Computer Assessment of Simulated Patient Interviews (CASPI): Psychometric Properties of a Web-Based System for the Assessment of Motivational Interviewing Skills

JOHN S. BAER, PH.D.,^{a,b,*} KELLY M. CARPENTER, PH.D.,^c BLAIR BEADNELL, PH.D.,^{d,e} SUSAN A. STONER, PH.D.,^c MICHELLE HANSTEN INGALSBE, M.S.W.,^b BRYAN HARTZLER, PH.D.,^b DAVID B. ROSENGREN, PH.D.,^{b,e} AND ZACH DRAGER, B.A.^c

^aDepartment of Psychology, University of Washington, Seattle, Washington

^bAlcohol and Drug Abuse Institute, University of Washington, Seattle, Washington

^cTalaria, Inc., Seattle, Washington

^dSchool of Social Work, University of Washington, Seattle, Washington

^ePrevention Research Institute, Lexington, Kentucky

ABSTRACT. Objective: Benefits of empirically supported interventions hinge on clinician skill, particularly for motivational interviewing (MI). Existing MI skill assessments are limited with respect to validity (e.g., self-report) and practicality (e.g., coding session tapes). To address these limitations, we developed and evaluated two versions of a web-based assessment of MI skills, the Computer Assessment of Simulated Patient Interviews (CASPI). **Method:** Ninety-six counselors from the community and 24 members of the Motivational Interviewing Network of Trainers (MINT) completed the CASPI ($N = 120$), in which they verbally responded via microphones to video clips comprising three 9-item vignettes. Three coders used an emergent coding scheme, which was compared with alternative MI skills measures. **Results:** CASPI demonstrated excellent internal consistency when averaging across two

or three vignettes (α 's = .86–.89). Intraclass correlations were above .40 for most items. Confirmatory factor analyses supported a correlated three-factor model: MI-consistent, resistance-engendering, and global change talk orientation rating. Means and factor loadings were invariant across forms (i.e., the two alternative versions of CASPI), and factor loadings were invariant across subgroup (i.e., community counselor or MINT member). Test-retest reliability was good for MI-consistent and resistance-engendering scores ($r = .74$ and $.80$, respectively) but low for change talk orientation ($r = .29$) unless coder was taken into account ($r = .69$). CASPI showed excellent construct and criterion-related validity. **Conclusions:** CASPI represents a promising method of assessing MI skills. Future studies are needed to establish its performance in real-world contexts. (*J. Stud. Alcohol Drugs*, 73, 154–164, 2012)

AS AN EMPIRICALLY SUPPORTED therapeutic approach for substance use disorders, motivational interviewing (MI) garners strong interest and dissemination activity. Nonetheless, published MI training evaluation studies suggest challenges for effective transfer of MI to clinical settings. Although intensive workshops produce skill gain, their retention likely requires adjunctive services such as the use of personalized feedback or coaching (Miller et al., 2004) or organizational support processes within agencies (Baer et al., 2009). There also appear to be individual differences in the learning of MI skills (Baer et al., 2004, 2009; Schoener et al., 2006). The substance-use-disorders treatment community needs reliable and valid tools for assessing clinician use of MI skills to identify training needs and responses to coaching.

Assessment that promotes dissemination of MI will be

Received: April 22, 2011. Revision: July 21, 2011.

This research was supported by a Small Business Technology Transfer grant from the National Institute on Drug Abuse (R42 DA020284).

*Correspondence may be sent to John S. Baer, S-116-ATC, VA Puget Sound HCS, 1660 S. Columbian Way, Seattle, WA 98108 or via email at: jsbaer@uw.edu. Information on the CASPI can be found at <http://www.talariainc.com>.

maximally useful if it is not only reliable and valid but also easily administered in real-world clinical settings. Unfortunately, currently available methods for assessing a clinician's MI skills present challenges that can be generally conceptualized as a trade-off between validity and ease of administration. Requiring the least amount of time and technological resources, clinician self-report is clearly efficient for large groups; consequently, it is commonly used to assess training needs and outcomes (Rubel et al., 2000; Saitz et al., 2000; Walters et al., 2005) despite its limited validity (Miller and Mount, 2001; Miller et al., 2004). Emphasizing validity, MI researchers typically record and review actual clinical encounters and have developed coding systems to quantify MI skillfulness, including the Independent Tape Rater Scale (Martino et al., 2008), the Motivational Interviewing Treatment Integrity (MITI) scale (Moyers et al., 2005), and the Motivational Interviewing Supervision and Training Scale (Madson et al., 2005). Although these approaches work well in tightly controlled research settings, they present numerous challenges in clinical settings, where it is difficult to record and review work samples and clinical case mixes are highly variable. Moreover, taping introduces practical dilemmas related to consent procedures, selection bias, and demand

characteristics, resulting in poor clinician compliance, even in research studies (Baer et al., 2004; Miller et al., 2004; Shafer et al., 2004).

An alternative approach, recording and review of standardized patient (SP) interviews (Baer et al., 2004; Luck and Peabody, 2002), allows for the presentation of relevant client stimuli while minimizing variance in client characteristics, clinician–client impression management, and client availability and confidentiality concerns. However, this method also requires recruiting and training actors to serve as SPs, clinical time for recording interviews, and resources for coding the resulting material. Various techniques have been used to make SP assessments more accessible and less costly, yet valid. Conducting SP interviews over the telephone or in chat rooms increases convenience for the clinician but still requires a live actor in real time. Using video via television or computer to simulate SP encounters simplifies logistics, lowers costs, and facilitates scoring by further standardizing the assessment.

Video-based SP examinations are increasingly used in clinician skill assessment (Funke and Schuler, 1998; Humphris and Kaney, 2000; Truxillo and Hunthausen, 1999). Recently, we developed and revised the Video Assessment of Simulated Encounters–Revised (VASE-R; Rosengren et al., 2005, 2008). This MI skill assessment uses video to present brief vignettes of actors portraying clients speaking to the camera about alcohol or drug use history, problems, and attitudes about change, enabling presentation of salient auditory and visual features (e.g., speech rhythm/intonation, body language, affect). After each video segment, respondents provide written responses corresponding to targeted MI skills (e.g., communicating understanding, asking a helpful question). Despite its strengths, practical limitations exist in using the VASE-R as with other video-based examinations. For example, respondents must access the examination via VHS or DVD. Because responses are written, they are dependent on writing skills and legibility and may be less nuanced than spoken language. Furthermore, time limits for writing responses may differentially hurry or delay respondents.

To build on the strengths of the VASE-R and address some of its limitations, we developed the Computer Assessment of Simulated Patient Interviews (CASPI), a web-based system for MI skills assessment. The content and format of the CASPI are based on the VASE-R, but the CASPI provides technology-based innovations in utility, accessibility, authenticity, ease of use, and feasibility. Accessible via the Internet, the CASPI enables clinicians, coders, supervisors, and trainers to use it wherever they have web-connected personal computers. To maximize authenticity of evaluated MI skills, respondents speak responses aloud (into a microphone) in real time. Skill evaluators benefit from consideration of inflection and tone of voice in rating respondent skills. The absence of additional equipment (e.g., tape recorders, questionnaires, television/videocassette recorder)

and schedule coordination (e.g., with an SP, client, clinical supervisor, or proctor) enhances feasibility.

The current report describes an initial evaluation of CASPI with two samples of clinicians: one sample of practitioners recruited without regard to training or experience in MI and a smaller sample consisting of MI trainers, whom we considered MI experts. We adapted content and scoring methods from the VASE-R. To facilitate utility of the measure when administered over repeated occasions, we developed two versions (i.e., forms) of the CASPI, with comparable stimuli. In this article, we present data with respect to item and scale reliability, test–retest reliability, consistency across forms, construct validity with respect to taped interviews with SPs, and comparisons between samples of clinicians.

Method

All procedures were approved by the University of Washington Human Subjects Division.

Development and formative evaluation

Each form initially consisted of three vignettes designed to measure two types of counselor skills thought to be necessary in MI. First are fundamental listening skills for engaging clients, including making empathic and accurate reflections and summaries, responding to sustain talk, and making affirming statements. Second are skills for guiding change talk, including identifying, responding to, and eliciting statements that promote change. The first of the three vignettes in Form A was taken from the VASE-R but portrayed by a new actor. We created the remaining five vignettes to be similar in length and complexity to those in the VASE-R. Each vignette portrays an individual with a substance use problem. Vignettes vary in gender, ethnicity, drug of choice, legal involvement, and degree of readiness to change.

Within each vignette, we developed nine test items. Items 1–4 are each preceded by a video snippet of the character making a statement about his or her substance use and then a prompt to the respondent: “Communicate to [Lisa] that you are listening” (Items 1 and 2, reflective listening), “Respond to [Lisa] in a way you think would be helpful” (Item 3, responding to sustain talk), and “Summarize for [Lisa] what you’ve heard [her] describe to this point” (Item 4, summarizing). The next item presents a still image of the character with the prompt, “Communicate to [Lisa] something you appreciate about [her] efforts” (Item 5, affirming). A multiple choice question follows: “Which of the following statements best suggests that [Lisa] may be considering making changes in [her] substance use?” (Item 6, identifying change talk). Then, CASPI repeats an earlier video clip ending with the prompt: “Respond to [Lisa’s] last statement in a way that reinforces [her] thinking about making a change in

drinking" (Item 7, responding to change talk). Next is a still image of the character with the prompt, "Respond to [Lisa] and attempt to elicit further statements in support of [her] changing [her] substance use" (Item 8, eliciting change talk). The last item prompts respondents to state a rationale for their prior response, "Briefly describe the reasons for your last response" (Item 9, rationale for change talk response).

Pilot testing with 12 participants and lab-based usability testing with six participants led to a set of detailed instructions for participants, trouble-shooting guidelines for the main study, and minor changes to the response interface. No changes were made to the content of CASPI vignettes or items based on usability or pilot testing.

Primary study procedures

We recruited community counselors through email notices sent to practitioners affiliated with the Pacific NW Node of the Clinical Trials Network, University of Washington Alcohol and Drug Abuse Institute, and University of Washington Department of Social Work. MI trainers volunteered in response to an email notice sent to the Motivational Interviewing Network of Trainers (MINT) LISTSERVE.

Study staff screened potential participants via the telephone to determine their eligibility and interest. For both groups, study eligibility required access to a broadband Internet-connected computer with recording capabilities, residence in the United States, and no prior exposure to the CASPI. Community counselors were required to work at least half time with clients on mental health or substance use issues. Eligible participants were sent a microphone headset, login credentials, and a link to the study website. Participants completed an online consent form and then tested recording capabilities. Staff members were available by phone to address questions and technical difficulties.

Community counselors completed two online assessments approximately 1 week apart. We randomly assigned counselors to complete either Form A or Form B of the CASPI at baseline and follow-up in a fully crossed 2×2 design. Assignment resulted in roughly equal numbers of participants ($n = 22$ – 26) completing A at both time points, B at both time points, A at baseline and B at follow-up, and B at baseline and A at follow-up. The baseline assessment took approximately 60 minutes and consisted of a demographics questionnaire, the Helpful Responses Questionnaire, the Short Understanding of Substance Abuse Scale, and the CASPI. The follow-up assessment took approximately 45 minutes and consisted of the Helpful Responses Questionnaire, the CASPI, and a CASPI satisfaction questionnaire. Research staff informed participants via email when their follow-up assessment was due. In between the two assessments, participants completed a 20-minute SP interview by telephone (see below). MI trainers completed only the baseline assessment and CASPI satisfaction questionnaire, followed by a

telephone SP interview. MI trainers completed either Form A or Form B of the CASPI, according to random assignment. Community counselors were paid \$100, and MI trainers were paid \$60 for their participation.

Participants

We completed screening with 157 potential participants based on inquiries. Of the 33 who did not participate, 7 were screened out (4 did not meet eligibility criteria and 3 were not interested), and 26 did not follow through. Of the remaining 124 who were randomly assigned, 120 completed all tasks and made recordings that were of acceptable volume and clarity for coding.

The analysis sample ($N = 120$) consisted of 96 community counselors and 24 MI trainers. MI trainers were all members of MINT. As shown in Table 1, MI trainers were typically older, married or domestic partnered, and earned a higher income than community counselors. They were more likely to report having a doctoral degree, extensive training and experience in substance use problems, more than 30 hours of MI training, and regular use of MI in clinical practice. Community counselors spent a greater number of hours per week providing substance use counseling; 22% were certified chemical dependency professionals.

Measures

Demographics questionnaire. A 24-item form gathered information such as age, gender, education, ethnicity, income, marital status, clinical experience with substance use, and amount of prior MI exposure.

The Helpful Responses Questionnaire (Miller and Rollnick, 1991). The Helpful Responses Questionnaire consists of six hypothetical client statements to which participants write how they would respond. Raters score responses using four codes: 1 = a response contrary to reflective listening ("roadblock"); 2 = a reflection in addition to a roadblock, or neither; 3 = a simple reflection ("rephrasing"); and 4 = reflection that infers appropriate meaning ("paraphrasing") or adds feeling or a metaphor/simile (Miller and Rollnick, 1991, modified by Rosengren et al., 2005). Two research assistants, blind to participant group membership, scored Helpful Responses Questionnaire responses. Rater intraclass correlations (ICCs) ranged from .80 to 1.00, based on a subset of 20 (of 218, 9%) randomly selected protocols.

Short Understanding of Substance Abuse Scale–disease model subscale (Humphreys et al., 1996). This seven-item subscale assesses beliefs about the origin and treatment of addictive behaviors rated on a 5-point Likert scale (1 = *strongly disagree* to 5 = *strongly agree*).

Computer Assessment of Simulated Patient Interviews satisfaction questionnaire. The satisfaction questionnaire consisted of 15 items using a 5-point Likert scale (1 = *strongly*

TABLE 1. Sample description ($N = 120$)

Characteristic	Community counselors ($n = 96$) %	MI trainers ($n = 24$) %	χ^2	df	p
Female	74	67	0.51	1	.47
Hispanic	7	0	1.88	1	.17
Race			2.21	5	.82
Asian	4	0			
Black	7	4			
White	80	83			
Mixed	5	8			
Other	3	4			
Age, in years			7.31	3	.06
25–34	33	8			
35–44	23	21			
45–54	23	33			
≥ 55	21	38			
Marital status			6.66	2	.04
Single	30	4			
Married/domestic partnered	55	75			
Divorced/separated	16	21			
Income			20.87	4	<.001
<\$20,000	2	0			
\$20,001–40,000	24	4			
\$40,001–60,000	33	13			
\$60,001–100,000	31	38			
$\geq$ \$100,000	11	46			
Education			29.64	5	<.001
High school	3	0			
Associate's degree	9	0			
Bachelor's degree	17	4			
Master's degree	66	50			
Doctoral degree	4	38			
Other	1	6			
State-certified chemical dependency professional			6.36	1	<.01
Yes	22	0			
No	78	100			
Training/experience with substance use problems			9.56	2	<.01
Minimal/slight	22	4			
Moderate	44	29			
Extensive	34	67			
Hours per week providing substance use counseling			20.90	3	<.001
0–3	17	58			
4–7	13	21			
8–15	24	8			
≥ 16	46	13			
Current use of MI in clinical practice			22.70	3	<.001
Not at all	14	17			
Infrequently, with a targeted client	23	4			
Occasionally, with some of my clients	40	8			
Regularly, with many or all clients	23	71			
Hours of MI training received			45.80	4	<.001
0	37	8			
1–10	27	0			
11–20	17	8			
21–30	9	8			
≥ 31	11	75			

Notes: Percentages may not add to 100 because of rounding. MI = motivational interviewing.

disagree to 5 = *strongly agree*). Examples include “the situations and characters presented in CASPI were realistic,” and “the video components of the program worked well.”

Standardized patient interviews. Experienced and supervised actors followed a detailed character description

developed by the investigators to portray a woman with substance use concerns. At the outset of the 20-minute telephone interview, we instructed study participants not to conduct an assessment but to interact with the SP in ways they thought would be helpful. SPs audiotaped interviews.

Denise Ernst Training and Consultation (i.e., not members of the investigative team) scored tapes for MI skillfulness using the MITI 3.1 coding system (Moyers et al., 2005, 2010). The MITI yields global ratings of MI skills (empathy, evocation, collaboration) as well as counts and ratios of specific kinds of utterances (e.g., open questions, reflections, MI-adherent interventions). There were four coders and one additional coder who established reliability only, all blind to participant group membership. All coders rated the same 14 (of 121, 12%) randomly selected protocols. ICCs for global scores ranged from .61 to .70, behavior counts from .47 (complex reflection) to .96 (closed questions), and summary measures from .55 to .94, all of which are fair to excellent according to Cicchetti's (1994) criteria.

Scoring the Computer Assessment of Simulated Patient Interviews

As noted, CASPI records verbal responses to standard patient statements. Three of the study investigators (two of whom are MINT members), blind to participant group membership and assessment order, conducted the CASPI coding. Using responses from pilot study participants, we initially used a coding system similar to that of the VASER, which assigned each CASPI response a single number from a 4-point range (0 = *resistance-engendering* to 3 = *highly skilled*) for each response. We discarded this approach because of poor interrater reliability and, given the many linguistic elements encountered, the need for complicated decision rules assuming priority of some elements over others.

A revised scoring system specified two or three dichotomous codes for each CASPI item. As seen in Table 2, responses to Items 1, 2, 3, 7, and 8 each received two codes: (a) the presence/absence of the MI-consistent element expected for that item (i.e., accurate reflection for Items 1 and 2) and (b) the presence/absence of a resistance-engendering element. We used two dichotomous codes to capture the MI-consistent element for Item 4 (summarizing—multiple concepts and strategic organization) and Item 5 (affirming—presence of affirmation and personal reference) in addition to a resistance-engendering code. Item 9 (rationale for prior response) received only one code.

Binary codes capture the presence of the targeted MI skill but do not reflect excellent or highly skilled responding. In an effort to capture more skilled MI responding, we added a vignette-level code for *change talk orientation* and defined it as the number of responses within a vignette that focus on prior client change statements or contain impetus to elicit further client change statements (e.g., an amplified reflection). We used a 5-point Likert scale ranging from 1 = *no responses focus on prior client change statements, nor do any contain impetus to elicit further client change statements*, to 5 = *most or all responses (6–7 responses) focus on client change statements or contain impetus to elicit further client change statements*.

Analysis strategy

We first created summed scores across the three vignettes for each item and then examined item variability and inter-

TABLE 2. Intraclass correlations for rater agreement of CASPI items

Item	Vignette			Summed across 3
	1	2	3	
1a. Reflective listening 1: Accurate reflection	.33	— ^a	.87	.79
1b. Reflective listening 1: Resistance engendering	.22	.59	.85	.74
2a. Reflective listening 2: Accurate reflection	.42	.49	.85	.63
2b. Reflective listening 2: Resistance engendering	.28	.53	.77	.74
3a. Responding to resistance: Rolls with resistance	.57	.56	.78	.78
3b. Responding to resistance: Resistance engendering	.53	.59	.61	.75
4a. Summarizing: Linkage of multiple concepts	.48	.37	.50	.65
4b. Summarizing: Strategic organization	.43	.50	.06	.46
4c. Summarizing: Resistance engendering	.14	.74	.60	.72
5a. Affirming: Presence of affirmation	.18	.38	.39	.33
5b. Affirming: Personalized reference	.38	.31	.35	.34
5c. Affirming: Resistance engendering	.59	.70	.43	.70
7a. Responding to change talk: Reflection, affirmation, or open question about change	.59	.49	.35	.66
7b. Responding to change talk: Resistance engendering	.74	.55	.63	.77
8a. Eliciting change talk: Open question or reflection about change	.68	.73	.74	.84
8b. Eliciting change talk: Resistance engendering	.59	.77	.63	.84
9a. Rationale for change talk response	.74	.73	.70	.87
10a. CTO rating	.39	.38	.50	.48

Notes: CASPI = Computer Assessment of Simulated Patient Interviews; CTO = change talk orientation. ^aVariance too low on the subsample coded by all three raters to calculate intraclass correlation (item was retained for analysis, given that it did have variability in the larger sample and had acceptable intraclass correlation when summed across vignettes).

rater reliability. Items were subjected to confirmatory factor analyses using Mplus Version 6.0 (Muthén and Muthén, 2010). After establishing a good-fitting model, we used MIMIC (multiple indicators/multiple causes) modeling (Muthén, 1989) to test for item loading invariance between community counselors and MI trainers and for item loading and mean invariance between the two CASPI forms. Based on these findings, we then calculated scale scores for examinations of test–retest reliability, construct validity, criterion-related validity, and practice effects. Finally, we evaluated consistency of vignettes by repeating the confirmatory factor analysis and MIMIC models described earlier for scores for each individual vignette and summed across each pair of vignettes.

Results

Computer Assessment of Simulated Patient Interviews satisfaction and use

Overall, participants gave positive ratings to the CASPI. Mean satisfaction ratings on a 1 (*strongly disagree*) to 5 (*strongly agree*) scale ranged from 3.9 to 4.7, with standard deviations typically less than 1. Ratings were above 4.5 on items related to ease of use (e.g., “I was able to use the voice recording tool” and “The CASPI voice recording tool instructions were easy to understand”). Participants enjoyed using the CASPI ($M = 4.0$) and found the situations and conversations to be realistic ($M = 4.4$). Over the course of the study, we documented 54 instances of technical difficulties from 44 different participants. All but 10 instances were resolved through consultation with study staff. We have since made further refinements to the user instructions.

Coding variability

All codes for CASPI items summed across vignettes showed acceptable variability. However, the multiple choice item for identifying change talk showed limited variability; only 11 (9.2%) and 3 (2.5%) participants answered incorrectly in Vignettes 2 and 3, respectively. This item was therefore excluded from further analysis. All Helpful Responses Questionnaire codes showed acceptable variability, as did MITI codes of SP interviews, again with one exception. The MITI global code for direction showed little variability, with 114 (94.2%) cases being scored a 4 or 5. We excluded this item from further analysis as well.

Computer Assessment of Simulated Patient Interviews interrater reliability

We randomly selected a subset of 30 protocols (14%) to be coded by all three raters. We calculated ICCs using an absolute agreement, mixed-model approach (Cicchetti, 1994). As can be seen in Table 2, ICCs were above .40 for

most items. Some exceptions occurred on individual vignette codes but were of less concern when the codes were summed across vignettes (e.g., 4c Vignette 1, summarizing: resistance engendering; 4a Vignette 2, linkage of multiple concepts code; and 4b Vignette 3, strategic organization code). In contrast, the codes designed to tap affirmations (affirming and personalized reference) showed low reliability on individual vignettes as well as across-vignette sums; the affirmation item and codes used to score it were thus excluded from further psychometric analysis.

Confirmatory factor analyses

We first tested a model with four factors, representing the two a priori skill areas (engaging and guiding change talk), as well as codes for resistance-engendering and our change talk orientation global score. Although conceptually related to guiding change talk, we chose to keep change talk orientation score separate because it was coded on the vignette level rather than the item level. Variables in the model were the codes summed across the three vignettes. The initial model had acceptable fit, $\chi^2(75) = 90.197$, $p = .111$ (comparative fit index [CFI] = .992; Tucker–Lewis index [TLI] = .988; root mean square error of approximation [RMSEA] = .041). However, high correlations occurred between the engaging and guiding change talk factors ($r = .88$, $p < .001$). We then estimated a three-factor confirmatory factor analysis, with the engaging and guiding change talk items loading on one rather than two factors, labeled as MI-consistent (Figure 1). Fit was acceptable, $\chi^2(78) = 96.346$, $p = .080$ (CFI = .990, TLI = .987, RMSEA = .044). Correlated errors between individual MI-consistent codes and their respective resistance-engendering counterparts (freely estimated in the model) ranged from as low as $-.03$ to as high as $-.57$ ($p < .001$). This final model was used for subsequent invariance testing across form (A vs. B) and group (community counselor vs. MI trainer).

Invariance testing

We estimated a MIMIC model by adding dummy variables for form (A vs. B) and group (community counselor vs. MI trainer), freeing paths to the latent variables, and fixing the paths to the individual codes to 0. Acceptable fit indicated factor loading invariance (no differential item functioning), $\chi^2(102) = 121.558$, $p = .09$ (CFI = .988, TLI = .984, RMSEA = .040). Additionally, findings were consistent with mean invariance for the two forms—that they would not produce different average scores—in that the path from the dummy variable for form to each latent construct was nonsignificant (β s = .13, $-.03$, and .13 to MI-consistent, resistance-engendering, and change talk orientation, respectively, all n.s.). The path from the dummy variable for group to each construct was statistically significant, with

FIGURE 1. Three-construct confirmatory factor analysis of Computer Assessment of Simulated Patient Interviews items ($N = 120$). MI = motivational interviewing; CTO = change talk orientation.

MI trainers scoring better than community counselors in the predicted directions: higher on change talk orientation and MI-consistency ($\beta = .29$ and $.42$, $p < .007$ and $.001$, respectively), and lower on resistance-engendering ($\beta = -.50$, $p < .001$).

Calculation of summary scores

For the remainder of our analyses, we calculated summary scores based on the above findings. For a CASPI administration, we first summed like codes across the three vignettes, then summed appropriate codes to create a score for MI-consistent responses and resistance-engendering responses. Three ratings of change talk orientation (1–5 scale) were summed to a single change talk orientation score (range: 3–15). Internal consistency for these items was acceptable (baseline and retest, respectively: Cronbach's $\alpha = .86$ and $.87$ for MI-consistent, $.87$ and $.89$ for resistance-engendering, and $.86$ and $.86$ for change talk orientation). We also computed an additional score to reflect overall CASPI performance, the CASPI summary score (CSS). This was calculated as MI-consistency score minus the resistance-engendering score.

Practice effects and test–retest reliability

Paired t tests comparing baseline and retest scores suggested that practice would not significantly alter scores: change talk orientation, $t(91) = -0.97$, *n.s.*; MI-consistent, $t(93) = 0.36$, *n.s.*; resistance-engendering, $t(92) = 1.77$, *n.s.*; and CSS $t(92) = 0.67$, *n.s.* Pearson correlations of baseline with retest scores were acceptable for MI-consistent and resistance-engendering scores ($r = .74$ and $.80$, respectively,

$ps < .001$). The correlation for change talk orientation scores was low ($r = .29$, $p < .001$). This appeared to be because of low reliability between coders. We separated the data into cases in which the same coder made both the baseline and retest ratings, and those where the ratings were made by different coders and found markedly different correlations. When the same coder rated both baseline and retest scores, the correlation reflected considerable stability over time ($r = .69$, $p < .001$); yet, when the coders were different, no stability was apparent ($r = .06$, *n.s.*).

Construct validity

As shown in Table 3, independent sample t tests indicated that MI trainers' scores were superior to community counselors', and between-group effect sizes were large (customarily defined as Cohen's $d \geq .80$). MI trainers, on average, missed few opportunities for MI-consistent responding, scoring on 22 of 23 possible items, and rarely made responses that were judged as likely to engender resistance ($M_{\text{CSS}} = 21.7$). Change talk orientation scores for MI trainers were usually about 4 (1–5 scale) for each vignette. For community counselors, MI-consistent scores averaged almost 19, with an average of three items thought to engender resistance ($M_{\text{CSS}} = 15.8$). Change talk orientation scores were more likely to be at a 3 on each vignette. Notably, standard deviations showed much greater variability among community counselors compared with MI trainers.

Scores from MI trainers should reasonably represent a proficiency standard for the CASPI. Our community counselors, however, appeared to be a heterogeneous group with respect to interest and experience with MI. We thus categorized community counselors based on self-reported prior MI

TABLE 3. Construct validity tested by comparing CASPI scores between MI trainers and community counselors with scores for community counselors grouped by amount of prior MI training

CASPI scale	MI trainers vs. community counselors, M (SD)		Community counselors, by prior MI training, M (SD) ^f			MI trainers vs. community counselors comparison			
	MI trainers ($n = 24$)	Community counselors ($n = 96$)	0 hours ($n = 37$)	1–15 hours ($n = 30$)	≥ 16 hours ($n = 28$)	t	df	p	Cohen's d^e
Three vignettes									
MI-consistent score ^a	22.1 (1.6)	18.8 (4.5)	17.7 (5.4)	19.0 (4.5)	19.9 (2.9)	5.94	106	<.001	0.81
Resistance-engendering score ^b	0.4 (0.7)	2.9 (3.5)	4.3 (4.3)	2.4 (2.9)	1.8 (2.4)	6.51	116	<.001	0.79
CTO rating	11.8 (2.5)	9.3 (3.2)	8.6 (3.1)	9.9 (3.5)	9.6 (3.1)	4.08	45	<.001	0.81
CSS	21.7 (1.8)	15.8 (7.5)	13.5 (9.0)	16.6 (7.2)	18.1 (4.7)	6.86	118	<.001	0.87
Vignettes 2 and 3									
MI-consistent score ^c	14.5 (1.1)	12.3 (3.1)	11.6 (3.7)	12.3 (3.0)	13.1 (2.0)	5.75	104	<.001	0.78
Resistance-engendering score ^d	0.3 (0.5)	2.2 (2.8)	3.3 (3.4)	1.8 (2.3)	1.2 (1.7)	6.41	114	<.001	0.75
CTO rating	7.6 (1.5)	6.0 (2.1)	5.5 (2.1)	6.3 (2.3)	6.4 (1.9)	4.27	51	<.001	0.80
CSS	14.3 (1.2)	10.1 (5.5)	8.4 (6.8)	10.5 (5.1)	11.9 (3.1)	6.80	116	<.001	0.85

Notes: CASPI = Computer Assessment of Simulated Patient Interviews; MI = motivational interviewing; CTO = change talk orientation; CSS = CASPI summary score. ^aEight items summed; range can be from 0 (*poor*) to 24 (*perfect*); ^bsix items summed; range can be from 0 (*perfect*) to 18 (*poor*); ^ceight items summed; range can be from 0 (*poor*) to 16 (*perfect*); ^dsix items summed; range can be from 0 (*perfect*) to 12 (*poor*); ^einterpretation of Cohen's d effect size: .20 small, .50 medium, .80 large; ^ffor both the two- and three-vignette scores, analysis of variance showed significant differences between the three groups on resistance-engendering and CSS scores but not MI-consistent and CTO scores.

TABLE 4. Criterion validity tested by zero-order Pearson correlations between CASPI scores with HRQ and MITI scores

Variable	CTO	MI consistent	Resistance engendering	CSS
MITI global score ^a				
Evocation	.47**	.55**	-.59**	.60**
Collaboration	.48**	.58**	-.60**	.62**
Autonomy	.43**	.55**	-.48**	.55*
Empathy	.50**	.56**	-.58**	.61**
MITI summary scores ^a				
% of questions that are open	.37**	.42**	-.43**	.45**
% of reflections that are complex	.28**	.24**	-.26**	.26**
Ratio of reflections to questions	.43**	.44**	-.44**	.47**
% MI adherent	.39**	.50**	-.57**	.57**
HRQ summary, baseline ^b	.46**	.53**	-.61**	.60**
HRQ summary, retest ^c	.43**	.55**	-.56**	.58**

Notes: CASPI = Computer Assessment of Simulated Patient Interviews; HRQ = Helpful Responses Questionnaire; MITI = Motivational Interviewing Treatment Integrity scale; MI = motivational interviewing; CTO = change talk orientation; CSS = CASPI summary score. ^aBaseline MITI score with baseline CASPI scores; ^bbaseline HRQ score correlated with baseline CASPI scores; ^cretest HRQ score correlated with retest CASPI scores.

* $p < .05$; ** $p < .01$.

training. Table 3 provides mean scores and variability for those reporting no prior training, those who report up to 15 hours of MI training, and those with 16 hours or more. As can be seen in Table 3, CASPI scores were positively correlated with prior MI training. For example, whereas community counselors on average achieved an overall CASPI score of 15.8, those with no prior MI training scored an average of 13.5.

CSSs were also related to counselor demographic and background variables. Counselors who were older ($r = -.45$, $p < .001$), carried a state Chemical Dependence Certification ($r = -.21$, $p < .05$), endorsed disease model beliefs on the Short Understanding of Substance Abuse Scale ($r = -.39$, $p < .001$), and spent more hours a week working in substance use treatment ($r = -.32$, $p < .001$) had lower CSSs compared with others. In contrast, those reporting higher educational attainment ($r = .37$, $p < .001$) and (consistent with Table 3) more hours of prior MI training ($r = .24$, $p < .05$) scored comparatively higher.

Criterion-related validity

Pearson correlations shown in Table 4 supported the hypothesis that CASPI scores would be significantly correlated with Helpful Responses Questionnaire and SP-MITI scores. In all cases, correlations were statistically significant, and, in most cases, moderate in magnitude.

Vignette analysis

To evaluate the stability of individual vignettes and establish if the CASPI could be shortened from its initial three-vignette design, we performed confirmatory factor analysis and MIMIC models (as above) separately for each of the

three vignettes and for scores summed across each pair of vignettes. We tested factor structure and mean and factor loading invariance. Generally, all models had similar results as described above: similar factor structures and factor loading invariance across forms and MI expertise. Also, models showed mean invariance across forms but, as hypothesized, mean differences because of MI expertise. Among these six models, each with many parameters being tested (15 factor loadings and 6 regression paths), there were two exceptions: The second vignette and the summed first/second vignettes' scores were higher for Form B versus Form A ($\beta_s = .21$ and $.18$, $ps \leq .05$).

These findings suggest that the use of two vignettes is a reasonable approach for situations requiring a briefer administration. We reviewed the content of each vignette and favored the combination of the second (which contains both alcohol and drug use scenarios) and third (designed to contain the most challenging scenarios). Internal consistency for items summed across these two vignettes, test-retest reliability, practice effects, and validity did not show appreciable changes from what we found with the three-vignette scores. Table 3 provides mean scores and variability estimates for MI trainers, community counselors, and subgroups of community counselors as a function of prior MI training for a two-vignette version of the CASPI.

Discussion

As dissemination of MI continues in substance use treatment (and many other areas of health care), it is necessary to assess skill needs, uptake, and maintenance in counselors. This study describes the development and psychometric properties of a web-based system to assess therapist MI skills. The CASPI is innovative in that it al-

lows test-takers to speak aloud to virtual clients, thereby increasing the realism and speed of the simulated encounter. The web-based system also greatly increases access—the examination can be taken at any time from any place with a broadband Internet connection and microphone. The CASPI stores audio recordings on a secure server, provides coding access through a computer interface, and calculates scale scores once coding is complete. We developed two versions (i.e., forms) to allow for assessment over time with novel stimuli. Overall, the CASPI appears to have robust psychometric properties. Internal consistency is excellent, particularly when averaged across two or three vignettes. There is little variability in scale structure across forms or vignettes. Construct validity also appears quite good; MI trainers scored significantly higher and with much less variability than a group of community counselors who reported a range of exposure to MI. CASPI scores were significantly correlated with a written measure of reflective listening (the Helpful Responses Questionnaire) and with a coding of recorded interviews with a standard patient using the MITI. Although developed with vignettes depicting problems with alcohol and other drug use, our data do not suggest that use of the CASPI should necessarily be limited to those working in substance use treatment. In fact, participants' hours working in treatment of substance use was a negative predictor of CASPI scores.

Dimensionality of the CASPI, as well as the operational definitions of specific MI skills, needs further study. Some concepts in MI appear more straightforward to code than others. Our affirmation scale had low interrater reliability and was eliminated from the total scores. In defining summaries, we tried to include an organizing element; although of technical interest, this aspect proved unreliable in coding. Our efforts at capturing change talk orientation with a vignette-level rating met with mixed results as well. The score correlated positively but not completely with other factors of the CASPI and with measures of construct validity. Yet agreement between coders was low and was reliable over time only within the same coder. Thus, these more general vignette-level ratings should be used with caution, and more work is needed to further develop and refine coding of affirmations and summary reflections. Despite these limitations, we have chosen not to remove items tapping affirmations and summaries from the CASPI protocol. These counseling techniques are potentially important in training MI (and other counseling approaches). Until we have a refined model of exactly what elements are necessary in learning MI, we prefer to allow trainers and coaches a variety of options for providing feedback in the learning process.

A similar logic applies to our confirmatory factor analyses, which supported three correlated factors: MI-consistent responses, resistance-engendering responses, and our global rating of change talk orientation. Although the CASPI could

be represented with one factor, we recommend a correlated factor model to support the utility of the instrument for training and coaching. Because we observed that the same CASPI response could contain both an MI-consistent element and a statement that violates the spirit of MI, trainers and supervisors should find it helpful to have the option to provide feedback on both aspects of a clinical response. Change talk orientation similarly is a scale that trainers can use to prompt attention to client motivation contained in the CASPI stimuli. Yet given limitations in the reliability of coding affirmations and change talk orientation scores, we have also removed them from scores that could be used to make assessment decisions.

In reviewing our analyses, we note that resistance-engendering codes were slightly more reliable than others, differed more consistently between community counselors and MI trainers, and were more strongly related to external criteria (e.g., MITI codes from SP interviews). This pattern of results implies that we were better able to identify responses that are contrary to MI than more subtle features of advanced MI skills. Indeed, this coding system, designed to provide resistance-engendering codes for most items, likely makes the CASPI sensitive in assessing a more directive or confrontational therapeutic style.

Our interest in developing the CASPI was, in part, to develop a system that would support the dissemination of evidence-based practice. We included content and design features to facilitate simplicity, ease of use, and accessibility. Of course, the length of the assessment procedure is also an important consideration for busy clinicians and supervisors. Depending on length and speed of response, it takes about 20 minutes to complete CASPI. Scoring time also varies based on length of response, but it is typically completed in about 15 minutes. We evaluated whether a shortened CASPI might retain similar psychometric properties, and our analysis suggests that a two-vignette version should function as well as a full three-vignette version. To support interpretation, we have provided mean scores in Table 3 representing both expert and several community comparisons.

One weakness of the present study is that the CASPI coders were well trained and had the opportunity to discuss coding procedures regularly. In general use, it is likely that the instrument will be coded by MI trainers or clinical directors with expertise in MI but without specific training in CASPI scoring. Although we strove to develop a scoring method that is easy to understand and apply, we do not know how reliable codes will be when coders are new to the system.

In sum, the characteristics of the CASPI create several advantages over other methods of MI skill assessment in terms of clinical realism, availability, ease of scoring, and training utility of subscale scores. It was also well liked by users. Because prior research shows that attending a training session does not necessarily equate to acquisition of MI skills, a tool such as the CASPI could have substantial utility

in assessing clinician skill before and after training and for efforts to support skill maintenance over time.

Acknowledgments

The authors thank Emily Downing, Kate Blumstein, Tasha Mikko, and KrisAnn Schmitz for their support of the project.

References

- Baer, J. S., Rosengren, D. B., Dunn, C. W., Wells, E. A., Ogle, R. L., & Hartzler, B. (2004). An evaluation of workshop training in motivational interviewing for addiction and mental health clinicians. *Drug and Alcohol Dependence*, 73, 99–106.
- Baer, J. S., Wells, E. A., Rosengren, D. B., Hartzler, B., Beadnell, B., & Dunn, C. (2009). Agency context and tailored training in technology transfer: A pilot evaluation of motivational interviewing training for community counselors. *Journal of Substance Abuse Treatment*, 37, 191–202.
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6, 284–290.
- Funke, U., & Schuler, H. (1998). Validity of stimulus and response components in a video test of social competence. *International Journal of Selection and Assessment*, 6, 115–123.
- Humphreys, K., Greenbaum, M. A., Noke, J. M., & Finney, J. W. (1996). Reliability, validity, and normative data for a short version of the Understanding of Alcoholism Scale. *Psychology of Addictive Behaviors*, 10, 38–44.
- Humphris, G. M., & Kaney, S. (2000). The Objective Structured Video Exam for assessment of communication skills. *Medical Education*, 34, 939–945.
- Luck, J., & Peabody, J. W. (2002). Using standardised patients to measure physicians' practice: Validation study using audio recordings. *British Medical Journal*, 325(7366), 679. Available at <http://www.bmj.com/content/325/7366/679.full>
- Madson, M. B., Campbell, T. C., Barrett, D. E., Brondino, M. J., & Melchert, T. P. (2005). Development of the Motivational Interviewing Supervision and Training Scale. *Psychology of Addictive Behaviors*, 19, 303–310.
- Martino, S., Ball, S. A., Nich, C., Frankforter, T. L., & Carroll, K. M. (2008). Community program therapist adherence and competence in motivational enhancement therapy. *Drug and Alcohol Dependence*, 96, 37–48.
- Miller, W. R., & Mount, K. A. (2001). A small study of training in motivational interviewing: Does one workshop change clinician and client behavior? *Behavioural and Cognitive Psychotherapy*, 29, 457–471.
- Miller, W. R., & Rollnick, S. (1991). *Motivational interviewing: Preparing people to change addictive behavior*. New York, NY: Guilford Press.
- Miller, W. R., Yahne, C. E., Moyers, T. B., Martinez, J., & Pirritano, M. (2004). A randomized trial of methods to help clinicians learn motivational interviewing. *Journal of Consulting and Clinical Psychology*, 72, 1050–1062.
- Moyers, T. B., Martin, T., Manuel, J. K., Hendrickson, S. M., & Miller, W. R. (2005). Assessing competence in the use of motivational interviewing. *Journal of Substance Abuse Treatment*, 28, 19–26.
- Moyers, T. B., Martin, T., Manuel, J. K., Miller, W. R., & Ernst, D. (2010). *Revised global scales: Motivational interviewing treatment integrity 3.1.1 (MITI 3.1.1)*. Albuquerque, NM: University of New Mexico, Center on Alcoholism, Substance Abuse, and Addictions (CASAA). Retrieved from http://casaa.unm.edu/download/MITI3_1.pdf
- Muthén, B. O. (1989). Tobit factor analysis. *British Journal of Mathematical and Statistical Psychology*, 42, 241–250.
- Muthén, L. K., & Muthén, B. O. (2010). Mplus (Version 6.0) [Computer software]. Los Angeles, CA: Author.
- Rosengren, D. B., Baer, J. S., Hartzler, B., Dunn, C. W., & Wells, E. A. (2005). The Video Assessment of Simulated Encounters (VASE): Development and validation of a group-administered method for evaluating clinician skills in motivational interviewing. *Drug and Alcohol Dependence*, 79, 321–330.
- Rosengren, D. B., Hartzler, B., Baer, J. S., Wells, E. A., & Dunn, C. W. (2008). The Video Assessment of Simulated Encounters–Revised (VASE-R): Reliability and validity of a revised measure of motivational interviewing skills. *Drug and Alcohol Dependence*, 97, 130–138.
- Rubel, E., Shepell, W., Sobell, L., & Miller, W. (2000). Do continuing education workshops improve participants' skills? Effects of a motivational interviewing workshop on substance-abuse counselor's skills and knowledge. *Behavior Therapist*, 23, 73–77.
- Saitz, R., Sullivan, L. M., & Samet, J. H. (2000). Training community-based clinicians in screening and brief intervention for substance abuse problems: Translating evidence into practice. *Substance Abuse*, 21, 21–31.
- Schoener, E. P., Madeja, C. L., Henderson, M. J., Ondersma, S. J., & Janisse, J. J. (2006). Effects of motivational interviewing training on mental health therapist behavior. *Drug and Alcohol Dependence*, 82, 269–275.
- Shafer, M. S., Rhode, R., & Chong, J. (2004). Using distance education to promote the transfer of motivational interviewing skills among behavioral health professionals. *Journal of Substance Abuse Treatment*, 26, 141–148.
- Truxillo, D. M., & Hunthausen, J. M. (1999). Reactions of African-American and White applicants to written and video-based police selection tests. *Journal of Social Behavior and Personality*, 14, 101–112.
- Walters, S. T., Matson, S. A., Baer, J. S., & Ziedonis, D. M. (2005). Effectiveness of workshop training for psychosocial addiction treatments: A systematic review. *Journal of Substance Abuse Treatment*, 29, 283–293.