Lifestyle Counselling Intervention to prevent Gestational Diabetes Mellitus

Jelsma, J.G.M.

2017

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

CHAPTER 7:

How to measure motivational interviewing fidelity in randomised controlled trials: practical recommendations

Judith G.M. Jelsma*, Vera-Christina Mertens*, Lisa Forsberg, Lars Forsberg

* shared first author

Published in Contemporary Clinical Trials 2015, 43: 93-99.

Many randomised controlled trials in which motivational interviewing (MI) is a key intervention make no provision for the assessment of treatment fidelity. This methodological shortcoming makes it impossible to distinguish between high- and low-quality MI interventions, and, consequently, to know whether MI provision has contributed to any intervention effects. This article makes some practical recommendations for the collection, selection, coding and reporting of MI fidelity data, as measured using the Motivational Interviewing Treatment Integrity code. We hope that researchers will consider these recommendations and include MI fidelity measures in future studies.

INTRODUCTION

"Motivational interviewing (MI) is a collaborative, goal-oriented style of communication with particular attention to the language of change. It is designed to strengthen personal motivation for and commitment to a specific goal by eliciting and exploring the person's own reasons for change within an atmosphere of acceptance and compassion"[215]. MI has been shown to be superior to no intervention and as efficacious as other evidence-based interventions in systematic reviews and meta-analyses across a variety of different problem behaviours and health care settings [41,143,187,188,282].

A randomised controlled trial (RCT) tests whether an intervention is efficacious in an ideal situation by maximizing internal validity through controlling all variables except the intervention to be tested. A controlled clinical trial tests instead whether an intervention is effective in a real life setting, maximizing the external validity to ensure generalizability [117]. In order for the move from an efficacy trial to an effectiveness trial to be successful, it is important to have identified the active mechanism of the intervention tested in the efficacy study [324]. In respect of behaviour change intervention research, the reporting of treatment fidelity is likely to improve the credibility of evidence that results from a trial [23]. Treatment fidelity refers to the "methodological strategies used to monitor and enhance the reliability and validity of behavioural interventions" (p.443) [23].

Treatment fidelity in MI has predictive validity in respect of patient behaviour following the intervention [14,183,231]. However, many research trials conducted have failed to assess treatment fidelity of the intervention that is being delivered. This makes it impossible to ascertain whether the result can accurately be attributed to the MI intervention, that is, whether we can in fact be sure that MI is the actual working mechanism that is "doing the job" [219]. Miller and Rollnick (2014) suggest that treatment fidelity should be assessed throughout a study, through a reliable assessment procedure ('coding'), and be reported in a manner that allows for comparison across trials [219].

The aim of this paper is to provide guidance to researchers in respect of assessing and reporting MI treatment fidelity. The practical recommendations offered are important to consider in designing, developing and conducting research, including in grant applications.

The Motivational Interviewing Treatment Integrity (MITI) code

The Motivational Interviewing Treatment Integrity (MITI) code is the most frequently used [219] instrument for assessing MI fidelity in RCTs [101,228,233]. The MITI has been derived from the Motivational Interviewing Skill Code (MISC) [214], and while reducing the MISC's complexity and length [190], the MITI focuses exclusively on the verbal behaviour of the practitioner, and does not take client responses into account [233]. The MITI is continuously revised and improved. Almost ten years ago, MITI version 2.0 was being used, and at present, the MITI 4.1 has just been published. Definitions of variables that measure important aspects

of MI practice are improved in each new version, with the aim to carefully follow and progress developments in MI research. Every previous version of the MITI instrument [228] has been shown to be reliable [101,103,233,256] and valid [101,228]. The recently published MITI 4.1 has been shown to have face validity, but the collection of data regarding its validity and reliability is still underway.

In the recent MITI 4.1, substantial changes have been made in comparison to MITI 3.1.1. The authors claim that the two versions are not comparable, and advise researchers to use the MITI 4.1 from now on. However, before MITI 4.1 may be used more widely, the instrument has to show predictive validity at least in respect of some problem behaviours and in some languages, and coders using the new instrument need to be able to achieve an adequate inter-rater reliability. In order to assist researchers both in conducting treatment fidelity assessment in future research, and in interpreting research conducted so far, both versions of the MITI are briefly discussed in this paper.

The MITI comprises two separate components: global variables and behaviour counts. In both versions of the MITI, a 20-minute segment is used both for the behaviour counts and for rating the global variables. The global ratings reflect the coder's overall impression of how well or poorly a practitioner performed in a certain aspect of MI practice, rated on a five point Likert scale. In MITI 3.1.1, the global scales are Empathy, Evocation, Collaboration, Autonomy/Support, and Direction. In both MITI 3.1.1 and MITI 4.1, the Empathy rating captures how well the practitioner understands the client's perspective, experiences, and feelings. In MITI 3.1.1, the global variable MI Spirit combines the ratings of Evocation, Collaboration and Autonomy/Support (by taking an average of the ratings of all three variables). In MITI 4.1, the MI Spirit variable is replaced by the variables Cultivating Change Talk, Softening Sustain Talk, and Partnership. This modification emphasizes the importance of the practitioner adapting her behaviour in response to client utterances. A further change in respect of the global variables in MITI 4.1 is that the Direction variable has been removed.

The behaviour counts are intended to capture specific practitioner verbal behaviours that are relevant to good practice of MI. The MITI 4.1 retains the behaviour count categories Giving Information, Simple Reflections, and Complex Reflections. However, some other behaviour count categories have been changed in the MITI 4.1. First, the two categories Open and Closed Questions have been combined into one single Questions category. Second, the category for MI Adherent practitioner behaviour has been divided into several categories for subtypes of such behaviour, each of which is given a separate code: Seeking Collaboration, Emphasizing Autonomy, and Affirm. Third, the category for MI Non-Adherent practitioner behaviour in MITI 3.1.1 has been split up into Confront and Persuade (with and without permission) in MITI 4.1. For a more detailed discussion of the different variables in the two different versions of the MITI we refer to the MITI 3.1.1 manual [228] and the MITI 4.1 manual [230].

ASSESSING TREATMENT FIDELITY IN MI SESSIONS

Prior to the start of an RCT where MI is one of the interventions being tested, it is important to consider the following three things: 1) which samples of MI practice (sessions) will be collected and selected for fidelity assessment; 2) who will do the assessment (coding) of these sessions; and 3) how will the results be reported. These three essential questions will be discussed in detail below.

Collection of audio-recorded sessions

Since audio-recorded sessions are used to assess treatment fidelity – that is, what really happened in the interaction between practitioner and client – it is important to audio record all, or as much as possible, of the conducted conversations. Recording all sessions allows the researcher to minimize selection bias, which is easily introduced if practitioners are permitted to select the sessions submitted for treatment integrity assessment themselves [180]. Approval by the relevant ethics review board, and the consent of clients and practitioners, have to be obtained prior to the audio recording. Informing clients and practitioners that the data will be anonymized might make them less reluctant to consent. In addition, providing practitioners with digital audio recording devices (and checking compliance throughout the study) could assist in obtaining the full spectrum of conversations.

Selection of samples for assessment

A random representative sample of the collected audio-recorded sessions should be selected. It will often not be possible to assess the treatment fidelity of all sessions, but coding multiple work samples from each practitioner provides a more accurate assessment of his or her proficiency [153]. So the question is, how large should this representative sample be, keeping in mind that studies have different design in respect of the number of participating practitioners, the number of sessions per client, and so forth.

In previously conducted RCTs where attempts have been made to assess treatment fidelity, between 11-32% of the total number of sessions were assessed (e.g. [12] (25%); [72] (16%); [33](11%); [329] (28%); [90] (25%); [236] (23%); [274] (32%)). However, the study of Smith et al. (2012) is an exception to this since 100% of recorded sessions were assessed, although the total number of sessions in this study only comprised 38 [305]. In studies where the intervention was delivered by more than one practitioner, 10-17 sessions per practitioner were selected for assessment ([294], (n=17); [33] (n=10)) to represent a reliable overview of the quality of the individual practitioner throughout the study period. El-Mallakh et al. (2012) assessed 18 sessions (25% of total sessions) [90], and McCarthy et al. (2014) assessed 4 sessions (20% of total sessions) [203] throughout the study period (both with only one practitioner delivering the intervention), providing an indication of the MI skill fluctuation in the practitioner delivering the intervention over time.

Some studies require a comparison of overall group results (average of multiple practitioners), for example, when usual care conditions containing an attention control intervention without an MI component and an MI intervention condition are compared (e.g. [305]), or when practitioners with different backgrounds/experience are compared (e.g. [236] (n=19)). In the study of Smith et al. (2012), one practitioner delivered both the intervention and the control arm of the study [305]. Here, it was examined if MI was more pronounced in the intervention group than in the control group by assessing 20 intervention sessions and 18 control sessions.

Audio-recorded session may vary in length between 10 minutes and over an hour. The MITI is used to assess a 20-minute segment of each session. It may be the case that sessions shorter in length than 20 minutes could not be reliably coded using the MITI [256]. For longer sessions, it may be hard to decide how to choose the segment submitted for assessment, in particular since the quality of a practitioner's MI practice might fluctuate throughout a session. This was, for example, found in the psycholinguistic study of Amrhein et al. (2003), in which there was an explicit requirement for practitioners to agree a change plan with the client at the end of a face-to-face-session. In cases where the client was not yet ready to agree to a change plan, this led practitioners to adopt a directive (rather than a collaborative) approach with the client towards the end of their sessions, when the change plan had to be discussed. This resulted in low MI fidelity, despite the beginning of the sessions being at an adequate level of MI fidelity [10].

Different approaches may be adopted to decide which segment of a session should be coded. These range from decisions being made based on the content of sessions, where the coding of segments that are off-topic or which do not focus on the target behaviour is avoided, to the selection at random of segments to be coded. To improve generalizability, the random sample should comprise some segments from the beginning of sessions, some from the middle, and some from the end of sessions, so that the whole spectrum of sessions' content is captured. However, it may be advisable to avoid coding the very beginning and the very end of audio-recordings, since these will often contain off topic material and talk not related to the targeted behaviour change (e.g. information about the trial or scheduling of new appointments), and might therefore introduce bias.

Most RCTs run over a year or more, due to the recruitment phase and subsequent follow-up. Considering that practitioners will experience changes during this time, and that their level of MI skill may fluctuate [102], it is important to select sessions from different points in time throughout the entire intervention period, to get a complete and representative picture of practitioners' MI skill. In particular, if practitioners receive ongoing supervision or additional training, improvements in MI proficiency could be expected [72,194]. Moreover, the client also influences the session; therefore it is desirable to include sessions with different clients, who are at different stages of change, in the sample. For example, Noordman et al. (2013) found that nurses applied more MI skills when their patients were in the so-called preparation phase of behaviour change [242].

In most research trials more than one intervention session is delivered to participants. To get an accurate picture of the quality of the delivered MI across the whole intervention period, not only the first session with a participant should be selected, but subsequent ones as well.

CODING

Coders

Several questions will need to be considered before it is possible to assess MI fidelity. The first question is: Who will code the sessions? To the best of our knowledge, there are only a few permanent coding labs consisting of a group of coders and with established procedures for maintaining inter-rater reliability. These have been founded for the purposes of assessing treatment fidelity in research, for carrying out quality assurance of routine clinical practice, and for providing feedback to participants as part of MI trainings. These coding labs are MIC Lab (Karolinska Institutet, Stockholm, Sweden, director: Lars Forsberg), KoRus (Bergen, Norway, directors: Nina-Elin Andresen and Solveig Storbakken), and the research group that developed the MITI and MISC instruments, based at CASAA, University of New Mexico (Albuquerque, United States, leader: Terri Moyers). To involve one of these in one's research project may be the best option, since these coding labs strive to maintain a gold standard for inter-rater reliability.

However, there may be several reasons (including language related ones) that it may not always be feasible to involve an established coding lab in one's research projects. As an alternative in cases where this may not be possible, individuals associated with the MI Network of Trainers (the MINT network) in many countries code sessions as part of their MI trainings, and who might more easily be able to learn to code reliably. There are also at any given time likely to exist several non-permanent coding labs, established for the purposes of assessing treatment fidelity in particular on-going research projects, but which tend to be dissolved once these projects come to an end.

Earlier research has shown that training inexperienced individuals (e.g. students) to use the MITI is possible [72,144,236,305,329]. However, training people to use the MITI is time consuming (more than 40 hours). It is necessary to create an environment where (prospective) coders can discuss questions and uncertainties and reach a consensus when faced with difficulties during the coding process. To minimize the risk of drift among coders and to promote adherence to the MITI manual, it is advised to assess coders' competence and calculate their inter-rater reliability prior to permitting coding of the study sample to begin, and to only allow coders to begin to code the study material once they have reached an adequate inter-reliability (see below for recommendations regarding what an adequate inter-rater reliability is).

Inter-rater reliability

Assessing and reporting coders' inter-rater reliability is absolutely crucial, but something that has often been neglected in MI research [219]. In order for reported results to be reliable, it is advised to have (at least) one second coder. At present, there are no official recommendations regarding the proportion of sessions that should be coded by two coders (double-coded) for the purposes of assessing inter-rater reliability. Most previously conducted RCTs report that 4%-32% of sessions were double-coded (e.g. [144] (32%; n=11); [294] (4%; n=6); [202] (20%; n=19); [329] (8%; n=115); [72] (30%; n=54); [44] (10%; n unknown); [180] (25%; n unknown); [194] (27%: n=15)). However, in some studies, all sessions were coded by at least two of the coders [30,82,90,305]. Having a large proportion, or indeed all sessions, double-coded will of course increase the validity of the results, but will not always be feasible were the total number of audio-recorded sessions is large.

It is important that coders rate the exact same segment of each of the recorded sessions when inter-rater reliability is tested. This requires that the exact start and end time of each segment, along with the first and the last sentence of the segment, is noted and made available to coders prior to their coding. It is also helpful to make a note of utterances that could be interpreted differently (e.g. the subtypes of MI Adherent, the subtypes of MI Non-Adherent, and Complex Reflections), along with the time at which they occurred, since this will facilitate subsequent discussion among coders. It is also helpful for coders to make a note of observations that are relevant to the global ratings, such as things that affected their rating of a particular global variable either positively or negatively. This makes it easier for coders to provide concrete examples to justify their rating.

The MITI 4.1 manual [230] (as previous iterations of the MITI) advises that only audio-recordings are to be used for the purposes of the coding process. It is fine to use transcripts of sessions for the purposes of training coders (and indeed, it is probably necessary to do so at least in the beginning of coder training, since coders need to be able to discuss in detail nuances in each utterance in order to develop an understanding of how the variables are distinguished, etc.), but the use of (only) transcripts for the purposes of assessing treatment fidelity in sessions is clearly against the recommendations. Using audio-recordings only, rather than transcripts, saves much time (no time needed for transcription). More importantly, however, audio-recordings allow for vital aspects such as voice intonation, which are lost when sessions are transcribed, to be taken into account. This is crucial in respect of, for example, discriminating between a Reflection and a Question, or when deciding whether an utterance should receive a Persuade or Confront (MITI 4.1) or an MI Non-Adherent (MITI 3.1.1.) code. A quiet environment without distraction is all that is needed for the coding of audio-recorded sessions. The only disadvantage of audio-recordings could be the unblinding of the coder to the identities of the practitioner or the client, if these are persons familiar to the coder.

Study specific situations

In research trials, the MI intervention is sometimes accompanied by other study specific requirements (e.g. if weight is a study outcome, measuring weight progression at each session might be required), or study specific manuals that need to be followed or used in the interaction. Although such requirements are intended to assist in promoting behaviour change, they might impact on the quality of practitioners' MI practice [143]. If certain study-specific situations are not covered by the existing MITI manual, coders should determine how to code these deviations prior to beginning to code the study material [228]. There may also be other aspects that might not be covered by the MITI. For example, client encounters that serve as attention-control interventions may be difficult to rate in respect of the Direction and Evocation variables (in MITI 3.1.1), if they are non-directive and/or do not relate to a specified target behaviour (both these variables require a target behaviour [228]). For such encounters, it may be preferable to omit them. In MITI 4.1, this problem may arise in respect of the variables Cultivating Change Talk and Softening Sustain Talk, which both require that the coder is aware of the designated target behaviour in the interaction.

REPORTING

Reporting MITI results

MITI results are reported in a variety of ways in the literature. Some studies report results in respect of all MITI variables, while others present results according to the standard way of reporting MITI results, by providing outcomes for the global variables and the behaviour count summary scores. For materials coded using MITI 3.1.1, the standard approach would report the ratio of Complex Reflections to Simple and Complex Reflections, the ratio of Open Questions to Open and Closed Questions, the ratio of MI Adherent to MI Adherent and MI Non-Adherent utterances, and the ratio of Reflections to Questions, along with the scores for the global variables Empathy, MI Spirit and Direction. In respect of these behaviour count summary scores and global variables, there are recommended thresholds for Beginning Proficiency and Competency, which are based on expert opinion and in need of validation [228]. In the MITI 4.1 this approach would report the global components Technical Global (Cultivating Change Talk and Softening Sustain Talk), Relational Global (Partnership and Empathy), and the summary scores percentage of Complex Reflections (of all Reflections), the ratio of Reflections to Questions, Total MI adherent (Seeking Collaboration, Affirm, and Emphasizing Autonomy) and Total MI Non-Adherent (Confront and Persuade) [230]. In MITI 4.1, the recommended thresholds are for Fair and Good MI practice.

Some studies report the average MITI results for different interventions [72,194,236,305], others for different (groups of) practitioners [202,328], and some for individual practitioners over time [90,328]. In one study, the MITI results of two practitioners were weighed based on the number of participants each practitioner had counselled to obtain an overall MI fidelity for the study intervention [33]. In the study of Ang et al. (2013), the results for individual MITI

variables were not reported, but the proportion of audio-recorded sessions that reached the recommended threshold for Beginning Proficiency (in all MITI variables) was reported for both the MI and the control condition [12]. These differences in the reporting MITI results depend in part on the research question asked in various studies (whether the researchers are interested in improvements in MI skill over time, or in the respective efficacy or effectiveness of different interventions, or in differences in MI skill across practitioners, etc.).

In MITI 4.1, the authors encourage the full reporting of all MITI scores in clinical trials where MITI is used to assess treatment fidelity, since this data, when related to clinical outcomes, could provide empirical support needed to confidently establish recommended thresholds for Fair and Good MI practice, in particular with regard to the MI Adherent and MI Non-Adherent variables, where data is currently lacking and in respect of which no thresholds have been recommended [230].

Reporting inter-rater reliability

A wish for the future is that inter-rater reliability scores are presented in all scientific articles. It is striking and puzzling that this is not done routinely at the moment. Reaching a sufficient degree of inter-rater reliability is difficult and requires intensive training, collaboration and ongoing discussion, which might be a reason that these scores are not reported frequently. In the literature to date, the inter-rater reliability reported for different MITI variables has varied tremendously, ranging from 'poor' to 'excellent' (see below for thresholds) (e.g. [72,194]).

It is important to report the inter-rater reliability in each of the variables [125]. So for example, both the inter-rater reliability of the Complex Reflections behaviour count and the Simple Reflections behaviour count should be reported, and rather than just the inter-rater reliability of the coding of the behaviour count summary score for the ratio of Complex Reflection to Simple and Complex Reflections. This is because measurement errors could be overestimated in the summary scores. In MITI 3.1.1, this was a problem, in particular in respect of some variables that did not occur frequently, most notably the MI Adherent and MI Non-Adherent variables. For example, when the summary score for the ratio of MI Adherent to MI Adherent and MI Non-Adherent behaviour is calculated (dividing the total of MI Adherent utterances by the total of MI Adherent and MI Non-Adherent utterances), it could be the case that one coder did not detect any MI Adherent utterances, and therefore has summary score of 0%, while another coder who has detected at least one MI Adherent utterances might get a score of 100%. In this case, the inter-rater reliability score would be low for the summary score, even though it would have been acceptable for the raw variables. In MITI 4.1, this problem may have been evaded in respect of these variables, since the MI Adherent and MI Non-Adherent scores are a summation instead of a percentage.

Judging from the existing literature, it may be easier to obtain sufficient inter-rater-reliability in respect of some MITI variables than others. For example, it seems that coders more easily code

the global Direction variable and behaviour counts Giving Information and Reflections (e.g. [194] [329]) reliably. The global variable Autonomy/Support and the MI Adherent and MI Non-Adherent behaviour counts seem considerably more difficult for coders to reach agreement on. In respect of these, along with the Reflections to Questions ratio, the reported inter-rater reliability has ranged 'poor' to 'fair' in several studies (e.g. [329] [194]).

It is important to note that some MITI variables may be more important with regard to predicting outcome, and therefore also more important to be able to code reliably. For example, high Empathy scores and low levels of MI Non-Adherent behaviour may be predictive of successful client outcome [14,234]. Reporting the inter-rater reliability for those variables is therefore vital.

If the inter-rater reliability between coders is high, it is fine to report the results of the first coder only. If the inter-rater reliability is medium or low, one might consider using a more pragmatic approach, such as presenting the results of each coder separately, or presenting an aggregated score of both coders' ratings. Coders could decide to discuss their respective ratings and reach consensus in respect of the global scores for each of the variables, and to use arithmetic averaging for the behaviour counts. This method offers a pragmatic and practical solution based upon two coders that provides an indication of the level of MI fidelity.

Calculation of Inter-rater reliability

The most common way to calculate inter-rater reliability scores for the MITI global variables and behaviour counts is by intra-class correlation coefficients (ICC), using a two-way mixed model with absolute agreement [295]. ICC scores are generally compared against the following benchmark values [61]: 0.40 = poor; 0.40-0.59 = fair; 0.60-0.74 = good, and 0.75-1.00 = excellent.

Hayes and Krippendorff (2007) argue that the calculation of inter-rater reliability scores for ordinal variables (the global MITI variables) should not be done using Cronbach's alpha or percent agreement [134]. Cronbach's alpha is a statistic for interval-level data, which is not sensitive to the level of agreement in judgment, but only acts as quantification method of judgment [134]. Percent agreement is limited to nominal or categorical levels, can only be calculated for two coders, and there is no correction for the minimal chance of agreement of scoring the same variable [174]. Furthermore, it is proposed that Krippendorff's alpha (KALPHA) [174] is used instead of ICCs for calculating the (ordinal) global MITI variables. This is more suitable where data might be missing [125]. Additionally, a restricted range, which the global MITI variables have, reduces the utility of the ICC for assessing inter-rater reliability [294]. At the moment, KALPHA is frequently used in content analysis, although not in the context of treatment fidelity assessment. The KALPHA can be used regardless of the number of coders and levels of measurement, and can deal with missing scores [134]. KALPHA takes into account the prevalence of answer categories (and not the amount of existing categories), meaning that the

rarity of categories will impact KALPHA. As norm for good reliability testing a KALPHA of 0.80 has been suggested [78].

It is worth noting that variables that are generally scored across a smaller range of the intended scale (e.g. the Direction global scale in the MITI 3.1.1), on which practitioners tend to score in the high end of the scale, and for which a score in the lower end is rare, will end up with a lower KALPHA when a rare event is not detected (or rated in the same way) by all coders, even if the general agreement between coders is otherwise high [78]. One of the reasons that KALPHA has thus far not been widely used is that it does not form part of standard software packages, such as SPSS. For this reason, a specific KALPHA macro for SPSS [134], which can be downloaded here: http://www.afhayes.com/spss-sas-and-mplus-macros-and-code.html, and a guideline written by De Swert [78] have been made available to assist researchers in its use.

DISCUSSION

The assessment of treatment fidelity is a prerequisite for being able to distinguish between behaviour change interventions where the delivered intervention was proficient 'state of the art' MI, and those where the delivered intervention was not competent MI. This is necessary in order to be able to know what conclusions may be drawn from the results of research trials. The present overview of practical recommendations on different aspects of treatment fidelity assessment in RCTs – collecting, selecting, coding, and reporting of MI fidelity – will assist researchers in assessing treatment fidelity in future studies.

However, determining whether an intervention is "MI" might not be as straightforward as it may sound. There are several reasons for this. First, the recommended thresholds for Beginning Proficiency and Competency (in MITI 3.1.1; in MITI 4.1, these are referred to as Fair and Good MI practice) are based on expert opinion, and in need of further research to establish their empirical support [228]. So far, they serve as guidance only. Second, it is frequently the case that some of these thresholds are reached, while others are not. Since certain aspects of MI practice (such as a high Empathy rating, and a low degree of MI Non-Adherent utterances, a high degree of client change talk, and a low degree of client sustain talk [219]) are perhaps more important than others, we may perhaps be justified in attaching greater weight to whether practitioners reach the recommended thresholds in respect of the MITI variables related to these aspect.

We may also have to think about how we should deal with situations in which some practitioners reach the recommended thresholds while others do not (even though all practitioners received the same amount of training). It is known that MI skill can vary substantially across practitioners (e.g. [102]). Such inter-practitioner variation in MI fidelity may also hamper discrimination between the MI-based intervention and the control condition in research trials. In the studies conducted so far, effects have been analysed according to the randomised groups independent of individual practitioners' level of treatment fidelity.

However, perhaps an adjustment for 'low fidelity practitioners' in the analysis is needed, followed by a sensitivity analysis with exclusion of low fidelity practitioners.

Treatment fidelity assessment can be used to monitor the fidelity within clinical trials, but also to evaluate and supervise skill development in clinical practice at the same time. Using the MITI for supervision purposes during an RCT will help to improve practitioners' MI skill, but this may also influence the level of overall fidelity. This is something to take into consideration while evaluating an RCT. Furthermore, it may be the case that one or more practitioners fail to reach an acceptable level of MI skill while the trial is ongoing. Should such practitioners be prevented from counselling participants and receive additional training until they reach a sufficient level of MI skill? This would of course be inconsistent with a research aim what was to deliver the intervention in a manner that as closely as possible resembled actual clinical practice, and would perhaps not even be possible in implementation studies.

If future (process evaluation) studies consider measuring the fidelity of MI sessions, this could elucidate the actual fidelity level needed for MI to work and its specific effect on the outcome variable. Mediation analyses could indirectly assess the effect of MI on some outcome variable through a proposed mediator, thereby helping to entangle the working mechanisms of MI. As a result, implementation studies using MI in specific settings and contexts could benefit optimally from the findings of RCTs by implementing and focusing only on the 'essential components'.

In this manuscript, we advocate the use of the MITI instrument to assess treatment fidelity in order to achieve uniform reporting across RCTs, which would facilitate comparison across studies. However, several other instruments have been developed for the purposes of measuring MI treatment fidelity [190]. If researchers are, for example, more interested in identifying the active mechanisms in MI, they might consider selecting (a combination of) the following measurement tools [83]: Global Rating of MI Therapist (GROMIT) [229], the Sequential Code for Observing Process Exchanges (SCOPE) instrument [198] or the MI Skill Code (MISC) [116,214,235]. Even if researchers choose to use an instrument other than the MITI for assessing treatment fidelity, many of the aforementioned recommendations will still apply.

Conclusion and recommendations

We have aimed to provide an overview of practical recommendations, available best practices, and pragmatic solutions to common problems that researchers might come across in the collection, selection, coding, and reporting of MI fidelity data. We recommend certain practices in order to facilitate comparisons across studies where MI is used. For a comprehensive overview of our recommendations and considerations see Table 7.1.

Table 7.1: Recommendations and considerations for assessing and reporting MI fidelity

Aspect		Recommendation	Potential considerations
Collecting and selecting	Collection	Audio record all sessions (anonymous)	Is permission of both the client and practitioner requested?
	Selection	Select a random representative sample throughout the whole study period. Preferably 20% of the study sample, at least 4 conversations for each practitioner involved or 20 conversations per intervention group when only overall results will be reported	Which MI fidelity instrument will be used to assess the quality of MI? Will the coding take place during or at the end of the RCT? Will the practitioner receive feedback on his/her performance? Will sufficient MI-fidelity be an entry requirement for the practitioner?
Coding	Coders	Arrange trained coders beforehand or facilitate training for (inexperienced) coders and weekly meetings and assess reliability prior to coding the study sample	Is there enough budget for employing (trained) coders? Is there enough time for inexperienced coders to acquire competency? Are there resources available for the training of inexperienced coders? Is comparison with a gold standard from an experienced coding lab considered?
	Inter-rater reliability	Double-code 20% of the sessions by a second coder	Is there enough budget for employing a second coder?
	Study specific situations	Report how specific trial aspects were dealt with	Does the intervention comprise specific requirements that might influence MI quality or are part of the trial's fidelity?
Reporting	Reporting MITI results	Report raw and summary MITI results	Does the research question focus on fidelity of (1) (intervention) group results, (2) all practitioners separately, or (3) (a) practitioner(s) over time? What level of MI is believed competent?
	Reporting inter-rater reliability	Report for the raw and summary variables inter-rater reliability scores	How will the results be handled of both coders in case of a low reliability score?
	Calculation inter-rater reliability	Calculate KALPHA for ordinal scores and ICC for behavioural counts	