Scandinavian Journal of Psychology, 2017, 58, 341–349 DOI: 10.1111/sjop.12375

Personality and Social Psychology

The design of the MBT-G adherence and quality scale

ESPEN J. FOLMO,[1] SIGMUND W. KARTERUD,[2] KJETIL BREMER,[1] KRISTOFFER L. WALTHER,[3] ELFRIDA H. KVARSTEIN[3,4] and GEIR A. F. PEDERSEN[3,5]

[1]*Norwegian National Advisory Unit on Personality Psychiatry, Oslo University Hospital, Oslo, Norway*
[2]*The Norwegian Institute for Mentalizing, Oslo, Norway*
[3]*Department of Personality Psychiatry, Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway*
[4]*Institute of Clinical Medicine, University of Oslo, Norway*
[5]*NORMENT, KG Jebsen Center for Psychosis Research, Institute of Clinical Medicine, University of Oslo, Norway*

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,* 341–349.

Few group psychotherapy studies focus on therapists' interventions, and instruments that can measure group psychotherapy treatment fidelity are scarce. The aim of the present study was to evaluate the reliability of the Mentalization-based Group Therapy Adherence and Quality Scale (MBT-G-AQS), which is a 19-item scale developed to measure adherence and quality in mentalization-based group therapy (MBT-G). Eight MBT groups and eight psychodynamic groups (a total of 16 videotaped therapy sessions) were rated independently by five raters. All groups were long-term, outpatient psychotherapy groups with 1.5 hours weekly sessions. Data were analysed by a Generalizability Study (G-study and D-study). The generalizability models included analyses of reliability for different numbers of raters. The global (overall) ratings for adherence and quality showed high to excellent reliability for all numbers of raters (the reliability by use of five raters was 0.97 for adherence and 0.96 for quality). The mean reliability for all 19 items for a single rater was 0.57 (item range 0.26–0.86) for adherence, and 0.62 (item range 0.26–0.83) for quality. The reliability for two raters obtained mean absolute G-coefficients on 0.71 (item range 0.41–0.92 for the different items) for adherence and 0.76 (item range 0.42–0.91) for quality. With all five raters the mean absolute G-coefficient for adherence was 0.86 (item range 0.63–0.97) and 0.88 for quality (item range 0.64–0.96). The study demonstrates high reliability of ratings of MBT-G-AQS. In models differentiating between different numbers of raters, reliability was particularly high when including several raters, but was also acceptable for two raters. For practical purposes, the MBT-G-AQS can be used for training, supervision and psychotherapy research.

Key words: Mentalization, Mentalization-Based Treatment, group therapy, Borderline Personality Disorder, reliability, generalizability theory.

Espen J. Folmo, Department for Personality Psychiatry, Oslo University Hospital, Ullevaal, PO Box 4956 Nydalen, 0424 Oslo, Norway.
E-mail: espfol@ous-hf.no.

INTRODUCTION

Over the past decades a number of evidence-based treatment approaches for Borderline Personality Disorder (BPD) have been developed (Stoffers, Völlm, Rücker, Timmer, Huband & Lieb, 2012). One of these, mentalization-based treatment (MBT), has been found efficient in several randomised controlled trials (Bateman & Fonagy, 2001; 2009; Rossouw & Fonagy, 2012), and favorable results have been replicated in naturalistic comparisons outside the United Kingdom (Bales, Timman, Andrea, Busschbach, Verheul & Kamphuis, 2015; Kvarstein, Pedersen, Urnes, Hummelen, Wilberg & Karterud, 2015).

MBT is an intensive, combined treatment approach that includes both individual and group therapy. The four structural pillars integrated within MBT are: (1) psychoeducation; (2) an individual dynamic MBT case formulation; (3) individual mentalization-based psychotherapy (MBT-I); and (4) mentalization-based group therapy (MBT-G; Karterud, 2015). MBT thus requires a collaborative team of therapists, and the importance of regular video-based therapy supervision for MBT teams is clearly emphasized (Bateman & Fonagy, 2016).

An adherence and competence scale for MBT-I (MBT-I-ACS) has previously been developed based on a Norwegian version of the MBT manual (MBT-I; Karterud & Bateman, 2010) and the reliability of the scale was found highly satisfactory (Karterud, Pedersen, Engen *et al*., 2013). The MBT-I-ACS has provided the possibility for documentation of model fidelity in studies of treatment outcomes (Kvarstein *et al*., 2015), and has also recently been used in a study relating outcomes to therapists' MBT interventions (Möller, Karlgren, Sandell, Falkenström & Philips, 2016).

Measures for treatment integrity are crucial when investigating whether the alleged "potion" is what is actually being delivered (Perepletchikova, Treat & Kazdin, 2007). Treatment integrity consists of two elements: (1) treatment adherence, i.e., "the extent to which a therapist used interventions and approaches prescribed by the treatment manual and avoided the use of interventions and procedures proscribed by the manual" (Waltz, Addis, Koerner & Jacobson, 1993, p. 620); and (2) the therapist's competence (quality), i.e., "the level of skill shown by the therapist in delivering the treatment" (Waltz *et al*., 1993, p. 620). By skill, we refer to the extent in which the therapist conducting the interventions took the relevant aspects of the therapeutic context into account and responded to these contextual variables appropriately. According to this definition, competence presupposes adherence, but adherence does not necessarily imply competence (McGlinchey & Dobson, 2003). The strong element of improvisation within dynamic psychotherapy implies that a certain competence is necessary to adhere to the ethos of the treatment. Nevertheless, such adherence can be performed with varying degrees of sophistication (timing, in-depth exploration,

integration, attunement, etc.). For the above reasons, we prefer the label 'quality' instead of competence.

Recently, both practical guidelines and manuals have been developed specifically for MBT-G (Bateman & Fonagy, 2016; Karterud, 2012, 2015). The MBT-G manual (Karterud, 2015) contains a 19-item adherence and quality scale for MBT-G (MBT-G-AQS; see Appendix).

There is a paucity of research on therapists' adherence and competence in group therapy. A review of the status of group therapy research by Burlingame, MacKenzie and Strauss (2004) issued a call for the development of group therapist intervention measures as a next step in the group treatment literature.

Documentation of treatment integrity requires manualized treatments and is essential when claiming effectiveness of specific psychotherapies (Perepletchikova *et al*., 2007). Wampold and Imel (2015, p. 233) highlight this by stating "It is now virtually required that clinical trials of psychotherapy assess and report adherence and competence." A main challenge, present in all dynamic group therapies, is the dialectical balance between "structuring" (e.g., item 2 "Regulating group phases"; see Appendix) interventions, explorations of current mental events and overall attunement to the dynamic process (Yalom & Leszcz, 2005). The MBT-G-AQS addresses this concern through nine group-specific items and 10 further items essentially common to MBT-I-ACS.

The primary aim of the present study was to investigate the reliability of the newly developed adherence and quality scale for MBT-G. Our research questions were: (1) Can trained MBT-G raters obtain adequate interrater reliability on (a) the full MBT-G-AQS, particularly the overall ratings, and (b) adherence and quality of the nine group-specific items within MBT-G-AQS? (2) What is the minimum number of MBT-G-AQS raters required to achieve adequate reliability?

MATERIAL AND METHODS

The study is based on video-taped recordings from regular treatment groups from the same clinical unit, Department for Personality Psychiatry (DPP), Oslo University Hospital. To maximize variance, groups belonging to different time periods (2006 and 2015) were chosen. All 16 session were rated with the MBT-G-AQS.

The group therapies and group members

In the first period (2006) DPP offered a psychodynamic, group-based treatment program. In the second period (2015) MBT was the principal treatment mode. The psychodynamic group therapy (PDG) was unmanualized, followed modified group analytic principles, and was influenced by object relations theory and self-psychology (Arnevik, Wilberg, Urnes, Johansen, Monsen & Karterud, 2009). The MBT followed manual requirements as previously described (Kvarstein *et al*., 2015).

All groups were conducted by two therapists and all group sessions lasted 1.5 hours. All groups were slow open, admitting new members whenever a place was vacant. Hence, the video material (both PDG and MBT) demonstrated patients who had attended groups for various lengths of time (range 2–36 months). Both programs combined individual and group therapy (Arnevik *et al*., 2009; Kvarstein *et al*., 2015).

Overall, approximately 85% of the group participants were female, age 20–30 years. The MBT groups primarily recruited BPD patients, while the PDG groups included a broader range of personality disorders (Arnevik *et al*., 2009; Kvarstein, 2015).

Group therapists

Fourteen group therapists from the same treatment unit (57% females) participated in the study. To minimize variance due to therapists' general competence we included two therapists who performed both PDG and MBT-G. Twelve were experienced clinicians and qualified group analysts. By profession there were five psychiatrists, one psychiatric resident, two clinical psychologists, one social worker, one psychology student, one physiotherapist and three psychiatric nurses. In 2015, all therapists, except the psychiatric resident, had also received MBT training.

Scale for MBT-G

The MBT-G-AQS is a 19-item scale developed for measuring therapist adherence and quality in MBT-G. See Table 1 and Appendix for the 19 items. The manual (Karterud, 2015) contains detailed description of the development of the scale.

Video-taped group sessions

The study includes a total of 16 video-taped group therapy sessions. Eight video-tapes show PDG group sessions from 2006 and eight show MBT-G sessions from 2015. Recordings were selected by convenience sampling, i.e., aiming to minimize the variance of general therapist competence in the two time periods, 2006 and 2015. Therapist pairs in MBT and PDG were matched with respect to formal level of education. This resulted in four groups being chosen from the 2006 material. Two consecutive sessions were then selected randomly within the specified 2006 group.

The total video material from 2006 included approximately 80 sessions for each of the four PDG groups. From this pool, two consecutive sessions with the same therapist pair were randomly selected for each PDG group. In 2015, therapist-pairs from four MBT groups provided videotaped recordings of two consecutive group sessions. Two consecutive sessions were preferred in order to minimize therapists' variance over time.

MBT-G-AQS raters

Five independent clinical research collaborators rated the available video material by MBT- G- AQS (no raters were among the rated therapists). These five raters were all trained MBT therapists and familiar with MBT-I rating procedures. Prior to the current study, four of the five raters had assessed at least 30 (range 30–91) sessions with the MBT-I-ACS as part of their work for the Norwegian MBT Quality Lab. Eight hours theoretical and practical training in the MBT-G-AQS preceded the current reliability study. The pre-assessment training included rating and discussion of two verbatim transcripts of MBT groups. Four of the raters were psychologists, and one a psychiatrist (author of the MBT manual).

Table 1. *Item descriptives, G-study results, and D-study results for different measurement designs*

	Adherence / frequency								Quality							
	Grand mean[a]		G-study		D-study				Grand mean[a]		G-study		D-study			
			Coefficients		Two raters		One rater				Coefficients		Two raters		One rater	
Item	Mean	SD	Rel.[b]	Abs.[c]	Rel.[b]	Abs.[c]	Rel.[b]	Abs.[c]	Mean	SD	Rel.[b]	Abs.[c]	Rel.[b]	Abs.[c]	Rel.[b]	Abs.[c]
1. Boundaries	6.18	3.34	0.89	0.86	0.77	0.71	0.63	0.54	4.13	1.00	0.90	0.89	0.79	0.76	0.65	0.61
2. Phases	3.98	3.60	0.93	0.90	0.85	0.78	0.73	0.64	3.34	1.85	0.96	0.96	0.90	0.90	0.82	0.82
3. Turntaking	5.48	4.61	0.95	0.95	0.88	0.87	0.78	0.78	3.35	1.86	0.95	0.95	0.89	0.88	0.80	0.79
4. External events	5.38	3.44	0.94	0.92	0.86	0.83	0.75	0.71	3.38	1.44	0.93	0.92	0.85	0.83	0.74	0.71
5. Events in group	3.31	2.78	0.85	0.85	0.69	0.69	0.53	0.53	3.09	1.42	0.90	0.87	0.78	0.73	0.64	0.57
6. Care for group	Not rated								4.36	1.23	0.91	0.91	0.80	0.80	0.67	0.66
7. Authority	Not rated								4.33	1.42	0.93	0.93	0.85	0.84	0.74	0.73
8. Group norms	2.36	2.64	0.83	0.83	0.67	0.67	0.50	0.50	2.81	1.88	0.78	0.78	0.59	0.58	0.42	0.41
9. Cooperation	1.53	1.93	0.86	0.84	0.70	0.68	0.54	0.51	2.10	1.71	0.96	0.95	0.91	0.89	0.84	0.80
10. Warmth	Not rated								4.50	1.15	0.92	0.91	0.81	0.81	0.68	0.68
11. Exploration	16.03	6.79	0.87	0.80	0.73	0.61	0.58	0.44	3.98	1.32	0.86	0.86	0.71	0.71	0.56	0.55
12. Unwarranted beliefs	2.48	2.64	0.88	0.88	0.75	0.74	0.60	0.59	3.00	1.35	0.84	0.82	0.68	0.64	0.51	0.47
13. Emotional arousal	Not rated								3.68	1.29	0.93	0.93	0.84	0.84	0.72	0.72
14. Acknowledging	1.64	1.84	0.77	0.77	0.58	0.57	0.41	0.40	2.69	1.56	0.86	0.85	0.70	0.69	0.54	0.53
15. Pretend mode	Not rated								2.23	1.58	0.67	0.64	0.44	0.42	0.28	0.26
16. Psychic equivalence	1.35	1.70	0.63	0.63	0.41	0.41	0.26	0.26	2.64	1.77	0.86	0.86	0.71	0.70	0.55	0.54
17. Affect focus	14.85	7.03	0.91	0.85	0.80	0.70	0.66	0.54	4.28	1.71	0.93	0.93	0.84	0.84	0.72	0.72
18. Stop and rewind	0.65	1.08	0.88	0.88	0.74	0.74	0.59	0.59	1.80	1.81	0.85	0.84	0.69	0.67	0.53	0.50
19. Relationship	5.00	5.29	0.94	0.93	0.86	0.84	0.75	0.73	3.30	1.63	0.88	0.88	0.75	0.74	0.60	0.59
Overall rating	3.76	1.76	0.97	0.97	0.92	0.92	0.86	0.86	3.80	1.67	0.96	0.96	0.91	0.91	0.83	0.83

Notes: [a]Grand mean and standard deviations of scores across therapists and sessions. [b]Generalizability coefficient (For relative decisions). [c]Dependability coefficient (For absolute decisions).

MBT-G-AQS rating procedures

The five raters rated all MBT-G-AQS items for all 16 sessions. Ratings were performed independently, but in the same room. After having fulfilled their ratings of each session and delivered their scoring sheets to the project coordinator, the raters met and discussed agreements and disagreements, a procedure also described in other research studies (Gutermann, Schreiber, Matulis, Stangier, Rosner & Steil, 2015; von Consbruch, Clark& Stangier, 2012; Weck, Weigel, Richtberg & Stangier, 2010). Ratings were not changed after this comparison. Ratings were not blind: the raters knew most of the therapists, and were therefore not blind to treatment modality.

Ratings of adherence and competence

A therapist intervention may receive an MBT-G-AQS rating or not. A single intervention may receive more than one rating. Non-MBT interventions may sound like: "When does school start this autumn?" or "I believe the group is paralyzed for the moment" or "when did he tell you that?" Adherence on the item level is assessed by counting the frequency. Five of the items ("care for the group and its members," "managing authority," "engagement, interest and warmth," "regulating emotional arousal" and "handling pretend mode") are not assessed for adherence/ frequency, as these interventions can be performed by indirect means. However, they are rated for quality. The adherence ratings equal the total number of counted interventions.

For the assessment of quality, all items are rated on a 1–7 Likert scale. The manual contains rating procedures as well as descriptions of what counts as low versus high quality. All items are displayed in the Appendix and described by their competence level of 4 ("good enough"). If the therapists fail to deliver clearly indicated interventions, the item can be rated low on quality (e.g., 2) even where there are no occurrences. Finally, the rater decides on the overall quality score, based on a global understanding of the session.

Data analysis

In the current research design two therapy sessions from each of eight therapist-couples were videotaped. This makes a total of 16 therapy sessions, and all five raters rated all 16 sessions. In the framework of G-theory (Shavelson & Webb, 1991), this implies a two facet partially nested "(s:t) x r" design, where sessions (s) are nested within therapists (t), and raters (r) are crossed over sessions within therapists. The design is partially nested because the effect of session (s) is both nested (within t) and crossed (over r). With respect to generalizations beyond this particular study, therapists, sessions and raters are considered randomly selected from the whole 'universe' of admissible therapists, sessions and raters. The object of measurement is therapist behavior, and the measurement design is balanced as all therapists are rated by the same number of raters. The two facets of observation give two differentiation variance components, the individual variance between therapists (t) and the systematic variance between sessions for each therapist (st). This makes three sources of instrumentation variance (error) that directly effects the reliability of the observed scores. These are; (1) the rater effect (r) indicating the consistency of how much 'behavior' the

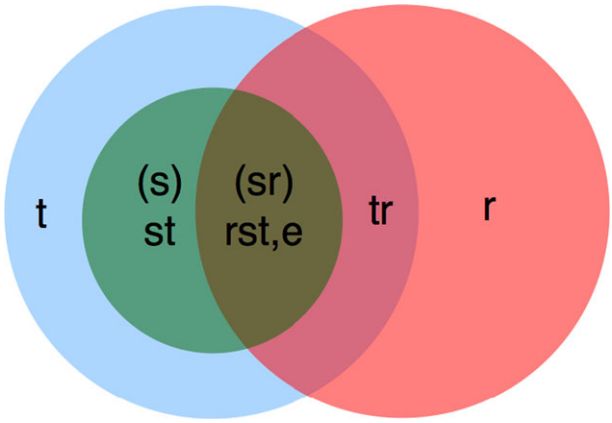

Fig. 1. Venn diagram of the variance components in the (s:t) x r design. The components are: The individual variance between therapists (t), the systematic variance between sessions for each therapist (st), the unique rater–therapist–session interaction plus other unknown error variance (rst, e), the interaction between raters and therapists (tr), and the rater effect (r). [Colour figure can be viewed at wileyonlinelibrary.com]

raters see, averaged over therapists and sessions; (2) the interaction between raters and therapists (tr), indicating the raters' different rank ordering of the therapists; and lastly (3) the unique rater–therapist–session interaction plus other unknown error variance (rst, + e) (see Fig. 1). Within this design, sessions (s) cannot be separated from therapist (t) and neither can the session–rater interaction (sr) be separated from the rater–session–therapist interaction.

Based on the sample data, the relative impact of different sources of variation is estimated by a G-study (Shavelson, Webb & Rowley, 1989), from which generalizability coefficients are computed. The G-coefficient (ρ^2) indexes the proportion of total variability in scores that is due to "universe scores" ($\rho^2 = \frac{\sigma^2(\tau)}{\sigma^2(\sigma^2(\tau)+\sigma^2(\delta)\tau)}$), where $\sigma^2(\tau)$ is the variance of the true score, and $\sigma^2(\delta)$ is the variance of the various error components. A low G-coefficient is due to a significant amount of error in measurement or to minimal variation across individuals, the measurement procedure, and the universe of generalization (Hagtvet, 1997). A G-coefficient below 0.4, is "Poor"; when it is between 0.4 and 0.59 it is "Fair"; between 0.6 and 0.74, is "Good," while a value above 0.75 is considered "Excellent" (Cicchetti, 1994).

Based on the obtained G-study components, the generalizability framework offers a subsequent study called D-study, or optimization study. Through the D-study it is possible to estimate how many conditions of each facet is necessary to obtain adequate generalizability, that is, how many raters are needed.

The intended use of the MBT-G-AQS concerns decisions of whether subjects are below or above some specific level of adherence or quality. Consequently, the most relevant reliability estimate is *absolute decisions* (i.e., absolute G-coefficients; see Karterud *et al.*, 2013 for a detailed discussion). The current G- and D-studies have been processed through the EduG program (Cardinet, Johnson & Pini, 2010; Swiss Society for Research in Education Working Group, 2010).

Ethics

After patients received a description of the study, they provided written, informed consent, as did the involved therapists. The PDG recordings were part of the UPP project, and were approved by the Data Inspectorate and the Regional Ethics Committee in Norway. The privacy ombudsman at Oslo University Hospital approved the MBT-G part of the study.

RESULTS

Table 1 presents the reliability for all five raters and estimated D-study coefficients for two and one raters. The mean reliability on item level for adherence was 0.86 (range 0.63–0.97) and 0.88 for quality (range 0.64–0.96) for five raters. For two raters it was 0.71 (range 0.41–0.92) for adherence and 0.76 (range 0.42–0.91) for quality, which, with some exceptions, are in the acceptable to high range. The mean reliability for one rater was 0.57 (item range 0.26–0.86) for adherence, and 0.62 (item range 0.26–0.83) for quality, which ranges from poor to acceptable estimates.

The reliability for overall ratings of adherence (0.97) and quality (0.96) were both excellent. The overall ratings for adherence and quality showed high to excellent reliability for all numbers of raters. The overall ratings are also the most "immune" items for a decreasing number of raters (see Table 2). Deleting the least reliable rater from the overall ratings would only slightly increase the reliability for these two items (+0.01). This signals that the overall rating was robust also when the number of raters decreased, and that the raters agreed strongly on the overall evaluation of a MBT-G session. Table 2 demonstrates which items are most affected by a decreasing or increasing number of raters.

There were only minor differences between the reliability coefficients for absolute and relative decisions (relative and absolute G-coefficients); i.e., raters agreed as much on exact scores as on the ranking of the interventions/sessions. Therefore, all results presented are based on the absolute G-coefficients.

The nine group specific items (item 1–9) displayed very high reliability for both adherence and quality. The four items least

Table 2. *Quality rating (G-coefficient) sorted by increasing difference (5R-1R) between five raters (5R) and one rater (1R)*

Item name	5R	1R	Difference (5R-1R)
Overall rating	0.96	0.83	0.13
02. Phases	0.96	0.82	0.14
09. Cooperation	0.95	0.8	0.15
03. Turntaking	0.95	0.79	0.16
07. Authority	0.93	0.73	0.2
04. External events	0.92	0.71	0.21
13. Emotional arousal	0.93	0.72	0.21
17. Affect focus	0.93	0.72	0.21
10. Warmth	0.91	0.68	0.23
06. Care for group	0.91	0.66	0.25
01. Boundaries	0.89	0.61	0.28
19. Relationship	0.88	0.59	0.29
05. Events in group	0.87	0.57	0.3
11. Exploration	0.86	0.55	0.31
14. Acknowledging	0.85	0.53	0.32
16. Psychic equivalence	0.86	0.54	0.32
18. Stop and rewind	0.84	0.5	0.34
12. Unwarranted beliefs	0.82	0.47	0.35
08. Group norms	0.78	0.41	0.37
15. Pretend mode	0.64	0.26	0.38

affected by the number of raters decreasing, and with the highest reliability on quality, were also group specific items: "Regulating group phases," "Cooperation with cotherapist," "Initiating and fulfilling turntaking" and "Managing authority." The three group specific items "Regulating group phases," "Engaging group members in mentalizing external events" and "Initiating and fulfilling turntaking" showed very high reliability for adherence (> 0.9). "Initiating and fulfilling turntaking" was also the only item where all five raters displayed a reliability above 0.9 on adherence.

For some items the reliability would increase slightly if one of the raters was omitted in the study. These findings indicate that some of the "disagreement" on specific items was due to one rater having a different view than the others. However, there was no indication of any systemic impact on the reliability for specific raters, i.e., different raters struggled with different items.

Table 1 reveals that items 16, 14, 11 and 8 proved difficult to rate for adherence (lowest reliability). We also observe that the quality ratings for items 15, 8, 12, 18, 14, 11 and 16 were more challenging than the other items to agree on. These items had lower reliability and were also more affected by a decreasing number of raters. However, the reliability of item 16, 15, 14 and 18 is very good considering their low variance.

The two items that displayed the lowest reliability across all number of raters were "Psychic equivalence" and "Pretend mode." "Psychic equivalence" had the lowest reliability for adherence, and "Pretend mode" had the lowest reliability for quality.

From a psychometric perspective, it is ideal with some variation between therapists (T), and within therapists from session to session (T:S). Further, it is favorable that the residual variance (RS:T), raters' ranking variation (TR), and disagreement between raters (R) is as low as possible. Overall rating for quality may serve as an example of a favorable result. The residual variance for the overall quality score was very low (17%). There was complete agreement (0% variance) among the raters on how much of the intervention was observed, and the ranking of therapists. Therapists varied a lot with respect to overall quality (62% variance), but less so from session to session (21%).

From Table 1 we see that item 16 "Handling psychic equivalence" had a high residual variance (40%). There was little systematic variance between therapists regarding the intervention (11%), and from session to session (15%). There was substantial variance in the raters' ranking order (34%), but no variance in how much of the behavior (the specific intervention) the raters observed.

Item 11 ("Exploration, curiosity and not-knowing stance") had a reliability coefficient of 0.80, which is high, but low compared to the rest, especially considering high variance and frequency (mean frequency = 16). Table 3 disentangles why this particular item proved difficult to rate. Item 11 had a moderate residual variance (29% variance), which implies that the item is relatively well defined. However, there was considerable disagreement among raters on how much of this intervention they observed (24% variance), although they did not deviate much in their ranking order of the therapists (3% variance). Different opinions on what counts as item 11 interventions may have large consequences for reliability if therapist variation (between therapists and between sessions) is low. In this case, all therapists used this item frequently, as variance between therapists was very low (7%), but they varied much from session to session (37%).

Table 3 displays a relation between low reliability and residual variance. The seven items with lowest reliability had a mean residual variance of 40.5, while the seven items with highest reliability had a mean residual variance of 27. The quality ratings displayed a similar, but slightly stronger, pattern. The reason for this connection is that high residual variance signals weak references for the raters as to how to rate these items. When the residual variance for an item is high, it may indicate that therapists do not know when and how to apply it, e.g., due to poor operationalization. Hence, the item is difficult to recognize for raters.

As half of the sessions were psychodynamic groups, half of the rated therapists were not trained in the items assessed, that is, they intervened in more unfocused ways. This may explain some of the residual variance for several items: The seven items with a quality rating below 3 had a mean residual variance on adherence of 44%, while the seven items with a quality rating above 3 had a

Table 3. *Sources of variation for adherence ratings for five raters (5R): percentages of total variation. Items sorted from low to high reliability (G-coefficients; "Abs G")*

	T: between therapist variation	R: variation in how much raters observe	S:T: therapist variation across sessions	TR: variation in raters ranking of therapists	RS:T: residual (including error) variance	Abs G
16. Psychic equivalence	11.3	0	14.5	34	40.2	0.63
14. Acknowledging	23.7	2.4	16	5.2	52.6	0.77
11. Exploration	6.9	24.2	36.9	2.7	29.3	0.8
8. Group norms	7.8	0	42.1	2.6	47.6	0.83
9. Cooperation	18.8	5.2	32.7	0	43.3	0.84
17. Affect focus	20.2	19.1	33.5	1.1	26.1	0.85
5. Events in the group	16	0	37	2.5	44.5	0.85
1. Boundaries	33.1	13.5	21.4	0	32.1	0.86
18. Stop and rewind	14.7	0.6	44.2	2.5	38	0.88
12. Unwarranted beliefs	25.5	2.3	33.4	0	38.8	0.88
2. Phases	54.2	13	9.6	1.4	21.8	0.9
4. External events	50.6	5.5	20.5	5.4	17.9	0.92
19. Relationship	0	2.4	73	0	24.5	0.93
3. Turntaking	65.8	0.4	11.8	5.9	16.1	0.95

mean residual variance on adherence of 24%. The same pattern was found in the quality ratings (40/25).

DISCUSSION

This is the first study to report psychometric properties for the MBT-G-AQS. It is also the first study of a scale for measuring therapists' interventions in group therapy since 2005. The results demonstrate that the MBT-G adherence and quality scale is a reliable instrument. This scale can be applied to document treatment integrity, and underpin the evidence-base for MBT.

The overall/global ratings for adherence and quality showed high to excellent reliability across all numbers of raters. The instrument can thus be used with only one rater for research purposes where the question of overall treatment fidelity needs to be documented, and where a detailed focus on the other items are of subordinate interest. This finding also supports that the MBT-G-AQS can be reliably applied to determine if a session qualifies as "good enough" MBT-G.

At item level, the reliabilities varied substantially. This is a common finding among rating scales (Barber, Liese & Abrams, 2003). With one rater some items had a satisfactory reliability, while others had low to very low reliability. With two raters, reliabilities ranged from fair to excellent.

As process studies based on a large number of raters are very expensive and difficult to achieve (Perepletchikova, Hilt, Chereji & Kazdin, 2009), acceptable reliability for the entire scale with just one rater is important for practical implementation of the scale. Due to more extensive training, calibration, and experience of the raters in the current study, it was expected to reveal higher reliability, particularly with one and two raters, than what was obtained in the MBT-I-ACS study by Karterud *et al.* (2013). Current results confirmed this expectation. However, the reliability for one rater was still below acceptable range for several of the items. This indicates a need for further calibration and training as well as more explicit definitions of the phenomena to be assessed.

One of the benefits of performing a G-study is that it allows for identifying items that individual raters view differently than others. The finding that different raters struggled with different items means that it is important for raters to calibrate (discuss) their ratings on a regular basis.

This is particularly true for more complex (abstract) items that display low frequency, which means that raters receive less practical training in rating them. For example, "Acknowledging good mentalizing" (item 14) and "Handling psychic equivalence" (item 16) both had low frequencies, and also proved more difficult to rate for adherence than other items (with low frequency). The results indicate a pattern that more "concrete items" (clearly defined and less abstract) such as "Cooperation with co-therapist" (item 9) and "Stop and rewind" (item 18; which also had low occurrence), had high reliabilities despite low frequencies. Item 9 can serve as example of an intervention easy to pinpoint, for example, if the rater notices some open communication between the therapists, this counts as an intervention. Items 14 and 16, unlike items 9 and 18, were more difficult to evaluate for quality as well as adherence.

Other items that were difficult to rate for both adherence and quality, and thus deserve careful attention, were "Stimulating discussions on group norms" (item 8) and "Exploration, curiosity and not-knowing stance" (item 11). Item 11 was used frequently, but it covers a wide range of interventions. The most central aspect of item 11 is to determine whether an open and curious question addresses mental states or not. For example, the intervention "When did he tell you that?" is not aimed at a mental state per se, but depending on the context, some raters may decide to count this as adherence to item 11 – for example if the question makes the patient rethink what really happened, and whether s/he wrongly perceived another person's mental state due to the timing of an utterance. It is difficult to define a clear cut-off without losing some of the flexibility crucial for attuned responsiveness.

We know from previous ratings of non-MBT psychotherapy sessions that non-MBT therapists might display high adherence on items such as "Exploration, curiosity and not-knowing stance" (item 11) and "Affect focus" (item 17). However, the way these therapists intervene is most often different from an MBT approach. They often receive a low quality rating, and raters might be bewildered by boundary occurrences (interventions that border on what might be called MBT). For item 11, the eight PDG sessions had a mean adherence rating on 14 (number of observed interventions), and three for quality. The eight MBT-G sessions had a mean adherence rating of 5, and a mean quality score of 5. The high frequency of low quality item 11-interventions in the rated PDG sessions may account for some of the observed difficulty in rating this item. Still, the manual should be more specific with respect to what counts as adherence and high versus low quality for this item.

From a psychometric perspective, items with low occurrence (e.g., items 9, 14, 16 and 18) may be seen as redundant. However, as underlined in the manual, these items are essential ingredients in a larger treatment "potion:" "The unique aspect of MBT lies less in each individual item per se, than in the overall 'package' of item design and context" (Karterud & Bateman, 2010, p. 26). The robust reliability of the overall ratings indicates that raters manage to capture (agree on) the overall flavor of MBT, even if they disagree on certain items.

Two items that proved difficult to rate were adherence for "Handling psychic equivalence" (item 16), and quality for "Handling pretend mode" (item 15). These two items are both central to the overall theory of mentalization and MBT. For item 16, the 8 PDG sessions had a mean adherence rating of 0, and 2 for quality. The 8 MBT-G sessions had a mean adherence rating of 1, and a mean quality score of 3. Item 15 is not rated for adherence, but both the PDG and MBT-G sessions had a mean quality rating of 2. Both items displayed low variance, high residual variance, and low reliability. In this case, it is unclear whether the group therapists delivered interventions for item 15 and 16 which were poor and/or unclear, or if the concepts of pretend mode and psychic equivalence were somewhat unclear for both therapists and raters. However, taking the small variance into account, the reliability is rather good for these items. Items 15 and 16 should be object for more research, and the manual made more "concrete" for both items.

Limitations

The generalizability of our findings is restricted by several limitations. Firstly, as mentioned above, the raters were not blind

to treatment modality (PDG or MBT-G), and this could have influenced the reliability. However, there were only minor differences between the two modalities and the combined reliability. In the current study, two therapists were rated four times (both in PDG and MBT-G). We cannot exclude the possibility that repeated ratings of the same therapists may have artificially increased inter-rater conformity. Thus, future studies should apply these scales to larger samples of both patients and therapists.

Utility

The MBT-G-AQS may contribute to future psychotherapy research by assuring internal validity and contribute to research on adherence and quality as possible moderators and mediators of treatment outcome. The scale can additionally be used for training and clinical purposes: assessing and providing feedback about therapeutic quality and adherence enables therapists and supervisors to stay on course.

CONCLUSION

The current results demonstrate that the MBT-G adherence and quality scale is a reliable instrument for rating adherence to and quality of mentalization-based group therapy with as few as two raters for the entire scale, and with one rater for overall/global assessment of MBT-G. Some items, especially "Handling pretend mode" and "Handling psychic equivalence" need more empirical attention, as our results indicate these items to be inadequately defined and understood. The scale can be applied for quality assurance, training, and supervision.

The research was conducted at The Department of Personality Psychiatry, Oslo University Hospital, Ullevål, Norway. Conflict of interest: Sigmund Karterud is administrative and professional director of the Norwegian Institute for Mentalizing and author of several books on mentalization-based treatment, including a manual of mentalization-based group therapy. Thanks to psychologist Christian Schlüter at the Norwegian National Advisory Unit on Personality Psychiatry for being part of the rater team in this study.

REFERENCES

Arnevik, E., Wilberg, T., Urnes, O., Johansen, M., Monsen, J. & Karterud, S. (2009). Psychotherapy for personality disorders: Short term day hospital psychotherapy versus outpatient individual therapy – A randomized controlled study. *European Psychiatry*, *24*, 71–78.

Bales, D. L., Timman, R., Andrea, H., Busschbach, J. J., Verheul, R. & Kamphuis, J. H. (2015). Effectiveness of day hospital mentalization-based Ttreatment for patients with severe borderline personality disorder: A matched control study. *Clinical Psychology & Psychotherapy*, *22*, 409–417.

Barber, J. P., Liese, B. S. & Abrams, M. J. (2003). Development of the Cognitive Therapy Adherence and Competence Scale. *Psychotherapy Research*, *13*, 205–221.

Bateman, A. & Fonagy, P. (2001). Treatment of borderline personality disorder with psychoanalytically oriented partial hospitalization: An 18-month follow-up. *American Journal of Psychiatry*, *158*, 36–42.

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*, 1355–1364.

Bateman, A. & Fonagy, P. (2016). *Mentalization based treatment for personality disorders: A practical guide*. Oxford: Oxford University Press.

Burlingame, G. M., MacKenzie, K. R. & Strauss, B. (2004). Small group treatment: Evidence for effectiveness and mechanisms of change. *Handbook of psychotherapy and behavior change*, *5*, 647–696.

Cardinet, J., Johnson, S. & Pini, G. (2010). *Applying generalizability theory using EduG*. New York: Routledge.

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

Gutermann, J., Schreiber, F., Matulis, S., Stangier, U., Rosner, R. & Steil, R. (2015). Therapeutic adherence and quality scales for Developmentally Adapted Cognitive Processing Therapy for adolescents with PTSD. *European Journal of Psychotraumatology*, *6*, 26632.

Hagtvet, K. A. (1997). The function of indicators and errors in construct measures: An application of generalizability theory. *Journal of Vocational Education Research*, *22*, 247–266.

Karterud, S. (2012). *Manual for mentaliseringsbasert gruppeterapi (MBT-G)*. Oslo: Gyldendal.

Karterud, S. (2015). *Mentalization-based group therapy (MBT-G): A theoretical, clinical and research manual*. London: Oxford University Press.

Karterud, S. & Bateman, A. (2010). *Mentaliseringsbaseret terapi. Manual og vurderingsskala*. Oslo: Gyldendal.

Karterud, S., Pedersen, G., Engen, M., Johansen, M. S., Johansson, P. N., Schlüter, C. *Et al*. (2013). The MBT adherence and quality scale (MBT-ACS): Development, structure and reliability. *Psychotherapy Research*, *23*, 705–717.

Kvarstein, E. H., Pedersen, G., Urnes, Ø., 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: Theory, Research and Practice*, *88*, 71–86.

McGlinchey, J. & Dobson, K. S. (2003). Treatment fidelity assessment in cognitive behavioral therapy. *Journal of Cognitive Psychotherapy: An International Quarterly*, *17*, 299–318.

Möller, C., Karlgren, L., Sandell, A., Falkenström, 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*, https://doi.org/10.1080/10503307.2016.1158433.

Perepletchikova, F., Hilt, L. M., Chereji, E. & Kazdin, A. E. (2009). Barriers to implementing treatment integrity procedures: Survey of treatment outcome researchers. *Journal of Consulting and Clinical Psychology*, *77*, 212–218.

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*, 829–841.

Rossouw, T. I. & Fonagy, P. (2012). Mentalization-based treatment for self-harm in adolescents: Arandomized controlled trial. *Journal of the American Academy of Child and Adolescent Psychiatry*, *51*, 1304–1313.

Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory. A primer*. Newbury Park, CA: Sage.

Shavelson, R. J., Webb, N. M. & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*, 922–932.

Stoffers, J. M., Völlm, B. A., Rücker, G., Timmer, A., Huband, N. & Lieb, K. (2012). Psychological therapies for people with borderline personality disorder. *Cochrane Database of Systematic Reviews*, *8*, https://doi.org/10.1002/14651858.CD005652.pub2.

Swiss Society for Research in Education Working Group (2010). *EDUG user guide*. Neuchatel: IRDP.

von Consbruch, K., Clark, D. M. & Stangier, U. (2012). Assessing therapeutic competence in cognitive therapy for social phobia: Psychometric properties of the Cognitive Therapy Competence Scale

for Social Phobia (CTCS-SP). *Behavioral and cognitive psychotherapy*, *40*, 149–161.

Wampold, B. E. & Imel, Z. E. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work*. New York, NY: Routledge.

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

Weck, F., Weigel, M., Richtberg, S. & Stangier, U. (2011). Reliability of adherence and competence assessment in psychoeducational treatment: Influence of clinical experience. *The Journal of Nervous and Mental Disease*, *199*, 983–986.

Yalom, I. D. & Leszcz, M. (2005). *Theory and practice of group psychotherapy*. New York: Basic books.

Received 6 November 2016, accepted 9 May 2017

APPENDIX

Rating scale for Mentalization-based Group Therapy

Rater ___ Rating date _____ Therapists _____ Group _____ Session date _____
Overall rating of MBT adherence _____ MBT quality _____
Running notes:

Item name	Adherence	Quality
1. Managing group boundaries		
2. Regulating group phases		
3. Initiating and fulfilling turntaking		
4. Engaging group members in mentalizing external events		
5. Identifying and mentalizing events in the group		
6. Care for the group and its members	No rating	
7. Managing authority	No rating	
8. Stimulating discussions on group norms		
9. Cooperation with co-therapist		
10. Engagement, interest and warmth	No rating	
11. Exploration, curiosity and not-knowing stance		
12. Challenging unwarranted beliefs		
13. Regulating emotional arousal	No rating	
14. Acknowleding good mentalizing		
15. Handling pretend mode	No rating	
16. Handling psychic equivalence		
17. Affect focus		
18. Stop and rewind		
19. Focus on the therapist – patient relationship		

Rating scale for Mentalization-Based Group Therapy quality

This is a table used for rating therapist's interventions during group therapy. The table describes the quality level 4 ("good enough"). For more detailed descriptions we refer to the manual.

Item name	Quality level 4 («good enough»)
1. Managing group boundaries	The group is functioning smoothly with respect to boundary issues. The therapists identify boundary relevant events and comment and deal with them in ways which seem appropriate and clarifying for the group as a whole.
2. Regulating group phases	At least two phases are addressed in a way that engages members to reflect upon the possibilities and choices they have.
3. Initiating and fulfilling turntaking	The therapists themselves take initiative and they also follow up patients' initiatives for turntaking. They contribute to the unfolding of the story and identification of relevant scenes, intervene in ways that facilitate a comprehensive narrative and keep a focus on emotions, mental states and interpersonal interactions.
4. Engaging group members in mentalizing external events	The therapists invite the other group members, implicitly or explicitly to clarify relevant events and engage members to participate in a collective exploration of the mental states involved therein.
5. Identifying and mentalizing events in the group	The therapists identify some important events in the group and engage group members in a collective exploration which seems meaningful and clarifying.

(continued)

Table (continued)

Item name	Quality level 4 («good enough»)
6. Care for the group and its members	At this level, the group process is on the even when it comes to care. The therapists seem to have an awareness regarding negative comments between group members and are quick to intervene in such situations.
7. Managing authority	The therapists seem calm and confident as MBT-G therapists. In theory and practise they stand up for the group's basic values.
8. Stimulating discussions on group norms	The therapists take initiative to norm discussions, engage in an interested way in spontaneous discussions and try to modify restrictive group solutions which are being made, if these are not challenged by other group members.
9. Cooperation with cotherapist	There seems to be a confident relationship between the therapists, their interventions are complimentary, and they communicate with each other with open, reflective comments.
10. Engagement, interest and warmth	The therapists appear genuinely warm and interested in each member and the group as a whole. The rater gets the impression that the therapists care in a positive way. Several interventions and their stance indicate this.
11. Exploration, curiosity and not-knowing stance	The therapists pose appropriate questions designed to promote exploration of the patients' and other's mental states, motives and emotions and communicate a genuine interest in finding out more about them.
12. Challenging unwarranted beliefs	The therapists confront and challenge unwarranted opinions about oneself or others in an appropriate manner.
13. Regulating emotional arousal	The therapists play an active role in terms of maintaining emotional arousal at an optimal level (not too high so that patients lose their ability to mentalize and not too low so that the session becomes meaningless emotionally).
14. Acknowleding good mentalizing	The therapists identify and explore good mentalizing and this is accompanied by approving words or judicious praise.
15. Handling pretend mode	The therapists identify pretend mode sequences and intervene to improve mentalizing capacity.
16. Handling psychic equivalence	The therapists identify psychic equivalence functioning and intervenes to improve mentalizing capacity.
17. Affect focus	The interventions focus primarily on emotions – more than on behavior. The attention is particularly directed at emotions as they are expressed in the here and now in the group, and particularly in terms of the relationship between patients and between patients and therapists.
18. Stop and rewind	The therapists identify at least one incident in which patients describe interpersonal events in a non-coherent and affected way, tries to slow down the pace and find out about the event step-by-step. In a similar way, the therapists halt events in the group that tend to be destructive and take initiative to explore the sequence together with the patients.
19. Focus on the therapist – patient relationship	The therapists comment on and attempt to explore, together with the patients, how the patients relate to the therapist during the session and stimulate reflections on alternative perspectives whenever appropriate. The therapists speak about their own feelings and thoughts, related to the patients, and by this they try to engage all parties in mutual exploration.