Advancing Methods for Reliably Assessing Motivational Interviewing Fidelity using the Motivational Interviewing Skills Code

Sarah Peregrine Lord, Psy.D. [a,*], Doğan Can, M.S. [b], Michael Yi, Ph.D. [c], Rebeca Marin, Ph.D. [a], Christopher W. Dunn, Ph.D. [a], Zac E. Imel, Ph.D. [d], Panayiotis Georgiou, Ph.D. [b], Shrikanth Narayanan, Ph.D. [b], Mark Steyvers, Ph.D. [c], David C. Atkins, Ph.D. [a]

[a] Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA
[b] Signal Analysis and Interpretation Laboratory, Departments of Electrical Engineering and Computer Science, University of Southern California, Los Angeles, CA, USA
[c] Department of Cognitive Sciences, University of California Irvine, Irvine, CA, USA
[d] Department of Educational Psychology, University of Utah, Salt Lake City, UT, USA

ABSTRACT

The current paper presents novel methods for collecting MISC data and accurately assessing reliability of behavior codes at the level of the utterance. The MISC 2.1 was used to rate MI interviews from five randomized trials targeting alcohol and drug use. Sessions were coded at the utterance-level. Utterance-based coding reliability was estimated using three methods and compared to traditional reliability estimates of session tallies. Session-level reliability was generally higher compared to reliability using utterance-based codes, suggesting that typical methods for MISC reliability may be biased. These novel methods in MI fidelity data collection and reliability assessment provided rich data for therapist feedback and further analyses. Beyond implications for fidelity coding, utterance-level coding schemes may elucidate important elements in the counselor–client interaction that could inform theories of change and the practice of MI.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

Motivational interviewing (MI) is a client-centered, collaborative style of counseling that attends to the language of change and is designed to strengthen personal motivation for and commitment to a specific goal (Miller & Rollnick, 2013). MI was originally developed to help clients prepare for changing addictive behaviors like drug and alcohol abuse (Miller & Rollnick, 1991, 2002) but has been shown to be effective across many populations for harmful behaviors including tobacco, drugs, alcohol, gambling, treatment engagement, and for promoting health behaviors such as exercise, diet, and safe sex (Lundahl, Kunz, Brownell, Tollefson, & Burke, 2010). As the basic efficacy and effectiveness of MI has been established, research has increasingly focused on how MI works (Magill et al., 2014) and how to practically measure MI counselor fidelity in real-world settings. Such research has typically used behavioral coding of MI sessions with fidelity assessment systems like the Motivational Interviewing Treatment Integrity (MITI) (Moyers, Martin, Manuel, Miller, & Ernst, 2010) and Motivational Interviewing Skills Code (MISC; Miller, Moyers, Ernst, & Amrhein, 2008). The MITI and MISC were designed to assess MI fidelity by independent raters (coders) identifying both relational and behavioral features of therapy sessions. Each utterance (i.e., complete thought) spoken by the counselor and client during the MI interview is assigned a behavioral code. Client behavioral codes include statements in favor of changing a problem behavior like expressing reasons for or commitment to change and are referred to as "change talk." Counselor behavioral codes also include speech patterns like open and closed questions and counseling techniques like reframing.

Research using these coding systems has explored hypothesized relationships between counselor and client speech, and between client speech and behavior change. Some research has shown that when counselors demonstrate high MI fidelity, clients are more likely to increase change talk and reduce statements away from change, called sustain talk (Moyers et al., 2007). The frequencies of client change and sustain talk, the type of change talk (e.g., commitment language) (Amrhein, Miller, Yahne, Palmer, & Fulcher, 2003; Baer et al., 2008), and where the change talk occurs in the session (Amrhein et al., 2003; Bertholet, Faouzi, Gmel, Gaume, & Daeppen, 2010) have been shown to independently predict behavior outcomes (Moyers et al., 2007). For example, commitment language such as, "I am going to stop drinking," at the end of a session predicts associated behavior change (Amrhein et al., 2003), even after accounting for severity of dependence, readiness and efficacy for change (Moyers, Martin, Houck, Christopher, & Tonigan, 2009). Other research has shown that the relationship between change talk and behavior change

is highly dependent on context, such as the presence of therapist MI consistent behaviors (Catley et al., 2006; Gaume, Gmel, Faouzi, & Daeppen, 2008). One study suggested that change talk was only predictive of behavior change when the MI session included a personalized feedback report (Vader, Walters, Prabhu, Houck, & Field, 2010). Other studies have found relationships between some but not all subtypes of client change talk and behavior change (e.g., Gaume, Gmel, & Daeppen, 2008).

Although coding systems like the MISC and MITI have become the standard for assessing MI counselor fidelity, there are challenges in how these coding data are typically collected that have implications for establishing reliability. Critically, behavioral codes are assigned to individual utterances, but data are typically collected or reported as the number of times a code was assigned across the entire session (i.e., a summary score). When the reliability of counselor and client behavior codes is assessed using summary scores, the true reliability of utterance-based codes is unclear. It is possible that coders had a similar total count of codes per session, but may have assigned different codes to the individual utterances. This distortion of reliability has implications for the accurate assessment of MI counselors and for analyses about the relationship between counselor and client speech.

The current paper presents novel methods for collecting MISC data and assessing reliability of behavior codes. It represents initial work of an interdisciplinary team of researchers applying quantitative linguistic tools such as speech signal processing (Narayanan & Georgiou, 2013) and text mining to MI and the MISC and MITI (Can, Georgiou, Atkins, & Narayanan, 2012). Prior to presenting the current research, we review the process and challenges associated with implementing the MISC.

1.1. MISC data collection and its mismatch with reliability assessments

The first step in assessing MI fidelity using the MISC is training a team of independent raters (coders) to utilize the coding system. The most commonly reported training method is a graded process wherein coders begin by parsing (i.e., deciding where an utterance begins and ends) and assigning behavioral codes to utterances in transcripts of MI sessions with an expert rater. The expert rater is someone who has had prior experience training MI coding teams and is a member of the Motivational Interviewing Network of Trainers (MINT) (Campbell, Adamson, & Carter, 2010; Gaume, Gmel, Faouzi, & Daeppen, 2008; Miller et al., 2008; Moyers, Martin, Catley, Harris, & Ahluwalia, 2003; Moyers, Miller, & Hendrickson, 2005).

To establish inter-rater reliability the coding team must reach agreement on what represents the beginning and ending of a complete thought or utterance (i.e., parsing reliability), and they must also reach agreement on what MI behavior code fits each utterance (i.e., behavior code reliability). Low parsing reliability can ultimately decrease the reliability of behavioral codes when the number of utterances, and thus number of codes, do not agree between raters. To reduce differences in coder parsing, studies have used expert raters or a separate coding team to pre-parse session transcripts (Barnett et al., 2012; Moyers & Martin, 2006; Moyers et al., 2003). Pre-parsing means that the boundaries of each thought unit or utterance are identified by one set of raters before a behavior code is assigned to the utterance by the coding team. This is the method of parsing employed in the MI-SCOPE (Martin, Moyers, Houck, Christopher, & Miller, 2005). Recent software developments also facilitate parsing sessions prior to coding (Glynn, Hallgren, Houck, & Moyers, 2012).

While pre-parsing has been utilized in recent research to increase reliability by separating the task of parsing from coding through different coding teams (i.e., one team parses while another conducts behavior coding), the MISC manual does not mention or recommend pre-parsing (Miller et al., 2008). Ideally one coding team should be trained to parsing and behavioral coding reliability such that the same

coders can consistently identify the boundaries of an utterance and label it with the appropriate code.

To assess coder reliability, the majority of studies use intraclass correlations (ICC) (Shrout & Fleiss, 1979) and Cicchetti's (1994) standards of agreement (i.e., below .40 = poor, .40–.59 = fair, .60–.74 = good, and .75–1.00 = excellent). Behavior code agreement varies notably across trials that use recent versions of the MISC (2.0–2.5) and most studies do not report ICCs for individual code agreement (Boardman, Catley, Grobe, Little, & Ahluwalia, 2006; Campbell et al., 2010; Catley et al., 2006; de Jonge, Schippers, & Schaap, 2005; Gaume, Gmel, & Daeppen, 2008; Gaume, Gmel, Faouzi, & Daeppen, 2008; Martin, Christopher, Houck, & Moyers, 2011). When ICCs for individual behavior codes are reported, many important codes, such as client change talk (represented by codes ending with a +) or sustain talk (represented by codes ending with a −), may not be reliably distinguished at a good to excellent level (see Table 1). This is even the case for studies that utilize utterance-level coding schemes, like the MI-SCOPE (e.g., Moyers et al., 2007).

Challenges related to assessing the reliability of MI coding systems are not new; however, there has been limited discussion in the literature about the fundamental mismatch between the process of assigning of codes to individual utterances and the assessment and reporting of reliability using summary ratings for the entire session. It is not clear whether ICCs that are based on session totals accurately reflect the true reliability of the utterance-based codes. The problem with low reliability for important behavioral codes is that much about what we can learn about how MI works hinges on the accurate identification of potential key ingredients for behavior change, like change talk. Furthermore, feedback to MI clinicians hinges on accurate fidelity assessment; if coding reliability is distorted it is difficult to gauge the level of clinical skill and provide appropriate feedback.

1.2. The current study

The present paper examines the preceding questions about how to collect MISC data and accurately assess its reliability. The primary aim was to compare the utterance-based reliability of MISC behavioral codes to traditional reliability based on summaries of an entire session. A team of raters coded transcripts of counselor–client MI sessions using the MISC by deciding on where each complete thought begins and ends (i.e., parsing the transcript) while exhaustively assigning MISC behavior codes to every utterance. Several different approaches were used to calculate utterance-based reliability given variations in parsing across the coding team and the approaches were compared to session-level assessments of reliability.

2. Methods

2.1. Study sample and setting

The present study drew from a collection of 985 MI-based audio-recorded therapy sessions from five different trials aimed at reducing drug and alcohol abuse (Krupski et al., 2012; Lee et al., 2013, 2014; Neighbors et al., 2012; Tollison et al., 2008). All studies were based in Seattle, Washington and all original trial methods were approved by the University of Washington Institutional Review Board prior to initiation. Four studies targeted either alcohol or marijuana abuse in college-aged students. The fifth study recruited from community primary care clinics where many of the clients were polysubstance users, sometimes concurrently abusing up to 5 or more types of drugs (Krupski et al., 2012). Approximately 15% ($n = 155$) of sessions were randomly selected for coding, and 148 were selected for final analyses after some session recordings were excluded due to recording or transcription error. Approximately 20% of sessions ($n = 31$) were selected for assessment of inter-rater reliability, where 63% ($n = 19$) included patients that reported abusing more than one substance.