AMERICAN PSYCHOLOGICAL ASSOCIATION

Society for HEALTH PSYCHOLOGY

Health Psychology

© 2021 American Psychological Association
ISSN: 0278-6133

2021, Vol. 40, No. 7, 439–449
https://doi.org/10.1037/hea0001064

Development of the Motivational Interviewing Coach Rating Scale (MI-CRS) for Health Equity Implementation Contexts

Sylvie Naar[1], Jason Chapman[2], Phillippe B. Cunningham[3], Deborah Ellis[4], Karen MacDonell[4], and Lisa Todd[4]

[1] Center for Translational Behavioral Science, Florida State University
[2] Oregon Social Learning Center, Eugene, Oregon, United States
[3] Department of Psychiatry and Behavioral Sciences, Medical University of South Carolina
[4] Department of Family Medicine and Public Health Sciences, Wayne State University

Objective: The field of implementation science emphasizes efficient and effective fidelity measurement for research outcomes and feedback to support quality improvement. This paper reports on such a measure for motivational interviewing (MI), developed with rigorous methodology and with diverse samples. ***Method:*** Using item response theory (IRT) methods and Rasch modeling, we analyzed coded (a) recordings ($n = 99$) of intervention sessions in a clinical trial of African American adolescents with obesity; (b) standard patient interactions ($n = 370$) in an implementation science study with youth living with HIV; and (c) standard patient interactions ($n = 172$) in a diverse community sample. ***Results:*** These methods yielded a reliable and valid 12-item scale on several indicators using Rausch modeling including single construct dimensionality, strong item-session maps, good rating scale functionality, and item fit after revisions. However, absolute agreement was modest. The 12 items yielded thresholds for 4 categories: beginner, novice, intermediate and advanced. ***Conclusions:*** The 12-item Motivational Interviewing Coach Rating Scale is the first efficient and effective fidelity measure appropriate with diverse ethnic groups, with interventions that are MI only or interventions that integrate MI with other interventions, and with adolescents and families as well as adults.

Keywords: fidelity, health equity, methodology

Supplemental materials: https://doi.org/10.1037/hea0001064.supp

Implementation science models emphasize the need to establish and sustain fidelity monitoring and a monitoring feedback system (Aarons et al., 2011). *Fidelity* refers to adherence to the intervention implementation plan (often measured via site visits or surveys of program staff), as well as competency in program delivery (Cross & West, 2011). Fidelity assessment is necessary to measure gaps between interventions delivered under research conditions and those delivered in routine care, both for outcomes assessment and for quality improvement. Yet, conventions for assessing intervention fidelity are not well established, measures are rarely developed with state-of-the-art methodology, little is known about the extent to which established methods can be used in real-world settings, and adherence is typically assessed more than competence (Schoenwald et al., 2011). Efficient competency measurement can aid sustainability by providing supervisors with easily used tools for ongoing quality assurance (Schoenwald et al., 2011). A measure of competency with strong established psychometric properties will not be used in real-world clinics if it is too costly or difficult to integrate into routine practice.

Motivational interviewing (MI) is a behavior change intervention approach that has been studied heavily in diverse settings (Lundahl et al., 2013) and, because of its strong evidence base, has been a focus in implementation science (Bauer et al., 2015; Madson et al., 2016; Midboe et al., 2011). MI is a collaborative, goal-oriented style of communication with particular attention to the language of change. It is designed to strengthen intrinsic motivation for and commitment to a specific goal by eliciting and exploring the person's own reasons for change within an atmosphere of acceptance and compassion (Miller & Rollnick, 2013).

Sylvie Naar https://orcid.org/0000-0001-6369-4685
Jason Chapman https://orcid.org/0000-0002-9396-5877
Phillippe B. Cunningham https://orcid.org/0000-0003-0814-2874

The instrument is copyrighted by Florida State University and licensed to a nonprofit agency, Behavior Change Consulting (behaviorchangeconsulting.org). Funding for this project came from National Heart, Lung, and Blood Institute U01HL097889, National Institute of Mental Health R34MH103049, and the Eunice Kennedy Shriver National Institute of Child Health and Human Development U19HD089875. Clinical Trials Registration: NCT01350531. Special thanks to Maurice Bulls and Sarah Martinez for providing their coding expertise.

Sylvie Naar served as lead for funding acquisition and project oversight, and contributed to the writing. Jason Chapman oversaw the design and analysis and contributed to the writing. Phillippe B. Cunningham contributed to the study conceptualization and design and supported the writing. Deborah Ellis contributed to study conceptualization and design and reviewed and edited the writing. Karen MacDonell supported study design and analysis and supported the writing. Lisa Todd provided supervision, contributed to study design, and reviewed and edited the writing.

Correspondence concerning this article should be addressed to Sylvie Naar, Center for Translational Behavioral Science, Florida State University, 2010 Levy Avenue, Building B, Suite 266G, Tallahassee, FL 32310, United States. Email: sylvie.naar@med.fsu.edu

However, the gold standard MI competency assessment (MI treatment integrity) requires highly trained professional raters, typically external to the implementation setting, to code recordings of audiorecorded sessions (Moyers et al., 2016). External coding is typically very costly and may interfere with the timeliness of feedback to staff on the quality of their MI implementation. Furthermore, these traditional MI competency measures were developed using mostly Caucasian samples in the context of substance abuse treatment (Moyers et al., 2016) and thus may not be appropriate for behavioral medicine interventions with minority groups. Developing competency measures with diverse samples that cannot only be used for measuring implementation outcomes, but can also be used by supervisors to provide rapid, accurate feedback, have a high likelihood of being sustained to support ongoing implementation. Moreover, as intervention researchers are pushed to design for dissemination (Brownson et al., 2013); it is critical to develop efficient and effective competency measures that target diverse populations early in the translational process (Phase 1 studies; Naar et al., 2018).

The present study used measurement development methods based in item response theory (Wolfe & Smith, 2007) to produce the Motivational Interviewing Coach Rating Scale (MI-CRS), an instrument for measuring the competency of provider MI delivery. The MI-CRS was developed based on coach ratings of providers' audiorecorded sessions or standardized patient interactions. These ratings were subsequently used not only as an implementation outcome measure but also to deliver feedback about performance and to focus the activities of the coaching sessions. Thus, the aims of the study were to (a) systematically develop the MI-CRS and compete an initial evaluation of its psychometric performance in a racial/ethnic minority sample; (b) assess the psychometric performance of the revised MI-CRS in two ethnically diverse independent samples (implementation study and community sample); and (c) use an objective standard-setting procedure to define criterion scores for MI competence to facilitate its use to provide feedback to providers. The goal is to present the design of a multistep study of competency assessment, to sample MI practice in both MI research samples and community learner samples and present a readable summary of a technically rigorous process that applied many best practices to develop the measure. See online supplemental material for more details.

Method

Study 1: Original Development in Research Setting—Fit Families Study

Overview of Parent Study Methods

Fit Families was a sequential multiple-assignment randomized trial (Naar, Ellis, et al., 2019) in which 181 African American adolescents (67% female), ages 12–17, with primary obesity were first randomized to office-based versus home-based behavioral skills treatment delivered from an MI foundation. All 30–60 min intervention sessions were delivered by community health workers (CHWs) trained in MI (2-day workshop followed by weekly individual or group supervision by a member of the Motivational Interviewing Network of Trainers). The study was approved by the Internal Review Board of Wayne State University.

Measurement Development Methods

The complete measurement development process is detailed in Appendix A in the online supplemental material. Briefly, the process involved three MI content experts and a measurement development expert, with the specific methods guided by the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999) and associated guidelines based in item response theory (IRT; Stone, 2003; Wilson, 2005; Wolfe & Smith, 2007). The process included defining the purpose of the instrument and intended use of the scores, defining key requirements and features of the observational coding system, defining components of MI corresponding to three levels of CHW expertise (i.e., novice, intermediate, expert), developing a coding manual with behaviorally defined rating scale categories for each item, training coders on the resulting instrument, and pilot testing the MI-CRS.

Coding Plan

The first 36 families enrolled in the clinical trial received treatment from seven CHWs, with a total of 52 sessions recorded and assigned. There were two coders, and of the assigned sessions, 47 (90.4%) were rated by both, which resulted in 99 rated sessions. The data were structured with sessions (Level 1) nested within families (Level 2) who were nested within CHWs (Level 3). Each session focused on one of three modules selected from the first 3-month period in the experimental obesity intervention.

Study 2: Revised Measure in an Implementation Trial—Adolescent Trials Network for HIV/AIDS Protocol 128 (ATN 128)

Overview of Parent Study Methods

ATN 128 was an implementation trial (Fortenberry et al., 2017) that partnered ATN study sites with Health Resources and Services Administration Ryan White sites to promote cross-agency collaboration and linkage to care. CHWs at each site were trained to connect with youth when testing positive for HIV and to support youth to link to HIV care and be retained in care. Longitudinal data were collected from 19 CHWs at 16 adolescent and young adult HIV health service organizations. CHWs were assigned to either the control group or an MI implementation group. CHWs in the MI group completed 10–20 min standard patient interactions by phone, which were audiorecorded. These recordings were coded with the MI-CRS (one preimplementation and 15 postimplementation per CHW over 2 years). Following a three-day workshop delivered by a member of the Motivational Interviewing Network of Trainers, CHWs received coaching sessions that begin with standard patient interactions coded in real time, followed by coaching in which the CHWs received immediate feedback and guided MI practice based on their coded scores. Control CHWs completed one preimplementation standard patient interaction, followed by monthly standard patient interactions for six months resulting in a total of seven MI-CRS scores. The study was approved by Wayne State University's Internal Review Board.

Measurement Revision

Results from the Fit Families sample led to the revision of two items and the addition of two new items, and this revised 12-item instrument was used in the ATN 128 sample. Details of the revisions are provided in the Results.

Coding Plan

There were four coders, with varied backgrounds from MI counselors to members of the Motivational Interviewing Network of Trainers, who coded 235 standard patient interactions of youth living with HIV and minority high risk youth. Of these, 117 (50.0%) were rated by two or more coders. resulting in 370 rated interactions. The 16 sites had only one or two CHWs, and the CHWs completed a moderate number of interactions (M = 11.8, SD = 3.8, Min. = 7.0, Max. = 16.0). Technically, the data were structured with repeated standardized patient interactions (Level 1) that were nested within CHWs (Level 2) who were nested within sites (Level 3). However, with CHWS and sites being largely singular, the data structure reduced to two levels, with interactions (Level 1) nested within CHWs/sites (Level 2).

Study 3: Community Implementation

Concurrent with the study above, three agencies received MI training and follow-up coaching that included 10- to 20-min standard patient interactions, real time coding using the MI-CRS, feedback, and practice. The study was approved as exempt by the Internal Review Board at Wayne State University.

Coding Plan

This sample was comprised of three sites (an urban children's hospital, HIV services in Jamaica, and a multidisciplinary federally qualified health center) each with a relatively large number of providers of varied backgrounds serving a diverse population (e.g., physicians, nurses, CHWs, mental health providers). However, these CHWs completed a smaller number of interactions than in previous studies (M = 2.5, SD = 1.7, Min. = 1.0, Max. = 7.0), with 172 total interactions that were rated by four coders. Of note, no interactions were double coded as this was part of community implementation efforts rather than research. Thus, the data were structured with repeated standardized patient interactions (Level 1) that were nested within providers (Level 2) who were nested within sites (Level 3). However, with the small number of sites, the data were reduced to a two-level structure with interactions (Level 1) nested within providers (Level 2).

Data Analysis Strategy

The study aims were addressed using a series of Rasch-based measurement models (Rasch, 1960; Bond & Fox, 2015; Wright & Mok, 2000). The Rasch model is a probabilistic measurement model and a special case of a single parameter model based in IRT. For the present study, the standard Rasch model was extended to accommodate other important features of the data, including the 4-point ordered categorical rating scale (Wright & Masters, 1982); ratings from multiple coders (Myford & Wolfe, 2003); and facets beyond items and CHWs (e.g., sessions, clients; Linacre, 1994a). The model results provide separate measures (i.e., "scores") for CHWs and items (and other facets), along with standard errors (SEs), fit indices, reliability estimates, and other indicators of psychometric performance. A defining feature of the model is that each item is assumed to be equally discriminating (i.e., to have a constant "slope"). Relative to traditional IRT models, this distinction is important—the Rasch model was one of several options, and the alternative models could potentially have better model fit. However, described next are the key features of the research aims that required use of the Rasch model, each of which is directly related to the challenges of measuring intervention competence in real-world evidence-based practice implementation efforts. The limitations of this approach are described in the Discussion.

Model Selection

The most consequential practical consideration was the number of observations required for modeling. For accurate estimation, IRT models, such as the 2PL or graded response model, require large samples of independent (i.e., nonnested) observations (e.g., >250; Embretson & Reise, 2000). The Rasch model, as a consequence of its restricted parameterization, has much more lenient requirements for acceptably precise estimates. Typically, 30–50 observations would be required for 95–99% confidence that SE estimates are stable within a 1 logit range (Linacre, 1994b; Wright & Stone, 1979). For many measurement efforts within real-world implementation contexts, large samples are simply not viable. This is further complicated by the nesting of data that is inherent to fidelity measurement—such as sessions within clients within CHWs within agencies—which reduces the effective sample size and introduces other challenges for psychometric evaluation. Another consideration is that the competency ratings are made by trained observational coders, which introduces the possibility of rater effects. Critically, the Many-Facet Rasch Model (MFRM; Linacre, 1994a) is specifically intended for rater data. Finally, the restricted form of the Rasch model confers practical benefits for measurement development, evaluation, and revision within real-world implementation efforts. The concept of a single, rather than variable, level of item difficulty has practical advantages for collaborating with stakeholders and community-based content experts to define item content to evaluate a broad range of practitioner fidelity. Likewise, for psychometric evaluation, items that do not fit the more restrictive model can be identified for revision; thus, an item that may be more discriminating in a traditional IRT model may not fit the Rasch model and can be revised accordingly. This is exemplified in the results of Study 1. Acknowledging its restrictions, the authors contend that the Rasch model provides pragmatic benefits for fidelity measurement in implementation efforts.

Measurement Models

To address the research aims, three types of models were performed: (a) Standard Rasch rating scale models were implemented in WINSTEPS software (Linacre, 2018b) with "facets" for items and sessions/interactions. This model used all data but did not specifically account for client, CHW, or coder. (b) MFRMs were implemented in FACETS software (Linacre, 2018a); with facets for sessions/interactions, clients, providers, coders, and items. However, because the data were nested, rather than cross-classified as is typical for MFRMs, not all results are reported. (c) To address the nested data structures directly, multilevel formulations of the Rasch measurement model were used, implemented as hierarchical generalized linear measurement models (HGLMMs; Beretvas & Kamata, 2005; Kamata, 2001; Adams et al., 1997). HGLMMs were performed in HLM software (Raudenbush et al., 2013). With this approach, a measurement model is added at the lowest level of the data structures previously described. Specifically, for Fit Families, item responses (Level 1) were nested within sessions (Level 2) that were nested within families (Level 3); for ATN-128, item responses (Level 1) were nested within interactions (Level 2) that were nested within CHWs/sites (Level 3); and for Community Implementation, item responses (Level 1) were nested within interactions (Level 2) that were nested within providers/sites (Level 3). In each case, dummy-coded indicators were used to differentiate the items and coders. The 4-point MI competence item ratings (i.e., poor, fair, good, excellent) were analyzed according to an ordinal outcome distribution with a logit link function. The resulting logit-based item estimates conform to Rasch item "difficulty" estimates, and the empirical Bayes residuals for sessions, families, or providers conform to Rasch "ability" estimates. For the series of measurement models just described, psychometric performance of the MI-CRS was evaluated across multiple indicators: dimensionality and local independence, rating scale functioning, item fit, reliability and separation, coder reliability, and item invariance. Each indicator is listed in Table 1 and described in detail in Appendix B in the online supplemental material.

Defining MI Competence Thresholds

The MI-CRS provides continuous scores; however, for implementation purposes and to be consistent with other measures in the literature (Moyers et al., 2016) we defined thresholds using a Rasch-based objective standard setting procedure (e.g., Stone, 2001), applied to a combined dataset from the ATN-128 and community implementation samples. Generally, this approach combines (a) item-by-item ratings from content experts, along with an overall rating, with (b) item parameter estimates from the Rasch measurement model. From this information, empirically based thresholds are defined. Variants of this approach are routinely used in educational applications to establish, for example, pass/fail or certification thresholds. In the present case, unique features included the 4-point rating scale structure for the MI-CRS components and the need for final threshold values to be based on raw scores (rather than logits). Because of the statistical focus, the details of each step are provided in the Results.

Results

Sample 1: Fit Families

Dimensionality and Local Dependence

A fundamental assumption of IRT-based measurement models is that the data are reasonably unidimensional. This was evaluated based on a principal component analysis of standardized Rasch item-person residuals. Specifically, after extracting the primary Rasch dimension, the analysis attempts to identify meaningful structure in the residual matrix. The results are evaluated based on theoretical considerations (Bond & Fox, 2015); the proportion of variance explained by the primary Rasch dimension (which ideally exceeds 60%) and by the magnitude of the eigenvalue for the first contrast in the residuals (which ideally is below 2.0; Linacre, 2018b). Of the total variance, 64.6% was explained by the Rasch item and person measures, and the eigenvalue for the first contrast was 1.7. However, for these results, each session was treated as an independent observation. To control for the nested data structure, a single session was randomly selected for each family. The dimensionality results indicated that 62.1% of the variance was explained, and the eigenvalue for the first contrast was 2.2. Although the eigenvalue exceeded the target of 2.0, the component most suggestive of dimensionality, "The counselor keeps the focus on the goal of the session," was misfitting (as detailed subsequently). When removed, the disattenuated correlations between the suspected dimensions were high, ranging from .86 to 1.00, indicating that dimensionality was trivial.

A related assumption is that the items comprising the MI-CRS are locally independent, which means that, after removing the primary Rasch dimension, there are no pairs of items with strongly related content. This was evaluated based on the magnitude of positive correlations for the residuals of each item pair (Linacre,

Table 1
Indicators of Psychometric Performance for the MI-CRS

Domain	Indicator	Guideline[a]
Dimensionality	Variance explained	>50%
	Eigenvalue of first contrasts	<2.0
Local dependence	Raw score residual correlations	≤.32
	Common variance	≤10%
Rating scale functioning	Percentage of ratings in each category	≥10%
	Category threshold spacing (logits)	≥1.4
	Category fit statistics	≤1.5
	Maximum category probability	≥60%
Item fit	Outlier-sensitive mean-square fit statistics	≤1.5
Reliability	Rasch reliability	≥.70
Separation	Person Separation Index	≥2.0
	Strata	≥3.0
Coder reliability	Absolute agreement	≥70%
	Coder facet reliability	≤.50
	Coder separation	≤2.0
	Krirppendorff's α	≥.60
Item invariance	Items within 95% confidence bounds	100%

Note. A detailed description of each indicator is provided in Appendix B in the online supplemental material.
[a] Each guideline reflects the general rule-of-thumb considered in the present evaluation but may not represent a suitable criterion for psychometric performance across all measurement contexts.

2018b). The results indicated that the average correlation was -.11, and the maximum value for any item pair was .22, which reflects only 5% shared variance between two items. With a single session for each family, the average was again -.11, with a maximum value of .35, reflecting 12% shared variance. This provides evidence that there was no meaningful local dependence for the MI-CRS items, and with no strong evidence of meaningful dimensionality, the analyses that follow are based on simultaneous analysis of the original 10 MI-CRS items.

Rating Scale Functioning

Based on the full sample of data, the rating scale categories of poor, fair, good, and excellent received 9%, 28%, 36%, and 27% of the ratings, respectively, indicating that each category was well-used by coders (i.e., a nontrivial percentage of responses in each category). Further, these ratings were not characterized by unpredictability, with outfit mean square fit statistics of .93, 1.03, 1.08, and .87 that fell below the threshold of 1.50. As used by the coders, the rating scale categories were ordered as intended (i.e., with higher ratings reflecting sessions with higher competence) and well-differentiated from the adjacent categories (i.e., with each category as the most probable response for a distinct segment of the competence continuum). This is reflected by the Andrich threshold values of -2.75, .16, and 2.59, which are the transition points between categories (i.e., the point where two neighboring categories have an equal probability of being endorsed). Combined, the results indicate that coders interpreted and used the rating scale as intended. The results were highly consistent for the model with one session per family and the MFRM.

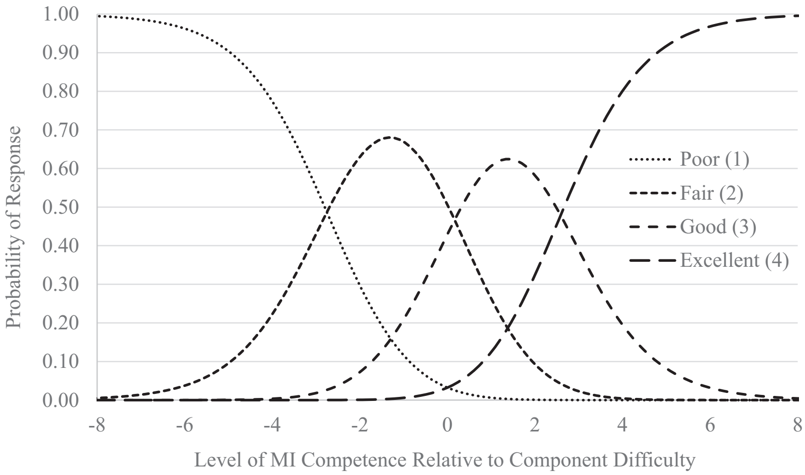
The performance of the rating scale is illustrated in Figure 1 When a rating scale functions as intended, each category—in this case, poor, fair, good, excellent—will form a distinct "hill." Such a pattern demonstrates that each category was not only the most probable rating for a segment of the underlying construct, it was also distinct from the adjacent categories. For instance, on the *x*-axis, the left side represents the combination of (a) a session with low overall competence and (b) an MI component that is advanced. For this combination, there was a high probability (*y*-axis) that the resulting rating would be poor. As the level of competence increased and the difficult of the MI component decreased, a rating of poor became less likely and a rating of fair became more likely. A rating of fair was then the most probable response for a range of the construct. As competence continued to increase, and the difficulty of the component continued to decrease, the probability of fair decreased, the probability of good increased, and so forth. Thus, the results and figure indicate that the rating scale performed as intended, and as such, no adjustments were necessary.

Item Fit

The item outfit mean square values are reported in Table 2. One component was significantly misfitting relative to the target cutoff of 1.50, with an outfit value of 1.62: "The counselor keeps the focus on the goal of the session." The misfit likely reflects content that is distinct from the overall MI competence construct. For instance, there could be a strong focus on the goal of the session in sessions that, overall, have either high or low levels of MI competence. This conclusion was consistent for the model based on one session per family and the MFRM. Because this component is

Figure 1

Rating Scale Category Probability Curves From the Rasch Measurement Model Based on the Fit Families Sample



Note. For each category, the curve reflects the probability of endorsement (y-axis) at each level of the motivational interviewing (MI) competency construct (x-axis). The x-axis reflects the difference between the level of competence for a session and the difficulty of the competence component being rated, from left to right, with more basic MI components and an increasing level of MI competence. If a session has a high level of competence and a basic component of MI is being rated, the probability of a rating in the highest category approaches 100%. Conversely, if a session has low competence and an advanced component of MI is being rated, the probability of the rating in the lowest category approaches 100%.

Table 2
Rasch Outfit M Square Item Fit Statistics by Sample

	Outfit mean square		
Variable	Sample 1[a]	Sample 2[b]	Sample 3[c]
1. Fosters collaboration	0.80	1.01	0.71
2. Supports autonomy	0.81	1.08	0.98
3. Evokes ideas	0.94	0.95	0.92
4. Keeps focus on goal	1.62	1.01	1.12
5. Uses reflective listening	0.88	0.91	0.96
6. Reinforces strengths	1.02	1.40	1.29
7. Uses summaries	0.93	1.43	1.45
8. Asks open-ended questions	1.24	0.68	0.86
9. Solicits feedback	0.68	0.91	0.76
10. Manages discord	0.97	1.08	1.61
11. Cultivates empathy		0.68	0.79
12. Uses reflections strategically		0.94	0.80

Note. Values > 1.50 reflect items that potentially degrade measurement.
[a]Fit Families.
[b]Implementation study.
[c]Community sample.

essential, it was retained and revised for the next version of the measure to be more consistent with an MI approach to collaborative agenda setting.

Targeting of Items to the Sample

A key summary of the instrument's performance is provided by the item-session map (Bond & Fox, 2015), which is illustrated in Figure 2 On the left is the distribution of sessions, with the highest levels of MI competence at the top and the lowest levels at the bottom. On the right is the distribution of items, with the least commonly occurring (i.e., most "difficult") at the top and the most commonly occurring (i.e., "easiest") at the bottom. For each item, there are three positions that reflect different regions of the rating scale. The left position is the item difficulty for a 50% chance of being rated in the lowest category (i.e., poor), the middle position reflects the average item difficulty, and the right position reflects the item difficulty for a 50% chance of being rated in the highest category (i.e., excellent). For each distribution, the sample mean and standard deviation are provided. For a well-performing instrument, there are several expectations: (a) The session and item distributions cover a wide range of approximately four or more logits, (b) the session mean and item mean are closely aligned, (c) there are no meaningful "gaps" in the item distribution (i.e., a > .5 logit range of the session distribution that is not assessed by any items; Linacre, 2018b), (d) there are items targeted to the full distribution of sessions, and (e) the ordering of items, from least to most difficult, should match theoretical expectations. The MI-CRS items met each of these conditions, which indicates that the items were well-targeted to the Fit Families sample.

Reliability and Separation

Rasch session reliability was high, .89, with a Cronbach's alpha-equivalent "test" reliability estimate of .93. Session separation reliability was 2.89, indicating that the sample of MI components was sufficient for providing at least two meaningful distinctions in the continuum of CHW MI competence. This translates into 4.19 strata, or four distinct levels of competence. To address the nested data structure, the HGLMM was performed using ratings from one randomly selected coder for each interaction. This was necessary due to software limitations. The results indicated that the reliability of session-level competence scores was .83, and the reliability of family-level scores was .54. When also controlled for differences across CHWs, the reliability of family-level scores decreased to .10, which indicates that the majority of variation in family-level scores was attributable to differences between CHWs. Of the total variance in MI competence ratings, 40%, 28%, and 31% was attributable to items, sessions, and families/CHWs, respectively.

Coder Reliability

The MFRM was used to compute the percentage of absolute agreements for sessions with multiple coders (i.e., two coders of the same item, session, client, and CHW). The rate of exact agreement was 40.8%, though when considering agreement in adjacent categories, this increased to 88%. The coder outfit statistics indicated that coders did not, overall, provide unpredictable ratings. Separation reliability was high at .85, as was the separation index at 2.4. However, for the coder facet, these values would ideally be low, as the large values indicate that coders can be reliably differentiated based on their ratings. Krippendorff's alpha for ordinal data (with 10,000 bootstrap samples) was also computed as a supplementary coder reliability estimate (Hayes & Krippendorff, 2007). The result of .46 indicated that coder reliability was relatively low.

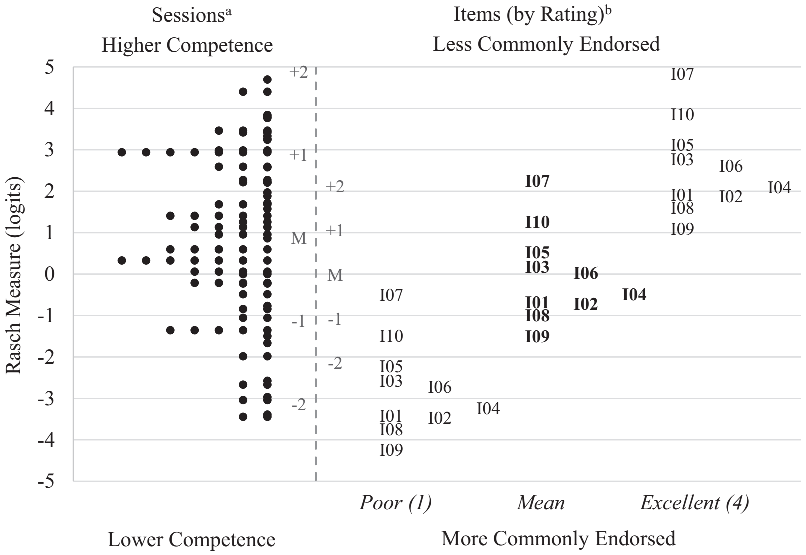
Samples 2 and 3: ATN-128 and Community Implementation

Instrument Revisions

As noted above, the misfitting item was revised to, "The counselor balances the client's agenda with focusing on the target behaviors." In addition, the resistance item was revised to reflect changes based on the newest conceptualizations of MI (Miller & Rollnick, 2013), "The counselor manages counter change talk and discord," and the rating protocol was revised for the item to be rated for each interaction instead of only with counter change talk (language against change) and discord (negative statements about treatment) are present. If there was truly no counter change talk or discord in the session, the score was a 4. Finally, two new items

Figure 2

Item-Session Map From the Rasch Measurement Model Based on the Fit Families Sample



Note. The item-session map illustrates the distribution of session measures (i.e., "scores") relative to the distribution of item measures. On the left side of the center dividing line are markers for a mean session (*M*) and Sessions 1 and 2 *SD*s above and below the mean. On the right side of this line are markers for the mean item difficulty location and Items 1 and 2 *SD*s above and below the mean. This leads to several criteria for evaluating performance: (a) The session and items distributions cover a range of approximately four or more logits, (b) the session mean and item mean are closely aligned, (c) there are no significant vertical "gaps" in the item distribution (i.e., a > .50 logit range of the session distribution that is not assessed by any items at the same location), (d) there are items targeted to the full distribution of sessions, and (e) the ordering of items, from least to most difficult, should match theoretical expectations.

[a] The session panel illustrates the distribution of session measures, with each circle representing the Rasch logit-based "score" for a single coded session. Sessions with the highest observed competence are at the top, and those with the lowest levels are at the bottom.

[b] The item panel illustrates the distribution of item measures, with less commonly endorsed items at the top and more commonly endorsed items at the bottom. The bolded center column reflects the location of the Rasch item difficulty estimate, the poor column is the difficulty estimate for a 50% probability of a rating in the lowest category, and the excellent column is the estimate for a 50% probability of a rating in the highest category.

were developed to fill perceived gaps in content: "The counselor cultivates empathy and compassion with client(s)" and "The counselor uses reflections strategically."

Dimensionality and Local Dependence

For the ATN-128 sample, the principal components analysis of Rasch item-person residuals indicated that there was likely not meaningful dimensionality, with 50.8% of the variance explained by the item and person measures and an eigenvalue of 1.6 for the first contrast. The same was true for the community implementation sample, with 53.8% of the variance explained and an eigenvalue of 1.7. For local dependence in the ATN-128 and community implementation samples, the maximum correlation between item pairs was .07 and .18, respectively, with an average of -.09 in both samples. As such, the results that follow are based on simultaneous analysis of all MI items.

Rating Scale Functioning

For the ATN-128 sample, the rating scale categories were well-used, with poor, fair, good, and excellent receiving 5%, 33%, 49%, and 13%, respectively, of the ratings and associated fit values of 1.05, 1.02, 1.02, and .95, respectively. The Andrich thresholds indicated that each category was well-differentiated, and the observed average person measures supported the intended ordering of the rating categories. For the community implementation sample, rating scale performance was highly consistent, with poor, fair, good, and excellent receiving 4%, 25%, 48%, and 23%, respectively, and associated fit values of .89, .91, 1.10, and 1.09, respectively.

Item Fit

For the ATN-128 sample, no items exceeded the outfit mean square threshold of 1.50. For the community implementation sample, one item was misfitting, "The counselor manages counter change talk and discord," with a value of 1.60. In the MFRM, no items were significantly misfitting.

Targeting of Items to the Sample

For both samples, the interaction and item distributions cover a wide range, the items and rating scale categories target the full distribution of interactions, and the ordering of items matches theoretical expectations. There are two potential concerns. First, in community implementation, the mean of the interaction distribution is somewhat higher than the mean for the items, indicating that the MI competence items were more likely to be rated highly in the sample of standardized patient interactions. Despite this, with the four-point rating scale, the full range of interactions is adequately assessed by the items. Second, for both samples, there is an apparent "gap" in the distribution of items; that is, a portion of the interaction distribution is not well-targeted by any items. For ATN-128, the gap is between the two least commonly endorsed items ("The counselor uses reflections strategically" and "The counselor uses summaries effectively"), whereas for Community Implementation, it is between the most commonly endorsed items ("The counselor demonstrates reflective listening skills" and "Counselor manages counter change talk and discord"). This indicates that in the respective samples, there could be gaps in the assessment of higher and lower, respectively, levels of competence. Although these concerns could suggest the need for additional items and/or coder training, the well-functioning rating scale largely removes this concern by affording strong assessment of nearly the entire range of CHWs.

Reliability and Separation

Rasch reliability and separation reliability for the interactions were high for both samples. For ATN-128, reliability and separation were .88 and 2.70, and the Cronbach's alpha-equivalent test reliability was .90. For Community Implementation, reliability and separation were .89 and 2.91, with a Cronbach's alpha-equivalent reliability of .91. To address the nested data structure, the HGLMM for ATN-128 was performed using ratings from one randomly selected coder for each interaction. Controlling for differences across coders, the results indicated that the reliability of interaction-level competence scores was .80 and the reliability of provider interaction scores was .87. Of the total variance in ratings, 56%, 25%, and 19% was attributable to items, interactions, and CHWs/sites, respectively. For community implementation, and also controlling for coders, the reliability of interaction-level competence scores was .75, and the reliability of provider interaction scores was .53. Of the total variance in ratings, 64%, 21%, and 16% was attributable to items, interactions, and providers/sites, respectively.

Coder Reliability

For the ATN-128 sample, the rate of exact agreement among coders was 44.6%, and for agreement in adjacent categories, this increased to 96%. The reliability of the coder facet was .88, with separation of 2.7, which indicates that coders could be differentiated based on their ratings. Krippendorff's alpha was low at .34.

Item Invariance

As noted previously, because of the nested data structures, HGLMMs were used to provide the item parameter estimates for evaluating item invariance. The estimate for each component was computed as described by Kamata (2001). Specifically, for each indicator, the estimated coefficient was added to the reference item (i.e., intercept), providing the logit-based item difficulty score. From Fit Families, these estimates were retained. For ATN-128 and Community Implementation, following the methods detailed by Bond and Fox (2015); the estimates were rescaled to be in the same frame of reference as Fit Families (i.e., by adding the difference in the two sample means). For each component, the estimated *SE* was used to compute 95% confidence intervals. With two exceptions, item estimates from the two independent samples—which included different CHWs, different interventions, different behavioral health problems, and a portion of different raters—were consistent within the 95% confidence regions. The exceptions, in each case, were Item 9 ("The counselor solicits feedback from clients") and Item 10 ("The counselor manages counter change talk and discord"). Specifically, Item 9 was rated more highly (yielding a larger negative item score) in Fit Families and Item 10 was rated more highly in ATN-128.

Threshold Definition

Using the continuous MI-CRS scores, the original aim was to differentiate three levels of competency consistent with other measures (Moyers et al., 2016). The content experts were 15 members of the Motivational Interviewing Network of Trainers. For each of the 12 items, each expert independently used the instrument's 4-point rating scale (ranging from low to high competence) and selected the minimum category that reflected Beginner competence (i.e., the expert would have considered the next lowest category to reflect Below competence). "Solid" competence was then defined as the next highest rating scale category. Across experts, the selected category for each component was combined with the results of a MFRM (Linacre, 1994a), specifically, the logit-based item parameter estimate and corresponding rating scale threshold. For each expert and item, this resulted in an item "difficulty" score, and the scores were then averaged across items. This was done separately for the beginner and solid competency thresholds. The resulting values were then adjusted for the experts' ratings of overall competency, ranging from 0% to 100%, that the experts defined as required for "somewhat acceptable" and "acceptable" competence. The resulting values were then averaged across experts, providing logit-based threshold values for beginner and solid competence, and these scores were converted to raw scores (using raw score conversion information from the MFRM) corresponding to the instrument's 4-point scale. The raw score thresholds for beginner and solid competency were 2.0 and 3.3, respectively. To evaluate validity evidence for the thresholds, they were applied to existing data, and the team reviewed the resulting percentages of observations in each category. This identified a large proportion of ratings in the beginner category, and based on (a) expert review and (b) the wide range from beginner to solid, the beginner category was divided into two parts, differentiating novice and intermediate levels of competence. Thus, the final categories and associated threshold scores were: beginner (<2.0), novice (2.0–2.6), intermediate (2.7-3.3), and advanced (>3.3).

Discussion

Efficient and effective competency measurement is critical to ensure fidelity in clinical trials, to assess implementation outcomes, and for establishing sustainable fidelity monitoring and feedback systems in real-world settings. The MI-CRS was developed for MI supervisors/coaches to listen to real or standard patient interactions and immediately rate 12 items in a "one pass" coding system. This reduces the cost of traditional coding systems (Moyers et al., 2016), allows for immediate feedback to implementers, and provides an efficient outcome assessment. This measure fills an important gap in MI competency literature as it was developed in an ethnically diverse sample and with adolescents and families. Also, MI competency measures have not typically been developed with IRT methods and evaluated for dimensionality, rating scale functioning, item fit, as well as reliability and separation. Likewise, MI competency measures have not typically used a formal IRT-based standard setting procedure to define empirically based competency thresholds.

The strong development approach and resulting psychometric properties suggest that the MI-CRS meets the call for efficient and effective fidelity measurement. The measure demonstrated strong psychometric properties. First, dimensionality results indicated that the MI-CRS appeared to measure a single underlying construct of MI competence as compared to other conceptions of MI skill as having at least two dimensions (Magill et al., 2018). Second, item-session maps were indicative of a well-performing instrument. Third, variance was primarily due to counselor versus client(s) or sessions, which is beneficial when rating provider competence as an implementation outcome or for quality assurance and feedback loops.

Fourth, rating scale functioning is often not formally evaluated in fidelity measurement research, but when it is, the rating scales often do not perform as intended and must be remedied (Bond & Fox, 2015). The 4-point rating scale showed excellent functionality in that each rating scale category was meaningfully distinct and discriminated in a consistent way by the coders. This is important from the perspective of both efficiency and precision. For each of the 12 items, the properly functioning four-point scale provides information equivalent to three dichotomous items. This, in turn, makes it possible for the relatively modest number of items to assess a wide range of competency, from beginner to advanced levels.

Item fit was generally good after revisions, with one item, managing counter change talk/discord possibly requiring further revisions. In the first version of the measure, coders were able to mark “Not Applicable” if there was no counter change talk or discord present. In the second version of the measure, coders were forced to submit a response, and were instructed to score a 4 if there was no counter change talk or discord present. The rationale was that a provider must be skilled if they avoided counter change talk or discord. However, the lack of counter change talk could be due to highly motivated clients rather than due to provider skill, and thus a 4 for managing counter change talk well could be very different than a 4 for no counter change talk present. In future iterations of the measure, coders could first indicate if counter change talk is present or not, and then code the item to avoid this confusion. One additional item, “Uses summaries effectively,” exhibited borderline levels of misfit in the ATN-128 and community implementation samples, suggesting the possible need for revisions to the coding manual or item content. Current items and descriptions are in Appendix C in the online supplemental material.

Despite multiple indicators of strong psychometric performance, a key concern was the modest level of absolute agreement among coders in the Fit Families and ATN-128 samples. Related to this, there was evidence that coders could be differentiated on the basis of their pattern of ratings. Generally, this reflected overall differences in leniency/severity in assigning ratings. As previously noted, the ideal scenario for fidelity measurement is that coders provide identical ratings across items for each interaction. Thus, with modest coder agreement and potentially distinct coding styles, there are several factors to consider. For instance, the rates of agreement identify a need for improved coder training and ongoing supervision, which should include greater attention to the nature of disagreements across coders. Likewise, it is important to consider that absolute agreement—particularly for the MI-CRS—is a particularly stringent criterion. Not only are ratings on a 4-point scale, all 12 of the components being rated are applicable in all sessions and they may occur dynamically throughout the interaction. As such, there is ample opportunity for disagreement. This stands in contrast to ratings of treatment adherence, where only a small subset of components may be delivered in a session, resulting in “automatic” agreement on components not delivered. Further, and perhaps most importantly, the rates of agreement emphasize the importance of using measurement models, such as the MFRM, that are specifically intended for rater data. Typically, a small subset of interrater cases is assigned, and following evaluation of reliability, the primary assignment is retained for scoring and analysis. However, with the challenges inherent to rater agreement, it is possible to leverage IRT-based measurement models designed for such data. In contrast to traditional raw scoring approaches—which ignore the influence of individual coders—these models use all available ratings and produce a best-estimated “fair” score for each session that is adjusted for each coder’s rating style. Information about the rating style is also used to adjust scores for sessions rated by single coders.

Another possible limitation was the use of the Rasch measurement model rather than the IRT graded response model. As detailed earlier, the main consideration was practical—the number of observations, particularly after considering the other data features, was not sufficient for the more complex model. Despite this, the more lenient sample size requirements of the Rasch model permitted thorough evaluations of the MI-CRS across three distinct samples. Further, the MFRM was particularly well-suited to rater data that are inherent to measuring CHW competence. That said, the main drawbacks are that, with a sufficient sample size, it would be possible to determine whether there were meaningful differences in item discrimination and the more complex model would most likely provide better model fit. At the same time, the MI-CRS was initially developed from a Rasch-based framework, and the subsequent revisions have been made with respect to the fit of the data to the Rasch model. Acknowledging the limitations of the simpler model, for new measurement development and evaluation in the context of real-world implementation efforts, the authors see some distinct benefits—practical and philosophical in nature—to the more constrained framework. Ultimately, a critical consideration is the impact on the resulting scores, and in this case, scores based on the Rasch model, GRM, and traditional raw scoring methods were nearly perfectly correlated (i.e., $\geq .99$).

Of note, coded standard patient sessions were more likely to yield higher ratings than coded real interactions. suggesting that the quality of MI was better with standard patients. For the purpose of evaluating competence in diverse, real-world implementation efforts, we believe that the use of standardized patient interactions—in addition to solving a pragmatic challenge—demonstrates the extent to which CHWs are capable of delivering key components of MI. In contrast, competence measurements from actual patient interactions demonstrate whether CHWs actually do deliver MI components. These are related, but different, aspects of fidelity measurement.

Finally, the IRT-based standard setting procedure defined three empirically-based thresholds to differentiate four levels of competence, in contrast to two thresholds and three levels used by other MI fidelity measures (Moyers et al., 2016). The resulting competency levels were named to be more affirming and supportive based on input from stakeholders in another community implementation study using the measure (Aarons et al., 2017): beginner, novice, intermediate, and advanced. Applied to competence scores from these samples, the percentage of scores falling in each category was consistent with expectations and anecdotal experience.

In summary, a 12-item measure of MI competency designed to be rated in one pass of real or simulated encounters targeting health behaviors showed excellent psychometric properties in diverse settings and samples. Head-to-head comparison of the MI-CRS to established measures of MI competency will confirm the relative advantage of the measure beyond efficiency. Future research is necessary to determine the sensitivity of the measure to implementations interventions designed to improve competence (Fortenberry et al., 2017), to test the use of the measure for trigger-based coaching (e.g., coaching triggered for scores below intermediate; Naar, MacDonell, et al., 2019), to compare real provider–patient interactions with standard patient interactions, to determine the properties of the instrument when coded by research assistants versus MI supervisors, and to test the measure in additional samples for generalizability of findings to other implementation settings.

References

Aarons, G. A., Ehrhart, M. G., Moullin, J. C., Torres, E. M., & Green, A. E. (2017). Testing the leadership and organizational change for implementation (LOCI) intervention in substance abuse treatment: A cluster randomized trial study protocol. *Implementation Science*, *12*(1), 29. https://doi.org/10.1186/s13012-017-0562-3

Aarons, G. A., Hurlburt, M., & Horwitz, S. M. (2011). Advancing a conceptual model of evidence-based practice implementation in public service sectors. *Administration and Policy in Mental Health*, *38*(1), 4–23. https://doi.org/10.1007/s10488-010-0327-7

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, *22*(1), 47–76. https://doi.org/10.3102/10769986022001047

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

Bauer, M. S., Damschroder, L., Hagedorn, H., Smith, J., & Kilbourne, A. M. (2015). An introduction to implementation science for the non-specialist. *BMC Psychology*, *3*(1), 32. https://doi.org/10.1186/s40359-015-0089-9

Beretvas, S. N., & Kamata, A. (2005). The multilevel measurement model: Introduction to the special issue. *Journal of Applied Measurement*, *6*(3), 247–254.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental measurement in the human sciences* (3rd ed.). Erlbaum. https://doi.org/10.4324/9781315814698

Brownson, R. C., Jacobs, J. A., Tabak, R. G., Hoehner, C. M., & Stamatakis, K. A. (2013). Designing for dissemination among public health researchers: Findings from a national survey in the United States. *American Journal of Public Health*, *103*(9), 1693–1699. https://doi.org/10.2105/AJPH.2012.301165

Cross, W. F., & West, J. C. (2011). Examining implementer fidelity: Conceptualizing and measuring adherence and competence. *Journal of Children's Services*, *6*(1), 18–33. https://doi.org/10.5042/jcs.2011.0123

DeMars, C. (2010). *Item response theory*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195377033.001.0001

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Erlbaum.

Engelhard, G. Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge. https://doi.org/10.4324/9780203073636

Fortenberry, J. D., Koenig, L. J., Kapogiannis, B. G., Jeffries, C. L., Ellen, J. M., & Wilson, C. M. (2017). Implementation of an integrated approach to the National HIV/AIDS strategy for improving Human Immunodeficiency Virus care for youths. *JAMA Pediatrics*, *171*(7), 687–693. https://doi.org/10.1001/jamapediatrics.2017.0454

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89. https://doi.org/10.1080/19312450709336664

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*(1), 79–93. https://doi.org/10.1111/j.1745-3984.2001.tb01117.x

Linacre, J. M. (1994a). *Many-facet Rasch measurement*. Mesa Press Chicago.

Linacre, J. M. (1994b). Sample size and item calibration stability. *Rasch Measurement Transactions*, *7*(4), 328.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*(1), 85–106.

Linacre, J. M. (2003). Size vs. Significance: Infit and outfit mean-square and standardized chi-square fit statistics. *Rasch Meausrement Transactions*, *17*(1), 918.

Linacre, J. M. (2018a). *FACETS. Rasch measurement computer program.* Winsteps.com.

Linacre, J. M. (2018b). *WINSTEPS. Rasch measurement computer program*. Winsteps.com.

Lundahl, B., Moleni, T., Burke, B. L., Butters, R., Tollefson, D., Butler, C., & Rollnick, S. (2013). Motivational interviewing in medical care settings: A systematic review and meta-analysis of randomized controlled trials. *Patient Education and Counseling*, *93*(2), 157–168. https://doi.org/10.1016/j.pec.2013.07.012

Madson, M. B., Villarosa, M. C., Schumacher, J. A., & Mohn, R. S. (2016). Evaluating the validity of the client evaluation of motivational interviewing scale in a brief motivational intervention for college student drinkers. *Journal of Substance Abuse Treatment*, *65*, 51–57. https://doi.org/10.1016/j.jsat.2016.02.001

Magill, M., Apodaca, T. R., Borsari, B., Gaume, J., Hoadley, A., Gordon, R. E. F., Tonigan, J. S., & Moyers, T. (2018). A meta-analysis of motivational interviewing process: Technical, relational, and conditional process models of change. *Journal of Consulting and Clinical Psychology*, *86*(2), 140–157. https://doi.org/10.1037/ccp0000250

Midboe, A. M., Cucciare, M. A., Trafton, J. A., Ketroser, N., & Chardos, J. F. (2011). Implementing motivational interviewing in primary care: The role of provider characteristics. *Translational Behavioral Medicine*, *1*(4), 588–594. https://doi.org/10.1007/s13142-011-0080-9

Miller, W., & Rollnick, S. (2013). *Applications of motivational interviewing. Motivational interviewing: Helping people change*. Guilford Press.

Moyers, T. B., Rowell, L. N., Manuel, J. K., Ernst, D., & Houck, J. M. (2016). The motivational interviewing treatment integrity code (MITI 4): Rationale, preliminary reliability and validity. *Journal of Substance Abuse Treatment*, *65*, 36–42. https://doi.org/10.1016/j.jsat.2016.01.001

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.

Naar, S., Czajkowski, S. M., & Spring, B. (2018). Innovative study designs and methods for optimizing and implementing behavioral interventions to improve health. *Health Psychology*, *37*(12), 1081–1091. https://doi.org/10.1037/hea0000657

Naar, S., MacDonell, K., Chapman, J. E., Todd, L., Gurung, S., Cain, D., Dilones, R. E., & Parsons, J. T. (2019). Testing a motivational interviewing implementation intervention in adolescent HIV clinics: Protocol for a type 3, hybrid implementation-effectiveness trial. *JMIR Research Protocols*, *8*(6), e11200. https://doi.org/10.2196/11200

Naar, S., Ellis, D., Idalski Carcone, A., Jacques-Tiura, A. J., Cunningham, P., Templin, T., Hartlieb, K. B., & Jen, K.-L. C. (2019). Outcomes from a sequential multiple assignment randomized trial of weight loss strategies for african american adolescents with obesity. *Annals of Behavioral Medicine*, *53*(10), 928–938. https://doi.org/10.1093/abm/kaz003

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. The Danish Institute of Educational Research.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2013). HLM 7: Hierarchical linear & nonlinear modeling (Version 7.01) [Computer software & manual]. Scientific Software International.

Schoenwald, S. K., Garland, A. F., Chapman, J. E., Frazier, S. L., Sheidow, A. J., & Southam-Gerow, M. A. (2011). Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health*, *38*(1), 32–43. https://doi.org/10.1007/s10488-010-0321-0

Schumacker, R. E., & Smith, E. V., Jr. (2007). Reliability: A Rasch perspective. *Educational and Psychological Measurement*, *67*(3), 394–409. https://doi.org/10.1177/0013164406294776

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195152968.001.0001

Smith, E. V., Jr. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, *3*(2), 205–231.

Stone, G. E. (2001). Objective standard setting (or truth in advertising). *Journal of Applied Measurement*, *2*(2), 187–201.

Stone, M. H. (2003). Substantive scale construction. *Journal of Applied Measurement*, *4*(3), 282–297.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Erlbaum.

Wolfe, E. W., & Smith, E. V., Jr. (2007). Instrument development tools and activities for measure validation using Rasch models: Part I - instrument development tools. *Journal of Applied Measurement*, *8*(1), 97–123.

Wright, B. D., & Masters, G. (1982). *Rating scale analysis*. MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. MESA Press.

Wright, B. D., & Mok, M. (2000). Rasch models overview. *Journal of Applied Measurement*, *1*(1), 83–106.

Received October 20, 2019
Revision received December 1, 2020
Accepted December 6, 2020 ■