Variability in Motivational Interviewing Adherence Across Sessions, Providers, Sites, and Research Contexts

Kevin A. Hallgren[1], **Aaron Dembe**[2], **Brian T. Pace**[2], **Zac E. Imel**[2], **Christine M. Lee**[1], and **David C. Atkins**[1]

[1]Department of Psychiatry and Behavioral Sciences, University of Washington

[2]Department of Educational Psychology, University of Utah

Abstract

Providers' adherence in the delivery of behavioral interventions for substance use disorders is not fixed, but instead can vary across sessions, providers, and intervention sites. This variability can substantially impact the quality of intervention that clients receive. However, there has been limited work to systematically evaluate the extent to which substance use intervention adherence varies from session-to-session, provider-to-provider, and site-to-site. The present study quantifies the extent to which adherence to Motivational Interviewing (MI) for alcohol and drug use varies across sessions, providers, and intervention sites and compares the extent of this variability across three common MI research contexts that evaluate MI efficacy, MI effectiveness, and MI training. Independent raters coded intervention adherence to MI from 1275 sessions delivered by 216 providers at 15 intervention sites. Multilevel models indicated that 57%–94% of the variance in MI adherence was attributable to variability between sessions (i.e., within providers), while smaller proportions of variance were attributable to variability between providers (3%–26%) and between intervention sites (0.1%–28%). MI adherence was typically lowest and most variable within contexts evaluating MI training (i.e., where MI was not protocol-guided and delivered by community treatment providers) and, conversely, adherence was typically highest and least variable in contexts evaluating MI efficacy and effectiveness (i.e., where MI was highly protocolized and delivered by trained therapists). These results suggest that MI adherence in efficacy and effectiveness trials may be substantially different from that obtained in community treatment settings, where adherence is likely to be far more heterogeneous.

Keywords

alcohol and drug use disorder treatment services; intervention integrity; motivational interviewing; treatment adherence; treatment fidelity

---

Corresponding author: Kevin A. Hallgren, Behavioral Research in Technology and Engineering (BRiTE) Center, Department of Psychiatry and Behavioral Sciences, University of Washington, 1959 NE Pacific Street, 98195, Box 356560, Seattle, WA 9195, United States; Phone: 206-616-2906; Fax: 206-543-9520; khallgre@uw.edu.

Motivational Interviewing (MI; Miller & Rollnick, 2013) is an evidence-based clinical method with strong empirical support for reducing problem alcohol and drug use (Lundahl et al., 2010). Research has evaluated MI's efficacy (i.e., impact of the intervention on client outcomes under tightly controlled conditions) and its effectiveness (i.e., impact of the intervention on client outcomes under "real world" conditions) through multiple randomized clinical trials (RCTs), and numerous training studies have evaluated methods for disseminating MI practices into community settings and enhancing providers' use of MI (Schwalbe, Oh, & Zweben, 2014). Efficacy, effectiveness, and training studies commonly assess providers' adherence to MI principles and techniques to understand the degree to which MI was implemented with fidelity. For example, in efficacy and effectiveness studies, investigators often measure adherence to assess the internal validity of the study and determine whether MI was being delivered in the manner intended by intervention developers. Likewise, training and implementation studies also commonly measure providers' MI adherence to assess the impact of training or implementation efforts on providers' intervention delivery.

In each of the above research contexts (efficacy, effectiveness, and training studies), ratings of providers' MI adherence demonstrate the extent to which they have skillfully executed MI. It is often assumed, but not explicitly tested, that MI skill varies considerably between providers, and that providers' overall adherence to MI varies based on the research context in which the intervention was conducted. For example, MI that is delivered in the context of clinical efficacy trials is often assumed to be delivered with high adherence and minimal variability, while MI that is delivered in the context of frontline community settings is assumed to be delivered with lower adherence and greater variability. Understanding the sources of variability can guide how training, supervision, and quality assurance procedures might better support and enhance MI adherence and intervention quality more generally. Existing research, described below, has shed some light on the extent to which providers differ from one another in MI adherence. In most cases these studies have evaluated associational relationships between setting- or provider-level covariates and MI adherence (e.g., to identify factors that predict adherence) or changes in MI adherence over time (e.g., to evaluate change from pre-training to post-training); however, actual variability in adherence has been examined less frequently. Moreover, there has been almost no work evaluating the extent to which this variability differs across intervention sites or across different research contexts.

Quantifying variability in provider adherence can have important implications for research and clinical practice. If MI adherence differs greatly between different research contexts, it would support the common assumption that the MI that is delivered and tested in more controlled research contexts (e.g., efficacy or effectiveness RCTs) may have minimal resemblance to and less variability than the MI that is delivered in less controlled clinical contexts (e.g., frontline community clinics). This may reveal additional limitations to the generalizability of many MI research findings across different research and clinical contexts. For example, much of the research on MI processes and mechanisms of behavior change is drawn from secondary analyses of efficacy RCTs (Magill et al., 2014; Pace et al., 2017), and the extent to which such research findings generalize to community treatment contexts may be questionable or unclear if the distributions of provider adherence in RCTs are

fundamentally different from those in community treatment agencies. Modeling differences in intervention adherence across these different research and clinical contexts also may be a stepping stone toward formally quantifying the size of the research-to-practice gap in delivery of evidence-based alcohol and drug use disorder interventions. This gap is often cited but rarely quantified, especially through the evaluation of intervention integrity by independent coder ratings. In addition, understanding the sources that contribute to the greatest variability in MI adherence (i.e., variability attributable to session-, provider-, or site-level differences) could also help highlight where intervention quality gaps originate, which may help pinpoint where to target efforts to improve MI adherence (i.e., targeting session-, provider-, or site-level factors that influence MI adherence).

Factors Influencing MI Adherence

Several provider-, session-, and site-level factors may influence MI adherence and therefore create variability in MI adherence. For example, at the provider level, higher adherence is associated with receiving formal MI training and follow-up coaching or feedback (Martino et al., 2016); however, community providers frequently fail to meet commonly-accepted proficiency cutoffs even after MI training (Hall, Staiger, Simpson, Best, & Lubman, 2016). Providers with less confrontational clinical styles and greater MI-related knowledge, MI-related self-efficacy, and positive beliefs about the clinical benefits of MI are often more likely to report intentions to learn and use MI (Walitzer, Dermen, Barrick, & Shyhalla, 2015).

Within-provider factors that influence MI have also been studied. Adherence is likely to improve after MI training (Schwalbe et al., 2014); however, MI adherence is unlikely to improve or may even deteriorate over periods of time after that (Dunn et al., 2016; Miller, Yahne, Moyers, Martinez, & Pirritano, 2004; Moyers et al., 2008; Schwalbe et al., 2014). MI adherence often varies considerably across different sessions administered by the same provider (Dunn et al., 2016; Imel, Baer, Martino, Ball, & Carroll, 2011), which may be due in part to mutual influence between provider behaviors and patients. For example, while provider adherence can influence clients' within-session behavior (Moyers, Houck, Glynn, Hallgren, & Manuel, 2017), client characteristics, such as severity of substance use, motivation for change, or interpersonal aggressiveness can influence provider adherence (Boswell et al., 2013; Imel et al., 2011).

Site-level factors are also associated with providers' MI adherence. Intervention sites that have affiliations with research institutions, more highly educated administrators, or better access to internet technology may have better MI adherence (Lundgren et al., 2011). Organizations with more favorable views of MI and greater support for using MI may have fewer barriers to adopting MI (Walitzer et al., 2015). Additionally, organizations that are more open to change and have less support for provider autonomy may be more likely to increase provider MI adherence after receiving MI training (Baer et al., 2009).

Barriers in Quantifying Variability in Adherence

Despite the reasons to suspect that providers' MI adherence varies across sessions, providers, intervention sites, and research contexts, there has been little to no research that quantifies the extent of these sources of variability. Research in this area has likely been hampered, in part, by the time-consuming nature of observational coding data to conduct meaningful analyses. Obtaining independent ratings of providers' MI adherence requires substantial coder training and imposes significant cost and labor burdens, which often limit the scale to which MI adherence is evaluated. For example, a recent meta-analysis that evaluated associations between coded patient and provider behavior in MI sessions identified only 19 studies with an average of 138 sessions per study (Pace et al., 2017). Other studies evaluating providers' use of MI have often used MI knowledge tests or self-reported use of MI, rather than ratings of MI adherence obtained by independent coders evaluating actual clinical sessions (Madson, Loignon, & Lane, 2009; Romano & Peters, 2015). Such measures are vulnerable to self-report biases that could limit their validity.

Aim and Hypotheses

The present study aimed to quantify the variability of MI adherence in substance use interventions between sessions, providers, and intervention sites across three types of research contexts that commonly evaluate MI adherence (i.e., efficacy, effectiveness, and training study contexts). We utilized MI adherence data, coded from independent raters, obtained from six studies, including: (a) three single-site RCTs and one two-site RCT evaluating MI efficacy, (b) one multisite RCT evaluating MI effectiveness in primary care clinics, and (c) one multisite MI training study in community addiction treatment agencies. The studies varied considerably in terms of their intended research goals, training methods, clients, and provider characteristics.

Informed by previous research (Dunn et al., 2016; Imel et al., 2011), we first hypothesized that for all three research contexts (efficacy, effectiveness, and training studies), variability in MI adherence would be greater within providers (i.e., greater variability between sessions within the same provider's caseload) than it would be between providers within the same context. We did not make specific hypotheses about the degree of variability in MI adherence between intervention sites (i.e., variability between sites within the same research context) due to a lack of previous research on the topic, and instead aimed to quantify and report this source of variability. Second, we hypothesized that MI adherence would be lowest – and most variable – in the less controlled research contexts, such as the MI training study (i.e., where providers had less MI training, less supervision, and more heterogeneous clients). We likewise hypothesized that MI adherence would, on average, be highest and least variable in more highly controlled research contexts, such as in MI efficacy studies (i.e., where providers were usually selected by study investigators based on MI knowledge or willingness to use MI, were given consistent supervision, and where patients were more homogenous). We hypothesized that MI adherence and variability in adherence in the effectiveness study would be intermediary to the levels observed in the efficacy and training studies.

Method

Data Sources and Participants

Data from this study were obtained from six previously published MI efficacy, effectiveness, and training studies. Efficacy studies included four RCTs evaluating single-session interventions with MI and personalized feedback, three of which were single-site studies that addressed high-risk college student drinking (Lee et al., 2014; Neighbors et al., 2012; Tollison et al., 2008) and one of which addressed college student marijuana use for students on two college campuses (Lee et al., 2013). The effectiveness study was a multisite RCT evaluating single-session MI delivered in primary care for reducing illicit drug use (Roy-Byrne et al., 2014). The training study was a cluster randomized trial comparing two methods for training community substance abuse treatment providers in MI (Baer et al., 2009). In total, these studies provided intervention adherence ratings from 1275 MI sessions and 216 providers[1]. The methodologies of these studies are summarized in Table 1 and described in more detail below. Several methodological features that could particularly influence MI adherence are summarized here, including the specific intervention elements in each study, client characteristics, and provider training and supervision. Additional methodological details are available in the parent study publications.

Efficacy studies—The four efficacy RCTs were designed to evaluate brief interventions with MI for alcohol or marijuana use within college student populations. Active intervention sessions were usually 50–60 minutes long and were protocol-guided. Protocols included several common brief intervention components, including personalized feedback regarding (a) clients' patterns of alcohol or marijuana use, including frequency, quantity, and intensity of drinking or marijuana use and consequences related to alcohol or marijuana use, (b) expectancies or motivations for use, (c) perceived social norms and risk factors (e.g., family history of use), (e) education about risks of use, and (f) protective behavioral strategies or other skills to help reduce drinking, marijuana use, and related consequences. Two studies targeted drinking intentions related to specific high-risk events (i.e., 21st birthday drinking or spring break trips; Lee et al., 2014; Neighbors et al., 2012), and two studies targeted drinking or marijuana use more generally (Lee et al., 2013; Tollison et al., 2008).

Clients in the efficacy studies were college students who were randomly selected from university registrar lists to complete initial screening assessments and who subsequently reported heavy drinking (i.e., 4+ drinks for women, 5+ drinks for men) in the past 30 days (Tollison et al., 2008), reported intentions to drink heavily with friends on upcoming 21st birthdays or spring break trips (Lee et al., 2014; Neighbors et al., 2012), or reported marijuana use on five or more days in the past month (Lee et al., 2013). Providers in these studies were typically doctoral-level professionals, clinical psychology graduate students, psychology undergraduates, and post-baccalaureates associated with the college or

---

[1]This total did not include additional sessions from non-MI study conditions or sessions that used standardized clients (i.e., actors). Six providers conducted sessions in two or more of the efficacy studies, and thus there were 38 unique efficacy-study providers in total. Multilevel models treating these providers as nested or non-nested within study (i.e., modeling these providers as if they were unique providers across studies versus modeling them as overlapping across studies) resulted in nearly identical estimates (all ICCs within ±0.02) with no substantive difference in results. We therefore treated these providers as non-nested for ease in modeling and interpretation.

university where the study took place. Providers in these studies were trained in MI and intervention protocols through workshops, role playing, mock sessions with videotaped review and MITI-coded feedback, MI competence evaluation, and ongoing weekly group supervision with videotape review of selected sessions and individual supervision as needed. Providers were required to meet or exceed established MI proficiency thresholds (Moyers, Martin, Manuel, Miller, & Ernst, 2010) before seeing clients.

Effectiveness study—A fifth study evaluated the effectiveness of single-session MI for illicit drug use in primary care sites within a large, university-affiliated medical center (Roy-Byrne et al., 2014). The MI condition utilized a single, 30-minute, protocol-guided MI session that included feedback, exploration of consequences, enhancement of self-efficacy, and explorations of options regarding drug use behavior change.

Clients in the effectiveness trial were recruited from waiting rooms of primary care clinics prior to their primary care medical appointments. All clients used illicit drugs or non-prescribed medications within the past 90 days and had not received substance use treatment within the past month. Much of the sample reported using more than one substance; the most commonly used substances were marijuana (75.6% of sample), alcohol (68.9%), stimulants (41.7%), and opiates (26.3%). A substantial portion of this patient population (30.3%) had been homeless for at least one night in the past 90 days. Co-occurring medical and psychiatric conditions were common (e.g., on average, more than six co-occurring conditions per client coded in medical charts).

Providers in the effectiveness trial were either social workers who were already employed within the clinical system or master's or bachelor's level providers whom study investigators employed into the clinics for the effectiveness trial. The study took place at seven primary care clinics within the same university medical system; however, most providers ($n = 9$) delivered MI sessions at multiple sites (median = 3 sites; range = 1–6 sites). As a result of this substantial degree of non-nesting, between-clinic effects in this study could not be differentiated from between-provider effects in the modeling procedure and all of the sessions were modeled as occurring within a single site in subsequent analyses. Providers received training through workshops and rehearsal with up to five standardized patients before seeing study participants, then received ongoing feedback from training supervisors. Providers continued to receive monthly in-person group supervision and were emailed with feedback on their intervention adherence ratings for approximately 25% of their clinical sessions.

Training study—The sixth study was a multi-site evaluation of two methods for training MI to frontline providers (Baer et al., 2009). In one condition, providers received a standard two-day MI workshop. In the other condition participants received context-tailored MI training, delivered over five 2–3-hour group training sessions scheduled two weeks apart. MI training sessions included practice interviews and feedback with standardized clients, discussion of challenging situations that are common in providers' specific clinical sites, and attempts to build agency-level support for using MI. Administrators asked providers in both conditions to integrate MI into their counseling style rather than deliver a protocol-specific intervention.

Study investigators recruited providers from eight community clinics, including six sites reported in Baer et al. (2009) and two pilot sites, plus an open enrollment group that was modeled here as a ninth site. Providers generally had little or no formal MI training prior to the study. Providers' education levels varied from graduate level to high school diploma or equivalent, and approximately half were licensed chemical dependency professionals. No patient data were collected in this provider training study. The parent study identified no differences in provider MI adherence for context-tailored MI training compared to training as usual, and therefore sites receiving each type of training were not distinguished in the present analysis.

Intervention Adherence Measure

MI sessions were rated by each original research team using the Motivational Interviewing Treatment Integrity (MITI) coding system (Moyers et al., 2005). Per the MITI protocol, 20-minute segments were randomly selected from each MI session then coded by one or more trained raters (in Baer et al., sessions were typically about 20 minutes and whole sessions were coded regardless of actual session length). Raters listened to the audio from these segments and provided global ratings encapsulating empathy (i.e., evidence of providers understanding clients' feelings and perspectives) and MI spirit (i.e., providers' evocation of reasons for change, fostering of power sharing, and support for clients' autonomy) across the entire session segments. Raters also tallied individual provider utterances as complex or simple reflections, closed or open questions, MI-adherent (e.g., emphasizing client control, affirming client choices), MI non-adherent (e.g., giving advice without permission, confronting), or MI-neutral (e.g., giving information). Utterances in each of these categories were summed to derive total behavior counts for each session segment, which were then used to compute four indices of MI technical proficiency described by the MITI: percentages of open (versus closed) questions (%OQ), percentages of complex (versus simple) reflections (%CR), ratios of reflections to questions (R:Q), and percentages of MI-adherent (versus non-adherent) behaviors. Higher values in all indices reflect better MI adherence.

Coder teams in each of the parent studies included multiple independent raters who were typically research assistants and/or undergraduate and graduate university students. Coder training typically included review of intervention and coding manuals, review of example audio-and video-recorded MI sessions, and graded coding practice requiring a specific criterion level of agreement before coding study sessions.

Five of the parent studies (Baer et al., 2009; Lee et al., 2013, 2014; Neighbors et al., 2012; Tollison et al. 2008) used the MITI version 2.0 (Moyers et al., 2005) and one parent study (Roy-Byrne et al., 2014) used the MITI version 3.1 (Moyers et al., 2010). Coding criteria in the MITI 2.0 and 3.1 overlapped considerably but included two differences of relevance to the present analysis. First, global codes in the MITI 2.0 used a 7-point Likert-type scale whereas the MITI 3.1 used a 5-point scale; therefore, we rescaled the MITI 3.1 global ratings from Roy-Byrne et al. (2014) onto a 7-point scale by multiplying each global rating by 1.4. Second, the MITI 3.1 yielded separate global ratings for evocation, collaboration, and autonomy/support that were then averaged to estimate the degree of MI spirit. In

contrast, the MITI 2.0 provided one overall MI spirit rating that instructed coders to account for evocation, collaboration, and autonomy/support in combination. Thus, we obtained MI spirit ratings from a single item in the MITI 2.0 and from the averages of three ratings in the MITI 3.1.

Inter-rater reliability estimates from original studies are summarized in Table 1. Reliability estimates were usually good to excellent for each MITI index (e.g., > 0.6; Cicchetti, 1994) but were occasionally in a range considered to be fair (between 0.41 and 0.60). Inter-rater reliability estimates for percentages of MI adherent behaviors were sometimes low (e.g., 0.13 in Roy-Byrne et al., 2014), which was attributable to restriction of range (i.e., few MI non-adherent behaviors) and these scores were therefore excluded from analysis in the present study due to low reliability.

Analytic Plan

We used multilevel modeling to evaluate differences in means and variances of MI adherence ratings across the efficacy, effectiveness, and training study contexts (Raudenbush & Bryk, 2002). Three-level models partitioned sources of variability for each MITI index as being attributable to differences between sites, differences between providers (within sites), and differences between sessions (within providers and sites). The following formula represented the three-level model:

$$\mathrm{MITI_{ijk}} = b_1 \mathrm{Efficacy}_k + b_2 \mathrm{Effectiveness}_k + b_3 \mathrm{Training}_k +$$
$$\nu_{\mathrm{Efficacy},k} + \nu_{\mathrm{Efficacy},k} +$$
$$u_{\mathrm{Efficacy,jk}} + u_{\mathrm{Effectiveness,jk}} + u_{\mathrm{Training,jk}} +$$
$$e_{\mathrm{Efficacy,ijk}} + e_{\mathrm{Effectiveness,ijk}} + e_{\mathrm{Training,ijk}}$$

In the above formula, *MITI$_{ijk}$* is an MI adherence score for session *i*, conducted by provider *j*, at site *k*. Three indicator variables (b_1, b_2, and b_3) were used to estimate means from the three research contexts (efficacy, effectiveness, and training studies, respectively). The suppression of an overall intercept allowed the estimates for b_1, b_2, and b_3 to reflect the estimated mean MI adherence for each type of research context. The random effects *v*, *u*, and *e* captured the variance components that reflect differences in adherence ratings between sites, providers, and sessions, respectively. Random effects were stratified by research context (efficacy, effectiveness, or training study), which allowed the model to estimate unique variance components that were specific to each of the three research contexts. In the efficacy studies, the three single-site RCTs and the two campuses in the two-site study were each treated as different sites (*k*-level variable); in the training study, each of the nine community treatment sites was considered a unique site. In the effectiveness study, between-site differences could not be estimated due to substantial non-nesting of providers within site for this study (see Data Source and Participants section above); therefore, we omitted site-level random effects regarding the effectiveness study. Random effects were assumed to be multivariate normally distributed and uncorrelated with means of zero.

Mean levels of MI adherence across the three research contexts were estimated based on fixed effect coefficients (b_1, b_2, and b_3) and 95% confidence intervals (CIs). Variability in

MI adherence was evaluated in two ways. First, intra-class correlation coefficients (ICCs; Raudenbush & Bryk, 2002) were computed to indicate the proportions of total variance within each research context that could be attributed to between-session, between-provider, and between-site variability. ICCs were computed by dividing the variance estimate at each level by the sum of the variance components across levels. Second, the magnitudes of variances in the random effect coefficients (v, u, and e, variances) were compared across research contexts to evaluate whether the overall amount of variability differed between the efficacy, effectiveness, and training study contexts. In this latter analysis, differences in variance estimates at each of the three levels (site, provider, client) were compared between research contexts via nested model comparison. Chi-square tests evaluated the incremental improvement in model fit for the full model (described above) compared to a restricted model that constrained variance terms from each combination of research contexts to be equal (e.g., $u_{Efficacy,jk}$ and $u_{Effectiveness,jk}$ constrained to be equal). A significant nested model comparison would indicate that variance estimates between the two research contexts were significantly different. Models were fit using *nlme* (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2016) and *lme4* (Bates, Machler, Bolker, & Walker, 2015) in R (R Core Development Team, 2015).

Results

Mean Differences in MI Adherence Across Research Contexts

Mean levels of MI adherence were compared between the three research contexts to evaluate the extent to which average MI adherence differed between the efficacy, effectiveness, and training studies. Figure 1 displays mean MITI ratings (and 95% CIs) for the three research contexts, estimated as fixed effect coefficients (b_1, b_2, and b_3) in multilevel models. In line with our hypotheses, the training study had significantly lower empathy and MI spirit ratings than the efficacy and effectiveness studies. The training study also had lower %OQ than the efficacy studies and lower R:Q than the effectiveness study. Contrary to our hypotheses, the effectiveness study had higher empathy, MI spirit, R:Q, and %CR than the efficacy studies.

Occasionally, there were no mean differences between research contexts for some adherence measures. Specifically, the training study was not significantly different from the efficacy studies on mean R:Q (difference = −0.17, SE = 0.14, p = .24) and mean %CR (difference = 6.99, SE = 4.27, p = .12). Additionally, the effectiveness study was not significantly different from the training study on mean %OQ (difference = −4.18, SE = 2.78, p = .14) or mean %CR (difference = −12.68, SE = 8.64, p = .18). However, in total, there was always at least one significant difference in mean levels of MI adherence between two or more of the three research contexts for each MITI index. Mean adherence ratings in the training study were never significantly higher than the thresholds that indicate basic competence in MI (i.e., confidence intervals entirely above the horizontal dotted lines in Figure 1); however, the efficacy and effectiveness studies occasionally exceeded the thresholds for basic competence.

Variability in MI Adherence

Variability within research contexts—The proportions of variability in MI adherence attributable to between-site, between-provider, and between-session differences were quantified for each research context. Table 2 provides ICC estimates that quantify the relative proportions of total variance in MI adherence that were attributable to between-site, between-provider, and between-session differences in each of the three research contexts. Across adherence measures, the amount of variability attributable to between-session differences was always greater than the variability attributable to between-provider and between-site differences (i.e., ICC values greater than 0.50 for between-session effects). Smaller proportions of variability were accounted for by between-provider differences (ICC=.03 to .26) and between-site differences (ICC=.001 to .28), which were considerably lower than the proportions of variability attributable to between-session differences (ICC=. 57 to .94).

Variability between research contexts—The magnitudes of the variance components reflecting between-session, between-provider, and between-site variability were estimated and compared across the three research contexts (efficacy, effectiveness, training). Figure 2 displays these variance component magnitudes grouped according to between-site, between-provider, and between-session components (different *x*-axis groupings) for each research context (differently shaded bars). The bar heights and 95% CI's illustrate the multilevel model-estimated variability in MI adherence expressed in standard-deviation units of the original MITI index (*y*-axis). Significant differences in these variability estimates between research contexts (i.e., within each grouping of bars) are indicated by asterisks.

Our hypothesis of lower variability in the efficacy studies, compared to the effectiveness and training studies, was mostly supported for the between-session variances. The efficacy studies had lower between-session variability compared to the effectiveness and training studies for measures of empathy, MI spirit, R:Q, and %OQ. The efficacy studies also had lower between-session variability than the training study for %CR, but the variability observed in efficacy and effectiveness studies did not differ for this measure. Our hypothesis that variability would be lowest within the efficacy studies was mostly unsupported for the between-provider variances: efficacy study providers had less between-provider variability in empathy than training study providers and less between-provider variability in R:Q than the effectiveness study, but there were no other differences in variability for the other MITI indices at the between-provider level.

Our hypothesis of higher variability in the training study was also mostly supported at the between-session level. The training study had higher between-session variability compared to both the efficacy and effectiveness studies for empathy, MI spirit, and %CR. As mentioned above, the training study also had higher between-session variability compared to the efficacy studies for R:Q and %OQ. However, in contrast with our hypotheses, the effectiveness study had the greatest between-session variability in R:Q and %OQ. Our hypothesis of greater variability in the training study was mostly unsupported at the between-provider level of analysis: as mentioned above, training study providers had higher between-provider variability in empathy than the efficacy study providers; however training

study providers also had lower between-provider variability in R:Q and had no differences in between-provider variability for the other MITI indices.

We also explored the amount of variability between intervention sites (within research context). Only some of the between-site variability estimates were significantly greater than zero, including R:Q and %OQ for the efficacy studies and MI spirit and %CR for the training study (indicated by 95% CI's above zero in Figure 2).

Figure 3 illustrates the distributions of provider- and site-level MITI ratings within each research context. The density plots in Figure 3 show the distributions of individual, provider-level estimates of MI adherence from fitted multilevel models, combining fixed effects with site-and provider-level random effects (i.e., model-implied estimates of providers' mean MI adherence ratings). The heights of the density plots represent the relative proportion of providers within each research context who are estimated to have a specific level of MI adherence (*x*-axis). Site-level variability is illustrated by the horizontal lines plotted above the distribution of each research context: the spans of these lines across the *x*-axis reflect the inner-quartile range (25th to 75th percentile) of MI adherence for providers within each research context's individual sites. MITI 2.0 thresholds for basic competence and proficiency are indicated by vertical dotted and dashed lines, respectively. As shown in the figure, there was considerable separation of provider-level and site-level MI adherence across the three research contexts. For example, there was almost no overlap in provider- and site-level distributions for the two global ratings, empathy and MI spirit, indicating a low probability that two providers from different research contexts would have similar mean levels of empathy or MI spirit. Providers in different research contexts were also likely to be differently categorized as being below or above MITI 2.0 thresholds for beginning proficiency and competency. There was greater overlap for the three behavioral codes, but still noticeable separation between research contexts for %CR and %OQ. Most training study providers and sites had similarly low R:Q ratios with little between-provider and between-site variability, whereas efficacy and effectiveness study providers had greater between-provider and between-site variability. In sum, the differences between providers within the same research context (ranges of distributions within each density curve) were often smaller than the overarching differences between research contexts (distances between density curves).

Discussion

In the present samples, MI adherence varied substantially and systematically by research context (efficacy, effectiveness, and training), as well as between clinical sessions, providers, and sites within these research contexts. To our knowledge, this study comprises the largest evaluation of variability in MI adherence and may be the first to evaluate variability for any behavioral intervention between sessions, providers, intervention sites, and research contexts.

Consistent with our hypotheses, MI evaluated within the context of a multisite training study often had lower overall adherence and greater between-session variability in adherence compared to the efficacy and effectiveness studies. Also, consistent with our hypotheses, MI

delivered in the context of efficacy studies often had lower between-session variability than that delivered in the effectiveness and training studies. However, in contrast with our hypotheses, the effectiveness study had higher mean levels of MI adherence than the efficacy studies on four of the five MITI indices that were tested. Only two MITI indices (empathy and R:Q) had significant differences in the degree of between-provider variability. We found no differences in the amount of between-site variability across research contexts; however, this finding should be couched by the relatively small number of intervention sites. Overall, the clear majority of variance in MI adherence could be attributed to between-session differences. While between-provider and between-site variability was relatively smaller, these often accounted for substantial proportions of variance and were often statistically non-negligible.

Reasons for Differences Between Research Contexts

Training study—The relatively lower mean levels of and greater variability in MI adherence within the training study could be attributable to several factors that complicate the delivery of high-fidelity MI in community treatment contexts. Providers in the training study were mostly chemical dependency treatment professionals with minimal prior MI training who may frequently use clinical styles that are inconsistent with MI (e.g., emphasizing confrontation of resistance, advising clients to change alcohol- and drug-related behaviors, and warning clients around the negative consequences of substance use; e.g., Perkinson, 2017). Like many other MI training studies, the protocol in the training study included here was also designed to help providers integrate MI into their existing clinical practices. Folding highly adherent MI into an existing clinical routine that consists of many other practices (e.g., case management) may result in lower MI adherence than MI delivered within a structured and stand-alone protocol. Thus, even though the MI trainings offered in the original training study significantly improved adherence (Baer et al., 2009), fidelity ratings were often still below thresholds indicating beginning proficiency. Although patient data was not collected in the training study, it is also likely that clients in the community treatment agencies frequently had co-occurring mental health, medical, legal, social, and financial concerns that may have competed with the delivery of more highly adherent MI for alcohol and drug use. While we lacked data on organizational factors that could contribute to MI adherence, it is also plausible that, at an organizational level, the training study sites could have had relatively less organizational support for MI, less immediate access to MI experts, and initially less favorable attitudes toward using MI compared to the efficacy and effectiveness studies, which could lead to greater difficulty implementing higher-fidelity MI (Lundgren et al., 2011; Walitzer et al., 2015).

Efficacy and effectiveness studies—Likewise, many design and contextual factors may have facilitated the relatively higher levels of (and lower between-session variability in) MI adherence that was often observed in the efficacy and effectiveness studies. Each of the efficacy and effectiveness studies utilized stand-alone, single-session intervention protocols delivered by a provider who was not responsible for delivering other forms of care or addressing other co-occurring issues during the observed intervention sessions, which may have helped them deliver MI with higher fidelity. Providers in the efficacy and effectiveness studies also obtained regular and ongoing supervision. The efficacy and effectiveness

studies' use of intervention protocols to specify how MI sessions should be delivered also likely helped provide a standard structure to those sessions, potentially facilitating higher MI adherence with less variability between sessions. For example, interventions in the four efficacy studies were heavily guided by a standardized personalized feedback report. However, this use of protocols did not lead to consistently lower variability between providers in the efficacy and effectiveness studies as we expected. Provider attitudes, buy-in, and support for MI were also likely more favorable in the efficacy and effectiveness research contexts, as MI may have been more likely to be championed as an effective approach for mobilizing changes in substance use by the researcher teams invested in testing its efficacy and effectiveness.

The higher mean levels of MI adherence in the effectiveness study, relative to the efficacy studies, was surprising. We expected the opposite result due to several factors that can degrade intervention adherence within frontline primary care clinics (versus university research labs). Patients in the effectiveness study analyzed here had high rates of co-occurring medical issues, polysubstance use, and frequent homelessness, and were therefore likely to be more heterogeneous in the scope of their substance use and related problems than the college student participants in the efficacy studies. Previous studies have demonstrated mixed levels of MI adherence when delivered in primary care; however, many existing studies that have shown lower MI adherence in primary care had the intervention delivered by primary care physicians and nurses (Mullin, Forsberg, Savageau, & Saver, 2015; Noordman, van der Lee, Neilken, & Van Dulmen, 2012; Östlund, Kristofferzon, Häggström, & Wedensten, 2015), whereas other studies with higher adherence, including the present study, had MI delivered by trained counselors (e.g., Aharonovich et al., 2017). These findings suggest that implementing high-fidelity MI in primary care at a level that is comparable to that observed in many efficacy studies may be achievable, particularly when it is delivered by specialists in counseling and social work professions. The delivery of alcohol and drug use interventions by such specialists in primary care clinics mirrors the design of many existing integrated care models for other mental health concerns (e.g., Collaborative Care model for depression; Sederer, Derman, Carruthers, & Wall, 2016), which may serve as models for the delivery of high-fidelity alcohol and drug interventions in primary care.

Variability between sessions, providers, and sites—The greatest degree of variability in MI adherence was attributable to between-session variability (versus between-provider and between-site variability). It was not possible to distinguish how much of the between-session variability could be attributable to differences between clients versus differences within therapists over time, but previous research has demonstrated that between-session variability due to both factors can be substantial (Dunn et al., 2016; Imel et al., 2011). The findings observed here are consistent with these previous findings and suggest that the preponderance of between-session variability is robust across research contexts. Additional research is warranted to identify factors that contribute to this large degree of within-provider variability, yet isolating the exact causes of this variance will likely be challenging due to a number of contributing factors that could vary both between and within clients (e.g., motivation for change, addiction severity, psychiatric symptom

severity; Imel et al., 2011) and between and within providers (e.g., size of caseloads, availability of time and resources; Coleman et al., 2017).

Implications

The findings of the present study lend themselves to practical concerns about MI delivery across different contexts. Most notably, the fidelity of MI that clients receive, operationalized here by MI adherence ratings, is likely to vary based on the context in which MI is delivered, and to a lesser but still substantial extent, the site at which they receive intervention and the provider with whom they work.

Improving MI service delivery—The profound differences in average levels of adherence between research contexts and strong degree of variability between clinical sessions suggests that future research and implementation efforts should aim to better understand and potentially ameliorate factors at these levels to improve the fidelity of MI service delivery. For example, although Miller and others (Miller & Moyers, 2016; Miller, Moyers, Arciniega, Ernst, & Forcehimes, 2005; Imel, Sheng, Baldwin, & Atkins, 2015) have argued that focusing hiring decisions based on pre-existing MI adherence could be one way to improve MI implementation, such efforts may be challenged by the substantial degree by which MI varies based on the context in which it is used and its high degree of variability within providers. It is possible that such decisions could also be based on information about a provider's variability in MI adherence over several sessions, and not merely an average score or single session sample. Substantial within-provider variability further suggests that evaluating only a single session of MI adherence is unlikely to yield a reliable estimate of a provider's typical level of adherence. Therefore, training, supervision, and implementation programs should acquire multiple ratings per provider, when possible, to more adequately characterize that provider's MI adherence. Developments in automated MI fidelity coding could support larger-scale assessment of provider adherence and facilitate performance-based feedback in closer temporal proximity to sessions (Tanana, Hallgren, Imel, Atkins, & Srikumar, 2016), which could potentially help providers maintenance adherence (Schwalbe et al., 2014).

Implications for MI process research—The results of the present study have implications for research on within-session processes and mechanisms of behavior change in MI. Such research often aims to understand how a provider's MI adherence impacts a clients' in-session behavior, which in turn is hypothesized to direct a client's longer-term alcohol and drug use outcomes. In light of the substantially different distributions in provider adherence between RCTs and community treatment clinics, it is unclear whether and to what extent existing research on within-session processes, which is most commonly generated from secondary analyses of efficacy and effectiveness study sessions, generalizes to the community treatment contexts. For example, provider MI adherence may predict client change talk in some efficacy-focused clinical trials (e.g., Moyers et al., 2009) but not in other trials with a more pragmatic focus (e.g., Palfai et al., 2016). In addition, the substantial variability in MI adherence within providers suggests that client-level factors likely contribute to providers' MI adherence, yet there is limited research exploring *which* client-level factors contribute to providers' within-session adherence and *how* this influence

occurs (e.g., Gaume, Gmel, Faozi, & Daeppen, 2008; Gaume et al., 2016; Imel et al., 2011). Future research may aim to better understand the factors that facilitate heterogeneity in within-session behavior.

Research-to-practice gap—Finally, the substantial differences in average MI adherence between research contexts (efficacy, effectiveness, and training studies) underscore potential gaps in the fidelity of evidence-based intervention delivery across treatment service settings. The fidelity of MI that is delivered in controlled efficacy and effectiveness studies is substantially different from that which is delivered by frontline providers with heterogeneous clients in community treatment contexts. While the assumption of this research-to-practice gap forms the basis of many training and implementation efforts, the present study provides support for this gap explicitly and provides quantitative measurements of the extent to which that gap exists. These differences in MI adherence between different intervention contexts complicate our understanding of MI's effectiveness. The impact of MI on clinical outcomes may be well-supported in many efficacy studies, yet non-generalizable to other real-world contexts due to substantial differences in how MI can feasibly be delivered (Hall et al., 2015).

Limitations

The present study has noteworthy limitations. The parent studies had heterogeneous participant samples, study designs, targeted substances of abuse, and specifications for how MI should be delivered. While this heterogeneity was a strength of the study, it also limited the extent to which specific reasons for different findings between studies could be attributable to a single factor. Client outcomes and session-, provider-, and site-level covariates that could affect treatment adherence also were measured differently between studies, or not measured at all, which precluded inclusion of those variables in the analysis. The extent to which coding practices and training varied between the parent studies also could not be evaluated: although the parent studies used nearly identical coding systems and training procedures, it is possible that the studies unknowingly emphasized different coding practices that could have resulted in between-study differences.

There were varying numbers of clients, providers, and sites across the three research contexts, which in turn affected the precision of the analysis. While multilevel models can yield unbiased estimates of such designs with unbalanced cluster sizes, this heterogeneity affected the likelihood of detecting differences in variability at different levels of analysis. For example, differences in sample size at each level led to standard errors that were smaller for between-session variance estimates (i.e., larger number of sessions yields higher precision of session-level variance estimates) and standard errors that were larger for between-site variance estimates (i.e., smaller number of sites yields lower precision of variance estimates).

Measurement error also poses limitations to the results obtained here. Measurement error (e.g., due to differences in ratings between coders) likely increased the degree of variability in MITI ratings and we were unable to remove error variance from our model estimates. While this limitation is present for most multilevel modeling research, it is worth noting that

this measurement error could lead to inflated variance estimates, particularly for session-level variances. Measurement error was particularly high for the MITI index reflecting percentage of MI adherent behaviors, which precluded our ability to include this index in our analyses.

Finally, these findings should be considered preliminary until further, larger-scale research can be conducted. Additional research is warranted to evaluate variability in MI adherence using a greater number of treatment sites and greater variability in populations and settings. For example, three of the efficacy studies used the same framework for delivering MI-based alcohol interventions with college students (i.e., Brief Alcohol Screening and Intervention for College Students, Dimeff et al., 1999), and the fourth efficacy study used a similar framework focused on marijuana use. Future work could extend this work to include efficacy studies with more diverse intervention frameworks and treatment samples. Similarly, the inclusion of only one effectiveness study and one training study also may have limited the extent to which variability could be observed within and between each of these research contexts, and including more studies with different populations of providers and clients or with different degrees of success in implementing MI into their clinical practice could have created greater variability than what was observed here. Larger-scale studies quantifying the degree of variability in MI adherence across a greater number of sites is therefore warranted.

Conclusion

MI adherence varies substantially from session to session and across research contexts. Factors that create this variability not yet well understood. The substantial influence of research context on provider behavior highlights gaps in translating research to practice and limits generalizability of findings across settings and contexts.

Acknowledgments

References

Aharonovich E, Sarvet A, Stohl M, DesJarlais D, Tross S, Hurst T, Urbina A, Hasin D. Reducing non-injection drug use in HIV primary care: A randomized trial of brief motivational interviewing, with and without HealthCall, a technology-based enhancement. Journal of Substance Abuse Treatment. 2017; 74:71–79. [PubMed: 28132704]

Baer JS, Wells EA, Rosengren DB, Hartzler B, Beadnell B, Dunn C. Agency context and tailored training in technology transfer: A pilot evaluation of motivational interviewing training for community counselors. Journal of Substance Abuse Treatment. 2009; 37(2):191–202. https://doi.org/10.1016/j.jsat.2009.01.003. [PubMed: 19339139]

Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. Journal of Statistical Software. 2015; 67(1):1–48. https://doi.org/10.18637/jss.v067.io1.

Boswell JF, Gallagher MW, Sauer-Zavala SE, Bullis J, Gorman JM, Shear MK, Barlow DH. Patient characteristics and variability in adherence and competence in cognitive-behavioral therapy for panic disorder. Journal of Consulting and Clinical Psychology. 2013; 81(3):443–454. https://doi.org/10.1037/a0031437. [PubMed: 23339537]

Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment. 1994; 6(4):284–290. https://doi.org/10.1037/1040-3590.6.4.284.

Coleman KJ, Hemmila T, Valenti MD, Smith N, Quarrell R, Ruona LK, Brandenfels E, Rossom RC. Understanding the experience of care managers and relationship with patient outcomes: The COMPASS initiative. General Hospital Psychiatry. 2017; 44:86–90. [PubMed: 27558105]

Dimeff, LA., Baer, JS., Kivlahan, DR., Marlatt, GA. Brief alcohol screening and intervention for college students: A harm reduction approach. New York: Guilford Press; 1999.

Dunn C, Darnell D, Atkins DC, Hallgren KA, Imel ZE, Bumgardner K, Roy-Byrne P. Within-provider variability in motivational interviewing integrity for three years after MI training: Does time heal? Journal of Substance Abuse Treatment. 2016; 65:74–82. https://doi.org/10.1016/j.jsat.2016.02.008. [PubMed: 27016875]

Gaume J, Gmel G, Faouzi M, Daeppen JB. Counsellor behaviours and patient language during brief motivational interventions: a sequential analysis of speech. Addiction. 2008; 103(11):1793–1800. [PubMed: 19032529]

Gaume J, Longabaugh R, Magill M, Bertholet N, Gmel G, Daeppen J. Under what conditions? Therapist and client characteristics moderate the role of change talk in brief motivational intervention. Journal of Consulting and Clinical Psychology. 2016; 84(3):211–220. [PubMed: 26727413]

Hall K, Staiger PK, Simpson A, Best D, Lubman DI. After 30 years of dissemination, have we achieved sustained practice change in motivational interviewing? Addiction. 2015; 111:1144–1150. [PubMed: 26216706]

Imel ZE, Baer JS, Martino S, Ball SA, Carroll KM. Mutual influence in therapy competence and adherence to motivational enhancement therapy. Drug and Alcohol Dependence. 2011; 115:229–236. [PubMed: 21227601]

Imel ZE, Sheng E, Baldwin SA, Atkins DC. Removing very low-performing therapists: A simulation of performance-based retention in psychotherapy. Psychotherapy. 2015; 52(3):329–336. DOI: 10.1037/pst0000023 [PubMed: 26301424]

Lee CM, Kilmer JR, Neighbors C, Atkins DC, Zheng C, Walker DD, Larimer ME. Indicated prevention for college student marijuana use: A randomized controlled trial. Journal of Consulting and Clinical Psychology. 2013; 81(4):702–709. https://doi.org/10.1037/a0033285. [PubMed: 23750464]

Lee CM, Neighbors C, Lewis MA, Kaysen D, Mittmann A, Geisner IM, Larimer ME. Randomized controlled trial of a spring break intervention to reduce high-risk drinking. Journal of Consulting and Clinical Psychology. 2014; 82(2):189–201. https://doi.org/10.1037/a0035743. [PubMed: 24491072]

Lundahl B, Burke BL. The effectiveness and applicability of motivational interviewing: A practice-friendly review of four meta-analyses. Journal of Clinical Psychology. 2009; 65(11):1232–1245. https://doi.org/10.1002/jclp.20638. [PubMed: 19739205]

Lundgren L, Chassler D, Amodeo M, D'Ippolito M, Sullivan L. Barriers to implementation of evidence-based addiction treatment: A national study. Journal of Substance Abuse Treatment. 2012; 42(3):231–238. https://doi.org/10.1016/j.jsat.2011.08.003. [PubMed: 22000323]

Madson MB, Loignon AC, Lane C. Training in motivational interviewing: A systematic review. Journal of Substance Abuse Treatment. 2009; 36(1):101–109. https://doi.org/10.1016/j.jsat.2008.05.005. [PubMed: 18657936]

Magill M, Gaume J, Apodaca TR, Walthers J, Mastroleo NR, Borsari B, Longabaugh R. The technical hypothesis of motivational interviewing: A meta-analysis of MI's key causal model. Journal of Consulting and Clinical Psychology. 2014; 82(6):973–983. https://doi.org/10.1037/a0036833. [PubMed: 24841862]

Martino S, Paris M, Añez L, Nich C, Canning-Ball M, Hunkele K, Carroll KM. The effectiveness and cost of clinical supervision for motivational interviewing: A randomized controlled trial. Journal of Substance Abuse Treatment. 2016; 68:11–23. [PubMed: 27431042]

Miller, WR., Rollnick, S. Motivational Interviewing: Preparing People for Change. 3. New York: Guilford; 2013.

Miller WR, Moyers TB. Asking better questions about clinical skills training. Addiction. 2016; 111(7): 1151–1152. https://doi.org/10.1111/add.13095. [PubMed: 26929059]

Miller WR, Moyers TB, Arciniega L, Ernst D, Forcehimes A. Training, supervision and quality monitoring of the COMBINE Study behavioral interventions. Journal of Studies on Alcohol. Supplement. 2005; (15):188–195. 169. [PubMed: 16223070]

Miller WR, Yahne CE, Moyers TB, Martinez J, Pirritano M. A randomized trial of methods to help clinicians learn motivational interviewing. Journal of Consulting and Clinical Psychology. 2004; 72(6):1050–1062. [PubMed: 15612851]

Moyers TB, Houck J, Glynn LH, Hallgren KA, Manuel JK. A randomized controlled trial to influence client language in substance use disorder treatment. Drug and Alcohol Dependence. 2017; 172:43–50. https://doi.org/10.1016/j.drugalcdep.2016.11.036. [PubMed: 28122270]

Moyers TB, Martin T, Houck JM, Christopher PJ, Tonigan JS. From in-session behaviors to drinking outcomes: A causal chain for motivational interviewing. Journal of Consulting and Clinical Psychology. 2009

Moyers TB, Martin T, Manuel JK, Hendrickson SML, Miller WR. Assessing competence in the use of motivational interviewing. Journal of Substance Abuse Treatment. 2005; 28(1):19–26. https://doi.org/10.1016/j.jsat.2004.11.001. [PubMed: 15723728]

Moyers, TB., Martin, T., Manuel, JK., Miller, WR., Ernst, D. Revised global scales: Motivational Interviewing Treatment Integrity 3.1.1 (MITI 3.1.1). 2010. Unpublished coding manual, accessed June 12, 2017 at https://casaa.unm.edu/download/miti3_1.pdf

Moyers TB, Miller WR. Is low therapist empathy toxic? Psychology of Addictive Behaviors. 2013; 27(3):878–884. https://doi.org/10.1037/a0030274. [PubMed: 23025709]

Mullin D, Forsberg L, Savageau JA, Saver BG. Challenges in developing primary care physicians' motivational interviewing skills. Family Systems & Health. 2015; 33(4):330–338.

Neighbors C, Lee CM, Atkins DC, Lewis MA, Kaysen D, Mittmann A, Larimer ME. A randomized controlled trial of event-specific prevention strategies for reducing problematic drinking associated with 21st birthday celebrations. Journal of Consulting and Clinical Psychology. 2012; 80(5):850–862. https://doi.org/10.1037/a0029480. [PubMed: 22823855]

Noordman J, van der Lee I, Nielken M, Van Dulmen AM. Do trained practice nurses apply motivational interviewing techniques in primary care consultations? Journal of Clinical Medicine Research. 2012; 4(6):393–401. [PubMed: 23226172]

Östlund A, Kristofferzon M, Häggström E, Wedensten B. Primary care nurses' performance in motivational interviewing: A quantitative descriptive study. BMC Family Practice. 2015; 16(89):1–12. [PubMed: 25608667]

Pace BT, Dembe A, Soma CS, Baldwin SA, Atkins DC, Imel ZE. A multivariate meta-analysis of motivational interviewing process and outcome. Psychology of Addictive Behaviors. 2017; Epub ahead of print. doi: 10.1037/adb0000280

Palfai TP, Cheng DM, Bernstein JA, Palmisano J, Lloyd-Travaglini C, Goodness T, Saitz R. Is the quality of brief motivational interventions for drug use in primary care associated with subsequent drug use? Addictive Behaviors. 2016; 56(8):8–14. [PubMed: 26779816]

Perkinson, RR. Chemical dependency counseling: A practical guide. 5. Thousand Oaks, CA: Sage; 2017.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., R Core Team. nlme: Linear and nonlinear mixed effects models [computer software]. 2016. Retrieved from https://CRAN.R-project.org/package=nlme

R Core Team. R: A language and environment for statistical computing [software]. 2015. Retrieved from https://www.R-project.org

Raudenbush, SW., Bryk, AS. Hierarchical linear models: Applications and data analysis methods. Thousand Oaks, CA: Sage; 2002.

Romano M, Peters L. Evaluating the mechanisms of change in motivational interviewing in the treatment of mental health problems: A review and meta-analysis. Clinical Psychology Review. 2015; 38:1–12. https://doi.org/10.1016/j.cpr.2015.02.008. [PubMed: 25779459]

Roy-Byrne P, Bumgardner K, Krupski A, Dunn C, Ries R, Donovan D, Zarkin GA. Brief intervention for problem drug use in safety-net primary care settings: a randomized clinical trial. JAMA. 2014; 312(5):492–501. https://doi.org/10.1001/jama.2014.7860. [PubMed: 25096689]

Sederer LI, Derman M, Carruthers J, Wall M. The New York state collaborative care initiative: 2012–2014. Psychiatric Quarterly. 2016; 87:1–23. [PubMed: 26040961]

Schwalbe CS, Oh HY, Zweben A. Sustaining motivational interviewing: A meta-analysis of training studies. Addiction. 2014; 109(8):1287–1294. [PubMed: 24661345]

Tollison SJ, Lee CM, Neighbors C, Neil TA, Olson ND, Larimer ME. Questions and reflections: The use of motivational interviewing microskills in a peer-led brief alcohol intervention for college students. Behavior Therapy. 2008; 39(2):183–194. https://doi.org/10.1016/j.beth.2007.07.001. [PubMed: 18502251]

Walitzer KS, Dermen KH, Barrick C, Shyhalla K. Modeling the innovation-decision process: Dissemination and adoption of a motivational interviewing preparatory procedure in addiction outpatient clinics. Journal of Substance Abuse Treatment. 2015; 57:18–29. https://doi.org/10.1016/j.jsat.2015.04.003. [PubMed: 25934460]

Highlights

- This is the first study to formally quantify variability in motivational interviewing (MI) adherence across sessions, providers, settings, and research contexts.

- We found lower and more variable MI adherence in community clinics compared to efficacy and effectiveness trials.

- The quality of MI one receives depends greatly on the context in which it is delivered.

- Findings underscore and quantify the gap between clinical research and frontline practice.
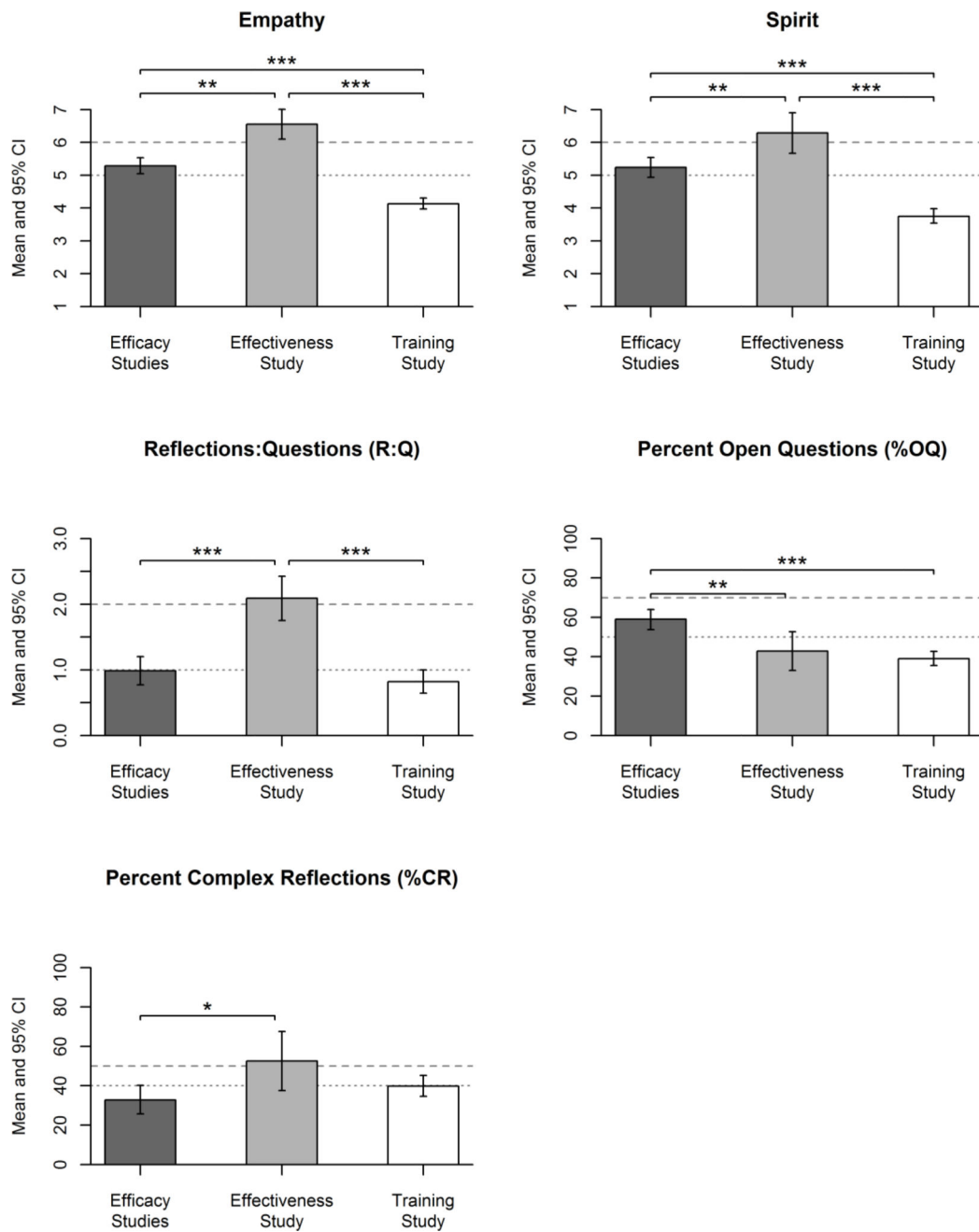
Figure 1.
Mean levels of MI adherence. Mean MITI scores for the three research contexts. Estimates correspond with fixed effects estimated in multilevel models. Vertical bars represent 95% CI's. Significant differences between research contexts are indicated by asterisks: * p < .05, ** p < .01, *** p < .001. Horizontal lines indicate MITI 2.0 thresholds for beginning proficiency (dotted lines) and competency (dashed lines).
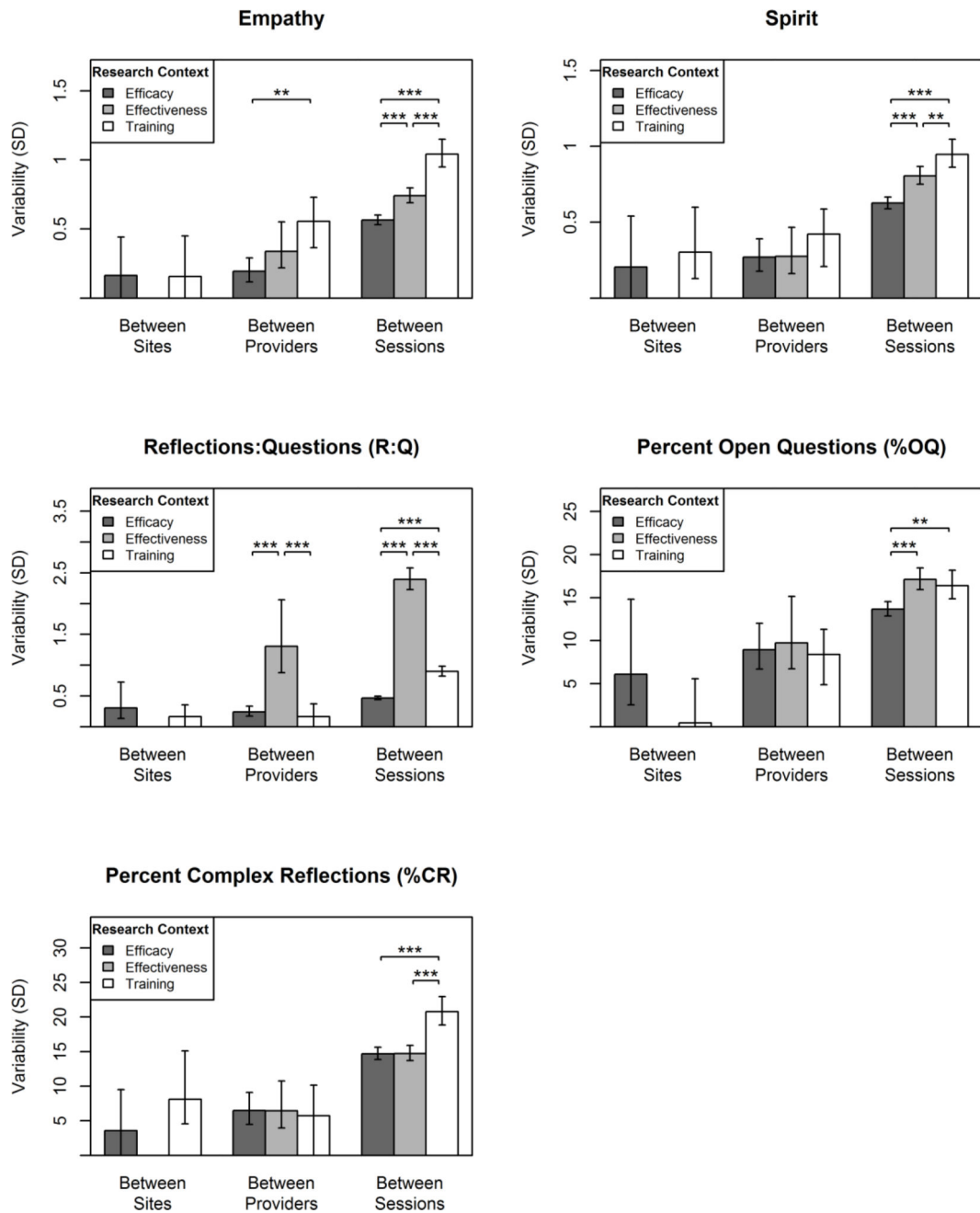
Figure 2.
Variability in MI adherence ratings attributable to between-site differences, between-provider differences (within sites), and between-session differences (within providers). Values are presented in standard deviation units based on random effect estimates in multilevel models. Vertical bars represent 95% CI's of the random effect estimates. Between-site variability could not be estimated in the effectiveness study (see text for explanation). Significant differences in variability between research contexts are indicated by asterisks: *p < .05, **p < .01, ***p < .001.
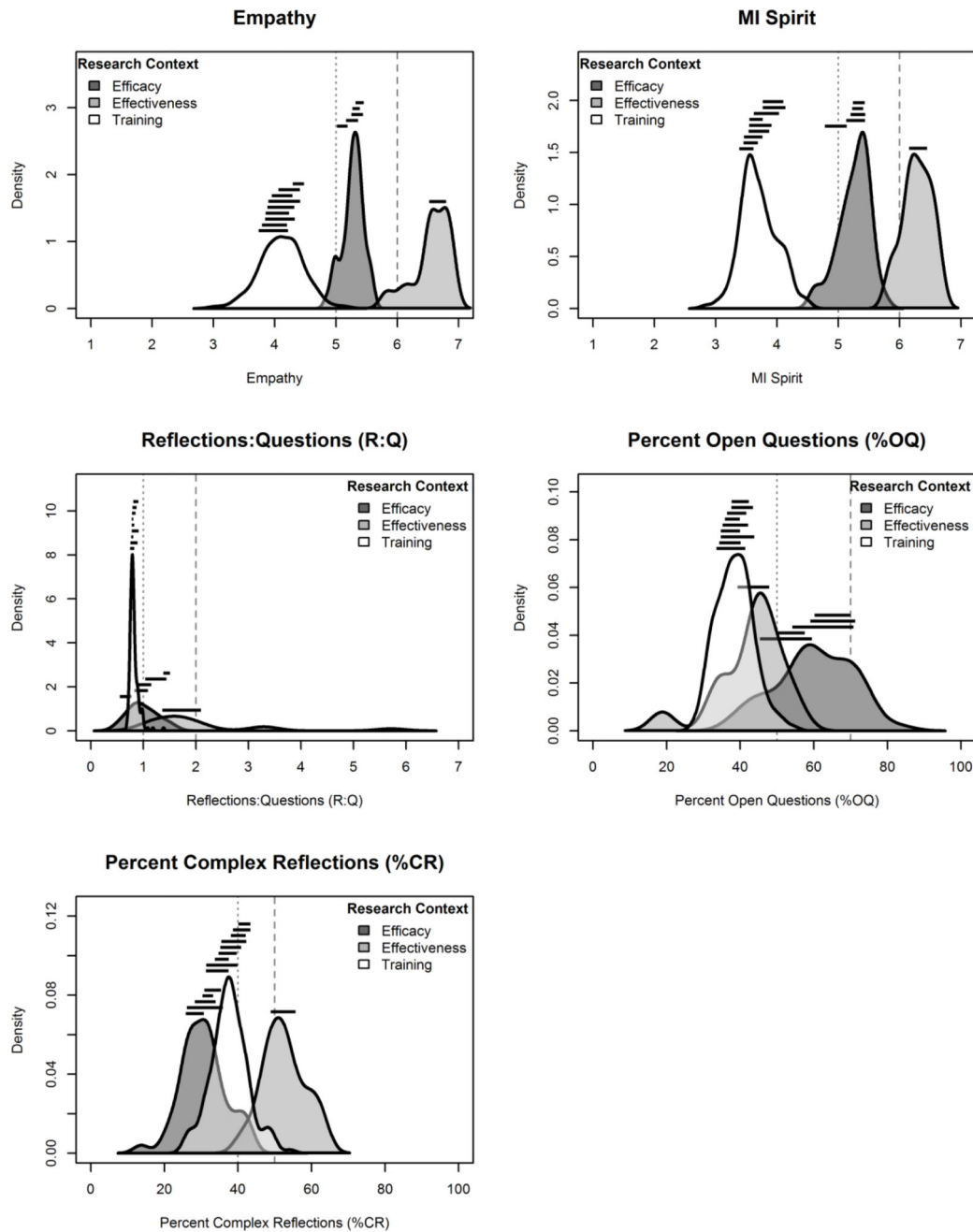
Figure 3.
Distributions of MI adherence across providers and sites. Distributions of provider MITI ratings are represented as density plots based on empirical Bayes estimates from random effects in multilevel models. The spans contained within the horizontal lines grouped above each research context designate the interquartile range (25th to 75th percentile) provider MI adherence within each site. Vertical lines indicate MITI 2.0 thresholds for beginning proficiency (dotted lines) and competency (dashed lines).

Table 1

Parent Study Information

	Study Aim	Provider Professions	MI Training, Supervision	Coder Reliability	N (providers)	N (sessions)
Efficacy Trials						
Lee et al. (2013)	Evaluation of MI + personalized feedback for reducing marijuana use in frequently using college students.	Doctoral-level professionals and clinical psychology graduate students.	Two-day workshop plus role-play and pilot sessions with competency evaluation, weekly group supervision and ongoing individual supervision, as needed/requested.	ICC=.43 to .95	9	53
Lee et al. (2014)	Evaluation of MI + personalized feedback for reducing spring-break alcohol use in college students.	Doctoral-level psychologists, graduate students in clinical psychology, masters students in social work, and post-baccalaureate students.	Two-day workshop plus role-play and pilot sessions with competency evaluation, ongoing weekly group supervision.	ICC=.86 to .96	20	231
Neighbors et al. (2012)	Evaluation of MI + personalized feedback for reducing 21st-birthday alcohol use in college students.	Doctoral-level psychologists, graduate students in clinical psychology, and post-baccalaureate students.	Two-day workshop plus role-play and pilot sessions with competency evaluation, ongoing weekly group supervision.	ICC=.78 to .98	12	215
Tollison et al., 2008	Evaluation of MI + personalized feedback for alcohol use reducing alcohol use in first-year college students with recent binge drinking episode.	Clinical psychology graduate and psychology undergraduate students.	Two-day workshop, additional weekly training for one academic quarter, pilot sessions with competency evaluation, weekly supervision.	α=.68 to .87	6	63
Effectiveness Trial						
Roy-Byrne et al. (2014)	Evaluation of brief intervention vs. enhanced care as usual in primary care within a large, publicly-funded hospital.	Social workers in primary care, bachelor's/master's level providers.	Training workshop, individual feedback on up to five standardized clients, monthly group supervision and ongoing e-mailed feedback for 25% of recorded sessions.	ICC=.60 to .91 (behavior codes); 95% and 87% agreement for global ratings *	15	372
Training Trial						
Baer et al. (2009)	Evaluation of two methods for training MI in 8 community substance abuse treatment facilities and 1 open-enrollment cohort.	Masters-, bachelors-, associates-, and GED-level community chemical dependency counselors.	Traditional two-day training workshop or 15 hours of training distributed over five 2-week intervals with practice interviews with feedback and attempts to build site-level support for MI.	ICC=.43 to .94	154	341
Total					216	1275

Notes. Three sites from Baer et al. (2009) are reported here but were not included in the parent publication, including two pilot sites and one open-enrollment group that was modeled here as an additional site. Coder reliability estimates reflect intra-rater agreement and are provided here as reported in parent publications. Ratings for percentage of MI adherent were low, ICC = 0.13, and therefore omitted from analysis and not included within the range presented in the table. Lee et al., 2013 reported "fair" reliability (ICC between 0.40 and 0.59) for empathy and MI spirit, Baer et al. 2009 provided a range for reliability estimates but did not report reliability estimates for specific MITI codes. BASICS = brief alcohol screening and intervention for college students, MI = Motivational Interviewing, ICC = intra-class correlation coefficient.

*
Reliability for Roy-Byrne et al. is described in Dunn et al., 2016.

Table 2

Proportions of Total Variance Explained by Between-Site, Between-Provider, and Between-Session Differences

MITI Index	Research Context	Between-Site	Between-Provider	Between-Session
Empathy	Efficacy	.06	.10	.84
	Effectiveness		.17	.83
	Training	.02	.22	.77
Spirit	Efficacy	.07	.14	.78
	Effectiveness		.10	.90
	Training	.08	.15	.77
Reflections-to-Questions ratio (R:Q)	Efficacy	.28	.15	.57
	Effectiveness		.23	.77
	Training	.03	.03	.94
Percent Complex Reflections (%CR)	Efficacy	.05	.15	.80
	Effectiveness		.16	.84
	Training	.12	.06	.81
Percent Open Questions (%OQ)	Efficacy	.13	.26	.61
	Effectiveness		.24	.76
	Training	.001	.21	.79

Note. Values are expressed as Intraclass Correlation Coefficients (ICCs). MITI = Motivational Interviewing Treatment Integrity.