Published in final edited form as:
Drug Alcohol Depend. 2008 September 1; 97(1-2): 130–138. doi:10.1016/j.drugalcdep.2008.03.018.

The video assessment of simulated encounters-revised (VASE-R): Reliability and validity of a revised measure of motivational interviewing skills

David B. Rosengren[a,*], **Bryan Hartzler**[a], **John S. Baer**[a,b], **Elizabeth A. Wells**[a,c], and **Christopher W. Dunn**[a,d]

a *Alcohol and Drug Abuse Institute, University of Washington, 98195-4805, United States*

b *Department of Psychology, University of Washington, and VA Puget Sound Health Care System, 98108-1597, United States*

c *School of Social Work, University of Washington, 98195-4900, United States*

d *Department of Psychiatry and Behavioral Sciences, University of Washington, 98195-6560, United States*

Abstract

The video assessment of simulated encounters-revised (VASE-R) is a video-based method, administered in individual or group settings, for assessing motivational interviewing (MI) skills. The 18-item instrument includes three video-based vignettes, in which actors portray substance abusers, with each vignette followed by questions that prompt examinees to write responses that are then scored against MI standards. The VASE-R was administered to two independent samples: (1) substance abuse practitioners participating in a study of MI training methods, and (2) MI training facilitators with a high level of MI skill and expertise. This multi-study report describes basic VASE-R psychometric properties – including scoring reliability, internal consistency, concurrent validity, and sensitivity to the effects of training – and then presents proficiency standards based on administration to a sample of MI training facilitators (MI Experts). The findings indicate excellent inter-rater reliability using intra-class correlations for the full-scale score (.85) and acceptable levels for subscales (.44 to .73). The instrument displayed strong concurrent validity with the Helpful Responses Questionnaire (HRQ) and a behavioral sample of clinician behavior with a standardized patient scored using the MI Treatment Integrity (MITI) system, as well as good sensitivity to improvement in MI skill as a result of training. The findings provide an empirical basis for suggesting VASE-R benchmarks for beginning proficiency and expert MI practice.

Keywords

Motivational interviewing; Training evaluation; Skill assessment; Evaluation instruments

*Corresponding author at: Alcohol and Drug Abuse Institute, University of Washington, 1107 NE 45th Street, Suite 120, Seattle, WA 98195-4805, USA. Tel.: +1 206 543 0937; fax: +1 206 543 5473. E-mail address: dbr@u.washington.edu (D.B. Rosengren).

Conflict of interest
None.

Contributors. All authors designed the study and wrote the protocol. David Rosengren managed the literature searches, summaries of previous related work and wrote the first draft. Bryan Hartzler completed the statistical analysis and preparation of the first draft of the Results section and the tables. All authors contributed to and have approved the final manuscript.

1. Introduction

Accurate, cost-effective measures for assessing clinician skillfulness and skill uptake from training are essential for transferring treatment research into practice (Brown, 2000; Dansereau and Dees, 2002; Simpson, 2002). Previously, we (Rosengren et al., 2005) reported preliminary development of one such instrument for assessing skills in motivational inter-viewing (MI), referred to as the video assessment of simulated encounters (VASE). Those initial findings generated revisions to scale structure and item content, culminating in a revised version of the instrument (VASE-Revised or VASE-R) since implemented in a dissemination trial testing MI training methods (Baer et al., 2008). This paper describes psychometric properties of the VASE-R.

MI is a method of clinical interaction that continues to receive significant interest in the addictions field and beyond (Carroll et al., 2006; Knight et al., 2006; Vasilaki et al., 2006). Defined as "*a client-centered, directive method for enhancing intrinsic motivation to change by exploring and resolving ambivalence* (p. 25)" (Miller and Rollnick, 2002), MI is grounded in the four principles of expressing empathy, developing discrepancy, rolling with resistance and supporting self-efficacy. It also relies on a collection of common communication skills, represented by the acronym *OARS* (i.e., *O*pen questions, *A*ffirmation of client strengths, *R*eflective listening, and *S*ummarizing). MI practitioners selectively attend to elements of a clinical interaction, eliciting and reinforcing client statements in favor of change (i.e., change talk) while diffusing resistance or client talk that sustains the status quo. The MI approach is infused with a collaborative, eliciting, and supportive spirit (Miller and Rollnick, 2002).

There are several assessment methods available for evaluating MI skillfulness. One example is the Helpful Responses Questionnaire (HRQ) (Miller et al., 1991), a 6-item paper-and-pencil questionnaire in which respondents generate written responses (e.g., 'what you would say next') to written clinical scenarios. The HRQ can be scored relatively quickly; however, its unidimensional summary score measures reflective listening primarily. Thus, other MI skills (e.g., responding to resistance, developing discrepancy) are not captured effectively. A more comprehensive but labor intensive approach to assessing MI skills involves reviewing and scoring audiotaped encounters; there are a number of fidelity systems available (Lane et al., 2005; Martino et al., 2006; Miller, 2000) for this purpose. The most commonly used system is the Motivational Interviewing Treatment Integrity (MITI) scale, which allows multi-dimensional evaluation of MI skills (Moyers et al., 2004). In addition to the considerable time this scoring requires, the MITI and similar instruments present difficulties in selection and recording of client sessions (e.g., need for recording equipment, confidentiality issues, and selection biases) or, alternatively, the costs and challenges of hiring and training standardized actors. Thus, we developed an assessment tool that provides the efficiency of group-based administration, a standardized MI challenge, and the ability to assess a variety of MI skills.

Video-based exams typically include videotaped clinical stimuli after which respondents are prompted to write short answers. The VASE-R presents videotaped vignettes approximating the presenting complaints, histories, and treatment expectations of real-world clinical encounters for patients presenting with substance use concerns. For the original VASE, we reported strong psychometrics for three sub-scales (Reflective Listening, Responding to Resistance, and Summarizing) and a need to revise four subscales (Identification of Change Talk, Use of Stage-Matched Questions, Stages of Change Assessment, and Development of Discrepancy). This initial study, completed with a small sample of practitioners ($N = 22$), did not provide information about the instrument's sensitivity to change in clinician behavior and did not develop proficiency standards for interpreting scores.

This multi-study report examines VASE-R psychometric properties when administered to substance abuse treatment practitioners as part of a dissemination trial comparing methods of MI training (Baer et al., 2008) and then reports a contrast sample obtained by administering the VASE-R to members of the Motivational Interviewing Network of Trainers (MINT), an international organization of training facilitators. The specific aims of these two studies were:

1. To evaluate the scoring reliability, internal consistency, concurrent validity, and sensitivity to change of the VASE-R.
2. To establish proficiency standards for the VASE-R.

2. Instrument revision

The original VASE presented three clinical vignettes, each with a brief description of the client and clinical context. A different actor portrayed each client, and offered a series of statements that reflected the client's concerns to which respondents generated written responses following a prompt (e.g., *write a response that indicates you are listening; write a response that you think would be most helpful in this situation*). Each vignette consisted of 6 such items, as well as two additional, multiple-choice items. The original VASE contained 24 items, developed by expert opinion and distributed across seven sub-scales (e.g., Reflective Listening, Summarizing, Rolling with Resistance, Developing Discrepancy, Identifying Change Talk, Stage-Matched Open Questions, and Stage of Change Assessment). Items were scored on a 3-point scale (e.g., 0–2) yielding a VASE total score that ranged from 0 to 48.

The VASE-R retains the same essential structure, with the original client scenarios and character statements intact (i.e., "Lisa," "Ulysses," and "Bailey"), but also includes a number of revisions suggested by our prior analyses. The stages of change assessment and stage-matched open-ended questions subscales were dropped due to poor internal reliability and item irregularity, respectively. The identifying change talk subscale also suffered from poor internal reliability, which led to the retooling of its items into a free-response format in which a prompt asks the respondent to generate a written response likely to '*elicit from the client statements that support making healthy changes*.' Accordingly, it was renamed Eliciting Change Talk. Revision of the Developing Discrepancy subscale was intended to increase item difficulty. Though the multiple-choice format was retained, these items now include just one (rather than two) correct response from a total of five response options. One response option for each item was rewritten.

There are now five subscales in the VASE-R: (1) Reflective Listening (RL—4 items), skill in active listening and formulating simple reflections; (2) Responding to Resistance (RR—5 items), skill in producing non-confrontational questions or statements; (3) Summarizing (S—3 items), skill in developing summary statements that incorporate client ambivalence and change talk; and (4) Eliciting Change Talk (ECT—3 items), skill in using reflections, questions or strategies likely to elicit client change talk. The fifth VASE-R subscale, Developing Discrepancy (DD—3 items), represents skill in identifying clinician utterances likely to enhance client motivation for change.

The revised instrument consists of six items per vignette and 18 items collectively. Fifteen items retain a free-response format, while the remaining three items are multiple-choice and contain an option for the respondent to provide a rationale for his or her choice. All items are scored using a 3-point system. For RL, RR and ECT the item scores are: 0 = confrontational or likely to engender resistance, 1 = neutral or inaccurately represents content of client's speech, and 2 = accurately reflects content of client's speech and represents intended MI skill. For the S scale the 1 equals a multiple idea statement or statements that include either ambivalence or client change talk, while a 2-point score includes *both* ambivalence and change talk; 0 responses are confrontational, likely to engender resistance or do not include multiple ideas. The DD

items also have different criteria: 0 = incorrect option paired with MI-inconsistent rationale, 1 = incorrect option paired with MI-consistent rationale, and 2 = correct option. The VASE-R yields a full-scale score ranging from 0 to 36. The scoring manual is available at: http://adai.washington.edu/instruments/VASE-R.htm.

3. Study 1

Substance abuse treatment personnel completed the VASE-R as part of training outcome assessments in a dissemination trial comparing MI training methods (Baer et al., 2008). The trial involved six community treatment programs (CTPs) at which interested staff participated in MI training and a series of MI skill assessments prior to, immediately after, and 3 months post training. This study allows examination of the psychometrics of the VASE-R, including its relationships to other instruments prior to training, as well as its sensitivity to training effects at post-training assessment.

3.1. Study 1 methods

3.1.1. Sample—The practitioner sample (N = 144) had a mean age of 46.7 years (S.D. = 11.7) and was predominantly female (69%). Distribution of reported ethnicity was as follows: 72% Caucasian, 11% Multi-Ethnic, 7% African-American, 3% Native American, 1% Pacific Islander, 1% Asian, 4% Other, and 1% chose not to identify. The sample reported a range from 0 to 38 years of prior clinical service, with an average of 9.4 years (S.D. = 8.5). Agency tenure ranged from 0 to 35 years, with a mean 5.1 years (S.D. = 6.2) at one's current agency. Participants reported prior education as the highest degree attained: 27% had a graduate degree (master's degree equivalent or above), 26% a bachelor's degree (e.g., B.A.), 33% an associate's degree (i.e., A.A.), and 13% a high school diploma or equivalent (e.g., G.E.D.). In addition, 45% of the sample endorsed being state-certified chemical dependency professionals (C.D.P.). Forty-five percent (45%) of the sample self-described as in recovery from substance abuse, 40% did not, and the remaining 15% declined to respond. In terms of prior MI exposure, 46% of the sample reported involvement in trainer-facilitated activities, and 48% indicated prior exposure via self-study methods.

3.1.2. Procedures—The University of Washington Institutional Review Board (IRB) approved all procedures. The investigators recruited participants sequentially from six substance abuse treatment facilities, and offered free MI training at these facilities as well as continuing education credits.

The MI Training compared two models that both provided approximately 15 h of training that covered core MI concepts (i.e., readiness to change, MI principles and spirit, "OARS" microskills, responding to resistance, and MI consistent strategies). The training models also included practice opportunities with trainer comment and feedback. The methods differed in the timing of training (grouped versus spaced) and in the use of standardized patients for skills practice and feedback. These training methods and outcome comparisons are discussed elsewhere (Baer et al., 2008), though overall findings suggest both training methods led to comparable increases in practitioner skills.

The study compensated agencies for time and space required for training. Interested staff provided informed consent, and agreed to complete all MI skill assessments in exchange for financial compensation ($30 per assessment). Pre-training skill assessments took place the week prior to training, and included: (1) self-report of demographics, (2) group administration of the VASE-R, (3) completion of the HRQ (Miller et al., 1991) and (4) audio-recorded interviews with a standardized patient (SP) scored using the MITI (Moyers et al., 2004). Reviewers, blind to condition and time point, scored HRQs and MITI-coded SP interviews, with acceptable inter- and intra-rater reliability according to published standards (Cicchetti,

1994). Using similar procedures for blinding, reviewers scored the VASE-R protocols; scoring reliability is detailed later in this report.

3.2. Study 1 results

3.2.1. Analytic plan—The analytic process consisted of the following sequence of computations: (1) inter-rater reliability; (2) descriptive statistics for individual items, subscales, vignettes, and full scale scores of the VASE-R; (3) internal consistency of VASE-R full scale scores and its individual subscales and vignettes; (4) inter-correlations between individual items and subscales, between individual items and the VASE-R full scale, and between subscales and the VASE-R full scale; (5) concurrent validity of the VASE-R full scale and its subscales with established measures of MI skillfulness (HRQ and MITI scores); (6) sensitivity to change of the VASE-R full scale and its subscales as a function of practitioner participation in MI training or practice. The description of each analytic step follows.

3.2.2. Scoring reliability—Three independent raters, blind to the timing of assessment and assigned training condition of the participant, shared in scoring of 424 VASE-R protocols. We combined participants for analyses because of a failure to find differences between training conditions. The investigators randomly selected a subset of 85 protocols (20%) to evaluate inter-rater reliability and compute intra-class correlations (ICCs); we interpreted the ICCs in relation to published guidelines (Cicchetti, 1994). ICCs varied among items, and between rater pairs, but values were acceptable (ICCs > .4) for all 18 individual VASE-R items. Scoring reliability was best for items comprising the Developing Discrepancy subscale (ICCs = .70 to .92), with greater variance in scoring reliability observed in items from other subscales (Reflective Listening, ICCs = .43 to .88; Responding to Resistance, ICCs = .41 to .74; Summarizing, ICCs = .46 to .74; and Eliciting Change talk, .43 to .83). Because all values met standards for interpreting ICCs (Cicchetti, 1994), all 18 VASE-R items were retained for subsequent analyses.

3.2.3. Item and scale description—Next, we computed descriptive statistics for VASE-Rs completed prior to training. Table 1 lists means and standard deviations for individual items, sub-scales, vignettes, and the VASE-R full scale score. (Table 1 also contains descriptive data about Trained and Expert Samples that will be discussed in greater detail later.) Results indicate variation in difficulty of individual items, subscales, and vignettes. For instance, raters recorded consistently higher scores for Reflective Listening items (M = 1.17–1.24) than for items corresponding to the summarizing subscale (M = .38 to .63). Vignette scores also differed (M = 5.41–6.59), though it is unclear whether this variance was attributable to varying difficulty of the three vignettes or to other factors such as respondent fatigue. The distribution of scores for the full VASE-R and its components appear to reflect considerable variation in MI skill among a community sample of practitioners prior to training.

3.2.4. Internal consistency and inter-correlation—Subsequently, we evaluated internal consistency of the full VASE-R scale, subscales, and vignettes using Cronbach's α coefficients. Internal consistency of the 18-item scale was .85, an improvement over the estimate of .73 for the original VASE (Rosengren et al., 2005). Internal consistency of vignettes ranged from .64 to .71, though values for subscales varied (see Table 2). As with the original VASE (Rosengren et al., 2005), internal consistency of particular subscales (Reflective Listening, Responding to Resistance, Summarizing) was stronger than for others (e.g., Eliciting Change Talk, Developing Discrepancy). Table 2 also lists subscale-full scale, item-subscale, and item-full scale correlations. All correlations were positive and showed contribution of individual items to their respective subscales, as well as item- and subscale-level contribution to the full scale.

3.2.5. Concurrent validity—The next step was evaluation of concurrent validity of the full VASE-R using bivariate correlations with the HRQ and SP interviews. The HRQ produces a single summary score, while the SP interviews, scored using the MITI, generate global indices of *empathy* and *MI spirit* (both rated on a 7-point Likert scale), as well as behavioral measures of the ratio of reflections to questions (R:Q), percentage of 'open' questions (%OQ), percentage of 'complex' reflections (%CR), and percentage of MI-Adherent behaviors (%MIA). Preliminary analyses suggested that MITI, HRQ, and VASE-R summary indices were normally distributed; thus Pear-son product-moment correlations were appropriate measures of association among these indices. Table 3 outlines a matrix of correlations between indices from the VASE-R and the HRQ and MITI-scored SP interviews. Results indicate that the full scale and subscale scores of the VASE-R correlated with the HRQ score, as well as with a number of global and behavioral MITI measures. The Reflective Listening, Responding to Resistance, and Eliciting Change Talk subscales yielded the strongest concurrent validity with other measures of MI skill. The Developing Discrepancy subscale was less consistently correlated with HRQ and MITI indices.

3.2.6. Sensitivity to effects of training—The last step was to assess the sensitivity of the VASE-R in measuring temporal trends in skill levels as a function of participation in structured MI training activities. Repeated-measures multivariate analysis of variance (MANOVA) assessed the extent to which full VASE-R scores and individual subscale scores differed when assessed prior to training, immediately following training, and after a 3-month follow-up. A multivariate time effect was detected for the full VASE-R score, $F(2, 109) = 54.01$, $p < .001$; with post hoc contrasts specifying a score increase from pre- to post-training, $F(1,110) = 63.21$, $p < .001$. The apparent score decline from post-training to follow-up was not statistically significant. Multivariate time effects for the five subscales were: Reflective Listening, $F(2,109) = 23.34$, $p < .001$; Responding to Resistance, $F(2,109) = 21.04$, $p < .001$; Summarizing, $F(2,109) = 34.58$, $p < .001$; Eliciting Change Talk, $F(2,109) = 12.57$, $p < .001$; and Developing Discrepancy, $F(2,109) = 9.32$, $p < .001$. Post hoc contrasts indicated significant score increases from pre- to post-training for all subscales (all p-values $< .001$). With one exception, these post hoc contrasts were nonsignificant from post-training to follow-up. Only the Summarizing subscale post hoc contrast indicated a significant decrease from post-training to follow-up, $F(1, 110) = 7.73$, $p = .006$. The overall pattern is inconsistent with temporal changes one might expect as a function of practice effects, and is congruent with patterns observed in the larger MI training literature (Baer et al., 2004; Miller et al., 2004). To additionally account for the potential influence of agency affiliation on these training effects, analyses were re-run with agency affiliation included as an independent variable. Agency affiliation was not a significant predictor of the temporal effects for the full VASE-R scale or any of its subscales (all F-values < 1.23, all p-values $> .26$), and parameter estimates for VASE-R subscales were comparable to initial models. Table 4 outlines the sample mean for the full VASE-R scale and its subscales at the three time points, as well as corresponding effect sizes for multivariate analyses.

4. Study 2

In an independent data collection, the VASE-R was administered at an annual meeting of the MINT. The 'train-the-new trainers' (TNT) process for MINT members requires prerequisite training and practice in MI skills, along with prior opportunity to develop skills as a training facilitator. Completion of a MINT-sponsored TNT is required for MINT membership. Thus, MINT members are presumed to be knowledgeable in MI concepts and skillful in their application. Consequently, the aim of Study 2 was to gather data to inform VASE-R proficiency standards.

4.1. Study 2 methods

4.1.1. Sample—MINT trainers participated in a group administration of the VASE-R at the annual MINT conference in Portland, ME, in September, 2004. A mean age of 46.4 years (S.D. = 8.5) was reported by participants, with a racial/ethnicity composition of 91% Caucasian, 6% Latino/Hispanic, and 3% Asian. Education level was as follows: 51% had earned a doctoral degree (e.g., M.D., Ph.D., or equivalent), 44% had earned a master's degree (e.g., M.A., M.S.W., M.P.H., or equivalent), and 5% had earned a bachelor's degree (e.g., B.A., B.S., or equivalent). All participants were training facilitators, with 80% identified as "experienced" (i.e., completion of their TNT at least one year prior) and the remaining 20% identified as "new" trainers (i.e., completed TNT in the days preceding this VASE-R administration).

4.2. Procedures

The University of Washington Institutional Review Board (IRB) approved study procedures. Investigators presented a brief description of the VASE-R during a plenary session at the MINT meeting and then invited attendees to participate in a group administration of the instrument for purposes of gathering norms to inform proficiency standards. Attendees returned their completed VASE-R response booklet (including the demographic data, but without names) to an investigator, if they were willing to have responses scored and included in data analyses. Alternatively, attendees could participate in the group administration, but retain their response booklet if they did not wish to provide data for analysis. Of the 80 MINT members registered for the meeting, 66 (82%) completed and returned their VASE-R response booklet. Though participants received no financial compensation for participation, all attendees were later provided a more comprehensive description of the instrument, information about the processes of its construction and revision, and instructions for how interested parties might receive copies of the VASE-R and associated materials (e.g., response booklets, scoring manual).

4.3. Study 2 results

Two of the three raters employed to score the Study 1 VASE-R protocols completed scoring of this smaller sample of Study 2 VASE-R protocols. As in Study 1, raters were blind to the background of VASE-R respondents. Due to the smaller Study 2 sample, both raters scored all 66 VASE-R protocols. Scoring reliability was again evaluated using ICCs, and referencing published psychometric guidelines (Cicchetti, 1994). As with Study 1, scoring reliability varied among items, but values were again acceptable (ICCs > .4) for all 18 VASE-R items. Scoring reliability was greatest for items comprising the Developing Discrepancy subscale (ICCs = .90 to .96), with strong but more variant reliability observed in items of other subscales (Reflective Listening, ICCs = .75 to .92; Responding to Resistance, ICCs = .51 to .78; Summarizing, ICCs = .50 to .63; and Eliciting Change talk, .55 to .82). As before, all eighteen VASE-R items were retained for subsequent analyses of the Study 2 sample.

We computed descriptive statistics for the VASE-R full scale, subscale, vignette, and individual item scores (see Table 1). The mean full-scale score was 30.88 (S.D. = 5.02), which exceeded the mean performance of substance abuse treatment practitioners in Study 1 at their baseline assessment by more than two standard deviations, and at their post-training assessment by more than one standard deviation. The contrast between these two samples appears to be consistent across subscales (MINT attendees scoring on average 1–2 points higher than trained practitioners) and vignettes (MINT attendees scoring on average 2–3 points higher than trained practitioners). There was also less variance in VASE-R scores among MINT members than among the sample of substance abuse practitioners in Study 1.

This sample, along with the data derived from Study 1, allow us to suggest reference points for understanding VASE-R results. We used scores from the MINT sample, along with baseline

scores from Study 1 practitioner sample, removing the fourteen participants that identified prior participation in a MI workshop at baseline, to develop three different benchmarks for the VASE-R: Untrained, Beginning Proficiency, and Expert Proficiency. The MINT trainers represent a group of individuals with considerable prior training and experience in MI, and therefore are used as exemplars of expert performance. We used a statistical approach to identify points between these two distributions of scores to define beginning proficiency.

Jacobson et al. (1999) and Jacobson and Truax (1991) developed methods to identify cut-points to designate when individuals receiving psychological services move from being members of a dysfunctional to a normal population. These methods can also be useful in identifying when clinicians move from being untrained to being trained at levels comparable to 'MI experts'. Accordingly, we employed the methods published by Jacobson et al. (1999) and Jacobson and Truax (1991), to develop cut scores for the full scale, each subscale, and each vignette, that would more likely place the individual within the distribution of experts' performance than an untrained, community sample performance. Conceptually, this is the point where the distributions of scores (mean and variability) of the two samples cross. The resulting scores were rounded to the nearest integer (for ease of use) and are presented in Table 5. The expert sample often scored near the top of VASE-R scales, thus producing skewed distributions. These negative skews likely yield estimates for cut scores that are higher than those derived with completely normal data, which suggests these cut scores are conservative. It is noteworthy that these scores are generally about one standard deviation below the MINT samples average score, and quite similar to average scores of the Study 1 sample post-training. Thus, these cut scores represent a reasonable measure of *beginning proficiency*—within the expert distribution but roughly one standard deviation below the mean, and consistent with scores for those completing an initial training in MI. Table 5 summarizes these findings.

5. Discussion

As the addictions field continues to press for dissemination of research findings by training community practitioners in empirically based practices, there remains a need for instruments that assess skills and skill gains. There is also a need for benchmarks to assess when proficiency has been attained. The results from this study suggest the VASE-R may serve as one such method for the evaluation of MI skills.

The current investigation indicates that the VASE-R can be reliably scored by different raters, though some ICC's were modest. The full-scale score also demonstrates excellent internal consistency. The Reflective Listening (RL), Responding to Resistance (RR) and Summarizing (S) subscales exhibit good levels of internal consistency as well. The new subscale (Eliciting Change Talk or ECT) and revised subscale (Developing Discrepancy or DD) demonstrate adequate internal consistency. The more challenging item set included in the DD scale may have led to lower levels of internal consistency. However, all items contributed in the direction expected, a clear improvement from the original VASE instrument.

As designed, Vignette 3 (Bailey) appeared to be the most difficult, but was also the last in a series of three vignettes. The intent in designing a more difficult client was to mirror treatment where more intransigent clients can challenge practitioner skills. While the response trend is in this direction, even for the expert sample, the contemporaneous effects of respondent fatigue cannot be ruled out as an alternative explanation. Subsequent evaluations should change the order of vignettes to address this question.

The scoring on the Summarizing subscale was lower than other subscales and displayed less concurrent validity with other measures, even among the MI experts. This finding may not reflect the ability of participants to develop a basic summary, but rather indicates an artifact

of the scoring criteria. In order for a response to receive full credit (a '2') a summary had to contain multiple client ideas, as well as both change talk and ambivalence evident in the 'client' statements. The manual included these last two criteria to differentiate basic summaries from those that more effectively targeted ambivalence and change, two elements considered central to the practice of MI. While inclusion of these elements can be justified as consistent with MI concepts, it is not an approach taught consistently by MI trainers and thus may result in artificially low scores. An alternative strategy may be to include *either* rather than *both* as the criterion for a full credit score.

In general, the VASE-R demonstrates good concurrent validity with two other forms (and formats) of MI measures. The full scale score and subscale scores demonstrate a strong positive relationship with the unidimensional HRQ score, as well as the MITI global measure of empathy and behavioral measure of reflection to question ratio (R:Q) from the SP interviews. Similarly, we found strong correlations between the MITI measure of MI Spirit and the VASE-R Full Scale, RL, RR and ECT sub-scales. Of equal interest is the finding that the percentage of open questions (%OQ) was correlated strongly with only the ECT sub-scale. It suggests that the more counselors used open questions, the better prepared they were to deploy a critical component of successful MI sessions, eliciting change talk (Amrhein et al., 2003). There was no association between percentage of complex reflections from SP interviews and the VASE-R. This finding is likely a function of the low base rates of complex reflections evident in the SP interviews, as well as the VASE-R's lack of differentiation between simple and complex reflections in scoring. The percentage of MI adherent behaviors (%MIA) showed more modest correlations to the VASE-R Full, RL and RR sub-scale scores. Findings support use of the VASE-R as a global index of MI skillfulness and use of its subscales to evaluate specific aspects of MI skills.

The VASE-R also appears to be sensitive to changes in MI skill level. The instrument detects changes at the global and subscale levels. While there are no repeated measures *without training* to rule out the effects of practice, the pattern of findings across time points suggest the results are unlikely to be a function of practice effects alone. That is, these scores would continue to improve if only exposure to the instrument alone and practice were producing the effect. The results are also consistent with findings reported in other studies of skill acquisition. There is little evidence to suggest that unguided practice results in improvement to targeted skills (Miller et al., 2004). Although additional research is required to rule out this threat to internal validity, it appears that the VASE-R may be used as a measure to assess skill change following MI training.

The availability of proficiency standards represents a significant advance in understanding MI skillfulness on the VASE-R. Findings from Study 2 suggest empirically based standards for full scale, subscale and vignette scores to determine expert levels in proficiency. Using a statistical approach (Jacobson et al., 1999) to determine the crossover point between these two samples (i.e., the point at which a score is more likely to fall in the Expert sample than the in the Untrained sample), estimates for cut-off scores for Beginning Proficiency identify results that exhibit improvement over untrained samples, but are still short of expert levels. Finally, mean scores from the Study 1 baseline sample provide an initial reference point for understanding the skill level of an untrained sample. In the previous VASE-R scoring manual we proposed using a 75% cut-off score as an initial marker of MI proficiency. It is noteworthy that, using this less precise cut-off, on average the MI experts from Study 2 would meet 75% proficiency standards for the Full Score and RL, RR, and DD subscales, while falling just below criterion ($M = 4.92$) for ECT. Thus, the 75% standard provides an alternative general estimate of VASE-R skill proficiency.

The findings indicate the VASE-R can be useful in assessing respondent skills and providing a preliminary, but data-based method to evaluate respondent scores. The revisions made from the original VASE appear to have improved the psychometric characteristics of the VASE-R instrument, especially with regards to the critical concept of eliciting change talk. The findings indicate the subscales could also allow for discrimination between areas of skill and areas in need of further work. The opportunity to use group administration, combined with low technology demands and standardized stimuli, allow for ease of use and comparison across individuals, as well as changes in behavior over time. These results suggest that video-based assessment may enhance the transfer of empirically based practices and the VASE-R, in particular, may be useful in assessing acquisition of MI skills. Finally, the VASE-R might prove helpful to researchers as means of establishing individual skill levels and to determine if a priori skill targets are met before permitting a MI clinician to begin a trial.

Acknowledgements

The authors appreciate the expert consultation and review of Blair Beadnell on the data analyses sections of this manuscript. The authors thank Andrew Slade, David Peterson and Avry Todd for their scoring of VASE-R protocols and coding of the MITI in reviews of the standardized patient encounters. The authors wish to express their appreciation for the collaborative efforts of the involved community treatment agencies and their staffs, as well as the MINT organization and individual MINT members for participating in this investigation. Finally, we wish to thank out VASE-R ‘clients’: Midge Strong-Beers, Phil Turner and Michelle Peavy.

Role of funding source: Research supported by NIDA Grant #R01 DA016360. NIDA had no further role in the study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the paper for publication.

References

Amrhein PC, Miller WR, Yahne CE, Palmer M, Fulcher L. Client commitment language during motivational interviewing predicts drug use outcomes. Journal of Consulting and Clinical Psychology 2003;71:862–878. [PubMed: 14516235]

Baer JS, Rosengren DR, Dunn C, Wells E, Ogle R, Hartzler B. An evaluation of workshop training in motivational interviewing for addiction and mental health clinicians. Drug and Alcohol Dependence 2004;73:99–106. [PubMed: 14687964]

Baer JS, Wells EA, Rosengren DB, Hartzler B, Beadnell B, Dunn C. Context-tailored training and technology transfer: evaluating MI training for community counselors, submitted for publication. 2008

Brown, BS. From research to practice: the bridge is out and the water’s rising. In: Levy, J.; Stephens, R.; McBride, D., editors. Emergent Issues in the Field of Drug Abuse: Advances in Medical Sociology. JAI Press; Stanford, CT: 2000. p. 345-365.

Carroll KM, Ball SA, Nich C, Martino S, Frankforter TL, Farentinos C, Kunkel LE, Mikulich-Gilbertson SK, Morgenstern J, Obert JL, Polcin D, Snead N, Woody G, CTN N. Motivational interviewing to improve treatment engagement and outcome in individuals seeking treatment for substance abuse: a multi-site effectiveness trial. Drug and Alcohol Dependence 2006;28:301–312. [PubMed: 16169159]

Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment 1994;6:284–290.

Dansereau DF, Dees SM. Mapping training: the transfer of a cognitive technology for improving counseling. Journal of Substance Abuse Treatment 2002;22:219–230. [PubMed: 12072166]

Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. Journal of Consulting and Clinical Psychology 1991;59:12–19. [PubMed: 2002127]

Jacobson NS, Roberts LJ, Berns SB, McGlinchey JB. Methods for defining and determining the clinical significance of treatment effects: description, application, and alternatives. Journal of Consulting and Clinical Psychology 1999;67:300–307. [PubMed: 10369050]

Knight KM, McGowan L, Dickens C, Bundy C. A systematic review of motivational interviewing in physical health care settings. British Journal of Health Psychology 2006;11:319–332. [PubMed: 16643702]

Lane C, Huws-Thomas M, Hood K, Rollnick S, Edwards K, Robling M. Measuring adaptations of motivational interviewing: the development and validation of the behavior change counseling index (BECCI). Patient Education and Counseling 2005;56:166–173. [PubMed: 15653245]

Martino, S.; Ball, SA.; Gallon, SL.; Hall, D.; Garcia, M.; Ceperich, S.; Farentinos, C.; Hamilton, J.; Hausotter, W. Northwest Frontier Addiction Technology Transfer Center. Oregon Health Sciences University; Salem, OR: 2006. Motivational interviewing assessment: supervisory tools for enhancing proficiency (MIA: STEP).

Miller, WR. Motivational Interviewing Skill Code: Coders Manual; University of New Mexico Center on Alcoholism. Substance Abuse, and Addictions. 2000 [accessed on 1/15/04]. http://casaa-0031.unm/edu/

Miller, WR.; Rollnick, S. Motivational Interviewing: Preparing People for Change. Guilford Press; New York: 2002.

Miller WR, Hedrick KE, Orlofsky D. The helpful responses questionnaire: a procedure for measuring therapeutic empathy. Journal of Clinical Psychology 1991;47:444–448. [PubMed: 2066417]

Miller WR, Yahne CE, Moyers TB, Martinez J, Pirritano M. A randomized trial of methods to help clinicians learn motivational interviewing. Journal of Counseling and Clinical Psychology 2004;72:1050–1062.

Moyers, TB.; Martin, T.; Manuel, JK.; Miller, WR. Motivational Interviewing Treatment Integrity (MITI) coding system. The University of New Mexico Center on Alcoholism, Substance Abuse, & Addictions. 2004 [accessed on 1/15/04]. http://casaa-0031.unm/edu/

Rosengren DB, Baer JS, Hartzler B, Dunn C, Wells E. 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 2005;79:321–330. [PubMed: 16102376]

Simpson DD. A conceptual framework for transferring research to practice. Journal of Substance Abuse Treatment 2002;22:171–182. [PubMed: 12072162]

Vasilaki EI, Hosier SG, Cox WM. The efficacy of motivational interviewing as a brief intervention for excessive drinking: a meta-analytic review. Alcohol & Alcoholism 2006;41:328–335. [PubMed: 16547122]

Table 1

Normative data for VASE-R scale and components for untrained, trained and expert samples

	Untrained practitioners		Trained practitioners		MI experts	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Reflective listening (4)	4.80	2.22	6.13	2.04	7.30	1.51
VASE-R #1	1.24	.78	1.53	.76	1.88	.41
VASE-R #2	1.17	.75	1.55	.72	1.83	.51
VASE-R #7	1.23	.80	1.59	.77	1.83	.54
VASE-R #9	1.17	.75	1.48	.71	1.76	.61
Responding to resistance (5)	5.69	2.57	7.41	2.49	9.06	1.77
VASE-R #3	.90	.88	1.36	.84	1.82	.49
VASE-R #8	1.37	.72	1.63	.65	1.83	.51
VASE-R #13	1.24	.81	1.53	.72	1.85	.47
VASE-R #14	1.14	.94	1.49	.72	1.76	.63
VASE-R #15	1.05	.75	1.42	.81	1.80	.53
Summarizing (3)	1.54	1.66	2.83	1.59	4.52	1.22
VASE-R #4	.53	.70	1.02	.72	1.83	.45
VASE-R #10	.63	.73	.93	.66	1.35	.64
VASE-R #16	.38	.65	.87	.73	1.33	.69
Eliciting change talk (3)	2.83	1.73	3.66	1.76	4.92	1.24
VASE-R #5	.96	.86	1.21	.84	1.79	.57
VASE-R #11	1.04	.87	1.19	.90	1.77	.60
VASE-R #17	.83	.71	1.26	.80	1.36	.74
Developing discrepancy (3)	3.24	1.90	4.10	1.61	5.08	1.14
VASE-R #6	1.31	.90	1.47	.88	1.88	.45
VASE-R #12	1.14	.94	1.39	.83	1.58	.70
VASE-R #18	.78	.94	1.25	.93	1.62	.76
Lisa vignette (6)	6.09	2.91	8.06	2.72	11.03	1.46
Ulysses vignette (6)	6.59	2.91	8.08	2.37	10.12	2.15
Bailey vignette (6)	5.42	2.96	7.67	3.15	9.73	2.52
Full VASE-R scale score (18)	18.09	7.56	24.13	6.74	30.88	5.02

Notes: Untrained practitioners were 144 substance abuse treatment practitioners *prior* to MI training. Trained practitioners were a subset (n = 111) of the above practitioners who completed 15 h of structured MI training activities and both pre- and post-assessments. MI experts were 66 members of the Motivational Interviewing Network of Trainers (MINT). Numbers in parentheses in column 1 reflect the number of items.

Drug Alcohol Depend. Author manuscript; available in PMC 2009 September 1.

Table 2

Internal consistency and inter-correlation of VASE-R scale and components

	Internal consistency	Corrected subscale-full scale correlation	Corrected item-subscale correlation	Corrected item-full scale correlation
Reflective listening	.69	.67		
VASE-R #1			.50	.49
VASE-R #2			.46	.45
VASE-R #7			.46	.52
VASE-R #9			.50	.61
Responding to resistance	.67	.64		
VASE-R #3			.26	.36
VASE-R #8			.41	.53
VASE-R #13			.49	.43
VASE-R #14			.56	.55
VASE-R #15			.41	.45
Summarizing	.73	.62		
VASE-R #4			.44	.51
VASE-R #10			.61	.52
VASE-R #16			.61	.59
Eliciting change talk	.49	.54		
VASE-R #5			.31	.39
VASE-R #11			.30	.37
VASE-R #17			.33	.45
Developing discrepancy	.44	.44		
VASE-R #6			.24	.37
VASE-R #12			.33	.26
VASE-R #18			.24	.36
Lisa vignette	.64	.66		
Ulysses vignette	.65	.74		
Bailey vignette	.71	.65		
Full VASE-R instrument	.85			

Notes: N = 144 substance abuse treatment practitioners *prior* to MI training. Internal consistency reflects Cronbach's α coefficients. Corrected subscale-scale metric reflects correlation (e.g., *r*) of the subscale with the full-scale instrument when the subscale is not included in the full-scale instrument score. Corrected item-subscale correlation reflects correlation of the item with the subscale score when the item is not included in the subscale score. Corrected item-scale correlation reflects correlation of the item with the full-scale instrument when the item is not included in the full-scale instrument score.

Table 3

Concurrent validity of VASE-R scale and components

	Full VASE-R scale	Reflective listening subscale	Responding to resistance subscale	Summarizing subscale	Eliciting change talk subscale	Developing discrepancy subscale
HRQ	.50***	.45***	.38***	.41***	.36***	.23**
SP interview empathy	.44***	.42***	.40***	.19*	.32***	.24**
SP interview MI spirit	.32***	.29**	.31***	.15	.24**	.13
SP interview R:Q	.35***	.33***	.25**	.22*	.31**	.18*
SP interview %OQ	.16	.14	.08	.12	.32***	−.04
SP interview %CR	.11	.08	.10	.02	.12	.10
SP interview %MIA	.26*	.30**	.31**	.15	.10	.01

Notes: N = 144 substance abuse treatment practitioners *prior* to MI training. All values are Pearson product-moment correlations. HRQ reflects summary score for Helpful Responses Questionnaire (Miller et al., 1991). SP interview reflects a 20-min audio-recorded encounter later reviewed and scored using the Motivational Interviewing Treatment Integrity scale (Moyers et al., 2004).

* $p < .05$,

** $p < .01$,

*** $p < .001$.

Table 4

Sensitivity of VASE-R scale and subscales to effects of training participation

	Pre-training		Post-training		3-Month follow-up		Multivariate effect size
	M	S.D.	*M*	S.D.	*M*	S.D.	
Full VASE-R score	18.21	7.41	24.13	6.74	23.08	7.53	.50
Reflective listening	4.88	2.26	6.13	2.04	5.97	2.25	.30
Responding to resistance	5.72	2.56	7.41	2.49	7.10	2.54	.28
Summarizing	1.53	1.67	2.83	1.59	2.41	1.62	.39
Eliciting change talk	2.79	1.69	3.66	1.76	3.52	1.77	.19
Developing discrepancy	3.29	1.86	4.10	1.61	4.07	1.57	.15

Notes: Data reflects 111 substance abuse treatment personnel who participated in formal MI training processes. All temporal changes reflect statistically significant difference at $p < .001$. Listed effect sizes represent partial η values.

Table 5

Untrained benchmark and proposed proficiency standards for VASE-R

	Untrained benchmark	Beginning proficiency	Expert proficiency
Full score VASE-R (range = 0–36)	18	26	31
Reflective listening (0–8)	5	6	7
Responding to resistance (0–10)	6	8	9
Summarizing (0–6)	1	3	5
Eliciting change talk (0–6)	3	4	5
Developing discrepancy (0–6)	3	4	5
Lisa vignette (0–12)	6	9	11
Ulysses vignette (0–12)	6	9	10
Bailey vignette (0–12)	5	8	10

Notes: Untrained benchmark (N = 129) are the scores from participants at pre-training in the community sample that had not received prior MI training via a workshop (lasting at least 1 day). Beginning Proficiency scores, are based on calculated cut scores considering means and variability of the untrained community sample and expert sample. Expert scores are based on means of 66 members of the MINT. All scores were rounded to the nearest whole number. As a result, sums of subscales may not match total scores.