Clinical Psychology Review 82 (2020) 101909

Contents lists available at ScienceDirect

Clinical Psychology Review

journal homepage: www.elsevier.com/locate/clinpsychrev

Review

Motivational interviewing quality assurance: A systematic review of assessment tools across research contexts

Margo C. Hurlocker[a,*], Michael B. Madson[b], Julie A. Schumacher[c]

[a] Center on Alcoholism, Substance Abuse, and Addictions, Department of Psychology, University of New Mexico, 2650 Yale Blvd SE, Albuquerque, NM 87106, USA
[b] School of Psychology, University of Southern Mississippi, 118 College Dr, Hattiesburg, MS 39406, USA
[c] Department of Psychiatry and Human Behavior, University of Mississippi Medical Center, 2500 North State St, Jackson, MS 39216, USA

HIGHLIGHTS

- Valid tools are needed to assure quality Motivational Interviewing (MI) delivery.
- Observer-, trainee-, and client-rated tools of MI skills/fidelity are available.
- Tools vary in empirical strength across research contexts.
- Certain tools are more appropriate for MI training versus outcome studies.

ARTICLE INFO

Keywords:
Motivational interviewing
Fidelity
Provider skills
Research contexts

ABSTRACT

The need for sustained skill development and quality assurance when executing behavioral interventions is best demonstrated in the empirical evolution of Motivational Interviewing (MI). As a brief behavioral intervention that identifies the therapeutic process as an active treatment ingredient, it is critical for researchers, trainers, and administrators to use psychometrically sound and theoretically congruent tools to evaluate provider skills and fidelity when executing MI. Yet, no prior work has evaluated the breadth of MI tools employed across research contexts. Therefore, this review identified MI fidelity and skill development tools across measurement, training and efficacy/effectiveness studies and evaluated their psychometric strength and fit with current MI theory. We identified 199 empirical studies that employed an MI fidelity/skill tool and we found 21 tools with varying degrees of empirical support and theoretical congruence. Specifically, we identified five observer-, two trainee- and one client-rated tool with strong empirical support, and nine observer- and two client-rated tools with preliminary empirical support. We detailed the empirical strength, including the extent to which tools were linked to trainee/client outcomes across research contexts and offer recommendations on which MI tools to use in training, efficacy, and effectiveness trials.

1. Introduction

Motivational Interviewing (MI) is an evidence-based intervention designed to explore and resolve client ambivalence around change (Schumacher & Madson, 2014). Stand-alone MI and Motivational Enhancement Therapy (MET; an adapted form of MI that integrates personalized normative feedback to facilitate change) were among the first empirically-supported MI interventions and MET is the most widely used adapted form of MI (Hettema, Steele, & Miller, 2005). In the 40 years since the original description of MI (Miller, 1983), several meta-analytic reviews and hundreds of empirical studies have supported the efficacy and effectiveness of MI (e.g., Burke, Arkowitz, & Menchola, 2003; Lundahl & Burke, 2009; Magill et al., 2018). A critical feature of MI's success is how the therapeutic process, rather than the intervention content facilitates client motivation to change. This requires a clear yet complex provider skillset to ensure quality delivery of MI (Miller & Rollnick, 2013). As Miller and Rollnick (2009) have noted, this complex skillset is not easily acquired. A recent evaluation of MI fidelity across several training, efficacy, and effectiveness trials found providers in training studies had lower overall adherence than those in efficacy/effectiveness studies (Hallgren et al., 2018). This is not surprising given that, with few exceptions (e.g., Schumacher et al., 2018), MI training studies, as compared to MI efficacy and effectiveness studies do not specify a priori benchmarks for therapist proficiency.

[*] Corresponding author.
E-mail addresses: mhurlocker@unm.edu (M.C. Hurlocker), Michael.madson@usm.edu (M.B. Madson), jschumacher@umc.edu (J.A. Schumacher).

https://doi.org/10.1016/j.cpr.2020.101909
Received 2 April 2020; Received in revised form 20 July 2020; Accepted 31 August 2020
Available online 06 September 2020
0272-7358/

Overall, the literature examining MI training methods is vast and more prominent than is typical of other behavioral interventions (Miller & Rollnick, 2014). In fact, the breadth of empirical work positions MI training to serve as a prototype for how to promote a 'cycle of excellence' in treatment providers when delivering behavioral interventions. Madson, Schumacher, Baer, and Martino (2016) outlined best practices for training in MI to underscore the importance of methodological quality in training MI, aligning with three components of the cycle of excellence: establish baseline skill level, participate in systematic, ongoing feedback, and engage in deliberate practice (Miller, Hubble, & Duncan, 2007). Further, four systematic reviews (Barwick, Bennett, Johnson, McGowan, & Moore, 2012; Madson, Villarosa, Schumacher, & Mohn, 2016; Madson, Villarosa-Hurlocker, Schumacher, Williams, & Gauthier, 2019; Söderlund, Madson, Rubak, & Nilsen, 2011) and two meta analyses (de Roten, Zimmerman, Ortega, & Despland, 2013; Schwalbe, Oh, & Zweben, 2014) have identified critical components for training providers in MI, including the need for individual feedback and coaching to ensure adequate skill development.These findings are highly reflective of the compnents of deliberate practice: identifying one's areas for development using expert feedback, reflecting on this feedback, and implementing a plan for impvoement (Rousmaniere, Goodyear, Miller, & Wampold, 2017). In accordance with the science of expertise (Ericcson, 2009), meaningful feedback on complex skills and accurate evaluation of the therapeutic process relies on psychometrically sound instruments of MI fidelity and skills.

There are several tools available to assess MI skills and fidelity. Two prior reviews have provided evaluative information about tools that largely focused on MI fidelity – therapist adherence to the tenets of MI (Madson & Campbell, 2006; Wallace & Turner, 2009). Recent work has also provided psychometric and administrative information on tools that assess MI competency in the context of MI training (Gill, Oster, & Lawn, 2019). Adherence/competence tools are an important part of training to develop MI proficiency (Schumacher, Madson, & Norquist, 2011), facilitating deliberate practice (Rousmaniere et al., 2017), and evaluating the benefit of MI on client outcomes (Miller & Rose, 2009). However, the conceptualization of MI has evolved since these two prior reviews, requiring increased attention to the composition of therapist skills needed for quality MI delivery across research contexts. Specifically, the theory of MI outlines two active ingredients – technical and relational – that, when integrated effectively, facilitate client motivation to make personal changes that are consistent with desired goals (Miller & Rollnick, 2013; Miller & Rose, 2009). Yet, no prior studies have identified and evaluated the myriad of skill assessments used across MI training, efficacy, and effectiveness studies. Such a critical analysis is needed given that (a) certain tools may be better suited for training (e.g., competency development) versus efficacy/effectiveness studies (e.g., therapeutic process), (b) selection of a given tool should depend on its empirical strength (i.e., psychometric strength and linked to client/trainee outcomes) and fit with MI theory, and (c) commonly used tools pose implementation challenges in clinical practice (e.g., objective tools require providers to record therapy sessions).

The current review is the first to identify and describe the breadth of MI measures used to evaluate provider fidelity and/or skill development across MI measurement, training and efficacy/effectiveness studies. In addition to updating the empirical strength of tools identified in prior reviews (Gill et al., 2019; Madson & Campbell, 2006; Wallace & Turner, 2009), we identify and describe tools that evaluated MI skills in diverse formats (e.g., machine-based models; trainee−/client-completed tools). The primary goal of this review is to evaluate the measurement of MI skills and fidelity and identify those tools with strong empirical evidence and that are congruent with current MI theory and practice.

2. Methods

2.1. Screening procedure

We reviewed several sources for eligible articles to include in our systematic review. First, we searched all articles that cited the Madson and Campbell (2006) or Wallace and Turner (2009) reviews (42 articles). We then conducted a literature search from January 2007 to December 2019 of the following electronic databases: psycINFO, Health and Psychosocial Instruments, and Medline. Keywords used in our literature search include motivational interviewing, motivational enhancement therapy, therapist fidelity, adherence and competence, therapist skills, therapist competence, and technology transfer. English-language articles were included if they directly stated that they used an objective and/or a trainee−/client-report measure to evaluate MI skills or fidelity. Given our attention to articles that evaluated *provider* MI skills in measurement, training, and efficacy/effectiveness studies, we excluded articles that (a) only described the methodological protocol (2 articles), (b) performed qualitative but not quantitative analyses (8 articles), (c) evaluated skills across multiple psychosocial interventions (15 articles), or (d) the measure was not used to evaluate provider skills (22 articles). Our search procedures resulted in 27 articles from the two prior reviews and 173 articles from our systematic search ($N = 200$ articles; see Fig. 1). Across these 200 articles, 21 measures fit the inclusion criteria and were evaluated for methodological quality and scientific rigor.

2.2. Evaluation procedure

The authors developed a measure rating form to evaluate the methodological approach and psychometric properties of each measure. Descriptors on the rating form were categorized based on article type: measurement, training, or efficacy/effectiveness. Across all article types, raters evaluated whether psychometric analyses were performed. Additional descriptors by article type included the description, items, and scoring of the measure (for measurement articles), the description of the MI training, how the measure was used, and if the measure was connected to trainee outcomes (for training articles), and the description of the MI intervention, how the measure was used, and whether the measure was connected to client outcomes (for efficacy/effectiveness articles). Two doctoral-level psychology students were trained by the first author (MH) to independently rate the MI skill/fidelity measures. Training entailed 24 h of didactic discussions and practice ratings. Trainers reached 89% consistency with the first author in rating practice articles prior to rating articles included in the systematic review. A random subsample of 19 identified articles (10%) was double-coded with the measure rating form to assess interrater reliability. Kappa was 0.72 indicating good reliability (Cicchetti, 1994).

3. Results

The majority of identified articles were efficacy/effectiveness studies ($n = 94$), followed by training studies ($n = 73$), and measurement development studies ($n = 33$). Although these studies vary in their reported purpose, we focused on whether authors (1) provided an adequate description of the measure and how it was used, and (2) reported psychometric properties (see Appendix, Table 1). Across studies, 16 observer-rated tools, 2 trainee-completed tools, and 3 client-rated tools were employed to evaluate MI skills or fidelity. Of note, 16 articles evaluated provider skills using multiple measures, and thus separate rating forms were completed for each measure. Table 1 describes how measures were used across training and efficacy/effectiveness studies. Below we described measures based on respondent type and empirical strength.

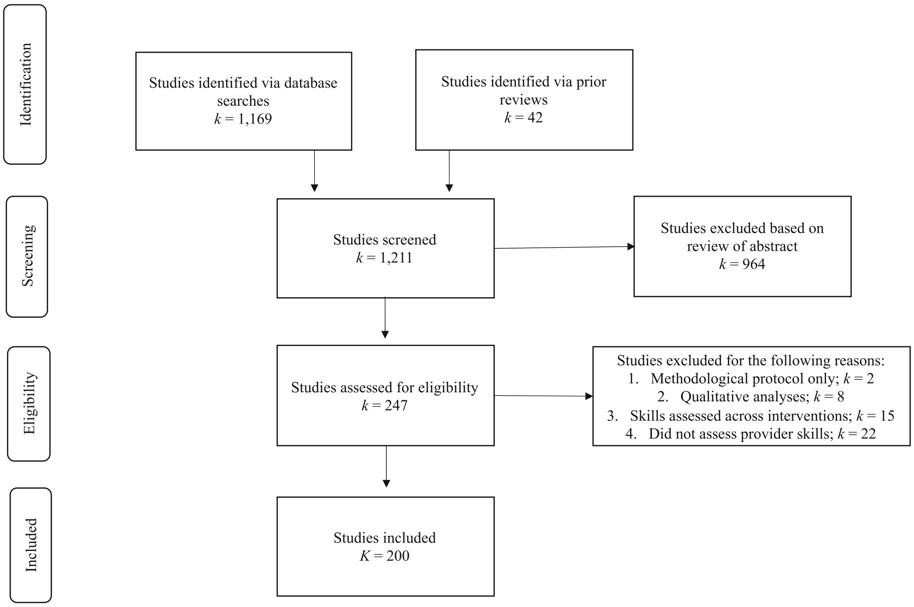
Identification
Studies identified via database searches
k = 1,169
Studies identified via prior reviews
k = 42
Screening
Studies screened
k = 1,211
Studies excluded based on review of abstract
k = 964
Eligibility
Studies assessed for eligibility
k = 247
Studies excluded for the following reasons:
1. Methodological protocol only; *k* = 2
2. Qualitative analyses; *k* = 8
3. Skills assessed across interventions; *k* = 15
4. Did not assess provider skills; *k* = 22
Included
Studies included
K = 200

Fig. 1. PRISMA diagram of article identification and inclusion procedure.
Note. K/k = number of studies.

3.1. Observer-rated tools with strong evidence

In accordance with best practices (Madson, Schumacher, et al., 2016), tools that use a trained rater to observe an MI session and rate provider MI skills were the most commonly employed across identified articles. The Motivational Interviewing Skills Code (MISC) and the Motivational Interviewing Treatment Integrity (MITI) system were the most frequently used observer-rated tools (MISC: 31 articles; MITI: 111 articles). Additional tools with strong empirical support include the Behavior Change Counseling Index (BECCI; 12 articles), the Yale Adherence Competence Scale (YACS; 10 articles), and the Independent Tape Rater Scale (ITRS; 9 articles). The reported reliability estimates of these tools are provided in Appendix, Table 2.

3.1.1. Motivational Interviewing Skill Code (MISC)

The MISC was the first behavioral coding system used to evaluate provider fidelity, client behavior, and the provider-client interaction during an MI session (Miller & Mount, 2001). The MISC has undergone several revisions and the most recent publicly available version is the MISC 2.5 (http://casaa.unm.edu/download/misc25.pdf) which integrated aspects of the MISC 2.1 (Miller, Moyers, Ernst, & Amrhein, 2008) and the MI-SCOPE (Moyers & Martin, 2006) to more accurately evaluate the subtleties of the MI process (Houck, Moyers, Miller, Glynn, & Hallgren, 2010). For provider behavior, the MISC 2.5 comprises six global dimensions (i.e., acceptance, empathy, direction, autonomy support, collaboration, and evocation) that are rated using a five-point scale and 17 behavior count categories that generally map onto prescribed (e.g. affirm) and proscribed behaviors (e.g., advise). Some of the categories have subcategories to better capture the nuances of MI (e.g.,

Table 1
Methodological descriptors of MI tools used in training and efficacy/effectiveness studies.

Scale	Training studies				Efficacy/Effectiveness studies			
	k	Training described	Feedback & coaching assessed	Linked to trainee outcomes	*k*	Intervention described	Behaviors assessed	Linked to client outcomes
MISC	3	Yes = 2	No = 3	Yes = 3	2	Yes = 25	Provider = 26	Yes = 21
		No = 1			6	No = 1	Client = 19	No = 5
MITI	47	Yes = 42	Yes = 21	Yes = 45	5	Yes = 49	Provider	Yes = 14
		No = 5	No = 26	No = 2	2	No = 3		No = 38
BECCI	9	Yes = 7	Yes = 4	Yes = 7	3	Yes = 3	Provider	Yes = 1
		No = 2	No = 5	No = 2				No = 2
ITRS	6	Yes = 6	Yes = 5	Yes = 6	3	Yes = 3	Provider	Yes = 2
			No = 1					No = 1
YACS	3	Yes = 3	Yes = 1	Yes = 3	7	Yes = 7	Provider	No = 7
			No = 2					
MI-SCOPE	–	–	–	–	4	Yes = 4	Provider = 4	Yes = 4
							Client = 4	
AMIGOS	–	–	–	–	1	Yes = 1	Provider	No = 1
DDMI	–	–	–	–	1	Yes = 1	Provider	Yes = 1
VASE	3	Yes = 3	No = 3	Yes = 3	–	–	–	–
HRQ	5	Yes = 5	No = 5	Yes = 5	–	–	–	–
TSR	–	–	–	–	1	Yes = 1	Provider	Yes = 1

Note. MISC = Motivational Interviewing Skills Code; MITI = Motivational Interviewing Treatment Integrity; BECCI = Behavior Change Counseling Index; ITRS = Independent Tape Rater Scale; YACS = Yale Adherence and Competence Scale; MI-SCOPE = Motivational Interviewing Sequential Code for Observing Process Exchanges; AMIGOS = Assessment of Motivational Interviewing Groups Observer System; DDMI = Dual Diagnosis for Motivational Interviewing; VASE = Video Assessment of Simulated Encounters; HRQ = Helpful Response Questionnaire; TSR = Therapy Session Report.

valenced reflections; Miller & Rollnick, 2013). While the MISC rating system allows for a comprehensive picture of the provider-client interaction, raters are required to listen to the entire MI session three separate times to complete the coding. Thus, the MISC is time-intensive and may be an impractical tool for MI training. In fact, our review suggests the MISC is often used to evaluate the MI process (80% of articles were efficacy/effectiveness studies).

3.1.1.1. Psychometric properties. In prior reviews, early versions of the MISC were evaluated, demonstrating variable reliability estimates (i.e., excellent estimates for global ratings and poor-to-good estimates for behavior counts), preliminary evidence of predictive validity, and little evidence of construct validity. In the current review, we identified 31 articles that used the MISC to evaluate provider MI skills: 2 were psychometric development studies, 3 were training studies, and 26 were efficacy/effectiveness studies. Unfortunately, none of the 7 studies that used the MISC 2.5 validated the tool. Despite the lack of construct validity, there is evidence to support the internal structure, inter-rater reliability, and predictive validity of the MISC. Consistent with recommended best practices (Madson, Schumacher, et al., 2016), most studies (81%, $k = 25$) reported some form of reliability estimate of the MISC.

3.1.1.2. Alternatives of the MISC. We identified three studies that validated a machine-learning approach to code sessions with the MISC. Atkins, Steyvers, Imel, and Smyth (2014) compared a labeled topic model coding method to human raters and demonstrated strong sensitivity and specificity of the topic model (AUC scores: 0.62–0.81) and poor-to-good reliability estimates with human raters (model ICCs: 0.10 [%CR]-0.68; human ICCs: 0.52–0.86). Tanana, Hallgren, Imel, Atkins, and Srikumar (2016) evaluated two language processing models – discrete sentence features (DSF) and recursive neural networks (RNN). The DSF model performed better than the RNN model, but both models had high consistency with human raters on therapist behaviors at the utterance-level (Ks > 0.50 for all behaviors except simple and complex reflections [0.30 < Ks < 0.50]) and the session-level (ICCs > 0.75 for all behaviors except for confront, structure, and advise with/without permission [ICCs < 0.40]). Imel et al. (2019) expanded on these two studies by employing a machine-learning feedback system to offer provider immediate feedback on specific MI skills. In comparison to human raters, the authors found poor-to-good reliability estimates (0.23 [empathy] < ICCs < 0.80).

3.1.2. Motivational interviewing treatment integrity (MITI)

The MITI is a behavioral coding system that evolved from the MISC to more efficiently assess MI fidelity. A 20-min segment of an MI session is rated with a focus on the provider fidelity to the foundational aspects of MI (Moyers, Martin, Manuel, Hendrickson, & Miller, 2005). Thus, the MITI addresses concerns about the practicality of the MISC (e.g., time commitment), particularly when training new providers in MI (Madson & Campbell, 2006). Specifically, trainers can use the MITI to evaluate changes in skills during an MI training and/or to facilitate the feedback and coaching portion of an MI training. The MITI has undergone several revisions and the current publicly available version is the MITI 4.2 (https://casaa.unm.edu/download/MITI4_2.pdf). The MITI comprises four global ratings (i.e., cultivating change talk, softening sustain talk, partnership, and empathy) that are rated using a five-point scale and 10 behavior count categories that are indicative of MI-adherence (e.g., emphasizing autonomy) and MI non-adherence (e.g., confront).

3.1.2.1. Psychometric properties. Prior reviews offered promising evidence that the MITI is psychometrically sound, highlighting fair-to-excellent reliability estimates and strong convergence with the MISC (Moyers et al., 2005). In the current review, the MITI was the most frequently used tool to evaluate fidelity across training studies (47 of 72 articles; 65.3%) and efficacy/effectiveness studies (52 of 99 articles; 52.5%). Given the MITI 4.2 was published in 2016, most identified articles used the MITI 2.0 ($k = 40$), the MITI 3.0 ($k = 28$), or the MITI 3.1 ($k = 39$) to evaluate MI fidelity. Although most training studies used the MITI to evaluate trainee outcomes, 15 studies (31.9%) did not report reliability estimates and 26 studies (55.3%) that had a feedback/coaching component did not use the MITI, or any other observational tool for this purpose. Among efficacy/effectiveness studies, 23 (44.2%) did not report reliability estimates and 37 (71.2%) did not link the MITI to provider or client outcomes. We also identified 12 studies (11.4%) that evaluated the psychometric properties of different MITI versions, three of which demonstrated support for adaptions of the MITI for Swedish (Forsberg, Berman, Kallmen, Hermansson, & Helgason, 2008; Forsberg, Kallmen, Hermansson, Berman, & Helgason, 2007) and German (Brueck et al., 2009) samples. With the MITI 4.2, we found two validation studies and two efficacy/effectiveness studies. The authors established content validity (e.g., expert panel) and demonstrated good-to-excellent reliability estimates for MITI 4.2 global ratings, behavior counts, and summary scores and strong criterion validity. The strong evidence for the MITI 4.2 is promising given that reliability estimates of prior MITI versions have been variable (see Appendix, Table 2).

3.1.2.2. Alternative of the MITI. We identified one study that validated a software to code MI interactions using the MITI 2.0 (Klonek, Quera, & Kauffeld, 2015). When comparing ICCs between the software and paper-pencil versions, the authors found software estimates that were either comparable (e.g., 0.72 and 0.70 for MIA behaviors) or superior (e.g., 0.91 and 0.53 for complex reflections). They also demonstrated convergent validity with the Rating Scales for the Assessment of Empathic Communication (REM). Finally, the software version captured an accurate estimate of behavior count frequencies of an entire session with a 10-min segment, as compared to a 20-min segment required of the paper-pencil version of the MITI.

3.1.3. Behavior change counseling index (BECCI)

Lane et al. (2005) developed the BECCI to assess practitioner competence in behavior change counseling (BCC) – an adaptation of MI intended for brief consultations in healthcare settings. Like MI, the goal of BCC is to talk with the patient about behavior change; however, BCC lacks the strategic aspects of MI such as eliciting change talk (Rollnick et al., 2002). The BECCI functions as a time efficient checklist and comprises 11 items that are measured on a 5-point Likert-type scale ranging from 0 (*not at all*) to 4 (*to a great extent*), depending on the frequency or strength of the given behavior. Example items include: *practitioner invites the patient to talk about behavior change*, and *practitioner and patient exchange ideas about how the patient could change current behavior*. In the current review, the BECCI was mostly used to evaluate training outcomes (9 articles; 75%) followed by efficacy/effectiveness studies (3 articles; 25%). Though most training studies connected the BECCI to trainee outcomes (7 studies; 77.8%), only 4 studies used the BECCI to provide feedback/coaching (44%) and 6 studies (66.7%) reported reliability estimates. Whereas all three efficacy/effectiveness studies reported reliability, no study linked the BECCI to client outcomes.

3.1.3.1. Psychometric properties. In this review we did not identify any articles that psychometrically validated the BECCI. Thus, the main source for evaluation is Lane et al. (2005) who found variable internal consistencies (0.63 < αs < 0.71) and good inter-rater reliability estimates (0.79 < *Rs* < 0.93). Though no studies in the current review evaluated validity, 8 studies (66.6%) reported interrater reliability, which ranged from adequate-to-excellent, or internal consistency, which ranged from poor-to-good (see Appendix, Table 1). Despite limited psychometric evaluation, the BECCI shows promise in being used reliably as a training tool for assessing treatment integrity of BCC.

3.1.4. Yale adherence and competence scale (YACS)

The YACS is a 50-item measure designed to evaluate clinician adherence and competence in implementing interventions common among most therapies, as well as interventions associated with specific therapy modalities (Corvino et al., 2000). The instrument includes three common interventions subscales (Assessment, General Support, and Goals for Treatment) and three modality-specific subscales (Clinical Management, Twelve-Step Facilitation, and Cognitive-Behavioral Management). For each item, raters judge both adherence to and quality of implementation. Frequency ratings range from 1 (*not at all*) to 7 (*extensive*), and ratings of quality range from 1 (*very poor* - therapist handled this in an unacceptable even toxic manner) to 7 (*excellent* - demonstrated real excellence and mastery in this area).

3.1.4.1. Psychometric properties. We identified 10 articles that assessed the utilization and psychometric properties of the YACS or an adapted version: 2 training studies and 8 efficacy/effectiveness studies. Though both training studies linked the YACS to trainee outcomes, neither study reported reliability estimates. However, one study demonstrated validity of an adapted YACS with the BECCI and MITI (Dray, Gilchrist, Singh, Cheesman, & Wade, 2014) and the other study used the YACS for feedback/coaching purposes (Marin-Navarrete et al., 2017). Across efficacy/effectiveness studies, three did not report reliability estimates and five did not link the YACS to client outcomes. Among studies that reported reliability estimates, most studies found adequate-to-excellent ICCs for adherence and competence (see Appendix, Table 1).

3.1.5. Independent tape rater scale (ITRS)

The ITRS was adapted from the YACS to evaluate therapist adherence and competence in delivering MI in real-world, community-based settings (Martino, Ball, Nich, Frankforter, & Carroll, 2008). The ITRS consists of 39 items that focus on therapist utilization of MI-consistent (e.g., open questions, affirmations) and -inconsistent (e.g., direct confrontation) techniques, as well as substance abuse counseling skills using a MI fidelity monitoring scale. Further, the 10 items that evaluate MI-consistent behaviors can be broken down into fundamental (e.g., reflections) and advanced (e.g., heightening discrepancy) MI skills. Evaluators rate each item using a seven-point Likert scale based on two dimensions including therapist adherence (1 = *not at all* to 7 = *extensively*) and competence (1 = *very poor* to 7 = *excellent*). Highlighting the complexity and impracticality of using some of the more established fidelity measures (i.e., MITI, MISC, and MISTS), the researchers wanted to create a tool that was more conducive to community-based programs. Specifically, few researchers have examined client outcomes using the more established tools, thus questioning their clinical utility in practice (Martino et al., 2008).

3.1.5.1. Psychometric properties. In the current literature review, nine articles were identified that assessed the utilization and psychometric properties of the ITRS: 6 training studies and 3 efficacy/effectiveness studies. Whereas all six training studies linked the ITRS to trainee outcomes, two studies did not report reliability and one study did not use the ITRS for feedback/coaching purposes. Alternately, all three efficacy/effectiveness studies reported reliability estimates but one did not link the ITRS to client outcomes. Three of the articles validated the ITRS using a confirmatory factor analysis and found strong construct validity for the subscales, one of which found strong support for a Spanish version of the ITRS (Santa Ana et al., 2009).The seven studies that reported reliability demonstrated fair-to-strong inter-rater reliability on both therapist adherence and competence across subscales (see Appendix, Table 1). Finally, six articles provided patient outcome data, which is less frequently reported for the YACS, highlighting the practical value of the ITRS.

3.2. Trainee- and client-completed tools with strong evidence

We identified two trainee-completed tools and one client-rated tool in the current review that have demonstrated empirical strength, particularly in the context of MI training. The reported reliability estimates of these tools are provided in Appendix, Table 2.

3.2.1. Helpful responses questionnaire (HRQ)

The HRQ is an open-ended, empirically validated tool of therapeutic empathy (Miller, Hedrick, & Orlofsky, 1991). The HRQ contains six separate vignettes of individuals disclosing a specific problem and participants provide a helpful response of no more than two sentences after each vignette. Responses are rated on a 5-point scale of depth and accuracy of reflection, as well as rated for the absence or presence of open-ended vs. closed-ended questions and communication roadblocks (Miller et al., 1991). A rating of “1” indicates a response that contains no reflection and at least one “communication roadblock” response and a rating of “5” indicates a response that contains a reflection of feeling or accurate metaphor/simile. Miller et al. (1991) found excellent interrater reliability and satisfactory internal consistency at pre- and post-training, but test-retest reliability was low.

3.2.1.1. Psychometric properties. We found six studies that used the HRQ for MI training purposes and offer additional support for its empirical strength. Though none of the studies used the HRQ for feedback/coaching purposes, all of the studies used the HRQ to link scores to trainee outcomes and reported some type of reliability. Across studies, the HRQ demonstrated adequate internal consistency and good-to-excellent interrater reliability estimates. While direct observation is the ideal approach to assessing therapeutic empathy, the HRQ shows promise as an alternative, empirically supported measure of therapeutic empathy (Miller et al., 1991). Childers et al. (2012) pointed out that, due to the written response format, it is difficult for the HRQ to effectively capture more complex conversational strategies and recommended that future researchers implement other validated measures of MI fidelity (e.g., MISC or MITI).

3.2.2. Video assessment of simulated encounters revised (VASE-R)

The VASE-R is an 18 item tool for assessing specific MI related skills such as reflective listening, responding to resistance, and sustain and change talk (Rosengren, Baer, Hartzler, Dunn, & Wells, 2005). Participants are asked to provide answers in an open response format to each prompt related to a client statement (e.g., “*write a response that indicates that you are listening*”). Responses are time limited (e.g., 60–90 s) in an attempt to simulate the response time in a typical clinical interaction. There are also multiple choice items in which participants “select the question or statement that you think would be most helpful to explore with Lisa right now, if you wanted to increase her motivation to change.” All responses are scored on a 3-point scale (0 = *MI-inconsistent responses* to 2 = *MI-consistent responses*) based on specific scoring rubrics as outlined in a rater manual. Higher scores indicate better use of MI consistent skills (Rosengren et al., 2005).

3.2.2.1. Psychometric properties. We identified four studies that employed the VASE-R in the current review: one psychometric validation study and three MI training studies. Whereas the training studies linked the VASE-R to trainee outcomes, only one of the training studies reported reliability, and none of the studies used the tool for feedback/coaching. However, the validation study used two independent samples and found acceptable-to-excellent interrater reliability estimates for the full-scale and subscale scores, unacceptable-to-adequate alpha coefficients, and strong concurrent validity with the HRQ total score and the MITI summary scores, except percent complex reflections. The authors also recommended score benchmarks for classifying trainees (i.e., untrained, beginning proficiency, and expert proficiency).

3.2.2.2. Alternatives of the HRQ and VASE-R. We identified one study that used the VASE-R and the HRQ to develop a, respectively, video- and a written-assessment of simulated encounters for school-based settings (VASE-SBA and WASE-SBA; Small, Lee, Frey, Seeley, & Walker, 2014). The authors outlined their iterative process in developing the tools, including feedback from an expert panel and pre−/pilot-testing. The authors found acceptable interrater reliability estimates ($0.54 > \text{ICCs} > 0.99$), adequate alpha coefficients ($0.71 > \alpha s > 0.81$), and sensitivity to training effects for the adapted HRQ and four of the six adapted VASE-R subscales. They also found strong concurrent validity between the two adapted tools. We also identified one study that used the VASE to develop a web-based program to assess trainee skills - Computer Assessment of Simulated Patient Interviews (CASPI; Baer et al., 2012). The authors found fair-to-good inter-rater reliability (ICCs: 0.46–0.84), test-retest reliability (*r*: 0.69–0.80), and criterion validity with the MITI ($r = -0.47$–0.62) and the HRQ ($r = 0.60$) as well as the ability to discriminate between those trained and not trained in MI. The CASPI is clinically and empirically useful given that it can be accessed from any internet-connected computer and it has two versions to allow for repeated administrations.

3.2.3. Client evaluation of motivational interviewing (CEMI)

The CEMI (Madson, Bullock, Speed, & Hodges, 2009) is an 11-item measure that assess client perceptions of clinician MI use in a session to assist in quality assessment and as a source of feedback in treatment and training settings and possibly as a measure of MI adherence in research. Clients complete the CEMI following a session where the intention was to use MI, or an adaptation of MI (e.g., motivational enhancement therapy). Clients are asked "During your most recent counseling session how much did your clinician [demonstrate each behavior]" using a four point Likert type scale (1 = *never* to 4 = *a great deal*). Behaviors rated include, "act as an authority on your life" and "help you talk about changing your behavior." Negative items are reverse scored and higher CEMI scores represent more MI consistent behavior. The CEMI is an emerging measure for assessing client perceptions of MI yet more evaluation is needed.

3.2.3.1. Psychometric properties. We identified five psychometric studies of the CEMI, including the original development of the CEMI (Madson, Bullock, et al., 2009), and one efficacy/effectiveness study. Across these studies, the number of items on the CEMI reduced from 35 to 11 items and four of the studies that performed factor analyses confirmed a two-factor structure of the CEMI: Relational and Technical. All studies found good internal consistency for the technical subscale and poor-to-good for the relational subscale. Further, one study demonstrated criterion validity of the CEMI subscales with the Working Alliance Inventory (Tracey & Kokotovic, 1989), and also found that the CEMI-Technical subscale partially mediated participants' improvement in readiness to change. However, no statistically significant relationships were found between the CEMI subscales and the MITI summary scores, though all of the relationships were in the expected direction.

3.3. Observer- and client-rated tools with little evidence

Several additional tools were identified in our search that demonstrated preliminary empirical evidence. Specifically, we identified nine observer-rated and two client-rated tools that we briefly describe below. A description of each tool and the existing psychometric support is provided in Table 2.

3.3.1. MI sequential code for observing process exchanges (MI-SCOPE)

The MI-SCOPE assesses transition probabilities between in-session therapist behavior and client language (Moyers & Martin, 2006). The MI-SCOPE assesses provider behaviors that are derived from the MISC 2.0 and is publicly available (https://casaa.unm.edu/download/scope.pdf). The original article reported adequate Kappa indexes for provider behaviors ($0.66 < K < 0.82$) but noted the lack of reliability among raters with parsing. We identified four studies that used the MI-SCOPE in efficacy/effectiveness studies. All studies reported psychometric properties and linked the MI-SCOPE to client outcomes. The authors found fair-to-excellent reliability estimates for parsing ($0.84 < K < 0.95$), utterance-to-utterance ($0.56 < K < 0.98$), behavior count frequency ($0.49 < \text{ICC} < 0.99$), and transition probabilities (ICC for Yule's $Q = 0.54$; $0.70 < k < 0.72$). Despite improved performance of the MI-SCOPE, more psychometric analyses on parsing and transition probabilities is needed.

3.3.2. Motivational interviewing assessment scale (MIAS) (originally the "Escala de valoracion de la entrevista motivacional" [EVEM])

The EVEM (in Spanish) was developed by an interdisciplinary group of physicians in Spain and designed as a brief, practical tool for general practitioners (Perula et al., 2012). The MIAS is an English version of the EVEM and is available to use free of cost. We identified one study that evaluated the psychometric properties of the MIAS. Using an iterative process to develop and validate the MIAS, the authors demonstrated a two-factor structure (directional and relational) and strong convergent validiy with the BECCI. The authors also found good internal consistency (0.97–0.99), variable Kappa indexes (e.g., 37.5% fair; 21.4% almost perfect), and variable ICCs (i.e., 31% poor-to-fair; 40% good-to-excellent). While initial psychometric properties are promising, additional research is warranted prior to implementing the tool in practice.

3.3.3. Content adherence scale (CAS) and global rating of motivational interviewing therapist (GROMIT)

We identified one article that developed the CAS and used the GROMIT (Moyers, 2004) to evaluate therapist adherence and competence in delivering a Brief Intervention for adolescents with alcohol abuse and violent behaviors. While the GROMIT was previously developed, no researchers have evaluated the GROMIT in an efficacy/effectiveness study. The CAS builds on the YACS (Carroll et al., 2000) and examines therapist accuracy in delivering intervention content. The one identified study found fair-to-good reliability for the CAS (*r: 0*.56–0.87) and the GROMIT (*r: 0*.42–0.89), and strong construct validity. Due to the novelty of these tools, more research is needed regarding their utility and psychometric soundness.

3.3.4. Rating scales for the assessment of empathic communication in medical interviews (REM)

We identified one article that developed the REM to examine physician empathic and confrontational behaviors during patient interactions. The researchers found support for a two-factor model of empathic communication ($\alpha = 0.93$) and confrontation ($\alpha = 0.87$) and strong convergent validity between the REM subscales and the MITI global scores (*r:* - 0.28 [MI spirit]-0.85 [MI empathy]) and the BECCI total score (*r*: 0.91 [REM-empathy]; −0.46 [REM-confrontation]). The researchers also found strong inter-rater reliability across three levels of empathy (*r:* 0.89 [high], 0.87 [medium], and 0.82 [low]).

3.3.5. Peer proficiency assessment (PEPA)

The PEPA was developed to evaluate peer counselor adherence to delivering MI in undergraduate students. Using the MITI as a framework, the PEPA focuses on the frequency of MI-consistent behaviors used among undergraduate peer counselors. The developer article (Mastroleo, Mallet, Turrisi, & Ray, 2009) was the only identified article in our review. The PEPA demonstrated good inter-rater reliability for open questions ($r = 0.97$) and complex reflections ($r = 0.89$) as well as construct validity with the MITI MI-adherent scores (*r*: −0.41 to −0.03 [closed questions]; *r:* −0.44 to −0.06 [simple reflections]). The PEPA also demonstrated predictive validity with client drinking outcomes ($r = 0.87$). While promising, more research is needed to validated the PEPA and examine patient outcomes.

Table 2
Description and psychometric properties of MI skill assessments with little empirical evidence.

	Description	Items (N)	Subscales	Reliability			Validity		
				α	ICC/K	*r*	Content	Construct	Criterion
Observer-rated									
MI-SCOPE	Tool to assess transition probabilities between therapist behavior and client language	30	1. MI-consistent 2. MI-inconsistent 3. Question 4. Reflect 5. Other	–	X	–	–	–	–
MIAS	Tool to assess practitioners' MI skills	14	1. Directional 2. Relational	X	X	–	X	X	X
CAS	Tools to evaluate therapist fidelity to content	15	1. Beginning session 2. Middle session 3. End session	–	X	–	–	X	–
GROMIT	Tool to evaluate therapist skill, responsiveness, and competence	16	1. Empathic style 2. Empowerment and Negotiation	–	X	–	–	X	–
REM	Tool to evaluate physician empathic communication	9	1. Empathic communication 2. Confrontation	X	X	–	–	X	X
PEPA	Tool to evaluate peer-counselor MI adherence	4	1. Question 2. Reflection	–	–	X	–	X	X
AMIGOS	Tool to evaluate provider MI skills in group format	20	1. MI group strategies 2. Group processes 3. Leader-related tasks	X	X	–	–	X	–
DDMI	Tool to evaluate fidelity to dual diagnosis MI intervention	–	1. Adherence (MIC, MIIN, Gen) 2. Competence (MIC, MIIN, Gen)	–	–	–	–	–	–
MD3 SBIRT	Tool to evaluate fidelity to SBIRT residency training program	23	1. SBIRT adherent 2. SBIRT nonadherent 3. Global empathy and collaboration	–	X	–	X	–	–
MISCHE	Tool to evaluate MI skills in brief health encounters	15	1. MI Philosophy 2. Health interviewing 3. Motivation 4. MI principles 5. Interpersonal processes	X	–	–	X	–	–
Client-rated									
PRF	Self-reported perception of MI adherence	8	1. Technical 2. Relational	X	–	–	–	X	X
TSR	Self-reported perception of MI fidelity	6	1. Therapeutic Bond 2. Helpfulness 3. Directiveness	–	–	–	–	–	–

Note. ICC = Intraclass Correlation Coefficients; K = Kappa; MI-SCOPE = Motivational Interviewing Sequential Code for Observing Process Exchanges; MIAS = Motivational Interviewing Assessment Scale; CAS = Content Adherence Scale; GROMIT = Global Rating of Motivational Interviewing Therapist; REM = Rating Scales for the Assessment of Empathic Communication in Medical Interviews; PEPA = Peer Proficiency Assessment; AMIGOS = Assessment of MI Groups Observer System; DDMI = Dual Diagnosis for Motivational Interviewing; MD3 SBIRT = MD3 Screening, Brief Intervention, and Referral to Treatment Coding Scale; MISCHE = Motivational Interviewing Skills for Health Care Encounters; PRF = Participant Rating Form; TSR = Therapy Session Report

3.3.6. Assessment of MI groups observer system (AMIGOS)

Wagner and Ingersoll (2013) developed the AMIGOS to evaluate provider skills when delivering MI in a group format. The AMIGOS has demonstrated construct validity and strong inter-rater reliability (Ingersoll & Wagner, 2014). In the current review, we identified one validation article and one efficacy/effectiveness article. Whereas no psychometric data was reported for one study, the other study demonstrated good internal consistency ($0.93 < \alpha s < 0.95$) and interrater reliability ($0.82 < \text{ICCs} < 0.88$) of the AMIGOS, as well as strong convergent validity with the MITI, the Therapist Empathy Scale, and the Group Climate Questionnaire.

3.3.7. MI skills for health care encounters (MISCHE)

The MISCHE was developed to evaluate provider MI skills in the context of brief health care encounters. The authors established content validity using an expert panel review with a particular focus on specific MI skills necessary for providers to better promote health and disease management. The authors found adequate internal consistency across domains ($0.75 < \alpha s < 0.80$), poor-to-excellent inter-rater reliability (0.21 [resists the righting reflect] < ICCs < 0.91), and good test-retest reliability ($0.61 < \text{ICCs} < 1.0$).

3.3.8. Additional observer-rated tools used for specific treatments

We identified two additional tools that were used to evaluate provider fidelity to a specific MI modality, including a dual diagnosis MI intervention (DDMI fidelity ratings) and an SBIRT residency training program (MD3 SBIRT coding scale). Whereas no psychometric data was provided on the DDMI, both validity and reliability was established with the MD3 SBIRT. Specifically, the MD3 SBIRT was developed through an interative process that involved piloting drafts and obtaining expert feedback. The authors found excellent inter-rater reliability ($0.85 < \text{ICCs} < 0.95$) and poor-to-excellent estimates for each behavior (0.30 [labeling, premature diagnoses, stereotyping] < ICCs < 0.96).

3.3.9. Additional client-rated tools used for specific treatments

We identified two client-rated tools that were used to evaluate client perceptions of provider skills to a specific MI modality. The Participant Rating Form (PRF) was used for a telephone-based brief MI intervention and the Therapy Session Report (TSR) was used in a study comparing MET and spirit-only MI. Both tools assess client perspective of therapist adherence to MI skills. Whereas no psychometric data was provided on the TSR, the PRF had good construct validity (Technical [$\alpha = 0.85$]; Relational [$\alpha = 0.74$] and predictive ability. Thus, there is preliminary support for the PRF but it is unclear if the TSR is a useful tool to assess MI fidelity from the client's perspective.

3.4. Fidelity/skill measures with no additional research

Importantly, we wanted to recognize two measures from these prior reviews that have not been further evaluated: the Motivational Interviewing Process Code (MIPC) and the Motivational Interviewing Supervision and Training Scale (MISTS).

4. Discussion

We systematically reviewed the literature from 2007 to 2019 to identify measures of MI skills and fidelity across research and training contexts and evaluate their empirical strength and congruence with MI theory and practice. Observer-rated tools were the most commonly employed measures across articles reviewed. Among these, earlier versions of the MITI had the most robust psychometric evaluation (and evidence for the newest version is emerging), and the MISC, BECCI, YACS, and ITRS also had strong evidence. Despite their strong psychometric properties, identified studies rarely linked these tools to client outcomes (except the MISC), signifying an important area for future research. Two trainee-completed tools, which depict clinical vignettes that therapists' respond to were utilized in training studies, and one client-completed tool, which assess client perspective of therapist MI skills, also demonstrated strong empirical support. We identified several additional tools that offer preliminary empirical evidence but warrant more investigation prior to employing in research or training.

Overall this review suggests that there are numerous strong and promising measures of MI skill and fidelity. The choice of measure will be driven by several factors. Observer-rated measures are typically considered gold-standard measures of MI skill and fidelity, and are likely preferred in most cases for assessing MI adherence in training, and MI efficacy/effectiveness studies. In fact, in the current review, the majority of articles identified were efficacy and effectiveness studies, which is promising given that early MI efficacy studies often reported limited information about MI fidelity (Madson, Campbell, Barrett, Brondino, & Melchert, 2005). Evidence of strong psychometric properties has been identified as important in MI outcomes research (Miller & Rollnick, 2014), and given that multiple observer-rated measures had such empirical strength, researchers have some choice about which measure will be best for their particular study. In some cases, however, an observer-rated measure may not be practically or financially feasible for an efficacy or effectiveness study.

Client-completed measures hold great promise for enabling assessment of MI quality in research studies. Although data on this class of tools is currently limited, the CEMI has a growing body of psychometric evidence. The CEMI is also likely to be preferred in many clinical practice and training settings where coding of MI sessions is impractical. Although subject to client bias in responding, these measures address important and ubiquitous barriers to assessment of MI quality, additional research on and development of client-completed MI skill measures is essential, as these measures can be relied upon to provide an indication of whether MI occurred in a particular session. Relatedly, certain trainee-completed measures appear helpful in evaluating provider skills in the context of MI training. Importantly, we excluded studies that used self-report measures of MI skills by therapists given that extant work finds that therapists overestimate their MI proficiency (Macdonald & Mellor-Clark, 2015; Martino, Ball, Nich, Frankforter, & Carroll, 2009; Miller & Mount, 2001; Tracey, Wampold, Lichtenberg, & Goodyear, 2014). However, the two trainee-completed measures we identified, the VASE-R and HRQ are distinct in that they require trainees to respond to clinical vignettes as a means to demonstrate and evaluate their MI skills. Thus, these two tools allow for quick and inexpensive collection of skill data, commonly within the context of a workshop or other group training setting, and all trainee-completed tools allow for an assessment of MI skill without relying on practice data at all.

There are several practical and empirical considerations that may influence the implementation of MI tools into clinical practice. Despite the empirical strength and large application of observer-rated tools in the current review, time constraints, limited resources, and ethnical concerns may prevent their uptake in clinical practice. Observer-rated tools are complex, with some tools requiring up to 40 h of rater training (e.g., Moyers, Rowell, Manuel, Ernst, & Houck, 2016) and between 85 and 120 min of coding time per MI session segment (e.g., Moyers et al., 2005). If outside companies are utilized, evaluating MI fidelity with these tools can be costly. Beyond the time and resources needed, providers may hold ethical and practical concerns around audio or video recording MI sessions. Empirically, observer-rated tools largely fail to link provider skills to client outcomes (e.g., Madson et al., 2019). In our review, the MITI was the most widely used tool in efficacy/effectiveness studies, and yet, most of these studies (73%) did not link MITI scores to client outcomes. This is an important limitation as evidence that a tool predicts client success is an important determinant of whether (a) MI will be implemented into practice and (b) a clinic will expend the time and resources to train their providers in MI. In the context of community-based training, utilization of trainee-completed or client-rated tools may be preferred (Schumacher et al., 2011). Whereas psychometric properties are strong for certain tools (e.g., VASE-R; CEMI), additional work evaluating the convergent validity of these tools with the well-established observer-rated tools can better justify their use in community-based settings.

Methods to overcome the challenges with implementing tools into practice are emerging. The recent technological approaches to evaluate MI fidelity and skills using observer-rated tools (e.g., Tanana et al., 2016) help mitigate some practical concerns. For example, the time to train raters and to rate MI sessions would be substantially reduced and the reliance on inter-rater reliability between human raters largely removed (Atkins et al., 2014). The financial benefits are also evidence, with Klonek et al. (2015) detailing an estimated savings of between $2000 and $20,000 to use a computer-based coding tool. Unfortunately, technological approaches do not overcome the ambivalence that providers or clients have about recording their MI sessions or the legal or ethical barriers to using actual provider-client interactions as the basis for assessment. If trainee-completed or client-rated tools are then preferred, ease of interpretation and established benchmarks are needed to determine provider skills when executing MI. Whereas Rosengren, Hartzler, Baer, Wells, and Dunn (2008) established benchmarks for the VASE-R, no other trainee-completed nor client-rated tools offered guidance on how to determine a therapist is adequately demonstrating MI skills. Relatedly, none of the trainee-completed or client-rated tools were used to provide feedback/coaching in MI training studies, a recommended best practice in MI training (Madson, Schumacher, et al., 2016). Given the practical issues with employing observer-rated tools, future work should evaluate the feasibility and impact of using trainee-completed or client-rated tools in the context of feedback/coaching. Finally, while a menu of tool choices is helpful to effectively implement MI into community-based programs, initial evidence suggests linking these programs to research networks can increase implementation (Rieckmann, Abraham, & Bride, 2016), and likely circumvent the myriad of barriers around using the best tools to evaluate fidelity/skills.

Another major objective of the proposed study was to determine how well existing MI tools aligned with the theoretical model of MI efficacy. Whereas some tools have been adapted to fit with the current conceptualization of MI (e.g., MITI; MISC), other well-validated tools have not taken such steps. For example, the BECCI and ITRS has not been adapted since the theory of MI was proposed by Miller and Rose (2009). However, many tools identified in the current review were developed to evaluate a specific type of MI intervention (i.e., BECCI for

BCC), provider skills across several interventions (i.e., YACS for behavioral substance use treatments), or in the context of community-based settings (i.e., ITRS; Martino et al., 2008). Despite some tools having multiple purposes, adapting these tools to align with skills represented in the two theorized active ingredients of MI (i.e., relational and technical; Miller & Rose, 2009) is critical to accurately evaluate provider MI skills and fidelity. The most recent versions of the MITI and the MISC offer specific behaviors as well as global indicators of MI in accordance with current theory. This may be a helpful starting point to adapt portions of tools that are focused on MI-specific behaviors without necessarily modifying the portions that pertain to a different treatment or general counseling skills. Across identified trainee-completed and client-rated tools, the CEMI is the only tool that was developed to align with the technical and relational components of MI. Although the PRF offered psychometric support on composite technical and relational scales, the technical scale comprised behaviors that were no longer congruent with MI theory (e.g., decisional balance; Miller & Rose, 2015). Developing interventions and employing tools that align with MI theory will help overcome concerns that some adapted MI interventions do not adhere to the fundamental tenets of MI (Miller & Rollnick, 2014). Further, employing tools that capture the theoretical elements of *how* MI works, and increasing investigations on the link between these tools and client outcomes can inform if specific or the composition of theoretical MI elements are essential to demonstrate MI efficacy. Finally, developing training protocols that adhere with deliberate practice and include tools that align with MI theory can assure providers acquire not only the knowledge but also the skills necessary to facilitate client change (Di Bartolomeo, Shukla, Westra, Ghashghaei, & Olson, 2020). In fact, recent efforts to train providers in MI using a deliberate practice workshop found that providers sustained MI skills longer than those who participated in a didactic workshop (Westra et al., 2020).

4.1. Tool recommendations across research contexts

Despite the array of implementation considerations when selecting a MI skill/fidelity tool, the current review offers guidance on which tools may be best for a research or training study (see Table 3). In the context of MI training, tools that evaluate provider skills and have been used for feedback and coaching are ideal. Given its purpose, psychometric strength, and congruence with MI theory, the MITI is arguably the best tool for MI training studies. We also identified the VASE-R as a top-tier tool for training studies, given its psychometric strength and established proficiency benchmarks. Alternately, the MISC appears to be the best tool to evaluate the therapeutic process, particularly in efficacy studies. Though training and evaluation of the therapeutic process is timely, the MISC permits temporal examination of the provider-client interaction (Gaume, Gmel, Faouzi, & Daeppen, 2009), a feature that is absent in all other tools identified with strong empirical support. An adequate determination that an intervention has passed the 'efficacy' test is to demonstrate the therapists' ability to employ MI skills and elicit client change language that leads to positive outcomes, all of which can be accomplished with the MISC. Though the MI-SCOPE also has this evaluative capability and was linked to client outcomes, more validation and empirical work is needed with this tool. We argue that, despite also recommending the MISC for effectiveness studies, the provider and contextual considerations of such studies warrants consideration of other tools. Thus, we also identified the ITRS as a top-tier tool, particularly in training and effectiveness studies. The ITRS was adapted from the YACS for the purpose of evaluating provider MI fidelity in community-based studies. We found that the ITRS was often linked to client outcomes, demonstrating its utility in outcome studies. Additionally, the ITRS distinguishes between fundamental and advanced MI skills, permitting training protocols to be adapted in community-based settings to fit with therapist current skill set. Finally, the empirical strength and conceptual fit of the CEMI suggests it may be beneficial across research contexts. Given the lack of evidence linking the CEMI to more established observer-rated tools (e.g., MITI), this tool would likely become top-tier if such research is completed and evidence of its congruence with more established tools is demonstrated.

5. Conclusions and future directions

Overall this review shows that researchers have generally been responsive to calls to assess and report quality in MI research (Madson, Schumacher, et al., 2016). In particular, commonly used observer-rated tools have large and generally robust bodies of psychometric evidence to support their use as measures of MI quality across research contexts. However, more work is needed on the predictive ability of these tools on client outcomes and the use of observer-rated tools both within and outside the research context is limited by factors such as lack of availability of work samples for coding (Schumacher et al., 2011), costs associated with training coders (Glynn, Hallgren, Houck, & Moyers, 2012), and barriers to accessing or utilizing treatment data to assess quality. Thus, future research must continue to focus on strategies that reduce costs associated with observed-coding based measures, such as

Table 3
Recommended tools across research contexts[a].

	Training studies	Efficacy studies	Effectiveness studies
Tier One			
Motivational Interviewing Skills Code		X	X
Motivational Interviewing Treatment Integrity	X		
Independent Tape Rater Scale	X		X
Behavior Change Counseling Index[b]	X		
Video-Assessment of Simulated Encounters-Revised	X		
Tier Two			
Yale Adherence and Competence Scale	X		
MI-Sequential Code for Observing Process Exchanges		X	X
Helpful Responses Questionnaire	X		
Client Evaluation of MI	X	X	
Tier Three			
MI Assessment Scale[b]	X	X	
Rating Scales for the Assessment of Empathic Communication			X
Peer Proficiency Assessment	X	X	
Assessment of MI Groups Observer System		X	X
Participant Rating Form		X	X

Note. MI: Motivational Interviewing.

[a] Tools that had no-to-minimal psychometric support and/or were not linked to trainee/client outcomes were excluded.

[b] Tools that are supported for use with Behavior Change Counseling, an adapted MI intervention commonly employed in primary care settings.

computer-based coding developed with machine-learning technologies (e.g., Atkins et al., 2014; Klonek et al., 2015), strategies that eliminate the need for coding, such as client-completed tools (e.g., Madson, Villarosa, et al., 2016), and strategies that eliminiate the need for utiliziation of any protected health information to assess MI quality, such as trainee-completed tools. Further, incorporating deliberate practice principles into MI training and outcome studies can assure appropriate tool selection and provider skill acquisition to effectively execute MI. Relative to observer-rated tools, these areas of MI quality are currently in their infancy.

Author contributions

MCH was involved in conceptualization, data curation, formal analysis, investigation, methodology, project administration, visualization, writing - original draft (i.e., introduction, methodology, specific instrument sections, and discussion) and review & editing (i.e., entire manuscript).

MBM was involved in conceptualization, methodology, project administration, validation, and writing – original draft (i.e., specific instrument sections) and review & editing (i.e., entire manuscript).

JAS was involved in conceptualization, methodology, validation, and writing - original draft (i.e., specific instrument sections) and review & editing (i.e., entire manuscript).

Funding

This work was supported by the National Institutes of Health [L30 AA027026].

Declaration of Competing Interest

None.

Acknowledgements

The authors wish to thank Hallie Jordan and Mallorie Carroll for their assistance rating a portion of the identified articles.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cpr.2020.101909.

References

Atkins, D. C., Steyvers, M., Imel, Z. E., & Smyth, P. (2014). Scaling up the evaluation of psychotherapy: Evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science, 9*, 49–60.

Baer, J. S., Carpenter, K. M., Beadnell, B., Stoner, S. A., Ingalsbe, M. H., Hartzler, B., ... Drager, Z. (2012). Computer assessment of simulated patient interviews (CASPI): Psychometric properties of a web-based system for the assessment of motivational interviewing skills. *Journal of Studies on Alcohol and Drugs, 73*(1), 154–164.

Barwick, M. A., Bennett, L. M., Johnson, S. N., McGowan, J., & Moore, J. E. (2012). Training health and mental health professionals in motivational interviewing: A systematic review. *Children and Youth Services Review, 34*, 1786–1795. https://doi.org/10.1016/j.childyouth.2012.05.012.

Brueck, R. K., Frick, K., Loessl, B., Kriston, L., Schondelmaier, S., Go, C., ... Berner, M. (2009). Psychometric properties of the German version of the motivational interviewing treatment integrity code. *Journal of Substance Abuse Treatment, 36*, 44–48.

Burke, B. L., Arkowitz, H., & Menchola, M. (2003). The efficacy of motivational interviewing: A meta-analysis of controlled clinical trials. *Journal of Consulting and Clinical Psychology, 71*(5), 843–861.

Carroll, K. M., Nich, C., Sifry, R. L., Nuro, K. F., Frankforter, T. L., Ball, S. A., ... Rounsaville, B. J. (2000). A general system for evaluating therapist adherence and competence in psychotherapy research in the addictions. *Drug and Alcohol Dependence, 57*(3), 225–238.

Childers, J. W., Bost, J. E., Kraemer, K. L., Cluss, P. A., Spagnoletti, C. L., Gonzaga, A. M. R., & Arnold, R. M. (2012). Giving residents tools to talk about behavior change: A motivational interviewing curriculum description and evaluation. *Patient Education and Counseling, 89*, 281–287.

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

Corvino, J., Carroll, K., Nuro, K., Nich, C., Sifry, R., Frankforter, T., ... Rounsaville, B. (2000). *Yale adherence and competence scale guidelines. Unpublished manuscript.* West Haven, CT: Yale University Psychotherapy Development Center.

Di Bartolomeo, A. A., Shukla, S., Westra, H. A., Ghashghaei, N. S., & Olson, D. A. (2020). Rolling with resistance: A client language analysis of deliberate practice in continuing education for psychotherapists. *Counselling and Psychotherapy Research*. https://doi.org/10.1002/capr.12335.

Dray, J., Gilchrist, P., Singh, D., Cheesman, G., & Wade, T. D. (2014). Training mental health nurses to provide motivational interviewing on an inpatient eating disorder unit. *Journal of Psychiatric and Mental Health Nursing, 21*(7), 652–657.

Ericcson, K. A. (2009). Enhancing the development of professional performance: Implications from the study of deliberate practice. *Development of professional expertise: Toward measurement of expert performance and design of optimal learning environments* (pp. 405–431). New York, NY: Cambridge University Press.

Forsberg, L., Berman, A. H., Kallmen, H., Hermansson, U., & Helgason, A. R. (2008). A test of the validity of the motivational interviewing treatment integrity code. *Cognitive Behaviour Therapy, 37*, 183–191.

Forsberg, L., Kallmen, H., Hermansson, U., Berman, A. H., & Helgason, A. R. (2007). Coding counsellor behaviour in motivational interviewing sessions: Inter-rater reliability for the Swedish motivational interviewing treatment integrity code (MITI). *Cognitive Behaviour Therapy, 36*, 162–169.

Gaume, J., Gmel, G., Faouzi, M., & Daeppen, J. (2009). Counselor skill influences outcomes of brief motivational interventions. *Journal of Substance Abuse Treatment, 37*, 151–159.

Gill, I., Oster, C., & Lawn, S. (2019). Assessing competence in health professionals' use of motivational interviewing: A systematic review of training and supervision tools. *Patient Education and Counseling, 103*(3), 473–483.

Glynn, L. H., Hallgren, K. A., Houck, J. M., & Moyers, T. B. (2012). CACTI: Free, open-source software for the sequential coding of behavioral interactions. *PLoS One, 7*(7), Article e39740.

Hallgren, K. A., Dembe, A., Pace, B. T., Imel, Z. E., Lee, C. M., & Atkins, D. C. (2018). Variability in motivational interviewing adherence across sessions, providers, sites, and research contexts. *Journal of Substance Abuse Treatment, 84*, 30–41.

Hettema, J., Steele, J., & Miller, W. R. (2005). Motivational interviewing. *Annual Review of Clinical Psychology, 1*, 91–111.

Houck, J. M., Moyers, T. B., Miller, W. R., Glynn, L. H., & Hallgren, K. A. (2010). Motivational interviewing skill code (MISC) 2.5. Retrieved from http://casaa.unm.edu/download/misc25.pdf.

Imel, Z. E., Pace, B. T., Soma, C. S., Tanana, M., Hirsch, T., Gibson, J., ... Atkins, D. C. (2019). Design feasibility of an automated, machine-learning based feedback system for motivational interviewing. *Psychotherapy, 56*(2), 318–328.

Ingersoll, K., & Wagner, C. C. (2014, June). Rating the fidelity of MI group sessions: The AMIGOS coding system. *4th International conference on motivational interviewing* (Amsterdam, NL).

Klonek, F. E., Quera, V., & Kauffeld, S. (2015). Coding interactions in motivational interviewing with computer-software: What are the advantages for process researchers? *Computers in Human Behavior, 44*, 284–292.

Lane, C., Huws-Thomas, M., Hood, K., Rollnick, S., Edwards, K., & Robling, M. (2005). Measuring adaptations of motivational interviewing: The development and validation of the behavior change counseling index (BECCI). *Patient Education and Counseling, 56*(2), 166–173.

Lundahl, B., & Burke, B. L. (2009). The effectiveness and applicability of motivational interviewing: A practice-friendly review of four meta-analyses. *Journal of Clinical Psychology, 65*, 1232–1245.

Macdonald, J., & Mellor-Clark, J. (2015). Correcting psychotherapists' blindsidedness: Formal feedback as a means of overcoming the natural limitations of therapists. *Clinical Psychology & Psychotherapy, 22*, 249–257.

Madson, M. B., Bullock, E. E., Speed, A. C., & Hodges, S. A. (2009). Development of the client evaluation of motivational interviewing. *Motivational Interviewing Network of Trainers Bulletin, 15*(1), 6–8.

Madson, M. B., & Campbell, T. C. (2006). Measures of fidelity in motivational enhancement: A systematic review of instrumentation. *Journal of Substance Abuse Treatment, 31*, 67–73.

Madson, M. B., Campbell, T. C., Barrett, D. E., Brondino, M. J., & Melchert, T. P. (2005). Development of the motivational interviewing supervision and training scale. *Psychology of Addictive Behaviors, 19*, 303–310.

Madson, M. B., Schumacher, J. A., Baer, J. S., & Martino, S. (2016). Motivational interviewing for substance use: Mapping out the next generation of research. *Journal of Substance Abuse Treatment, 65*, 1–5.

Madson, M. B., Villarosa, M. C., Schumacher, J. A., & Mohn, R. S. (2016). Evaluating the validity of the client evaluation of motivational interviewing scale in a brief motivational intervention for college student drinkers. *Journal of Substance Abuse Treatment, 65*, 51–57.

Madson, M. B., Villarosa-Hurlocker, M. C., Schumacher, J. A., Williams, D. C., & Gauthier, J. M. (2019). Motivational interviewing training of substance use treatment professionals: A systematic review. *Substance Abuse, 40*(1), 43–51.

Magill, M., Apodaca, T. R., Borsari, B., Gaume, J., Hoadley, A., Gordon, R. E. F., ... Moyers, T. (2018). A meta-analysis of motivational interviewing process: Technical, relational, and conditional process models of change. *Journal of Consulting and Clinical Psychology, 86*(2), 140–157.

Marin-Navarrete, R., Horigian, V. E., Medina-Mora, M. E., Verdeja, R. E., Alonso, E., Feaster, D. J., ... de la Fuenta-Martin, A. (2017). Motivational enhancement treatment in outpatient addiction centers: A multisite randomized trial. *International Journal of*

Clinical and Health Psychology, 17, 9–19.

Martino, S., Ball, S. A., Nich, C., Frankforter, T. L., & Carroll, K. M. (2008). Community program therapist adherence and competence in motivational enhancement therapy. *Drug and Alcohol Dependence, 96*, 37–48.

Martino, S., Ball, S. A., Nich, C., Frankforter, T. L., & Carroll, K. M. (2009). Correspondence of motivational enhancement treatment integrity ratings among therapists, supervisors, and observers. *Psychotherapy Research, 19*(2), 181–193. https://doi.org/10.1080/10503300802688460.

Mastroleo, N. R., Mallet, K. A., Turrisi, R., & Ray, A. E. (2009). Psychometric properties of the peer proficiency assessment (PEPA): A tool for evaluation of undergraduate peer counselors' motivational interviewing fidelity. *Addictive Behaviors, 34*(9), 717–722.

Miller, S. D., Hubble, M. A., & Duncan, B. L. (2007). Supershrinks: Learning from the most effective practitioners. *Psychotherapy Networker, 31*, 26–35.

Miller, W. R. (1983). Motivational interviewing with problem drinkers. *Behavioural Psychology, 11*, 147–172.

Miller, W. R., Hedrick, K. E., & Orlofsky, D. R. (1991). The helpful response questionnaire: A procedure for measuring therapeutic empathy. *Journal of Clinical Psychology, 47*, 444–448.

Miller, W. R., & Mount, K. A. (2001). A small study of training in motivational interviewing: Does one workshop change clinician and client behavior? *Behavioral and Cognitive Psychotherapy, 29*, 457–471.

Miller, W. R., Moyers, T. B., Ernst, D., & Amrhein, P. (2008). *Manual for the motivational interviewing skills code (MISC) version 2.1*. Unpublished manualCenter on Alcoholism, Substance Abuse, and Addictions, The University of New Mexico.

Miller, W. R., & Rollnick, S. (2009). Ten things motivational interviewing is not. *Behavioural and Cognitive Psychotherapy, 37*, 129–140.

Miller, W. R., & Rollnick, S. (2013). *Motivational interviewing: Helping people change* (3rd ed.). New York, NY: Guilford Press.

Miller, W. R., & Rollnick, S. (2014). The effectiveness and ineffectiveness of complex behavioral interventions: Impact of treatment fidelity. *Contemporary Clinical Trials, 37*, 234–241.

Miller, W. R., & Rose, G. S. (2009). Toward a theory of motivational interviewing. *American Psychologist, 64*, 527–537.

Miller, W. R., & Rose, G. S. (2015). Motivational interviewing and decisional balance: Contrasting responses to client ambivalence. *Behavioural and Cognitive Psychotherapy, 43*(2), 129–141. https://doi.org/10.1017/S1352465813000878.

Moyers, T. B. (2004). History and happenstance: How motivational interviewing got its start. *Journal of Cognitive Psychotherapy, 18*, 291–298.

Moyers, T. B., & Martin, T. (2006). Therapist influence on client language during motivational interviewing sessions. *Journal of Substance Abuse Treatment, 30*, 245–251.

Moyers, T. B., Martin, T., Manuel, J. K., Hendrickson, S. M. L., & Miller, W. R. (2005). Assessing competence in the use of motivational interviewing. *Journal of Substance Abuse Treatment, 28*, 19–26.

Moyers, T. B., Rowell, L. N., Manuel, J. K., Ernst, D., & Houck, J. M. (2016). The motivational interviewing treatment code (MITI 4): Rationale, preliminary reliability and validity. *Journal of Substance Abuse Treatment, 65*, 36–42.

Perula, L. A., Campinez, M., Bosch, J. M., Brun, N. B., Arbonies, J. C., ... Fontan, J. B. Collaborative Group Dislip-EM. (2012). Is the scale for measuring motivational interviewing skills a valid and reliable instrument for measuring the primary care professionals motivational skills?: EVEM study protocol. *BMC Family Practice, 13*, 112–119.

Rieckmann, T. R., Abraham, A. J., & Bride, B. E. (2016). Implementation of motivational interviewing in substance use disorder treatment: Research network participation and organizational compatibility. *Journal of Addiction Medicine, 10*(6), 402–407. https://doi.org/10.1097/ADM.0000000000000251.

Rollnick, S., Allison, J., Ballasiotes, S., Barth, T., Butler, C., & Rose, G. (2002). Variations on a theme: Motivational interviewing and its adaptations (pp. 251–269). In W. R. Miller, & S. Rollnick (Eds.). *Motivational interviewing: Preparing people for change* (pp. 251–269). (2nd ed.). (Guilford).

Rosengren, D. B., Baer, J. S., Hartzler, B., Dunn, C. W., & Wells, E. A. (2005). The video assessment of simulated encounters (VASE): Development and validation of a group-administered method for evaluating clinician skills in motivational interviewing. *Drug and Alcohol Dependence, 79*(3), 321–330.

Rosengren, D. B., Hartzler, B., Baer, J. S., Wells, E. A., & Dunn, C. W. (2008). The video assessment of simulated encounters-revised (VASE-R): Reliability and validity of a revised measure of motivational interviewing skills. *Drug and Alcohol Dependence, 97*(1–2), 130–138. https://doi.org/10.1016/j.drugalcdep.2008.03.018.

de Roten, Y., Zimmerman, G., Ortega, D., & Despland, J. (2013). Meta-analysis of the effects of MI training on clinicians' behavior. *Journal of Substance Abuse Treatment, 45*, 155–162.

Rousmaniere, T., Goodyear, R. K., Miller, S. D., & Wampold, B. E. (2017). *The cycle of excellence: Using deliberate practice to improve supervision and training.* Hoboken, NJ: John Wiley and Sons.

Santa Ana, E. J., Carroll, K. M., Anez, L., Paris, M., Ball, S. A., Nich, C., ... Martino, S. (2009). Evaluating motivational enhancement therapy adherence and competence among Spanish-speaking therapists. *Drug and Alcohol Dependence, 103*, 44–51.

Schumacher, J. A., & Madson, M. B. (2014). *Fundamentals of motivational interviewing: Tips and strategies to address common clinical challenges.* New York: Oxford University Press.

Schumacher, J. A., Madson, M. B., & Norquist, G. (2011). Using telehealth technology to enhance motivational interviewing training for rural substance abuse treatment providers: A services improvement project. *The Behavior Therapist, 34*, 64–70.

Schumacher, J. A., Williams, D. C., Burke, R. S., Epler, A. J., Simon, P., & Coffey, S. F. (2018). Competency-based supervision in motivational interviewing for advanced psychology trainees: Targeting and a prior benchmark. *Training and Education in Professional Psychology, 12*, 149–153.

Schwalbe, C. S., Oh, H. Y., & Zweben, A. (2014). Sustaining motivational interviewing: A meta-analysis of training studies. *Addiction*. https://doi.org/10.1111/add.12558.

Small, J. W., Lee, J., Frey, A. J., Seeley, J. R., & Walker, H. M. (2014). The development of instruments to measure motivational interviewing skill acquisition for school-based personnel. *Advances in School Mental Health Promotion, 7*(4), 240–254.

Söderlund, L., Madson, M. B., Rubak, S., & Nilsen, P. (2011). A systematic review of motivational interviewing training for general health care practitioners. *Patient Education and Counseling, 84*, 16–26. https://doi.org/10.1016/j.pec.2010.06.025.

Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., & Srikumar, V. (2016). A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of Substance Abuse Treatment, 65*, 43–50.

Tracey, T. J., & Kokotovic, A. M. (1989). Factor structure of the working alliance inventory. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1*, 207–210. https://doi.org/10.1037/1040-3590.1.3.207.

Tracey, T. J. G., Wampold, B. E., Lichtenberg, J. W., & Goodyear, R. K. (2014). Expertise in psychotherapy: An elusive goal? *American Psychologist, 69*(3), 218–229. https://doi.org/10.1037/a0035099.

Wagner, C. C., & Ingersoll, K. S. (2013). *Motivational interviewing in group.* New York, NY: Guilford.

Wallace, L., & Turner, F. (2009). A systematic review of psychometric evaluation of motivational interviewing integrity measures. *Journal of Teaching in the Addictions, 8*, 84–123.

Westra, H. A., Norouzian, N., Poulin, L., Coyne, A., Constantino, M. J., Olson, H. K., & Martin, M. (2020). Testing a deliberate practice workshop for developing appropriate responsivity to resistance markers. *Psychotherapy*. https://doi.org/10.1037/pst0000311.