Lessons learned from measuring fidelity with the Motivational Interviewing Treatment Integrity code (MITI 4)

Kramer Schmidt, Lotte; Andersen, Kjeld; Nielsen, Anette Søgaard; Moyers, Theresa B.

Published in:
Journal of Substance Abuse Treatment

DOI:
10.1016/j.jsat.2018.11.004

Publication date:
2019

Document version
Publisher's PDF, also known as Version of record

Document license
CC BY-NC-ND

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim. Please direct all enquiries to puresupport@bib.sdu.dk
Lessons learned from measuring fidelity with the Motivational Interviewing Treatment Integrity code (MITI 4)

Lotte Kramer Schmidt\textsuperscript{a,}\textsuperscript{*}, Kjeld Andersen\textsuperscript{a,b}, Anette Søgaard Nielsen\textsuperscript{a,c}, Theresa B. Moyers\textsuperscript{d}

\textsuperscript{a} Unit of Clinical Alcohol Research, University of Southern Denmark, J.B. Winsløw's vej 20, entrance 220B, 5000 Odense C, Denmark
\textsuperscript{b} Department of Mental Health, Region of Southern Denmark, Denmark
\textsuperscript{c} OPEN Odense Patient data Explorative Network, Denmark
\textsuperscript{d} Center on Alcoholism, Substance Abuse, and Addictions, MSC11 6280, 1 University of New Mexico, Albuquerque, NM 87131, United States

\textbf{ARTICLE INFO}

Keywords: Motivational interviewing, Integrity, Fidelity, MITI4, Reliability, Cognitive behavioral therapy

\textbf{ABSTRACT}

\textbf{Background and aims:} The Motivational Interviewing Treatment Integrity code (MITI) measures fidelity to, and the quality of, Motivational Interviewing (MI), and can also be used when MI is combined with other treatment methods. The current study presents a fidelity measurement with the MITI 4.2.1, in both Motivational Enhancement Therapy sessions and the combined Community Reinforcement Approach-Senior (CRA-S).

\textbf{Method:} The MITI 4.2.1 was used to evaluate treatment sessions provided in the Elderly Study, a multi-national randomized trial evaluating treatment for alcohol use disorders in the elderly. Following expert recommendations, training was conducted at two international sites as well as at the Danish site. Twenty percent of the sessions at the Danish study site were rated. Twelve percent were multiply rated by all raters. Interrater reliability was assessed by the Intraclass Correlations Coefficient (ICC).

\textbf{Results:} Mean ICC of the 52 sessions rated by all raters was 0.78 (95% CI: 0.70; 0.86). The rare measures confront and emphasize autonomy, and the global measure soften sustain talk only reached fair levels of ICC, while the remaining measures were good or excellent. In the sessions of MI combined with other treatment approaches in the CRA-S, the MITI 4.2.1 has a similar reliability as in MET sessions only, except for the measure persuade with permission.

\textbf{Conclusion:} The MITI 4.2.1 is a reliable instrument for measuring fidelity to Motivational Interviewing elements, also in the context of Community Reinforcement Approach Senior. However, in softening sustain talk, the rare measures, and persuade with permission it has proved more difficult to reach high levels of interrater reliability.

\section{1. Introduction}

\subsection{1.1. Background}

Although Motivational Interviewing (MI) is well-established in behavior change treatment (Ekong & Kavookjian, 2016; Lundahl et al., 2013; Smetslund et al., 2011), the effects of MI vary substantially across studies, sites, and even therapists (Miller & Rollnick, 2014; Moyers, Rowell, Manuel, Ernst, & Houck, 2016). Some of the variability may be explained by how MI is delivered (Miller & Rollnick, 2014; Moyers, Rowell, et al., 2016). In other words: what therapists are trained and expected to do, and what they actually do, is not always the same (Hogue, Dauber, Lichvar, Bobek, & Henderson, 2015; Miller & Rollnick, 2014).

Lack of treatment fidelity may result in outcome studies of MI not necessarily reflecting the true impact of MI. Furthermore, lack of treatment fidelity in daily practice may lead to patients not receiving the evidence-based treatment as intended. Thus, fidelity to the MI protocol is of crucial importance and measurement of treatment fidelity has become a key part of treatment research in general, including studies of MI (Miller & Rollnick, 2014; Moyers, Rowell, et al., 2016).

\subsection{1.2. Treatment fidelity}

Treatment fidelity, also called treatment integrity, is defined as the extent to which treatment is implemented as intended.

\textbf{Abbreviations:} AUD, alcohol use disorder; CBT, cognitive behavioral therapy; CRA-S, community reinforcement approach senior; ICC, intraclass correlations coefficient; MI, motivational interview; MET, motivational enhancement therapy; MITI, Motivational Interviewing Treatment Integrity Manual

* Corresponding author at: Unit of Clinical Alcohol Research, University of Southern Denmark, J.B. Winsløw's vej 20, entrance 220B, 5000 Odense C, Denmark.

\textbf{E-mail addresses:} lkramer@health.sdu.dk (L. Kramer Schmidt), kjeld.andersen@rsyd.dk (K. Andersen), annsnielsen@health.sdu.dk (A.S. Nielsen), tmoyers@ unm.edu (T.B. Moyers).

https://doi.org/10.1016/j.jsat.2018.11.004

Received 29 June 2018; Received in revised form 12 October 2018; Accepted 18 November 2018
The previous version of the MITI, version three (MITI 3), demonstrated acceptable overall validity and reliability, but struggled to attain sufficient reliability on a couple of measures (Forsberg, Berman, Kallmen, Hermansson, & Helgason, 2008; Forsberg, Kallmen, Hermansson, Berman, & Helgason, 2007; Moyers, Martin, Manuel, Hendrickson, & Miller, 2005). In addition, the principles behind motivational interviewing have evolved over time, in particular, as demonstrated by the 3rd edition of Motivational Interviewing by W.R. Miller and Steve Rollnick (Miller & Rollnick, 2012). In consequence, MITI 3 was modified and transformed into the current version, MITI 4.2.1 (Moyers, Manuel, & Ernst, 2014). The MITI 4.2.1 seeks to solve the difficulties inherent in attaining satisfactory reliability on all measures (Forsberg et al., 2007; Moyers, Rowell, et al., 2016). Furthermore, MITI 4.2.1 focuses even more sharply on MI-specific elements by tallying MI-adherent therapist behavior separately, by for instance, distinguishing behaviors such as affirming, seeking collaboration, and emphasizing autonomy, as well as introducing global measures of therapist competence in the technical elements of MI, such as evoking change talk and softening sustain talk (Moyers, Rowell, et al., 2016).

1.3. Studies on reliability of MITI 4.2.1

The MITI 4.2.1 was developed in 2014 and is still a relatively new instrument. So far, three studies have both tested for and found acceptable reliability of the measures of MITI 4.2.1 (Copeland et al., 2018; Moyers, Rowell, et al., 2016; Owens, Rowell, & Moyers, 2017), except for one study which reports low reliability on some of the global measures (Copeland et al., 2018). In addition, two of them provide information on the construct and predictive validity of the MITI 4.2.1. Moyers, Rowell, et al. (2016) showed satisfactory results when comparing the new global score in MITI 4.2.1 (cultivating change talk) with a gold standard, namely the Motivational Interviewing Skill Code (MISC) (Moyers, Rowell, et al., 2016). Owens et al. (2017) found associations between the global measures of the MITI 4.2.1 and client motivation for change by the end of treatment (Ovens et al., 2017).

However, not all elements of MITI 4.2.1 have been extensively investigated, and so far, only a few studies have reported on its reliability and validity. Reports on the reliability of MITI 4.2.1 in different settings and countries are essential not only for the continuous process research of Motivational Interviewing, and, in particular, for studies investigating outcome of therapies that are grounded in Motivational Interviewing, but also for the use of MITI 4.2.1 as a tool in supervision and feedback when training therapists in MI. A reliable instrument to facilitate implementation of MI would also benefit MI-practice and, in consequence, the accessibility of evidence-based MI for clients.

1.4. MI as a single therapy versus MI combined with other therapies

It has been suggested that the effectiveness of MI is even greater when combined with Cognitive Behavioral Therapies (CBT) or other specific treatment methods (Hettema, Steele, & Miller, 2005; Spoelstra, Schueller, Hilton, & Ridenour, 2015; Westra, Aviram, & Doell, 2011; Westra, Constantino, & Antony, 2016), than when delivered as a single treatment approach. In many studies of therapy, MI is used during treatment start and thereafter combined with other therapeutic elements. This is for instance the case in project COMBINE (Anton et al., 2006), and in the Elderly Study (Andersen et al., 2015); in both studies, MI and CBT in the version of the Community Reinforcement Approach (Meyers, Roosen, & Smith, 2011) were combined (Andersen et al., 2015; Anton et al., 2006).

What happens to fidelity to MI when it is combined with other therapies? As treatment approaches where MI is combined with other therapies are gaining ground (Gates, Sabioni, Copeland, Foll, & Gowing, 2016; Hogue, Henderson, Becker, & Knight, 2013; Riper et al., 2014), fidelity measurements of the MI delivered in these combined settings are important. Only few studies report on this. Based on experiences from the COMBINE Study, Moyers and Houck (2011) raised a concern about competing recommendations in therapist behavior in some situations (Moyers & Houck, 2011). Their point was supported by Aviram, Westra, Constantino, and Antony (2016) and Boyer, MacKay, McLeod, and van der Oord (2018). On the fidelity measures, findings are mixed and based on previous versions of the MITI. Westra et al. (2016) reported a low to moderate fidelity to CBT in a treatment combining MI and CBT, while fidelity to MI remained high. Similarly, Boyer et al. (2018) reported high fidelity to MI in a combined setting with CBT while, Magill et al. (2018) reported relatively lower fidelity to MI in a combined intervention of MI and CBT.

Reliability of the previous versions of MITI in earlier studies of MI combined with CBT has been satisfactory. Two studies utilized the MITI 3 for fidelity measurement of an integrated MI and CBT treatment, with good reliability of the elements measured (Boyer et al., 2018; Magill, Graves, et al., 2018): elements such as empathy, giving information, and reflections are also represented in the MITI 3 (Moyers, Martin, Manuel, Miller, & Ernst, 2010). Westra et al. (2016) measured fidelity to MI combined with CBT with the MITI 2 and also reported a high level of interrater reliability. We therefore assume that MITI 4.2.1 can also be applied to measure the extent to which therapy is delivered according to MI principles, even when elements of other treatment approaches are included, but we do not yet know how valid and reliable MITI 4.2.1 is when MI is combined with other therapies. The present study reports on the reliability of the MITI 4.2.1 in both a combined treatment and a treatment approach consisting of MI and feedback only (MET).

1.5. Aim

Training and supervision of therapists are often described in studies of MI, but the literature rarely describes how the process of training the raters who measure treatment fidelity is conducted. Consequently, there is a need for greater transparency in rating procedures and fidelity measurement that would enable us better to train and test our rating teams, (Mesters, Keulen, de Vries, & Brug, 2017).

The aim of this study is to present the process and lessons learned from measuring treatment fidelity with the MITI 4.2.1 at the Elderly Study’s Danish site (Andersen et al., 2015). Specifically, we seek: 1) to increase transparency in rater training, while maintaining and measuring competence of raters of the MITI 4.2.1, 2) to discuss reliability of the instrument in relation to the individual measures, and finally, 3) to report the reliability of MITI 4.2.1 when applied in measuring fidelity to MI in a combined condition.

2. Method

2.1. Project

This fidelity measurement was performed at the Danish site for the Elderly Study, an international, multisite randomized clinical trial (Andersen et al., 2015). Briefly, in the Elderly Study, people aged 60+ years, with alcohol use disorder according to DSM 5 criteria, were randomized to one of two outpatient treatments: 1) Four weeks with four sessions of Motivational Enhancement Therapy (MET), or 2) Four weeks with four sessions of MET, plus an add-on of up to eight weeks with eight sessions of the Community Reinforcement Approach Senior (CRA-S) (Andersen et al., 2015).

The underlying counseling approach in both MET and MET+CRA-S was MI. (Please find links to both manuals beneath the reference list below.) The MET condition was a slightly modified version of the COMBINE and Project MATCH protocol: a manualized form of Motivational Interviewing, supplemented with personalized feedback.
Treatment with CRA involves coping skills training and facilitates social and recreational activities (Smith, Meyers, & Miller, 2001). The eight additional sessions in the extended MET + CRA-S condition comprised an individually planned combination of the following modules: coping with craving, urges and social pressure to drink, building a sober network, mood management training, and social and recreational therapies (Andersen et al., 2015). In essence, similar to the modules in the COMBINE protocol, but tailored to a population of elderly. Since the participants were 60+, they were also offered a module designed specifically for the Elderly Study named Coping with Aging (Andersen et al., 2015; Sogaard Nielsen et al., 2016). The Danish site for the Elderly Study involved three outpatient alcohol treatment centers. Seven therapists were trained and supervised prior to the study and over the course of the study in MET and CRA-S. All sessions were audio-recorded.

2.2. Sample

The sample for the fidelity measurement presented in this paper was drawn from the recordings made at the Danish site. Since MI fidelity may vary across the course of a session, the overall study period, and the type of session (Jelsma, Mertens, Forsberg, & Forsberg, 2015), the following randomization process was performed. First, 20% of all the sessions were randomly chosen to be assessed for fidelity. The sample of sessions consisted of both MET sessions and CRA-S sessions. Thereafter, within each session, a 20-minute window was randomly picked by means of a mathematical equation in Excel.

2.3. The MITI 4.2.1 instrument

The MITI 4.2.1 uses global ratings to assess technical and relational components of MI, in addition to explicitly counting MI adherent, MI non-adherent behavior, and supposedly neutral therapist behavior such as giving information and persuading with permission. (Moyers et al., 2014; Moyers, Bowell, et al., 2016). Table 1 provides an overview of the MITI 4.2.1. The elements are assessed during a randomly selected 20-min window of the session. The global scores are measured on a five point Likert scale: five reflecting a high level of proficiency (Moyers et al., 2014). The English and the Danish versions of the MITI 4.2.1 were used interchangeably in the process of training, as described below. The English version of the MITI 4.2.1 manual was translated into Danish by Anette Søgaard Nielsen, one of the authors of this article. Several people from the Motivational Interviewing Network of Trainers in Scandinavia reviewed the manual to ensure correct translation. Internet links to the Danish and English versions are listed in the references.

2.4. Training – International team

To ensure agreement between raters at all sites in the Elderly Study, stress was laid on the importance of measuring the fidelity of MI uniformly across sites. That way, a foundation would be created for measuring impact of treatment while also taking treatment fidelity at the sites into account. The training of raters and fidelity measurement were supervised by Theresa B. Moyers (TM), one of the authors of the MITI 4.2.1. Training of raters was completed in a five-month period prior to the actual fidelity measurement. A significant challenge of this study was ensuring that raters across three cultures and languages were consistently giving similar scores in similar situations. To accomplish this: 1) The leaders of the German, American, and Danish rating teams met via videoconferencing approximately every two weeks (international meetings), and 2) The leaders of the ratings teams had face-to-face meetings with their respective raters at the local sites (local meetings).

Between each international videoconference meeting, at least two English-language recordings (both audio and transcripts) were rated by all the local teams. A short summary of expert recommendations on training in the MITI 4.2.1 is listed in Table 2. It follows the principles generally used (Brueck et al., 2009; Moyers et al., 2014; Seng & Lovejoy, 2013). As recommended (Table 2), rating competence was developed gradually, measure by measure. During the process of training, reliability was assessed by checking for 80% observed agreement without using statistical programs. This is a crude assessment of agreement: it does not consider factors such as the risk of raters reaching agreement by chance (Hallgren, 2012). The team of raters followed roughly the order for including new measures suggested in Table 2. The discussions of ratings at the international meetings visualized the importance of cultural interpretations and how these may be difficult to explain across cultures. As an example, and as might be expected (Bhatara et al., 2016), some of the utterances that led to discrepancies between raters were caused by the tone of voice of the therapist. The non-native English speaker found it more difficult to differentiate between the different tones used by the therapists in the English recordings and would more readily rate the measure persuade. We did not collect any data on this.

2.5. Training – The Danish team

The Danish team consisted of two medical undergraduate students, one medical PhD student – not trained in MI – and two experienced MI therapists and trainers. The age range was between 22 and 64 years, four women and one male. The Danish rating team held weekly meetings of 90–120 min during the 5 months of training. Initially, the group rated English recordings prior to each international meeting. Between 15 and 20 English-language recordings were rated for the international meetings. Later, training was based on Danish recordings of sessions from the Elderly Study and other studies. To attain project-specific competence (Seng & Lovejoy, 2013), raters were given an overview of the assessment instruments used as feedback in MET sessions, the manuals for MET and CRA–S, and a short introduction to MI in general.

The Danish team's training followed the guidelines in Table 2. During the initial stages of training, reliability was assessed by checking for 80% observed agreement; later intra class correlations (ICC) were calculated on 8–10 recordings at a time and feedback was given to the raters accordingly. Prior to commencing the actual fidelity measurement of the recordings at the Danish sites of the Elderly Study, all raters in the Danish rating team attained excellent ICC (calculated as described in the “Interrater reliability” section) for all measures on the basis of ten recordings from the Elderly Study. To secure interrater reliability over the course of the 10 months of the fidelity assessment, we did the following: 1) out of every five recordings there would be at least one which was multiply rated by all raters; 2) the raters were blinded to which recordings were multiply rated; 3) weekly meetings were held to discuss the ratings and compare already multiply rated recordings; 4) to maintain rater competencies and still be able to correct raters’ deviations before too many ratings were rated, the raters were encouraged to rate at least three, but no more than ten, recordings per week; 5) if a rater stopped rating for more than two weeks, they had to compare interrater reliability on a minimum eight recordings with the rest of the group before they could continue; 6) disagreements among the raters were reduced by making specific decision rules: examples are provided in Table 3; and 7) helping tools providing guidance on and anchors for different measures were supplied to the raters (available from the first author). It should be noted that rule 5 was never used, since no rater stopped rating for two weeks or more during the rating process.

The discussions in the Danish rating team turned especially on differentiating between simple and complex reflections, as well as on the measures persuade, seeking collaboration, and persuade with permission. Recordings which provoked significant disagreement between raters had to be rated jointly by the rating team. By discussing each rating, one by one, agreement on the tricky measures in the team slowly increased. To maintain reliability in these measures, they needed to be
Table 1
Overview of therapist behavior measured by the Motivational Interviewing Treatment Integrity code (MITI 4.2.1) and descriptive values extracted from the 423 sessions measured for fidelity at the Danish site of the Elderly Study.

<table>
<thead>
<tr>
<th>Behavioral measures</th>
<th>Median and interquartiles (25%,75%)</th>
<th>Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Giving information</td>
<td>5 (3;7)</td>
<td>5.57 (3.90)</td>
</tr>
<tr>
<td>Persuade</td>
<td>0 (0;1)</td>
<td>0.72 (1.39)</td>
</tr>
<tr>
<td>Persuade with permission</td>
<td>0 (0;1)</td>
<td>0.52 (0.96)</td>
</tr>
<tr>
<td>Question</td>
<td>11 (6;17)</td>
<td>12.07 (7.37)</td>
</tr>
<tr>
<td>Simple reflection</td>
<td>8 (5;13)</td>
<td>9.25 (6.32)</td>
</tr>
<tr>
<td>Complex reflection</td>
<td>12 (8;17)</td>
<td>12.62 (6.32)</td>
</tr>
<tr>
<td>Affirmation</td>
<td>0 (0;1)</td>
<td>0.73 (1.06)</td>
</tr>
<tr>
<td>Seeking collaboration</td>
<td>1 (0;2)</td>
<td>1.24 (1.43)</td>
</tr>
<tr>
<td>Emphasize autonomy</td>
<td>0 (0;0)</td>
<td>0.20 (0.50)</td>
</tr>
<tr>
<td>Confront</td>
<td>0 (0;0)</td>
<td>0.02 (0.17)</td>
</tr>
<tr>
<td>Cultivating change talk</td>
<td>4 (3;5)</td>
<td>3.85 (0.98)</td>
</tr>
<tr>
<td>Softening sustain talk</td>
<td>4 (4;4)</td>
<td>4.04 (0.51)</td>
</tr>
<tr>
<td>Partnership</td>
<td>4 (4;5)</td>
<td>4.07 (0.83)</td>
</tr>
<tr>
<td>Empathy</td>
<td>4 (4;5)</td>
<td>4.32 (0.65)</td>
</tr>
</tbody>
</table>

Table 2
Recommendations on training for fidelity measurement with the MITI 4.2.1.

To get proficient at rating with the MITI takes at least two hours of training a day over four weeks of training.
Do not think too hard about the rating, no one else will be able to follow your line of thought.
At least 10% of the rated material should be double rated.
The raters should reach 80% agreement before rating the recordings for the project.
Only a difference of more than one on the Likert scale is considered disagreement on the global scores.
It is not possible to agree 100% on the ratings.
Start training and thereby learning the MITI bit by bit, gradually adding more and more behavioral rates from the MITI in the order suggested below:
1 Parsing: practice identifying utterances and volleys
2 Questions
3 Reflections without differentiation between simple and complex reflections
4 Giving information
5 Differentiation between simple and complex reflections
6 Use of persuasion
7 Affirmations, Seeking collaborations, Persuading with permission, and Emphasizing autonomy
8 Confront
9 The global scores - initially without the behavior codes, but ultimately including all elements of the MITI

discussed frequently within the rating team.

2.6. Interrater reliability

Intraclass Correlations Coefficient (ICC) has different variants depending on aim and design (Hallgren, 2012). The decision on which version of ICC to apply was based on recommendations made by Hallgren (2012) as follows: 1) When the raters remain the same throughout the study and are not selected at random from a larger group of raters, the two-way model takes into account systematic divergencies of the individual raters. 2) One rater may consistently be a bit lower on the scale than another rater and the ICC variant “consistency of agreement” allows for this variance by rank-ordering the ratings. 3) When the recordings are being measured by multiple raters, we want to assess reliability on the average of multiple raters, as opposed to measuring how well raters are aligned with a single “true” rater. We did not have a “true value”-rater, but the average of the five raters. In this case, ICC based on average measures is recommended. 4)
If the subjects included in the rating are random, but the raters are constant, a mixed-effects model is recommended. The mixed-effects designation has no consequences for the ICC value, but highlights the fact that this rating is specific for a fixed set of raters, as opposed to a randomly selected sample of raters (Hallgren, 2012). In summary, the ICC in the present study was assessed by a two-way model, consistency of agreement, average-measure, and mixed effects ICC. This allows for systematic differences between the raters and is therefore a less stringent version of the ICC than that which uses the absolute agreement and single measure. For comparison between studies, though, absolute agreement may be more appropriate; therefore, ICC using absolute agreement was also calculated.

Benchmark values for agreement of ICCs are: 0.00 = poor, 0.40–0.59 = fair, 0.60–0.74 = good, 0.75–1.00 = excellent (Cicchetti, 1981). ICC will be reported for each MITI 4.2.1 variable (Jelsma et al., 2014; Dunn et al., 2007; Moyers, Rowell, et al., 2016) and a large number of multiply rated sessions in general leads to more solid estimates of ICC (Forsberg et al., 2013). Moreover, some of the raters could rate up to ten recordings per week. To visualize variability, descriptive values of the MITI 4.2.1 scores are listed in Table 1. In the 52 sessions rated by all five raters, low variability was observed in some measures, especially softening sustain talk: it was rated as a four in 205 (5 × 52) ratings for interrater reliability. Some of the measures were also quite rare: confront was only rated eight times (three across the same CRA-S session) and emphasize autonomy was rated 27 times (seven across three different CRA-S sessions) in the 260 ratings for interrater reliability. By comparison, persuade with permission, persuade, affirm, and seeking collaboration were rated 126, 150, 204, and 320 times, respectively, in the 260 ratings.

### 2.7. Number of raters and multiple rated sessions

Prior to the study it was estimated that approximately 50 recordings should be multiply rated. Since a larger number of multiply rated sessions in general leads to more solid estimates of ICC (Forsberg et al., 2007; Moyers, Rowell, et al., 2016) and a large number of multiply rated sessions is cost-expensive and may not be feasible, we sought to illustrate the interrater reliability of the MITI 4.2.1 analyzed on the basis of only 11 sessions and with only two raters. Therefore, every fifth of the recordings utilized for the reliability measurement was randomly picked in order to perform an ICC on a lower number of ratings. Finally, to report on whether raters’ experience with MI matters, ICC on the basis of the rating data of the two MI-trained raters in this project was calculated and difference between mean ICC-levels were tested using t-tests.

### 3. Results

#### 3.1. Number of recordings

The treatment courses of the 341 participants resulted in 2127 audio-recorded sessions, of which 423 were randomized for fidelity measurement. The number of recordings measured by each rater varied because the five raters did not enter the fidelity measurement project at the same time: 68, 102, 142, 157, and 162 ratings, respectively. Moreover, some of the raters could rate up to ten recordings per week while others only rated three per week. To visualize variability, descriptive values of the MITI 4.2.1 scores are listed in Table 1. In the 52 sessions rated by all five raters, low variability was observed in some measures, especially softening sustain talk: it was rated as a four in 205 (5 × 52) ratings for interrater reliability. Some of the measures were also quite rare: confront was only rated eight times (three across the same CRA-S session) and emphasize autonomy was rated 27 times (seven across three different CRA-S sessions) in the 260 ratings for interrater reliability. By comparison, persuade with permission, persuade, affirm, and seeking collaboration were rated 126, 150, 204, and 320 times, respectively, in the 260 ratings.

#### 3.2. Interrater reliability at the Danish site of the elderly study

A total of 52 audio recordings (12%) were rated in parallel by all five raters: the ICCs are presented in Table 4. Agreement on nine measures in the present study is excellent, two measures are good, while that on softening sustain talk, confront, and emphasize autonomy is fair. Intra-class correlation coefficients calculated as absolute agreement did not differ significantly (data not shown). The reliability measurement based on a randomly picked one fifth of the 52 recordings (n = 11) is also presented in Table 4. When basing the analysis on a fewer number of sessions or raters, the ICC-level in the measures emphasize autonomy and confront fell below the threshold for fair ICC.

The overall mean ICC of all 14 measures when including all raters was 0.78 (95% CI: 0.70; 0.86); whereas the range of the mean (95% CI) ICC of all 14 measures for each pair of raters was between 0.49 (0.32; 0.66) and 0.78 (0.70; 0.86) (n = 52). ICC on the basis of 11 recordings rated by the rater pair with the highest and the lowest mean ICC level is also displayed in Table 5. Several of the measures were unstable in this analysis with wide confidence intervals, and ICC on
empathy fell to poor interrater reliability. In detail, though, the rater pair with the highest mean ICC never differed by more than one on the Likert scale. As explained in the Method section, this is acceptable agreement in empathy. All the ICCs based on pairs of raters were lower than the ICC calculated from five raters. The ICCs were not affected by the rater's level of experience with MI. The mean ICC between the two raters who had experience with MI was 0.54 (95% CI: 0.40; 0.67) (n = 52); moreover, they did not have a significantly higher reliability than the remaining. On the contrary they had a significantly lower mean ICC than the rater pair with the highest mean ICC (p < 0.01). Additionally, they did not have a significantly higher level of agreement than the other rater pairs on mean global scores or MI-adherent behavior.

3.3. Interrater reliability in the CRA-S

Since MITI has not previously been used to assess therapist behavior in CRA-S sessions, we particularly analyzed the ICC in the CRA-S sessions. Persuade with permission fell below the threshold for fair reliability, but only the measure confront differs significantly in reliability when compared to the reliability in the MET sessions, only (Table 4).

4. Discussion

4.1. Summary of results

Overall, this assessment of Motivational Interviewing (MI) fidelity to the English and Danish versions of MITI 4.2.1, and based on the training course described above, demonstrated high interrater reliability between five raters. As expected, it was more difficult to reach high and stable levels of ICC between two raters compared to five raters, and only one pair of raters reached the same high mean level of ICC as that achieved by five raters. Additionally, most of the elements of the MITI 4.2.1 were reliably measured in an intervention where MI was combined with the Community Reinforcement Approach (CRA).

4.2. Reliability of the individual measures of the MITI 4.2.1

Intraclass correlations on the behavioral counts of emphasize autonomy and confront only reached fair levels when including all ratings and all five raters, and lost all reliability when ICC was calculated based only on only 11 ratings from only two raters. Both measures were rare. Although these measures attained a higher level of reliability in the study by Moyers, Rowell, et al. (2016), the rare measures are difficult to measure reliably (Moyers, Rowell, et al., 2016). We do not have any specific data on this, but we noticed some resistance in the weekly meetings among the raters to use the behavioral measure confront. It is important to emphasize in rater training that the MI non-adherent measures persuade and confront are not bad as such - just not adherent to MI.

In earlier versions of MITI, confront and persuade were considered MI non-adherent in a combined measure; additionally, emphasize autonomy, seeking collaboration, and affirmation were more or less measured by means of the combined measure MI adherent (Moyers et al., 2010). In the previous versions of MITI, these combined measurements had problems attaining good reliability (Beckman et al., 2017; Dunn et al., 2015; Eno Persson et al., 2016; Forsberg et al., 2007; Mesters et al., 2017; Palfai et al., 2016). Thus, separating the elements of MI non-adherent as well as MI adherent has improved reliability of the MITI overall and made it possible to measure some of these behaviors reliably. In turn, this enables more specific process research and training in MI.

The global score of softening sustain talk did not reach high interrater reliability. Problems with attaining higher levels of interrater reliability on this measure is not new (Copeland et al., 2018; Owens et al., 2017). There is a decision rule in the MITI 4.2.1 to rate this global as a four on the Likert scale if the client does not present sustain talk (Moyers et al., 2014). This was the case with many of the sessions rated in the present study: 205 of the 260 multiply measured recordings were rated as four. Like the rating of empathy in the analysis of ICC based on 11 sessions and only two raters, the ICC takes into account the high chance of the raters agreeing on a four; therefore, the ICC is low even when raters agreed on four in most of the cases. Difficulties in attaining high ICCs without the whole range of the Likert scale being utilized is not new (Brueck et al., 2009; J. E. Hettema et al., 2018; Seng & Lovejoy, 2013; Spohr et al., 2015). The ICC is simply not a good measure of agreement in such cases. ICC may be improved if softening sustain talk is simply not rated when sustain talk does not occur, coding it by default instead as a four on the Likert scale. Such a change in the rating system could also provide information on the level of sustain talk in the population of the study.

Although it is a crude approach and only feasible when utilizing two raters, reporting reliability of the global measures by how often the raters differed by more than one on the Likert scale may be a better expression of reliability of the global measures than the ICC – especially with lower numbers of multiple-rated tapes. It may be valuable to report this in addition to reporting ICC, in particular, in the rather common cases with low ICC for the global measures (Copeland et al., 2018; de Gee et al., 2014; Dunn et al., 2015).

Another consideration would be to revise some of the global measures. One of the strengths of the MITI is the measures of global therapist competence; competencies that relate to client behavior. Overall, the description of anchors, as requested for the relational global measures (Forsberg et al., 2007; Madson & Campbell, 2006), has successfully increased their reliability. However, part of the technical global measure (i.e. the average of softening sustain talk and cultivating change talk) appear to be more difficult to measure reliably (Serrano et al., 2017), in spite of the anchors. Moreover, the softening sustain talk measure has questionable validity (Moyers, Rowell, et al., 2016). Even though the information is valuable, it may be that assessing the technical “gestalt” of the session is just too difficult to measure reliably and a more pragmatic approach could be needed. Both sustain talk and change talk appear to be related to outcome and especially the proportion of change talk in relation to change and sustain talk is of importance (Magill et al., 2018; Pace et al., 2017). As the behavior counts do better in attaining good reliability, a quantitative separate counting of therapist statements evoking change talk and of statements softening sustain talk may be an alternative measure of these technical elements. This might also enable process research on the proportion of these statements in relation to one another; moreover, it may aid disentangling the relational and technical elements of MI as requested by Owens et al. (2017).

4.3. Reliability in the community reinforcement approach-senior (CRA-S)

In this study, most of the ICC levels did not differ between sessions of MET and CRA-S, indicating that MITI 4.2.1 may be used as a fidelity measurement of the MI elements applied in the CRA-S condition. The variations in the behavioral measures of confront and emphasize autonomy may be explained in the same terms as the variations in general, as described above: confront and emphasize autonomy were only rated three times across the same session and seven times representing three sessions in the CRA-S sessions, respectively. This underlines the importance of choosing a representable sample when assessing fidelity to MI in a combined treatment (Moyers, Rowell, et al., 2016).

Research using previous versions of the MITI in combined treatments has introduced new aspects, such as the timing of switching between MI and CBT (Aviram et al., 2016); how the structure of the other treatment approach affects the importance of fidelity to MI (Boyer et al., 2018); and measured elements of MI in treatments where MI is not part of the prescribed treatment (Magill, Graves, et al., 2018). Future studies are needed to report on what happens to fidelity to MI in a
combined treatment approach, what happens to fidelity to the other treatment approach, how MI and other treatment elements are integrated, and how these are related to outcome.

Although it was not significantly different from the reliability of the MET sessions, the interrater reliability of persuade with permission in the CRA-S sessions was rather low. This indicates that this measure may not be suitable in this setting, or that consistency when rating this behavior is generally unstable. The latter view is supported by Moyers, Rowell, et al. (2016) when assessing persuade with permission in sessions of MI only. Since the measure persuade with permission has been included as a separate measure in the MITI 4.2.1 to accommodate therapists’ giving recommendations without compromising the spirit of MI, in particular when MI is combined with other treatment methods (Moyers, Rowell, et al., 2016), this has implications for the use of the measure. The lack of manualized guidance in how explicitly to combine expert guidance with MI in the CRA-S context could also explain the differences in reliability in the use of persuade with permission: the more standardized a setting, the better the reliability (Imel et al., 2014). Therefore, the differences in how therapists interpret the integration of MI with a treatment that is more directional may have affected raters’ ability to rate this measure reliably. Revision and guidance on when to use the measure persuade with permission, especially when measuring MI in a combined setting, may be necessary.

4.4. Lessons learned by the rating team

Prior to the measurement process, the rating team had achieved excellent ICC levels on all measures of the MITI. We noted a decline in ICC between training and actual fidelity measurement, similar to the experience had by Seng and Lovejoy (2013). Speculations on this decline point to a possible lack of sufficient training in the rare measures and to intra- and interrater variation in the rating team. Inter-rater variation took the form, for instance, of differences in the time permitted to complete ratings between the raters; therefore, the recorders rated by all raters were rated at different time points during the 10 months of the fidelity measurement. Others have solved this problem by rating all sessions within a short timeframe (Forsberg, Forsberg, Lindqvist, & Helgason, 2010). Furthermore, some of the raters seemed to pay more attention to the tone of voice of the counselor than others, and were more apt to rate, as an example, the measure persuade. We do not have any data on this, but there is some support for this notion: prosodic (tone and melody of voice) perception ability has been found to vary across normal subjects, moderated by demographic factors (Uskul, Paulmann, & Weick, 2016), and associated with empathic abilities (Aziz-Zadeh, Sheng, & Gheytnachi, 2010). To offset this, frequent meetings and discussions, where the raters’ individual benchmarks are addressed, and mutual alignment is emphasized are key. The commitment to consistency in rating the recordings over almost a year and to regular attendance at training sessions would appear to be important (Forsberg et al., 2010). We noticed the most significant instances of drift when a rater did not rate or attend supervision to the extent planned. Another study has noted the importance of weekly or biweekly meetings (Mesters et al., 2017); furthermore, weekly or biweekly meetings appear to be associated with higher levels of interrater reliability on empathy (Beckman et al., 2017; Eno Persson et al., 2016; Lindqvist, Forsberg, Enebrink, Andersson, & Rosendahl, 2017; McCambridge, Day, Thomas, & Strang, 2011; Moyers, Houck, Rice, Longabaugh, & Miller, 2016; Palfai et al., 2016; Pierson et al., 2007; Spohr et al., 2015), compared to less frequent meetings (Brueck et al., 2009; McMaster & Resnicow, 2015; Mesters et al., 2017; Pollak et al., 2014). This finding is supported by a study in which fidelity was assessed by three trained raters using MITI 3 (McMaster & Resnicow, 2015): the raters had no contact prior to or during the fidelity measurement and came out with very low levels of interrater reliability – this in spite of being highly skilled in rating with the MITI (McMaster & Resnicow, 2015).

In the present study, the calculation of ICC took into consideration the tendency of raters to consistently score higher or lower on a scale. Raters may, however, vary inconsistently due to emotional state. We did not assess emotional status of the raters in this study, but we speculate that especially the measures where the tone of voice of the therapist is important, such as persuade and empathy, are ones that may be affected by the rater’s intrapersonal emotional status. There is some support for this notion: Paulmann, Furnes, Bokenes, and Cozzolino (2016) found that stressed listeners found it more difficult to decode emotional prosody of the speakers of recordings. As the within-therapist empathy varies (Moyers, Houck, et al., 2016; Chris Dunn et al., 2016), the ability to detect empathy may also vary between raters (Aziz-Zadeh et al., 2016; Uskul et al., 2016) and within the rater (Paulmann et al., 2016).

In summary, to keep the team of raters on the same page throughout the fidelity measurement, the rating team coordinator needs to be very clear and to stick to the benchmarks set at the beginning of the fidelity measurement process. Moreover, frequent training and supervision may be required to tune in on the idiosyncratic details. We recommend raters not to rate when tired or emotionally distressed. Shortening the time frame of the fidelity measurement to a minimum could be an advantage. To optimize reliability of the MITI 4.2.1., future studies could include raters’ emotional state, empathic abilities and sociodemographic characteristics. Finally, systematically reviewing studies describing rater training and ICC levels would shed light on this area.

4.5. MI experience of the raters

It is possible to train non-MI-trained people to rate with the MITI (Jelsma et al., 2015; Moyers, Rowell, et al., 2016); furthermore, this may inject further objectivity since the trained MI therapist may be disposed to rate therapist intent as opposed to the tallies of therapist behavior (Moyers, Rowell, et al., 2016). Intent refers to the purpose or meaning of the therapist’s demeanor and a MI therapist may interpret this demeanor in the light of MI theory (Moyers, Rowell, et al., 2016). This may diminish the reliability of the instrument (Moyers, Rowell, et al., 2016). The tallies of therapist behavior rated with the MITI 4.2.1 is, in theory, the objective behavior, without judgement and interpretations from the individual rater, and this should more readily enable raters to agree on the item (Moyers et al., 2014).

However, as there may be more therapist intent related to global measures, Pierson et al. (2007) suggested that formal therapeutic training in MI is needed to assess the global measures in the earlier versions of MITI. As demonstrated by the fidelity measurement done by Moyers, Houck, et al. (2016), it is possible to train people, who are not trained MI therapists, to rate MI sessions by means of the global measures of the MITI 4.2.1. The present study supports this – the two trained MI therapists on the rating team in the present study did not agree more than did other rater pairs on global or any of the other measures. Nor, previously, have high levels of MI training and experience led to high levels of interrater reliability (McMaster & Resnicow, 2015). Furthermore, previous studies attained comparable levels of interrater reliability when measuring fidelity by means of the MITI, irrespective of whether they included training in MI (Pollak et al., 2014; Spohr et al., 2015) or the level of training in MI was not described (Brueck et al., 2009; Palfai et al., 2016; Seng & Lovejoy, 2013; Woodin, Sotskova, & O’Leary, 2012). This difference may be arbitrary, though, since learning to rate with the MITI in itself provides knowledge of MI.

4.6. Limitations and strengths

Some limitations of study comparisons are that this study only included Danish speaking data; the population in this study were all aged 60+ and, and the additional rules of Table 3 are specific to the Danish site of this project. As discussed above, raters will never be completely objective; thus, the interpretation of scores of the MITI will always be
somewhat affected by rater bias. Although there was 80% observer agreement between the sites at the international meetings, based on recordings in English, we do not have interrater reliability data between the sites. Hence, we do not have any data on the fidelity to the fidelity measure. Moreover, we do not know how fidelity to the MITI 4.2.1 may have changed during the fidelity measurement, since we did not compare rated sessions with the international team after the training period was over. Partly compensating for this, though, the multiply-rated tapes were not rated at the same time by all the raters at the Danish site.

Furthermore, this is an efficacy study, which may have increased interrater reliability, not attainable in a real-world setting. Efficacy elements include frequent meetings with feedback and training in the rater team over an extended period of time, high levels of supervision provided for the therapists, and highly manualized treatments. Future studies are needed in real world settings to further substantiate reliability of the MITI 4.2.1, as for instance, when used for ongoing supervision of therapists.

In spite of the efficacy elements of this study, and in combination with findings from a recent study where reliability in previous versions of the MITI did not differ between training, efficacy, and effectiveness conditions (Hallgren et al., 2018), the high level of interrater reliability among non-MI-trained raters supports the accessibility of the MITI 4.2.1 as a structured tool that provides guidance to the individual therapist, the supervisor, and colleagues in improving training in the use of MI.

This is the first time the MITI 4.2.1 has been applied to an approach where MI is used in combination with other therapy styles, putting focus on the use of this instrument in a combined setting. This is also the first time it has been used as a fidelity measure in a Danish and elderly setting. Since the version four is still new, so far only a few studies have published data on the use and reliability of the MITI 4.2.1 internationally. Finally, this study supports the reliability of the Danish version of the MITI 4.2.1.

4.7 Conclusion

With a high level of interrater reliability on almost all measures, the MITI 4.2.1 may be used in a non-English speaking setting and in contexts where MI is combined with other treatments. Moreover, rating with the MITI 4.2.1 may be carried out by people not trained in MI as such. Training and weekly meetings throughout the course of the fidelity measurement is vital to sustain high levels of interrater reliability. Finally, the global measure softening sustain talk may need revision.

Funding

The Lundbeck Foundation, the Region of Southern Denmark, and the University of Southern Denmark have financially supported the study unconditionally.

Conflicts of interest

None.

Acknowledgements

The authors wish particularly to thank participants who without financial compensation participated in the sessions and the research follow-ups. We are grateful for the support and cooperation of staff at the alcohol outpatient treatment centers in Aarhus, Odense and Copenhagen, and for the dedication and persistence of the raters, besides the primary author, in this project: Louise Bundgaard, Ayse Corap, Birgit Jensen, and Peter Gabelgaard Brolund. Finally, Anders Bo Bojesen and Anna Mejdlad have provided statistical support.

References


Hettema, J. E., Cockrell, S. A., Reeves, A., Ingersoll, K. S., Lum, P. J., Saitz, R., ... Carrejo, V. A. (2016). Development of interrater and differentiability of three questions regarding alcohol use that include varying doses of motivational interviewing. Addiction Science


