not necessarily imply competence (Waltz, Addis, Koerner, & Jacobsen, 1993).

[1]Stolpegaard Psychotherapy Centre, Mental Health Services, Gentofte, Capital Region of Denmark
[2]Psychiatric Clinic Roskilde, Region Zealand Psychiatry, Roskilde, Denmark
[3]Department of Psychology, University of Southern Denmark, Odense, Denmark
[4]Psychiatric Research Unit, Region Zealand, Slagelse, Denmark
[5]Norwegian National Advisory Unit for Personality Psychiatry, Section for Personality Psychiatry, Oslo University Hospital, Oslo, Norway
[6]The Norwegian Institute for Mentalizing, Oslo, Norway

© 2018 The Editors of Nordic Psychology

Reports of treatment integrity in psychotherapy research are rare, even in randomized controlled trials (Perepletchikova et al., 2007; Prowse, Nagel, Meadows, & Enticott, 2015), which results in questionable internal and external validity. Only a few adherence rating scales have been developed, and even fewer have been investigated to determine their function outside the narrow context of highly specialized research communities. This is rather unfortunate. Besides being relevant tools for evaluating findings from psychotherapy research, treatment integrity scales can be useful educational tools for the training and supervision of psychotherapists in clinical settings. In addition, using treatment integrity scales can be particularly useful during a time when psychotherapy is increasingly being audio-recorded or videotaped, creating unprecedented opportunity for supervisors to assess treatment fidelity (Anderson, Crowley, Patterson, & Heckman, 2012). However, if treatment integrity scales are to be used more broadly for educational purposes outside specialized research communities, it is important to obtain knowledge about their utility in clinical settings and to examine threats to reliability.

MBT is an empirically supported manualized treatment for borderline personality disorder (BPD) that structurally contains three major components; psychoeducation, individual- and group psychotherapy (Bateman & Fonagy, 1999, 2009). Furthermore, the model is potentially efficacious in the treatment of other psychiatric disorders, e.g., eating disorders (Robinson et al., 2016) and antisocial personality disorder (Bateman, O'Connell, Lorenzini, Gardner, & Fonagy, 2016).

The Mentalization-Based Treatment Adherence and Quality Scale (MBT-AQS) was developed by Karterud and Bateman (2011) to measure the treatment integrity of individual MBT. The MBT-AQS is designed as a 17-item scale with separate ratings of adherence and quality for each item, and an overall scale score for both adherence and quality. Adherence is assessed as the frequency of interventions belonging to an item, e.g., "Acknowledging good mentalizing" (Item 7). Five items are not rated for adherence as they reflect more overarching strategies and behaviors, such as being warm and genuine. Hence these five items target *other issues than the frequency* of a specific behavior, e.g., "Engagement, interest and warmth" (Item 1), or mainly target *phenomena that are handled indirectly* by therapists, e.g., "Pretend mode" (Item 8) or "Adjustment to level of mentalizing" (Item 4). The therapist's "Engagement, interest and warmth" is an important common-factor which colors all interventions. Furthermore, some interventions serve many functions simultaneously: Item 2 (a curious explorative question) with a slightly cognitive quality can be performed as "Regulation of arousal" (Item 5) with a patient who is becoming overly emotional. Sometimes such indirect interventions are very difficult to identify and rate reliably. Quality is rated for all 17 items on a 1–7 Likert scale. A mean overall score of 4 or above qualifies the session as adequate MBT (see Table 1). The rater's starting point should be at 4, reflecting the basic assumption that the therapist is "good enough". The rater should then consider positive or negative deviations from this starting point. The manual for the MBT-AQS contains procedures to guide the rater in determining the degree of deviation from a "good-enough" practice. If an item/intervention is not performed, but should have been, according to the rater, this is considered a "missed opportunity" and the quality score for that item is 1–3, depending on an overall judgment of the clinical significance of the missed opportunity. Examples of missed opportunities would be when a patient has an unwarranted belief that is not challenged or when the therapist fails to acknowledge a patient for good mentalizing. The consequences of such missed opportunities can be viewed as constituting a continuum from relatively unimportant (quality score of 3) to have a negative impact on the entire session, for example, when a patient is obviously

 © 2018 The Editors of Nordic Psychology

Table 1. The 17 items of the MBT Adherence and Quality scale and description of level 4.

Item name	Good enough quality level (4)
1. Engagement, interest and warmth	The therapist appears genuinely warm and interested. The rater gets the impression that the therapist cares. Several concrete comments communicate this positive attitude
2. Exploration, curiosity and a not-knowing stance	The therapist poses appropriate questions designed to promote exploration of the patient's and others mental states, motives and affects and communicate a genuine interest in finding out more about them
3. Challenging unwarranted beliefs	The therapist confronts and challenges unwarranted opinions about oneself or others in an appropriate manner
4. Adjusting to mentalizing capacity	The therapist seems to have adapted to the patient's mentalizing level and the interventions are for the most part short, concise and unpretentious
5. Regulating arousal	The therapist plays an active role in terms of maintaining emotional arousal at an optimal level (not too high so that the patient loses his or her ability to mentalize; not too low so that the session becomes meaningless emotionally)
6. Stimulating mentalization	The aim of the interventions clearly seems to be to stimulate the mentalizing of experiences of self and others in an ongoing process and is less concerned about content and interpretation of content in order to promote insight
7. Acknowledging positive mentalizing	The therapist identifies and explores good mentalization and this is accompanied by approving words or judicious praise
8. Pretend mode	The therapist identifies pretend mode and intervenes to improve mentalizing capacity
9. Psychic equivalence	The therapist identifies psychic equivalence functioning and intervenes to improve mentalizing capacity
10. Focus on affects	The interventions focus primarily on affects, more than on behaviour. The attention is directed at affects as they are expressed in the here and now, and particularly in terms of the relationship between patient and therapist
11. Focus on interpersonal affects	The therapist connects emotions and feelings to recent or immediate interpersonal events
12. Stop and rewind	The therapist identifies at least one incident in which the patient reacts in a maladaptive way to an interpersonal event, then tries to slow down the pace and find out about the incident step-by-step
13. Validating feelings	The therapist expresses a normative view on the warranted nature of the patient's emotional reactions after these are sufficiently investigated and understood
14. Relation to therapist	The therapist comments on attempts to explore – together with the patient – how the patient relates to the therapist during the session and stimulates reflections on alternative perspectives whenever appropriate
15. Counter-transference	The therapist actively utilizes his/her own feelings and thoughts about the relationship to the patient and attempts by this to stimulate an exploration of the relationship between them
16. Validating understanding	The therapist checks out his/her understanding of the patient's state of mind and to what extent this corresponds with the patient's understanding. Then he/she lets his/her own understanding be influenced by the patient's understanding and openly admits to any misunderstanding whenever they occur
17. Integrating group experiences	The therapist stimulates exploration of the patient's experiences from the group therapy sessions and helps to integrate the material so that the treatment as a whole is coherent

© 2018 The Editors of Nordic Psychology

annoyed with the therapist but the therapist fails to address this as prescribed in MBT (quality score of 1). After rating each item individually, the rater decides on an overall score for the whole therapy session for both adherence and quality, respectively. The overall score is given, not on the basis of an average score for the 17 items, but on the basis of a clinical judgment of the entire therapy session and with special emphasis on the items concerning the "Exploration, curiosity and a not-knowing stance" (Item 2), "Stimulating mentalization" (Item 6), "Focus on affects" (Item 10), and "Focus on interpersonal affect" (Item 11) (Karterud & Bateman, 2011).

In a study of the structure and reliability of the scale, Karterud et al. (2013) found initial support for high reliability coefficients for both overall adherence (.84) and quality (.88). At item level, however, several problems were identified. These problems included low reliability of ratings of items that concerned core MBT constructs, e.g., "Dealing with pretend mode" (Item 8). In a reliability study of a newly developed adherence and quality scale for mentalization-based group therapy (MBT-G-AQS) (Folmo et al., 2017) the items "Handling pretend mode" and "Handling psychic equivalence" were again found to be the least reliable items, supporting the need for more empirical attention to these items. The MBT-AQS has been applied in two outcome studies (Kvarstein et al., 2015; Möller, Karlgren, Sandell, Falkenstrom, & Philips, 2016). Findings from the latter study indicate that the therapist's adherence to MBT principles and competence in the performance of MBT stimulated in-session patient mentalization. Thus, the MBT-AQS can potentially be a beneficial instrument, both for the education of therapists and in studies focusing on treatment mediators of enhancement of patient's mentalization skills.

However, with the exception of one randomized controlled study by Möller et al., 2016; all previous studies using MBT-AQS have employed raters from the Norwegian Quality Lab for MBT (The MBT-lab). These highly specialized and certified MBT raters also participated in the research group that developed the scale. If the scale is to be used more broadly, e.g., for supervision, training, and educational purposes in clinical contexts, it is important to investigate whether it is possible to train raters outside the MBT-lab and other highly specialized research settings. Thus, the main purpose of this study was to explore whether it was possible for experienced MBT clinicians to replicate acceptable reliability of ratings within a clinical setting outside the MBT-lab following a brief, one-day training course (Loeb et al., 2005; Schanche, Høstmark Nielsen, McCullough, Valen, & Mykletun, 2010).

Method

The study was carried out in two Danish outpatient clinics specialized in MBT for personality disorders (PDs): *Stolpegaard Psychotherapy Center, Mental Health Services, Capital Region of Denmark* and *Psychiatric Clinic Roskilde, Region Zealand Psychiatry.*

Patients

The patients had been diagnosed with PDs, mainly of borderline type, according to ICD-10 criteria, and were already included in an MBT treatment program at the two clinics (see Simonsen, Heinskou, Sørensen, Folke, and Lau (2017) for a detailed explanation of patient characteristics in such specialized outpatient treatment facilities.) Thirteen patients, who had given their informed consent, were included in the study.

 © 2018 The Editors of Nordic Psychology

Therapists

Thirteen therapists across the two outpatient clinics participated in the study. All therapists were experienced in MBT treatment of PDs. By profession there were one psychiatrist, one MD, four clinical psychologists, one social worker, one physical therapist, and five psychiatric nurses. The therapists were informed about the project in staff meetings, and 13 therapists volunteered to participate in the study.

Training of raters

Across the two trial sites, six experienced Danish MBT therapists completed a one-day training course on how to apply the MBT-AQS in individual MBT, provided by two of the members of the MBT-lab. By profession there were three PhD-level psychologists, one clinical psychologist, one social counselor, and one psychiatrist. All raters were certified MBT practitioners, and at the time of data collection two were certified MBT supervisors. Written instructions were sent to the raters prior to the one-day training course along with a transcript of one full therapy session, which they were asked to rate prior to the training course. Aside from thoroughly reviewing the transcript, the training course included ratings of two video-recorded MBT sessions using the MBT-AQS. The Norwegian experts and the Danish raters then compared and discussed their ratings of the transcript and the two training videos in order to identify systematic rating mistakes and to strengthen a shared understanding of items and rules. The two training videos were not included in this study.

Procedure

The six Danish raters independently rated all MBT-AQS items for all 13 individual MBT sessions. Ratings were not blind, as the raters knew most of the therapists. The same 13 video-recorded sessions were then sent to the MBT-lab using encrypted USB keys, and the sessions were rated, blind to the Danish ratings, by two members of the MBT-lab separately. The two Norwegian raters decided on a consensus rating which served as the "gold standard". All ratings were handled and entered by a secretary and a student assistant in a research unit to ensure that all the Danish raters were blind to each other's results. Statistical analyses were performed by one of the raters after data completion.

Statistical analyses

Interrater reliabilities were assessed by means of the intraclass correlation coefficient (ICC), 2.1 (Shrout & Fleiss, 1979). The ICC is an appropriate measure of variables on a common scale, where the variables share both their metric and variance (McGraw & Wong, 1996). We choose a two-way random effect model because we want to generalize our reliability results to raters who have similar characteristics as the raters in our study, certified MBT practitioner or supervisor. The single-measure ICC is appropriate when the instrument will be rated by only one rater, which is how the scale is mostly used in training and supervision contexts (Das, de Ruiter, Doreleijers, & Hillege, 2009). In accordance with the general recommendations from Shrout and Fleiss (1979) we consider absolute agreement to be more important than consistency in this type of study.

ICCs vary from 0 to 1. The following guidelines were used for evaluating the observed interrater reliability: when ICC is below .40, the level of clinical significance is poor; when it is between .40 and .59, the level of clinical significance is fair; when it is between .60 and .74, the level of clinical

 © 2018 The Editors of Nordic Psychology

significance is good; and when it is between .75 and 1, the level of clinical significance is excellent (Cicchetti, 1994).

Correlations between Danish ratings and ratings by the MBT-lab were measured with Pearson's *r*. Pearson's *r* is used for measuring variables that share neither their metric nor variance (McGraw & Wong, 1996). Ratings where an intervention was appropriately not performed (rated 0) were excluded from analysis. When an intervention was not performed, but according to raters should have been (rated 1–3), ratings were included in analysis.

Results

Tables 2 and 3 display the mean scores and standard deviations for both the Norwegian MBT-lab and the Danish raters, the correlation between these, and the reliability coefficients for both single and average rater for adherence and quality, respectively. The correlations (Pearson's *r*)

Table 2. Mean item scores, correlations between MBT-Lab and danish raters, and interrater reliability of MBT-AQS items (*N* = 13): Adherence.

Item description	*M* lab	*SD* lab	*M* raters	*SD* raters	*r*	ICC single rater	ICC 95% confidence interval
1. Engagement							
2. Exploring	18.4	7.7	22.5	9.8	.87**	.55	.32–.79
3. Challenging	4.3	4.2	3.5	2.4	.87**	.54	.32–.79
4. Adjustment							
5. Regulating arousal							
6. Stimulating mentalization							
7. Acknowledging positive mentalizing	2.5	2.0	1.2	.82	.59*	.42	.20–.71
8. Pretend mode							
9. Psychic equivalence	2.6	2.1	1.7	1.6	.76**	.37	.15–.66
10. Focus on affects	10.1	5.7	5.8	1.9	.64*	.15	.10–.45
11. Focus on interpersonal affects	6.3	4.2	8.6	2.9	.43	.19	.20–.50
12. Stop and rewind	1.5	.6	.71	.80	.74**	.57	.34–.80
13. Validating feelings	4.0	3.1	3.8	1.8	.15	.50	.27–.76
14. Relation to therapist	5.1	3.5	2.8	4.0	.87**	.90	.80–.96
15. Counter-transference	2.5	2.3	1.2	.85	.73**	.25	.06–.56
16. Validating understanding	8.2	5.8	13.8	8.9	.84**	.59	.36–81
17. Integrating group experiences	17.8	9.1	9.8	6.6	.93**	.86	.74–95
Overall	4.2	1.3	4.2	1.2	.72**	.67	.46–.86

Notes: MBT-AQS = Mentalization-based therapy – adherence and quality scale; ICC = intraclass correlation coefficient.
*$p < .05$, two-tailed.
**$p < .01$, two-tailed.

© 2018 The Editors of Nordic Psychology

between the Norwegian MBT-lab and the Danish raters were high for both adherence (.72) and quality (.86) for the overall ratings. Statistically significant Pearson's *r* correlations varied from .59 to .93 at item level for both adherence and quality. Two items were not significantly correlated for adherence: "Focus on interpersonal affects" (Item 11) and "Validating of emotional reactions" (Item 13). Two items were not significantly correlated for quality: "Stop and rewind" (Item 12), and "Integrating group experiences" (Item 17).

The reliabilities for overall ratings of adherence (.67) and quality (.62) for the average rater were both good, according to the guidelines provided by Cicchetti (1994). Tables 2 and 3 reveal a large reliability variation among the different items. Interventions rated with the highest reliability on both adherence and quality were "Relation to the therapist" (Item 14) and "Integration of group experiences" (Item 17). Reliability for these items ranged from .65 to .90. With regard to reliability of the adherence ratings most items fall into the fair range. The lowest adherence reliability were "Focus on affects" (item 10) and "Focus on interpersonal affects" (item 11) with reliabilities of rat-

Table 3. Mean item scores, correlations between MBT-lab and danish raters, and interrater reliability of MBT-AQS items ($N = 13$): quality.

Item description	*M* lab	*SD* lab	*M* raters	*SD* raters	*r*	ICC single rater	ICC 95% confidence interval
1. Engagement	4.2	1.3	4.7	.87	.77**	.35	.14–65
2. Exploring	4.5	1.5	4.5	.93	.72**	.49	.27–.76
3. Challenging	3.5	1.2	3.8	1.1	.65*	.32	.12–.63
4. Adjustment	3.4	1.4	4.1	.95	.73**	.48	.25–.75
5. Regulating arousal	3.5	1.5	3.6	.91	.90**	.52	.37–.82
6. Stimulating mentalization	3.4	1.7	4.0	1.2	.89**	.59	.29–.77
7. Acknowledging positive mentalizing	3.4	1.1	3.5	1.1	.68*	.57	.34–.80
8. Pretend mode	3.0	.92	3.3	.87	.74**	.21	.03–.51
9. Psychic equivalence	2.6	1.5	3.5	1.0	.91**	.26	.07–.57
10. Focus on affects	3.9	1.5	4.0	1.1	.69**	.57	.34–80
11. Focus on interpersonal affects	3.9	1.6	4.2	.90	.72**	.39	.17–.68
12. Stop and rewind	2.6	.52	3.4	.86	.63	.18	.02–.48
13. Validating feelings	3.3	1.1	4.1	.76	.65*	.31	.11–.62
14. Relation to therapist	2.8	1.3	3.6	1.2	.82**	.65	.43–85
15. Counter-transference	3.2	1.6	3.4	.95	.90**	.33	.12–.63
16. Validating understanding	3.9	.86	4.3	.74	.76**	.34	.13–.64
17. Integrating group experiences	5.0	1.4	4.0	.90	.54	.70	.50–87
Overall	3.7	1.6	4.0	1.2	.86**	.62	.41–83

Notes: MBT-AQS = mentalization-based therapy – adherence and quality scale; ICC = intraclass correlation coefficient.
*$p < .05$, two-tailed.
**$p < .01$, two-tailed.

© 2018 The Editors of Nordic Psychology

ings of .15 and .19, respectively. With regard to reliability of the quality ratings, roughly half of the items fall into the poor range, indicating that absolute agreement for single ratings of item quality might be difficult to obtain with little training in clinical settings. Furthermore, inspection of the 95% confidence intervals similarly reveals that although the reliabilities of overall ratings are fairly robust, this is generally not the case at the item level. Also evident is the large variation in the frequency of item ratings as shown in the mean adherence ratings ranging from 22.5 for "Exploring" (Item 2) to only .80 for "Stop and rewind" (Item 12) and .85 for "Counter-transference" (Item 15).

The individual therapists varied with respect to their overall adherence and quality and at item level. The overall A/Q for the least adherent (MBT) therapist was 1.5 (Therapist 13), while the overall A/Q for the most adherent therapist (Therapist 1) was 7, as rated by the MBT-lab (the gold standard). Tables 4 and 5 display a ranking of the 13 therapists' overall A/Q ratings by the MBT-lab and the associated reliability ratings of both adherence (Table 4) and quality (Table 5) by the six Danish

Table 4. Ranking of therapist and the corresponding single-rater reliability of the Danish raters (N = 13): overall adherence.

Ranking MBT lab overall adherence	Reliability 6 Danish raters overall adherence
7 (Therapist 1)	.81
5 (Therapist 9)	.77
5 (Therapist 11)	.75
5 (Therapist 10)	.71
5 (Therapist 6)	.70
5 (Therapist 4)	.69
5 (Therapist 7)	.67
4 (Therapist 2)	.75
4 (Therapist 12)	.72
3 (Therapist 8)	.72
3 (Therapist 5)	.71
2 (Therapist 3)	.87
2 (Therapist 13)	.72

Table 5. Ranking of therapist and the corresponding single-rater reliability of the Danish raters (N = 13): overall quality.

Ranking MBT lab overall quality	Reliability 6 Danish raters overall quality
7 (Therapist 1)	.65
6 (Therapist 9)	.64
5 (Therapist 7)	.48
4 (Therapist 2)	.57
4 (Therapist 6)	.71
4 (Therapist 10)	.47
4 (Therapist 12)	.47
3 (Therapist 4)	.48
3 (Therapist 11)	.47
2 (Therapist 3)	.50
2 (Therapist 5)	.51
2 (Therapist 8)	.21
1 (Therapist 13)	.36

raters. Regarding the adherence ratings (Table 4), the reliabilities of the six Danish raters were good to excellent for both high- and low-ranked therapists. However, regarding the quality ratings (Table 5), it is evident that cases where therapists were ranked highest by the MBT-lab were also the ones that obtained the highest reliability scores from the six Danish raters. Conversely, in cases where the MBT-lab gave very low quality ratings, the reliabilities of the Danish ratings were poor.

 © 2018 The Editors of Nordic Psychology

Discussion

In this study, we have presented data on the inter-rater reliability of the Mentalization-Based Therapy Adherence and Quality Scale (MBT-AQS) in a clinical setting, where 13 videotaped individual MBT sessions, mainly for BPD, were rated. Our findings demonstrated that it is generally possible for experienced MBT therapists to reliably rate the sessions overall adherence and quality of MBT, as also reported in the original study (Karterud et al., 2013). The reliabilities of six raters were high (.62/.67). This study, thus, provides empirical support for the use of MBT-AQS for clinical purposes outside the MBT-lab. However, important challenges to reliability on item level were identified. The main sources of unreliability included low frequency of items, ambiguity of constructs, and low therapist quality. In the general treatment integrity literature, challenges to reliability are primarily dealt with through more refined item descriptions (Bellg et al., 2004). Results from this study indicate a large reliability variation across the different items. This is a common finding in adherence rating scale studies (Barber, Liese, & Abrams, 2003; Karterud et al., 2013). Some items showed satisfactory reliability, while others showed low reliability. Items with the lowest reliabilities were "Focus on affects", "Focus on interpersonal affects", "Counter-transference", and "Psychic equivalence". Firstly, then, this study suggests a potential challenge for clinical raters to judge whether the therapist's intervention is concerned with the patient's general affects or with interpersonal affects. The manual should be more concise in clarifying this issue. Furthermore, some items were rarely used ("Counter-transference" and "Psychic equivalence"), which creates a situation where small deviations in the raters' understandings have a large negative impact on reliability. Thus, consistent with the recommendations proposed by Karterud et al. (2013) and Folmo et al. (2017), the manual should be more refined and specific, especially in regard to what counts as high versus low quality.

The individual profiles revealed large differences among therapists. From a psychometric point of view, this is favorable, as it attests to the discriminatory power of the scale. The scale makes it possible to distinguish the excellent performance from the poor. High overall therapist adherence and quality (as established by the MBT-lab) was found to be associated with high reliability of ratings. However, low therapist quality (as established by the MBT-lab) was associated with low reliability of ratings. In short, raters had a higher reliability when they rated sessions with high adherence and quality, and disagreed more when they evaluate low rated MBT. This is an important finding, as it has implications for the use of the MBT-AQS for educational and supervisory purposes. It is often recommended that therapists bring audio- or video-recorded sessions with difficult cases to supervision, allowing the supervisor to examine the manual adherence and therapeutic skills appropriately. However, supervisors may have difficulty obtaining satisfying levels of reliability for therapy sessions in which the overall quality of interventions is low. We did not study the exact source of this lack of reliability, but speculate that even though the item quality is not based on the effects of the intervention, lack of agreement may have to do with raters differing in how much they attribute poor effects to the therapist or to the patient. What if the therapist "does the right thing", but the patient responds aversively? When Waltz et al. (1993) rigorously defined adherence and competence, they realized that the context of therapy-characteristics of the client was important: "When clients like their therapist and improve substantially, it is easier for therapists to look competent" (p. 624). Further, patterns of attribution may be affected by well-known social cognitive biases, e.g., halo effect and confirmation bias that may make the rating more difficult for a rater or supervisor who knows the therapist well,

 © 2018 The Editors of Nordic Psychology

compared -with a rater who is an outsider. Problems with patterns of attributions in clinical contexts can be further studied by collecting more data on the raters' knowledge and preconceived ideas (e.g., years of experience, amount of MBT training, reputation etc.) about the clinicians and relating such data to reliability estimates. Threats to reliability and possible remedies need further investigation and are important if adherence and quality scales are to be used more widely in education and psychotherapy.

There are several limitations to this study. First, a systematic patient sampling was not used to ensure a larger variance in the patients' personality pathology or therapeutic phases. all patients were clinically diagnosed with PDs, mainly BPD, but no standardized clinician-administered tests were conducted to establish a reliable BPD-diagnosis for patients according to DSM-V criteria.

Thus, we were unable to study the association between therapist adherence and the patient's psychiatric severity as indicated by pretreatment or concurrent levels of personality pathology (Barber & Critis-Christoph, 1996). For example, in the initial study of the scale, Karterud et al. (2013) found low reliabilities for the item Stop and rewind (Item 12) for both adherence and competence (quality). In contrast, we found a pattern of good adherence reliability but poor quality reliability for that item (.57/.18). This may be due to the differences in our samples' psychiatric severities, and how far along they were in the therapeutic process. A "Stop and rewind" intervention is more appropriate when the patient is highly aroused. Most of the patients in the initial study were in the "middle phase" of therapy, and none displayed an acute suicidal risk. Thus, the item was rated infrequently, potentially resulting in low reliability of ratings. However, in the present study, some patients were in the "initial phase" of treatment, and the item was used more frequently, resulting in higher reliabilities. Thus, an ideal patient sample covering a wider range of psychiatric severity and therapy phases would have been favorable. Second, there may have been some unknown biases related to the fact that the Danish raters knew most of the therapists. Data about such knowledge and biases among raters was not systematically collected and may have affected reliability either positively or negatively. Lastly, this is a small study with only six raters and thirteen observations. Therefore, findings should be interpreted very conservatively and larger studies are needed before any robust conclusions about the MBT-AQS and its clinical applicability can be made.

In conclusion, the findings reported in this study should be considered as offering further empirical support for the use of the 17-item version of the MBT-AQS for education and supervision purposes in clinical contexts outside laboratory conditions. However, consistent with earlier studies, the reliabilities at item level should be enhanced further. In addition, sessions with low overall quality seem to be more difficult to judge. This is potentially a challenge for the wider use of adherence and quality scales in clinical practice, as low-quality sessions are more likely to be used as case material, especially for supervision.

Disclosure statement

No potential conflict of interest was reported by the authors.

REFERENCES

Anderson, T., Crowley, M. E., Patterson, C. L., & Heckman, B. D. (2012). The influence of supervision on manual adherence and therapeutic processes. *Journal of Clinical Psychology, 68*(9), 972–988. doi:10.1002/jclp.21879

 © 2018 The Editors of Nordic Psychology

Barber, J., & Critis-Christoph, P. (1996). Development of a therapist adherence/competence rating scale for supportive-expressive dynamic psychotherapy: A preliminary report. *Psychotherapy Research, 6*(2), 81–94. doi:10.1080/10503309612331331608

Barber, J., Liese, B. S., & Abrams, M. J. (2003). Development of the cognitive therapy adherence and competence scale. *Psychotherapy Research, 13*(2), 205–221. doi:10.1093/ptr/kpg019

Bateman, A., & Fonagy, P. (1999). Effectiveness of partial hospitalization in the treatment of borderline personality disorder: A randomized controlled trial. *American Journal of Psychiatry, 156*(10), 1563–1569. doi:10.1176/ajp.156.10.1563

Bateman, A., & Fonagy, P. (2009). Randomized controlled trial of outpatient mentalization-based treatment versus structured clinical management for borderline personality disorder. *American Journal of Psychiatry, 166*(12), 1355–1364. doi:10.1176/appi.ajp.2009.09040539

Bateman, A., O'Connell, J., Lorenzini, N., Gardner, T., & Fonagy, P. (2016). A randomised controlled trial of mentalization-based treatment versus structured clinical management for patients with comorbid borderline personality disorder and antisocial personality disorder. *BMC Psychiatry, 16*, 304. doi:10.1186/s12888-016-1000-9

Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., & Czajkowski, S. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the nih behavior change consortium. *Health Psychology, 23*(5), 443–451. doi:10.1037/0278-6133.23.5.443

Cicchetti, D. V. (1994). Guidelines, criteria, an rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290.

Das, J., de Ruiter, C., Doreleijers, T., & Hillege, S. (2009). Reliability and construct validity of the dutch psychopathy checklist: Youth version. *Assessment, 16*(1), 88–102.

Folmo, E. J., Karterud, S. W., Bremer, K., Walther, K. L., Kvarstein, E. H., & Pedersen, G. A. F. (2017). The design of the MBT-G adherence and quality scale. *Scandinavian Journal of Psychology, 58*(4), 341–349. doi:10.1111/sjop.12375

Karterud, S., & Bateman, A. (2011). *Mentaliseringsbaseret terapi, manual og vurderingsskala* (Vol. 1). København: Hans Reitzels Forlag.

Karterud, S., Pedersen, G., Engen, M., Johansen, M. S., Johansson, P. N., Schluter, C., & Bateman, A. W. (2013). The MBT adherence and competence scale (MBT-ACS): Development, structure and reliability. *Psychotherapy Research, 23*(6), 705–717. doi:10.1080/10503307.2012.708795

Kvarstein, E. H., Pedersen, G., Urnes, O., Hummelen, B., Wilberg, T., & Karterud, S. (2015). Changing from a traditional psychodynamic treatment programme to mentalization-based treatment for patients with borderline personality disorder – Does it make a difference? *Psychology and Psychotherapy, 88*(1), 71–86. doi:10.1111/papt.12036

Loeb, K. L., Wilson, G. T., Labouvie, E., Pratt, E. M., Hayaki, J., Walsh, B. T., & Agras, W. S. (2005). Therapeutic alliance and treatment adherence in two interventions for bulimia nervosa: A study of process and outcome. *Journal of Consulting and Clinical Psychology, 73*(6), 1097–1107.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46.

Möller, C., Karlgren, L., Sandell, A., Falkenstrom, F., & Philips, B. (2016). Mentalization-based therapy adherence and competence stimulates in-session mentalization in psychotherapy for borderline personality disorder with co-morbid substance dependence. *Psychotherapy Research,* 1–17. doi: 10.1080/10503307.2016.1158433

Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology, 75*(6), 829–841. doi:10.1037/0022-006X.75.6.829

Prowse, P. T. D., Nagel, T., Meadows, G. N., & Enticott, J. C. (2015). Treatment fidelity over the last decade in psychosocial clinical trials outcome studies: A systematic review. *Journal of Psychiatry, 18*, 258.

Robinson, P., Hellier, J., Barrett, B., Barzdaitiene, D., Bateman, A., Bogaardt, A., & Fonagy, P. (2016). The nourished randomised controlled trial comparing mentalisation-based treatment for eating disorders (MBT-ED) with specialist supportive clinical management (SSCM-ED) for patients with eating disorders and symptoms of borderline personality disorder. *Trials, 17*(1), 549. doi:10.1186/s13063-016-1606-8

Schanche, E., Høstmark Nielsen, G., McCullough, L., Valen, J., & Mykletun, A. (2010). Training graduate students as raters in psychotherapy process research. *Nordic Psychology, 62*(3), 4–20. doi:10.1027/1901-2276/a000013

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.

 © 2018 The Editors of Nordic Psychology

Simonsen, S., Heinskou, T., Sørensen, P., Folke, S., & Lau, M. (2017). Personality disorders: Patient characteristics and level of outpatient treatment service. *Nordic Journal of Psychiatry,* 1–7. doi: 10.1080/08039488.2017.1284262

Waltz, J., Addis, M. E., Koerner, K., & Jacobsen, N. S. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology, 61*(4), 620–630.

© 2018 The Editors of Nordic Psychology