Psychological Assessment

Estimation of Equable Scale Scores and Treatment Outcomes From Patient- and Clinician-Reported PTSD Measures Using Item Response Theory Calibration

Antonio A. Morgan-López, Lissette M. Saavedra, Denise A. Hien, Therese K. Killeen, Sudie E. Back, Lesia M. Ruglass, Skye Fitzpatrick, Teresa López-Castro, and Julie A. Patock-Peckham

Online First Publication, December 5, 2019. http://dx.doi.org/10.1037/pas0000789

CITATION

Morgan-López, A. A., Saavedra, L. M., Hien, D. A., Killeen, T. K., Back, S. E., Ruglass, L. M., Fitzpatrick, S., López-Castro, T., & Patock-Peckham, J. A. (2019, December 5). Estimation of Equable Scale Scores and Treatment Outcomes From Patient- and Clinician-Reported PTSD Measures Using Item Response Theory Calibration. *Psychological Assessment*. Advance online publication. http://dx.doi.org/10.1037/pas0000789

Psychological Assessment

© 2019 American Psychological Association
ISSN: 1040-3590

http://dx.doi.org/10.1037/pas0000789

Estimation of Equable Scale Scores and Treatment Outcomes From Patient- and Clinician-Reported PTSD Measures Using Item Response Theory Calibration

Antonio A. Morgan-López and Lissette M. Saavedra
RTI International, Research Triangle Park, North Carolina

Denise A. Hien
Rutgers, The State University of New Jersey

Therese K. Killeen
Medical University of South Carolina

Sudie E. Back
Medical University of South Carolina and Ralph H. Johnson Veterans Affairs Medical Center, Charleston, South Carolina

Lesia M. Ruglass
Rutgers, The State University of New Jersey

Skye Fitzpatrick
York University

Teresa López-Castro
The City College of New York

Julie A. Patock-Peckham
Arizona State University

Across multiple RCTs, discrepancies between patient and clinician reports of PTSD symptoms are at least a partial contributing factor to large discrepancies between treatment outcome effect sizes from self-report and clinician reports within the same patients. Using secondary data from the NIDA-funded Women and Trauma Study, we demonstrated Common Persons Item Response Theory (IRT) Calibration for calibrating self-reported and clinician-reported PTSD severity scores in a manner similar to the process used to produce equated scores across multiple forms of standardized tests (e.g., SAT, GRE). Under IRT calibration, treatment effect sizes between the CAPS and MPSS-SR did not differ, while with the noncalibrated measures, the CAPS effect size was 85% larger than the MPSS-SR. Further, across the range of a combined CAPS/MPSS-SR gold standard, IRT-calibrated CAPS and MPSS-SR individual scores did not differ; for uncalibrated individual scores, MPSS scores were higher than CAPS scores at higher levels of PTSD severity while the reverse was true at lower levels of severity. The use of IRT calibration approaches for calibrating self-report and clinical interview measures of PTSD will allow treatment researchers to reflect the treatment effect on PTSD as a construct (regardless of reporter) as opposed to being limited to reporting treatment effects that may be discrepant within patients and specific to the particular assessment measure being employed.

Antonio A. Morgan-López, Behavioral Health Research Division, Substance Use Prevention and Evaluation Research (SUPER), RTI International, Research Triangle Park, North Carolina; Lissette M. Saavedra, Center on Social Determinants, Risk Behaviors, and Prevention Science Mental Health, Risk, and Resilience Research Program, RTI International; Denise A. Hien, Center of Alcohol and Substance Use Studies, Graduate School of Applied and Professional Psychology, Rutgers, The State University of New Jersey; Therese K. Killeen, Department of Psychiatry and Behavioral Sciences Addiction Sciences Division, Medical University of South Carolina; Sudie E. Back, Department of Psychiatry and Behavioral Sciences and DART Research Training Program, Medical University of South Carolina, and Ralph H. Johnson Veterans Affairs Medical Center, Charleston, South Carolina; Lesia M. Ruglass, Department of Psychology and Center of Alcohol and Substance Use Studies, Graduate School of Applied and Professional Psychology, Rutgers, The State University of New Jersey; Skye Fitzpatrick, Department of Psychology, Faculty of Health, York University; Teresa López-Castro, Department of Psychology, The City College of New York; Julie A. Patock-Peckham, Social Addictions Impulse Lab, Arizona State University.

The work presented in this article was supported by grants from the National Institute on Drug Abuse (NIDA Clinical Trials Network protocol 0015, Denise A. Hien, PI) and the National Institute on Alcohol Abuse and Alcoholism (Grant R01AA025853, Denise A. Hien and Antonio A. Morgan-López, MPIs). Secondary data for this study are registered under ClinicalTrials.gov (NCT00078156; Women's Treatment for Trauma and Substance Use Disorders) http://www.clinicaltrials.gov/ct2/show/NCT00078156. We are grateful to Sara Kass for editorial assistance in the preparation of this article.

Correspondence concerning this article should be addressed to Antonio A. Morgan-López, Behavioral Health Research Division, RTI International, 3040 Cornwallis Road, P.O. Box 12194, Research Triangle Park, NC 27709-2194. E-mail: amorganlopez@rti.org

Public Significance Statement

We demonstrated the application of Common Persons Item Response Theory (IRT) Calibration of self-reported and clinician-reported PTSD severity scores using data from the NIDA-funded Women and Trauma Study. The use of IRT approaches to calibrate self-report and clinical interview measures of PTSD will allow treatment researchers to better reflect treatment effects on PTSD as a construct, regardless of reporter.

Keywords: differential item functioning, latent variables, measurement invariance, posttraumatic stress disorder

Supplemental materials: http://dx.doi.org/10.1037/pas0000789.supp

Proper characterization of changes in PTSD symptomatology in treatment outcome studies requires repeated assessments throughout and following treatment. Yet the assessment of changes in PTSD by trained clinicians, particularly in treatment outcome studies, poses concerns regarding inefficiencies in the assessment process. Semistructured clinical interviews such as the Clinician-Administered PTSD Scale (CAPS; Blake et al., 1995), the gold standard in assessing PTSD symptoms and diagnosis, are often difficult to conduct repeatedly in treatment outcome studies because of the repeated time and respondent burden that comes with multiple assessments. Further, there are also clinical concerns about interference of the clinical interview with treatment process, particularly in trials using exposure therapies (Back et al., 2019). In response, treatment outcomes researchers have used brief self-report PTSD measures to capture changes over time in PTSD symptomatology, which are particularly advantageous given the differences in assessment time (e.g., 5–10 min for self-report vs. 40–60 min for clinical interviews; Foa, Riggs, Dancu, & Rothbaum, 1993).

Often, the implicit (if not explicit) intent of self-report PTSD measures such as the PTSD Checklist (PCL/PCL-5; Weathers, Litz, Huska, & Keane, 1994; Weathers et al., 2013), the Posttraumatic Stress Scale-Self Report (PSS-SR; Foa et al., 1993), and the Impact of Events Scale (IES; Weiss & Marmar, 1997) is to not only facilitate more rapid identification of a probable diagnosis of PTSD than is possible with clinical interviews but, in some contexts, to replace the structured interview (Sijbrandij et al., 2013). As such, there appears to be an assumption in the literature that there is at least sufficient cross-reporter validity between patient and clinician responses (despite their different perspectives on patient symptomatology) to conclude there is a reasonable level of similarity between self-report and clinician-interview scales in measuring the construct of PTSD (American Psychological Association, 2018). However, several systematic factors may impact whether the clinician and the patient would be equally likely, or not, to endorse the same symptom frequency or severity, independent of differences in clinician/patient perspectives (i.e., method bias; Podsakoff, MacKenzie, & Podsakoff, 2012). These include factors that vary across patients (i.e., type of trauma, age, gender, race/ethnicity) as well as factors that differ between self-report and clinical interview measures themselves such as differences in question stems, item response scaling, lack of clarifying prompts that redirect patients to the target traumatic event, and the manner in which questions are converted from frequency/severity items to symptoms and *DSM* diagnoses (Engelhard et al., 2007; Lunney, Schnurr, & Cook, 2014; Monson et al., 2008; Palmieri, Weathers, Difede, & King, 2007; Sijbrandij et al., 2013; Weathers, Ruscio, & Keane, 1999).

The factors that can impact differences in the measurement of PTSD between patients and clinicians have likely contributed to inconsistencies in reported within-study effect sizes on PTSD outcomes based on clinical interviews versus self-report measures (Ramaswamy et al., 2017). For example, in a number of studies that used both the CAPS and PCL for treatment outcomes analysis, some studies report larger effect sizes with the CAPS compared to the PCL (Blanchard et al., 2003; Monson, Schnurr, Stevens, & Guthrie, 2004) while others report larger effect sizes with the PCL than the CAPS (Monson et al., 2008; Schnurr et al., 2007; Schumm, Monson, O'Farrell, Gustin, & Chard, 2015); interestingly, the two studies of motor vehicle accident victims (Blanchard et al., 2003) had steeper decreases in PTSD scale scores on the CAPS than PCL. In contrast to the Blanchard studies, female veterans showed steeper decreases on the PCL, which further underscores variation in cross-reporter measurement that vary across populations and has been noted outside of the treatment outcomes context (Lunney et al., 2014). Similar inconsistencies in treatment effect sizes have been noted for comparisons between the CAPS and PSS-SR. For example, Gerardi, Rothbaum, Ressler, Heekin, and Rizzo (2008) found larger effect sizes with the PSS-SR, while others have observed larger effect sizes with the CAPS (Hien et al., 2009; Price et al., 2015).

Although as a field we appear to take these effect size differences across reporters for granted, this phenomenon should be disconcerting because these effect sizes are representative of what should presumably be the same underlying level of severity of PTSD coming from the same patients, particularly if the PTSD research community subscribes to the underlying belief that patient reports are sufficiently representative of the ground truth of their PTSD. Further, these differences would need to be reconciled in any attempts to summarize findings across PTSD treatment outcome studies (e.g., Petrakis & Simpson, 2017; Roberts, Roberts, Jones, & Bisson, 2015; Torchalla, Nosen, Rostam, & Allen, 2012). Although, the latter concern can be reasonably dealt with in, for example, any meta-analysis by accounting for the measure as a study-level factor that informs variation in effect sizes, the former distinguishes an estimate of the underlying effect size for treatment of the disorder versus an effect size that is specific to the measure that, to date, has gone unaddressed.

An additional concern that may contribute to discrepancies in treatment effects observed across self-report and clinical interview measures is the reliance on total scores, such as sum scores in which the sum of the response categories is taken across items, or mean scores in which the sum score is divided by the number of items (Curran et al., 2008). The above-summarized literature on factors that could impact item responses speaks to the likelihood that the measurement properties of self-report measures and clinical interview measures differ at the symptom-level, yet total scores reflect an equal weighting of how each symptom manifests itself as a reflection of the construct a) within-measure across symptoms and b) across reporters. For example, PTSD symptoms such as trouble concentrating would get equal weight (i.e., a de facto "weight" of 1) to other symptoms such as flashbacks or nightmares and also be equally weighted across reporters when they likely should not (Brewin, Lanius, Novac, Schnyder, & Galea, 2009). As a result, if the symptom "weights" (e.g., factor loadings in factor analysis, discrimination parameters in IRT) actually differ across symptoms and/or across reporters but are treated as though they do not differ, as is the case with total scores, the resulting scores will be biased, even if content validity remains equally salient across time, reporters or populations (Cappelleri, Jason Lundy, & Hays, 2014). These biased scores can then introduce differences across groups or across time that do not exist or obscure actual differences (Beaujean & Osterlind, 2008), and influence the accuracy and generalizability of PTSD measures and the treatment effect estimation that they reflect.

Calibration via Item Response Theory (IRT)

It was noted earlier that, ideally, self-report measures would reflect an accurate representation of what would have been reported from a clinical assessment in a fraction of the time. Although clinician-assessed and self-reported PTSD scale scores are often inconsistent, statistical approaches exist that allow for estimating PTSD severity scores that can be calibrated for comparability even in the presence of clinician/patient discordance through item response theory (IRT) calibration (Hanson, Harris, Pommerich, Sconing, & Yi, 2001; Kim & DeCarlo, 2016). A framework for estimating comparability of self-report and clinical interview measures under IRT is taken from Dorans (2007), which has been adapted for use from educational testing for health outcomes measurement in order to "adjust scores obtained on different instruments measuring the same construct so that they are comparable" (p. 86).

The general process of making scores comparable across measures is called score linking (Kolen & Brennan, 2004; Nunes Baptista, Primi, de Francisco Carvalho, Gomes Oliveira, & Elhai, 2017; Yu & Popp, 2005), used primarily in educational testing in scoring multiple test forms to combat student cheating (Cizek, 2001; Cizek & Wollack, 2017), but it is beginning to see wider use in health outcomes research (Edelen & Reeve, 2007). Dorans (2007) and others (e.g., Kolen & Brennan, 2004; Linn, 1993; Mislevy, 1992) describe a set of five requirements for the strongest form of score linking: namely score equating. These requirements are that a) the instruments to be linked measure the same construct, b) both instruments be equally reliable, c) the linking function for equating scores from instrument A to instrument B is the inverse of the linking function for equating scores from instrument B to instrument A (i.e., symmetry), d) the patient could be assessed with instrument A and their score would not deviate from what they would have scored on instrument B (and vice versa; i.e., equity) and e) the equating function used to link scores from instruments A and B should not vary across populations. When all five requirements are met, the resulting scores from instruments A and B are said to be "closely equable" (Hanson et al., 2001); these requirements are more characteristic of, for example, multiple forms of the Scholastic Aptitude Test (SAT).

The next best alternative for linking scores from instruments capturing the same construct under different measurement conditions, such as self-report versus clinical interview measures of PTSD, is calibration, and specifically, common persons (or single group; Chen, Huang, & MacGregor, 2009) calibration, referring to contexts where all participants have both measures. For calibration, from a score linking perspective, only two of the five requirements listed above need be met: that the instruments measure the same construct (even if under different conditions) and that symmetry is met (Dorans, 2007; Hanson et al., 2001). Thus, for two sets of scores to be calibrated (but not equated), the two instruments need not be equally reliable, nor does the linking function need to be equivalent across populations (e.g., can accommodate measurement noninvariance/DIF). Such scores would not be "closely equable", though calibrated scores can be "weakly equable" (Morris, 1982). Calibration, through a multiple step estimation process, would simultaneously estimate scales scores and item parameters using IRT for two or more measures of the same construct during the IRT estimation itself (Chen, Revicki, Lai, Cook, & Amtmann, 2009; Dorans, 2007). The two sets of calibrated scores could then be assessed statistically with regard to whether the calibrated scores have means and variances that do not deviate from each other (i.e., first- and second-order equity, respectively; Hanson et al., 2001; Kim & DeCarlo, 2016). Other linking approaches that do not require the same construct be measured across instruments fall under the umbrella of linking for nonequable scores, scores that are never meant to be exchangeable, such as the linking of SAT scores to grade point average (Chen et al., 2009; Dorans, 2007).

The purpose of this study is to demonstrate the potential for using IRT calibration to estimate comparable scale scores and, subsequently, equable treatment effects and effect size across the CAPS and the MPSS-SR. We first describe common persons IRT calibration and then show treatment effect estimation with IRT-estimated PTSD scale scores under three scenarios: a) a combined CAPS/MPSS-SR measure; b) a CAPS score calibrated to be on a comparable scale with the MPSS-SR and an MPSS-SR score calibrated to be on a comparable scale with the CAPS; and c) a raw score analog for the CAPS and MPSS-SR (but in an IRT-standardized metric) that is included for direct comparison which, as an almost ubiquitous finding across multiple RCTs, have been shown to differ substantially. This study aims to establish guidelines for the joint use of self-report and clinical interview measures in clinical trial contexts such that the effect sizes from treatment RCTs truly reflect the construct of interest. As Dorans (2007) notes: "If the outcome scores of different health assessment instruments are not properly linked, inferences based on them could be flawed, and might have serious consequences such as misperceptions about the efficacy of a treatment" (p. 86).

Method

Participants

Data were drawn from the Women and Trauma Study (Hien et al., 2009, 2012), which compared two behavioral interventions for the treatment of PTSD and co-occurring substance use disorders (SUDs). Participants ($N = 353$, 100% women) received community-based, outpatient substance abuse treatment from [locations blinded]. The RTI International Institutional Review Board determined that the project was exempt from IRB review because the project used "existing data that are publicly available where subjects cannot be identified or, through identifiers, linked to subjects." Inclusion criteria were at least one lifetime traumatic event and *DSM–IV* diagnosis of either full or subthreshold PTSD in the past 30 days. Subthreshold PTSD met criteria B (reexperiencing the trauma) and had to meet criteria C (avoidance of trauma reminders) or D (hyperarousal) instead of both. Other inclusion criteria were: a) being a female, b) being 18–65 years of age, c) using alcohol or substances within the past six months, and d) meeting a current (within the prior year) *DSM–IV* diagnosis of drug or alcohol abuse or dependence. Exclusion criteria were: a) impaired mental cognition, b) significant risk of suicidal/homicidal behavior (plan or attempt in the past 6 months), and c) history of schizophrenia-spectrum diagnosis or active (past two months) psychosis.

Recruitment and randomization. Seven community-based substance abuse treatment programs (CTPs) participated in the study, with the number of participants randomized at each site ranging from 7 to 106. The site randomizing 7 participants dropped from the study due to slow recruitment but did complete assessments, randomization, and treatment as prescribed in the protocol. The sites were a mixture of urban ($n = 5$) and suburban ($n = 2$) settings, located geographically in the Western ($n = 1$), Midwestern ($n = 1$), Northeastern ($n = 2$), and Southeastern ($n = 3$) United States. All participating programs offered a combination of outpatient individual and group treatment components, reflecting varying orientations and philosophies of addiction treatment.

The study was advertised via brochures, fliers, newspaper, and other print media, as well as through referrals from CTP treatment staff. A potential participant who was not already in treatment at the CTP and who responded to an advertisement needed to enroll in outpatient treatment at the CTP in order to participate. Interested participants ($N = 1963$) completed a brief in-person or telephone screen to ascertain likely eligibility, followed by an in-person screening assessment to confirm eligibility. All participants who completed a screening assessment ($N = 1212$; 751 ineligible) first signed an informed consent which included appropriate Health Insurance Portability and Accountability Act (HIPAA) language. Finally, a third (baseline) interview was completed ($N = 370$; 671 no-shows, 171 ineligible), with additional study consent, to further assess substance use, PTSD, and social characteristics. Baseline interviews lasted approximately 2.5 to 3 hr. Independent assessors who remained blind to randomization assignment performed all baseline and posttreatment assessments.

After the baseline assessment, eligible participants were randomized ($N = 353$; 17 refusals and no-shows) to receive Seeking Safety (Najavits, 2002) or an active control condition (i.e., Women's Health Education, Miller, Pagan, & Tross, 1998). Women in both conditions completed weekly self-report measures for PTSD symptoms, substance use, and service utilization during treatment and were reassessed using the clinician-administered PTSD battery after treatment at 1-week, 3-, 6-, and 12-months post treatment. For this study, only the baseline and 1-week posttreatment assessments were used.

Therapists. Therapists and therapist supervisors from each site were selected based on willingness to be randomized after submitting an audiotaped therapy session exemplifying their ability to deliver a cognitive–behavioral style of therapy. All counselors were women. About 6% had less than a bachelor's degree, 39% held a bachelor's degree, and 56% had a master's degree or greater. Half of the counselors were white, 28% black, and 22% Latina. After signing informed consent, two counselors and two local supervisors per site were randomized to deliver one of the two study interventions. All counselors and supervisors attended a 3-day workshop, and supervisors received another half day of training focused on how to implement supervision. An expert from the lead training team rated the videotaped certification sessions for adherence to the manual and competency in the delivery of the interventions. The local supervisors obtained interrater reliability with the lead expert trainers on the adherence scores using the certification sessions.

All intervention sessions were videotaped, and a proportion of the tapes were rated by local supervisors (≥50%). Throughout the study, therapists met weekly with local supervisors for supervision, and if adherence fell below competency criterion, additional supervision was provided. The lead node experts rated a randomly selected quarter (29%) of the therapist session tapes reviewed by the local supervisor, comparing their ratings with the local supervisors' ratings to assure supervisor fidelity and interrater reliability. For both interventions during the study, supervisor fidelity was determined by whether or not lead node experts and site supervisor ratings were in agreement on fidelity at a 70% level using specific adherence measures for each treatment.

Measures

CAPS. Clinician-rated PTSD was assessed via the Clinician-Administered PTSD Scale (CAPS; Blake et al., 1995), a structured interview that measures traumatic life events and frequency and intensity of signs and symptoms of PTSD in the past 30 days. Typically, frequency and severity ratings for the CAPS are summed to yield a total score ranging from 0 to 136 (for the CAPS under *DSM–IV*; see Discussion for implications of this work for *DSM–IV* vs. *DSM–5*). However, given the previously noted criticisms of total scores (e.g., Curran et al., 2008) and the emphasis on using *DSM* symptoms as input for the estimation of scores under IRT, the focus was first on converting frequency and severity ratings to binary measures of symptom endorsement. Symptoms were coded 0/1 for presence or absence of the particular symptom based on the Blake et al. (1995) symptom endorsement rule of experiencing the symptom at a) a frequency of at least once in the previous month and b) a symptom intensity that was at least moderate. Internal consistency was .72 at pre and .86 at post.

Modified PSS-SR. Self-reported PTSD was assessed via the 17-item Modified Post Traumatic Stress Disorder Symptom Scale-Self Report (MPSS-SR; Falsetti, Resnick, Resick, & Kilpatrick, 1993), which extends the PSS-SR (Foa et al., 1993) for separate ratings of frequency and severity of PTSD symptoms. Typically, frequency and severity ratings for the MPSS-SR are summed to yield a total score ranging from 0 to 119. The MPSS-SR symptoms were coded 0/1 for presence or absence of the symptom over the previous 7 days based on a frequency of at least once in the previous week and of moderate severity, the analog to the Blake et al. (1995) criteria for converting frequency and severity to symptoms in the CAPS. Internal consistency was .86 at pre and .90 at post. Symptom endorsement and symptom level concordance rates (i.e., number of patients and clinicians who agree on presence/absence divided by the total N) for the CAPS and MPSS-SR are shown in Table 1.

Data Analysis Strategy

IRT common persons item calibration. SAS Proc IRT (Wicklin, 2013) was used to fit a series of multiple-group (MG) IRT models under marginal maximum likelihood for the generation of expected a posteriori (EAP) IRT scores. First, an item calibration sample was selected such that for each patient, a single observation was randomly selected (pre or post) using SAS Proc SURVEYSELECT, consistent with the recommendations of Bauer and Hussong (2009) for item calibration and scale scoring in repeated measures contexts. Next, with the calibration sample, a series of single factor MG-IRT models were fit where a) time was used as a grouping variable (0 = pre, 1 = post), but because each patient only provided one-and-only one observation in the calibration sample, observations were independent across time, and b) the single-factor consisted of 34 items: the 17 CAPS symptoms and the 17 MPSS-SR symptoms. A defining feature of common persons calibration, where all subjects have both measures (Kolen & Brennan, 2004; Nunes Baptista et al., 2017; Yu & Popp, 2005), is that PTSD symptoms from both the CAPS and MPSS-SR measures are estimated on a single factor, so that item parameters are simultaneously calibrated on both measures as generated by the same PTSD severity factor score. It is the modeling of within-person difference (if differential item functioning [DIF] exists) in item/symptom parameters across measures that differentiates common persons calibration from contexts where multiple measures are used across subjects, but subjects either have only one measure within each study (e.g., integrative data analysis; Witkiewitz, Hallgren, O'Sickey, Roos, & Maisto, 2016) or only have minimal overlap in items within the same person across two or more groups of people (calibration with "anchor" items; Yu & Popp, 2005).

An initial base IRT model with varying slopes and thresholds across the 17 symptoms (2-parameter logistic IRT model) was fit to the calibration sample where item parameters for each symptom were constrained to equality both across measures and across time; for example, equal slopes and thresholds were estimated for intrusive recollections as measured by the CAPS and the MPSS-SR both at pre and post in the base model. The factor mean was set to 0 and the variance was set to 1 for the full set of observations across pre and post (in a online supplementary material, for model fit assessment and obtaining statistics such as the root mean square error of approximation (RMSEA) and the comparative fit index (CFI), the same model was fit under the nonlinear factor analysis framework with Delta parameterization in Mplus Version 8 (Muthén & Muthén, 1998–2017)).

Next, a series of 17 models were fit where slopes and intercepts were allowed to vary for both the CAPS and MPSS-SR measures at both timepoints for each of the 17 symptoms to test for symptom-specific DIF. Models were compared to the base model via likelihood ratio tests (LRT) where −2 times the log-likelihood values from the base and the focal model was taken and evaluated against a 1 *df* χ^2 distribution. The parameters from the final calibration model were then used to score all observations by fitting a model to the full sample that constrained all item parameters to fixed values from the DIF tests and estimating PTSD symptom severity factor scores under three scenarios of interest: a)

Table 1
Symptom Percentages and Concordance

Symptom	Pre			Post		
	CAPS	MPSS-SR	Concordance	CAPS	MPSS-SR	Concordance
Intrusive recollections	67.00	62.70	61.90	29.40	37.60	72.00
Dreams	33.00	42.70	73.00	17.20	23.10	81.50
Flashbacks	18.60	36.10	69.90	7.20	20.40	82.30
Psychological cues	55.00	72.20	58.20	25.80	44.30	64.20
Physiological cues	44.40	63.90	58.70	16.70	31.70	74.20
Thought avoidance	66.20	50.90	57.60	24.40	30.30	70.60
Activity avoidance	44.70	30.40	57.00	23.50	17.70	77.00
Inability to recall	38.10	41.30	53.90	22.60	22.10	71.50
Diminished interest	55.80	55.80	66.20	15.40	30.30	76.00
Detachment	73.40	49.50	58.50	28.60	27.50	77.40
Restricted affect	70.20	44.40	54.70	28.10	25.80	73.30
Foreshortened future	28.10	69.60	48.10	9.10	44.90	57.00
Sleep	72.80	55.50	59.30	44.80	37.60	61.10
Irritability	60.50	62.50	58.50	28.50	38.50	65.60
Concentration probs	69.10	53.90	55.60	29.40	29.90	77.80
Hypervigilance	64.20	48.70	62.80	22.60	25.30	79.20
Startle	37.50	49.60	62.80	14.50	28.10	74.20

Note. CAPS = Clinician Administered PTSD Scale; MPSS-SR = Modified PTSD Symptom Scale-Self Report. Concordance is percent agreement between CAPS and MPSS-SR.

scoring based on all 34 symptoms across both the CAPS and MPSS-SR measures, b) scoring based on the CAPS symptoms only, and c) scoring based on the MPSS-SR symptoms only. For comparison to standard scoring models, we also scored the CAPS and MPSS-SR separately under a model where all slopes were constrained to equality across symptoms and time while item thresholds were allowed to vary across symptoms but equated across time; this is essentially a 1-parameter logistic (i.e., "Rasch") IRT model with no DIF across time. From a practical scoring perspective, this is analogous to taking the raw total symptom score but scaled under IRT to standard deviation units (symptoms equally weighted and assuming no DIF over time; Andrich, 1978).

Tests of first- and second-order equity. Equity across two measures is defined as a condition where, for the same true underlying PTSD severity, the distribution of scores for Form A (i.e., clinician ratings) is the same as the distribution of scores from Form B (i.e., self-report) that have been equated (or calibrated) to Form A. However, Lord (1980) showed that this level of equity can only be met when both forms are perfectly reliable. Morris (1982) proposed a form of so-called "weak" equity, where scores for two forms are expected to have similar means (i.e., first-order equity; FOE) and similar variances (i.e., second-order equity; SOE) given for the same true underlying disorder severity. While originally proposed for equating (Hanson et al., 2001), FOE and SOE have been extended to the assessment of score similarity across forms via model-based calibration within IRT (Dorans, 2007; Wolf, 2014). However, as Hanson et al. (2001) warn, a "linkage could result in a high degree of comparability of score *distributions* while not [necessarily] providing a high degree of score comparability for some individuals" (p. 8).

Failing to meet FOE would have the practical consequence of the same individual having a different expected value for different measures of the same construct (Wolf, 2014). Failing to meet SOE (i.e., unequal variances across calibrations) when FOE is met would have the consequence of one measure having a significantly smaller sampling distribution, which (if means were equal across measures) would artificially advantage the measure with the smaller variance, yielding a larger observed effect size (Hanson et al., 2001; Wolf, 2014). Boos and Brownie's (1989) bootstrap approach to FOE and SOE was used; 1000 bootstrap samples were generated in SAS Proc SURVEYSELECT, and empirical confidence intervals were formed to test whether the FOE CIs contained 0 (i.e., no mean differences between CAPS and MPSS-SR IRT calibrated scores) and SOE CIs contain 1 (ratio of CAPS-to-PSS-SR IRT score variances equals 1).

Treatment Effect Estimation

Three-level hierarchical linear models were structured in SAS Proc MIXED under maximum likelihood estimation to accommodate clustering of both repeated measures among patients and clustering of patients within treatment site; the modeling of within-site clustering was more for correction of patient-level standard errors than modeling site-level variation per se. Unstructured variance components (i.e., intercept, linear slope, covariance between intercept and slope) were estimated at both the site and patient-levels, with the within-person ("Level-1") residual variance fixed to 0 as a necessary model restriction for growth models with two timepoints. The treatment condition indicator (SS vs. WHE) was centered so that the pre-post change estimate would be interpreted as the sample average estimate (instead of the pre-post estimate for the treatment arm coded as the comparison condition). The focal effect of interest for this investigation is the average pre-post change in PTSD severity scores across the five approaches to scale score estimation, given the previously noted differences in treatment effect size across measures; in previous intent-to-treat analyses of these data, there were no differences in changes over time between SS and WHE (Hien et al., 2009), though differences did emerge favoring SS when taking into account variation in treatment efficacy across attendance patterns (Hien et al., 2012; Morgan-Lopez et al., 2013, 2014).

Results

IRT Common Persons Item Calibration

An initial base model that had symptom/item parameters constrained to equality across measures and across time was fit in Mplus under robust weighted least squares estimation. This model fit the data moderately well; while the RMSEA was well below the .05 threshold for good fit, RMSEA = .037 (90% CI [.029, .044]), the CFI was slightly below the .90 threshold for good fit (CFI = .890), suggesting a model assuming measurement equivalence across measures and/or across time could be improved upon by incorporating measure- and/or time-specific DIF. This same base model was fit in SAS Proc IRT for establishment of the baseline log likelihood value (−10,486.783) against which tests of DIF would be conducted.

Next, a series of 17 IRT models were fit where slopes and intercepts were allowed to vary for both the CAPS and MPSS-SR measures at both timepoints for each of the 17 symptoms to test for symptom-specific DIF. Models were compared to the base model via likelihood ratio tests (LRT) where −2 times the log-likelihood values from the base and the focal model was taken and evaluated against a 1 *df* χ^2 distribution. These results for symptom-specific DIF are shown in Table 2 with a) only 4 out of the 17 symptoms showing no DIF of any sort and b) the majority of DIF shown between corresponding symptoms from the CAPS and MPSS-SR (as opposed to within-measure DIF across time); put another way, there was very little difference in measurement properties across pre and post within each measure but large differences in the measurement properties between the MPSS-SR and the CAPS across symptoms. This result is not surprising considering previously reported differences in total score concordance between the CAPS and MPSS-SR (Ruglass et al., 2014).

The final calibration model, with varying item parameters across time and/or measure as appropriate, is shown in Table 3. Compared to the base model, the final calibration model constituted a highly significant improvement in model fit, $\chi^2(52) = 610.794$, $p < .001$. The parameters from the final calibration model were then used to score all observations by fitting a model to the full sample that constrained all item parameters to the values in Table 3. Severity scores were generated under: a) scoring based on all 34 symptoms across both the CAPS and MPSS-SR measures (the "gold standard"), b) scoring based on the CAPS symptoms only (CAPS-IRT), and c) scoring based on the MPSS-SR symptoms only (MPSS-IRT). We also scored the CAPS and PSS-SR under a

Table 2
DIF Summary Table

Symptom	CAPS versus MPSS slope	Pre versus post slope	CAPS versus MPSS intercept	Pre versus post intercept
Intrusive recollections				
Dreams			12.29	
Flashbacks	35.01		53.01	
Psychological cues	7.81		51.04	
Physiological cues	9.52		9.90	
Thought avoidance	11.41			7.11
Activity avoidance	7.30		6.55	
Inability to recall			7.15	
Diminished interest			7.41	
Detachment			18.72	
Restricted affect		5.97	27.89	
Foreshortened future	23.09		46.85	
Sleep				
Irritability				
Concentration probs				
Hypervigilance				7.21
Startle	15.08		21.55	

Note. DIF = differential item functioning; CAPS = Clinician Administered PTSD Scale; MPSS = Modified PTSD Symptom Scale. $\chi^2(1)$ values that are significant at $p < .05$ are shown.

Rasch model with no DIF over time to mimic CAPS and MPSS-SR raw scores but keep scores in and IRT-type metric (referred to as CAPS-raw and MPSS-raw).

Scoring descriptives. Scoring descriptives for the gold standard measure, CAPS-IRT, MPSS-IRT, and the CAPS and MPSS-SR raw score analogs are shown in Table 4. Several key features emerge from these findings. First, the mean differences between the gold standard measure from pre-to-post (−.73) are nearly equated to the pre–post mean differences for the CAPS-IRT scores (−.65) and the MPSS-IRT (−.60); the corresponding pre-post differences between the CAPS raw score analog (−.82) and the MPSS-SR raw score analog (−.62) are a distance apart in magnitude that would be analogous to a small Cohen's *d* effect size difference. This is the case despite the fact that a) key correlations between the CAPS and MPSS-SR measures are virtually identical for the IRT calibrated scores or the raw score analogs (e.g., rs = .626 vs. .627 at baseline) and b) correlations between the scores for each measure under calibration versus raw score analog exceed .99 (e.g., CAPS calibrated and CAPS raw score analog).

FOE among scale-score means. The mean difference between the model-calibrated CAPS-IRT and MPSS-IRT scores was nonsignificant, $\Delta_{\text{FOE}} = .008$ (95% CI [−.047, .070]), suggesting the average difference in IRT-calibrated scores between the CAPS and MPSS-SR did not differ from 0, thereby meeting first-order equity. Also, the mean difference between the raw score analogs to the CAPS and MPSS-SR scores was nonsignificant, $\Delta_{\text{FOE}} = -.051$ (95% CI [−.108, .012]), suggesting the raw score analogs also met first-order equity.

SOE among scale-score variances. The ratio of variances between the CAPS-IRT and MPSS-IRT scores did not differ significantly from 1, SOE = .987 (95% CI [.890, 1.08]), suggesting the IRT calibration also met second-order equity. However, the ratio of variances between CAPS-RAW and MPSS-RAW scores did differ significantly from 1, SOE = .647 (95% CI [.578, .717]), suggesting the raw score analogs failed to meet second-order equity, with a smaller variance for CAPS raw score analogs than MPSS-SR raw score analogs.

Treatment Effects

Gold standard scores. For fixed effects tests using the CAPS + MPSS-SR IRT scores (used as the gold standard in this study), the average pre–post change in PTSD severity was significant and had a longitudinal Cohen's *d* effect size analog (Feingold, 2009, 2015) that exceeded what would be considered large (b = −.729 (.106), *t* = −6.89, *p* < .0001, *d* = −.87); the differences in change over time between SS and WHE were nonsignificant (*p* = .76).

CAPS-only IRT scores. For fixed effects tests using the CAPS-only IRT scores, the average pre–post change in PTSD severity was significant with a large effect size (b = −.661 (.105), *t* = −6.31, *p* < .0001, *d* = −.84); the differences in change over time between SS and WHE were nonsignificant (*p* = .93).

MPSS-SR-only IRT scores. For fixed effects tests using the MPSS-SR-only IRT scores, the average pre–post change in PTSD severity was significant with a large effect size (b = −.622 (.086), *t* = −7.27, *p* < .0001, *d* = −.76); the differences in change over time between SS and WHE were nonsignificant (*p* = .63).

CAPS raw score analogs. For fixed effects tests using the CAPS raw score analogs, the average pre–post change in PTSD severity was significant with a large effect size (b = −.807 (.079), *t* = −10.16, *p* < .0001, *d* = −1.10); the differences in change over time between SS and WHE were nonsignificant (*p* = .83).

MPSS-SR raw score analogs. For fixed effects tests using the PSS-SR-only IRT scores, the average pre–post change in PTSD severity was significant with a large effect size (b = −.623 (.079),

Table 3
IRT Parameters

		CAPS Pre	MPSS Pre	CAPS Post	MPSS Post
Item	Parameter	Estimate	Estimate	Estimate	Estimate
Intrusive recollections	Difficulty	−.71824	−.12126	.28161	−.07531
	Slope	.77927	1.85818	1.85796	2.87523
Dreams	Difficulty	1.02425	.54280	.97478	.53375
	Slope	1.17771	1.60747	1.50828	1.90841
Flashbacks	Difficulty	2.10685	.81070	1.90383	.55737
	Slope	.91416	1.43417	1.38629	2.44319
Psychological cues	Difficulty	−.00199	−.50102	.48841	−.27172
	Slope	.76361	1.67668	1.61436	2.30929
Physiological cues	Difficulty	.58077	−.10271	.83393	.11432
	Slope	.87052	2.53416	1.93116	2.76898
Thought avoidance	Difficulty	−.44465	.26101	.56670	.23697
	Slope	1.09080	1.61035	1.56476	1.91883
Activity avoidance	Difficulty	.55863	1.09835	.67744	.89710
	Slope	.89987	1.28234	1.39861	1.61531
Inability to recall	Difficulty	1.34555	.55538	1.29373	.49062
	Slope	.47806	2.11839	.79199	2.37974
Diminished interest	Difficulty	.04464	.10565	.97097	.18219
	Slope	1.17707	1.87900	1.75567	2.41865
Detachment	Difficulty	−.78223	.30562	.31778	.32771
	Slope	1.13403	1.79155	1.75965	2.13155
Restricted affect	Difficulty	−.96930	.50970	.36570	.44690
	Slope	.74612	1.28949	1.70953	1.77278
Foreshortened future	Difficulty	1.72354	−.53301	1.83392	−.25279
	Slope	.71174	1.25153	1.25611	1.44520
Sleep	Difficulty	−.22111	−.22111	−.22111	−.22111
	Slope	1.16269	1.16269	1.16269	1.16269
Irritability	Difficulty	−.01345	−.01345	−.01345	−.01345
	Slope	1.30930	1.30930	1.30930	1.30930
Concentration	Difficulty	.04651	.04651	.04651	.04651
	Slope	1.52579	1.52579	1.52579	1.52579
Hypervigilance	Difficulty	−.48360	.33124	.85892	.53080
	Slope	.84822	2.08366	1.17337	1.54521
Exaggerated startle	Difficulty	.85824	.30589	1.18520	.23913
	Slope	1.06005	1.64157	1.45051	2.69199

Note. CAPS = Clinician Administered PTSD Scale; MPSS = Modified PTSD Symptom Scale; IRT = item response theory. Parameters repeat across measure and time for sleep, irritability, and concentration problems.

$t = -7.84$, $p < .0001$, $d = -.59$); the differences in change over time between SS and WHE were nonsignificant ($p = .66$).

Comparison of effect size sampling distributions. In the interest of assessing whether the treatment effects observed with the gold standard measure would have been preserved across the various single-measure calibration approaches, the effect size confidence intervals were compared. Effect size confidence intervals were formed across all five measures (see Figure 1), with the effect size from the combined CAPS/MPSS-SR calibration treated as the gold standard. The horizontal dotted line in Figure 1 shows whether the gold standard effect size is contained in the sample confidence intervals across the other scale score estimation meth-

Table 4
Scoring Descriptives

Variable	*N*	*M*	Std Dev	Sum	Minimum	Maximum
Pre						
Gold standard	349.00	.29	.84	100.66	−2.12	2.68
CAPS-IRT	349.00	.28	.79	96.30	−1.72	2.28
MPSS-SR-IRT	349.00	.25	.82	86.86	−1.58	1.91
CAPS raw	349.00	.17	.73	60.19	−1.66	2.12
MPSS-SR raw	349.00	.15	1.04	51.92	−2.00	2.12
Post						
Gold standard	221.00	−.46	.94	−100.64	−1.90	2.20
CAPS-IRT	221.00	−.39	.86	−85.28	−1.42	2.10
MPSS-SR-IRT	221.00	−.37	.84	−81.53	−1.42	1.77
CAPS raw	221.00	−.65	.61	−143.33	−1.45	1.13
MPSS-SR raw	221.00	−.48	.75	−106.59	−1.45	1.51

Note. CAPS = Clinician Administered PTSD Scale; MPSS = Modified PTSD Symptom Scale; IRT = item response theory; MPSS-SR = Modified PTSD Symptom Scale-self report.

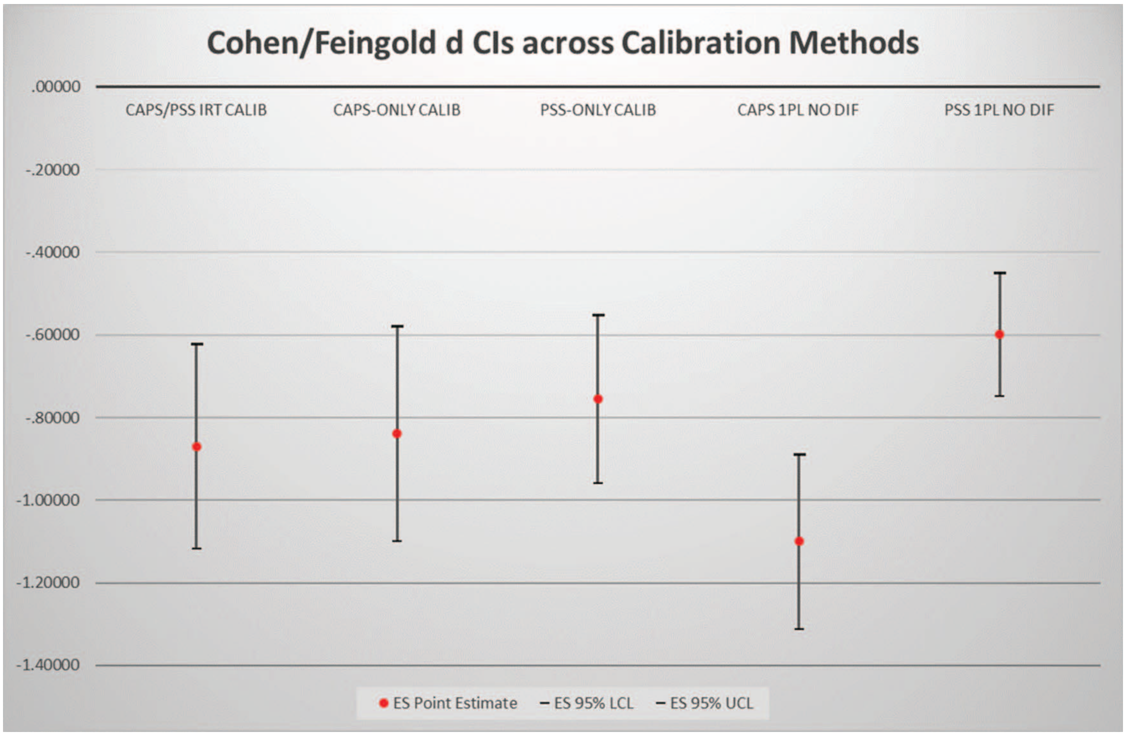

Figure 1. Cohen/Feingold *d* confidence intervals across methods. DIF = differential item functioning; PSS = PTSD Symptom Scale; CAPS = Clinician Administered PTSD Scale. See the online article for the color version of this figure.

ods, analogous to the examination of confidence interval coverage where a known population parameter is contained within a sample confidence interval, as is standard practice in statistical simulations (e.g., Muthén & Muthén, 1998–2017, p. 410–414; Wicklin, 2013). While the gold standard ES is within the sample confidence interval for the CAPS-only and MPSS-SR-only IRT calibrated measures, the raw score analogs for both measures a) do not contain the gold standard ES value in the confidence intervals and b) reproduce almost the identical finding from [masked for blind review] where the CAPS effect size was 86% larger than the MPSS-SR effect size (85% with actual raw scores).

Comparability of Individual Scores

It was noted earlier that a linkage between two instruments under IRT calibration would yield comparable score distributions while not necessarily providing comparability of calibrated scores for some individuals (Hanson et al., 2001; Wolf, 2014). Since typical treatment outcomes analysis focuses on inferences regarding treatment group differences between outcome distributions, comparability of calibrated distributions should generally yield comparable treatment effect sizes (such as were observed in this study) across calibrated instruments. However, researchers and practitioners may be interested in the comparability of individual scores after calibration for at least two specific contexts where the individual is paramount: a) assessment of functioning (e.g., initial screening, confirmation of diagnosis) and b) treatment outcomes contents where the interest is in modeling clinically significant change at the patient-level (Jabrayilov, Emons, & Sijtsma, 2016; Jacobson & Truax, 1991; Saavedra et al., 2019). Two approaches to evaluating the adequacy of individual scores—or at least scores at a given point on the latent severity distribution—are a) local reliability and b) local first-order equity.

Local reliability. Under classical test theory (CTT), the most common measure of reliability (at least as defined here as internal consistency) used in practice is, of course, Cronbach's alpha. Under CTT, α presumes that the reliability of a score is constant throughout the range of the construct, which is often unrealistic in practice. In IRT, the concept of reliability is "local", or specific to different levels of the construct (Embretson & Reise, 2000); for health outcomes research, a measure is ideally at its maximum reliability at the level of the construct at which a diagnostic decision is made (i.e., clinical cutoff; Chiesi et al., 2017). In order to calculate and graph local reliability (LR), test information function (TIF) values are output, where TIF values are the expected value of the inverse of the error variances for each estimated value of the latent construct score; these values can be requested as output from SAS Proc IRT, Mplus, or any IRT-capable software. Then, the TIF values are converted to LR values using 1 – (1/(TIF)) for each specific value of the latent construct score. Figures 2a and 2b show LR plots for the gold standard measure, the IRT-calibrated CAPS, and IRT-calibrated MPSS-SR separate for baseline and posttest. For the baseline LR, the apexes of the reliability curves for GS (~.93), IRT-calibrated MPSS (~.91) and IRT-calibrated CAPS (~.74) roughly coincide with the baseline mean for the GS scale scores (.29), indicating the maximum reliability for all three sets of scores is observed at the point

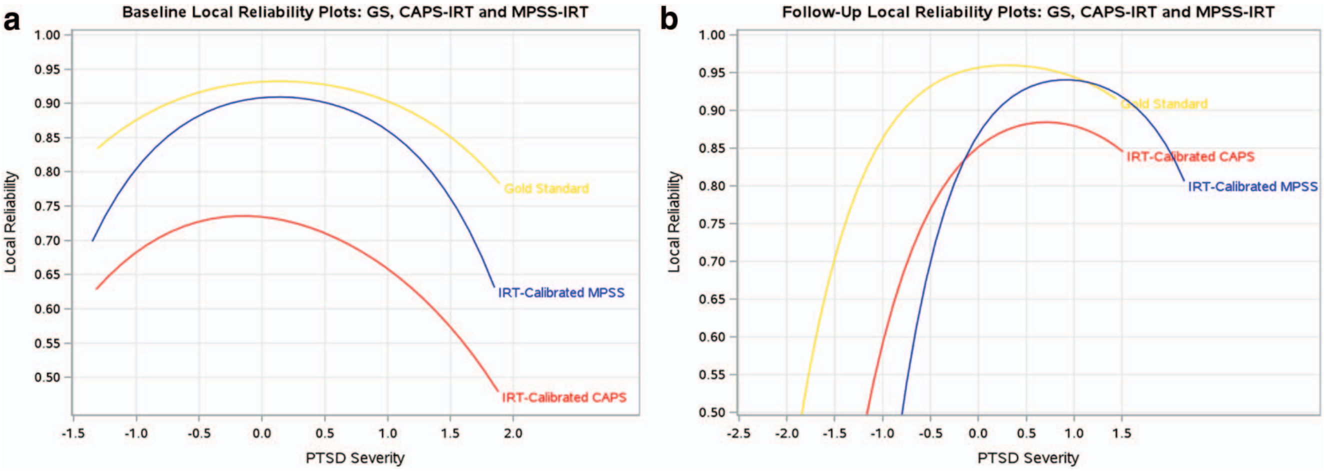

Figure 2. Baseline and post-test local reliability plots. CAPS = Clinician Administered PTSD Scale; IRT = item response theory; MPSS = Modified PTSD Symptom Scale. See the online article for the color version of this figure.

of the scale that is very close to the estimated mean for the GS distribution (i.e., the average level of latent PTSD severity for baseline). For the posttest LR, the reliability curves for GS, IRT-calibrated MPSS and IRT-calibrated CAPS are each no lower than ~.85 at the point of the posttest mean for the GS scale scores (−.46), though reliability is even higher for all three sets of scores at around .75 SDs above the mean.

Local first-order equity. Local reliabilities have great utility in and of themselves, but in this study they were also used to identify the key points on the gold standard scale at which we should expect the smallest discrepancies between calibrated scores; Lee, Lee, & Brennan (2012) note that discrepancies between IRT-calibrated scale scores will be at their smallest when scores are jointly and maximally reliable. Though no work to our knowledge has examined this principle in the relation between higher local reliability and smaller localized scoring discrepancies as measured by FOE, this principle was used to guide the evaluation of local FOE (Kim & DeCarlo, 2016) for a) the IRT-calibrated CAPS and MPSS and c) the CAPS and MPSS raw score analog counterparts.

Figures 3a and 3b correspond to localized confidence intervals for first-order equity. Recall that the global test for FOE was nonsignificant Δ_{FOE} = .008 (95% CI [−.047, .070]) suggesting that, on average, calibrated scores from the CAPS and MPSS-SR did not differ. Across the range of ±1 *SD* for the GS score distribution at baseline (−.55, 1.13), and posttest (−1.40, .48) Figures 3a and 3b show that the local confidence intervals include 0. Recall also that the uncalibrated CAPS and MPSS-SR score distributions did not differ on average Δ_{FOE} = −.051 (95% CI [−.108, .012]). However, local FOE as shown in Figures 3c and 3d suggest that across the same ±1 *SD* range of GS scores, there are points where the local FOE confidence intervals are above 0 at lower ranges of the GS scores and below 0 at higher ranges of the GS scores, particularly at baseline. This indicates that, for patients at low levels of PTSD severity, their uncalibrated CAPS scores will be significantly higher than their MPSS-SR scores while at higher levels of PTSD severity, uncalibrated MPSS-SR scores will be higher than uncalibrated CAPS scores. Tests for local second-order equity require, among other factors, both high and equal reliabilities for both sets of scores (Lee et al., 2012) and thus were not conducted.

Discussion

Establishing parallel scoring between patient self-report instruments, versus instruments that require clinician administration can create significant efficiencies in measurement of change in treatment outcomes evaluation. But the establishment of parallel scores under the strongest form of score linking (i.e., equating) across multiple instruments (Dorans, 2007; Kolen & Brennan, 2004) requires conditions that likely will not be met outside of educational testing contexts (e.g., construct measured under the same conditions, equal reliabilities, population invariance of item parameters), where instrument developers have unlimited latitude to write, test, score, drop and rewrite item content until these conditions are met. In PTSD treatment research contexts, despite general agreement that patient and clinician-reported instruments capture the same construct (APA, 2018; Sijbrandij et al., 2013), this very high standard for justification of score equating between patient and clinician reports of symptoms is unlikely to be met, given general differences in patient/clinician perspectives and factors that influence differences in item responses between self-report measures and clinical interview measures (e.g., Engelhard et al., 2007; Monson et al., 2008; Weathers et al., 1999). Multiple investigators who have contributed to the score linking literature within educational testing (Chen et al., 2009; Kolen & Brennan, 2004; Linn, 1993; Mislevy, 1992) have described score calibration as the next best alternative to score equating when some but not all of the assumptions of the score linking framework are met; Dorans (2007), in particular, has explicated this framework for health outcomes research and touted potential for utility for score calibration outside of educational testing.

Using data from the NIDA-funded Women & Trauma Study, we demonstrated the use of Common Persons IRT calibration for the

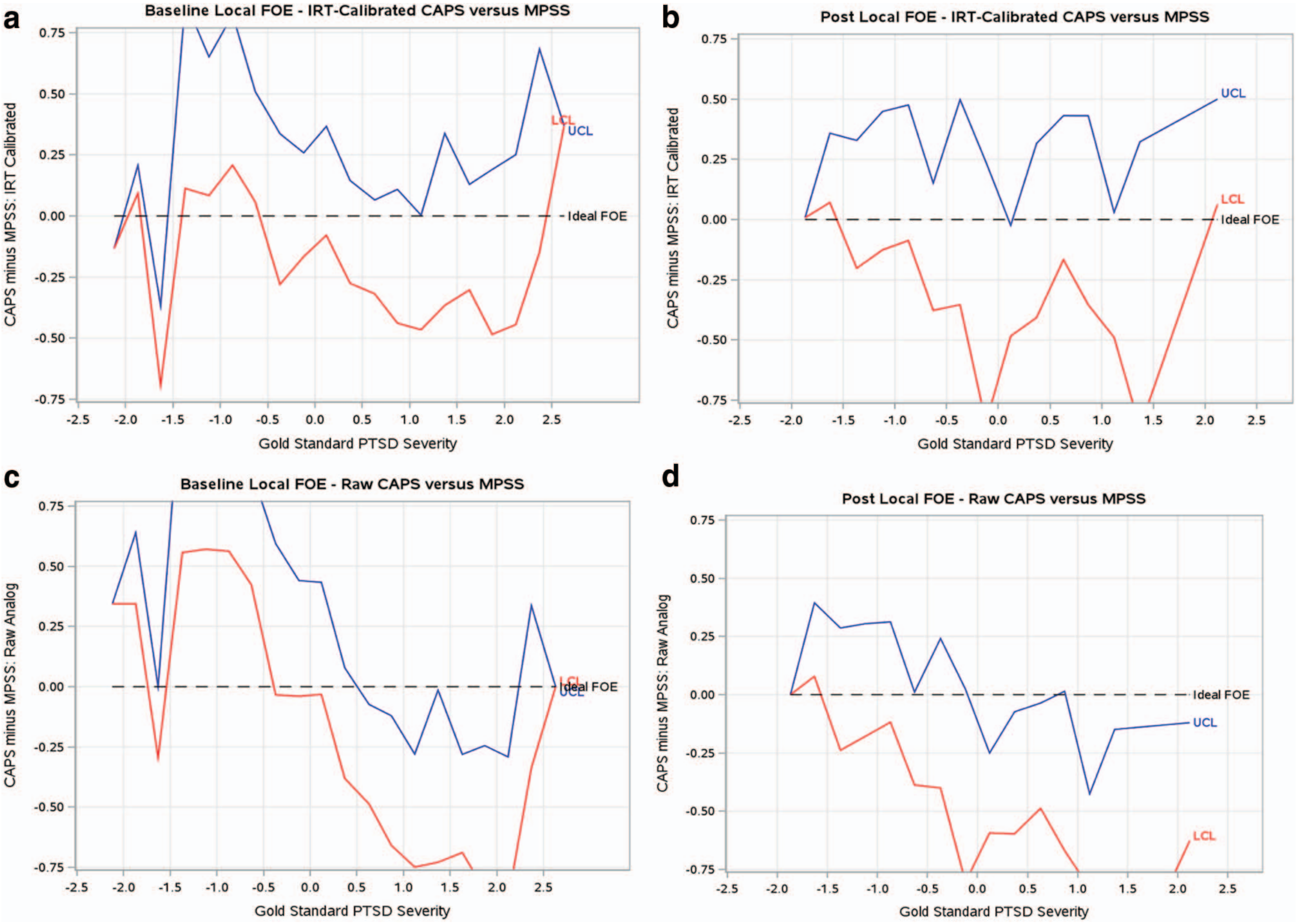

Figure 3. Baseline and post-test local first-order equity confidence intervals. CAPS = Clinician Administered PTSD Scale; IRT = item response theory; MPSS = Modified PTSD Symptom Scale; FOE = first-order equity. See the online article for the color version of this figure.

generation of equable scale score distributions from the CAPS and MPSS-SR. We showed how CAPS and MPSS-SR scores can be calibrated against a combined gold standard such that, regardless of whether a combined self-report/clinical interview measure is used or a calibrated measure from either reporter is used, statistical equity of score distributions and treatment effect size sampling distributions is preserved. This is in contrast to raw score analogs for the CAPS and MPSS-SR that are uncalibrated, which remained far apart from each other and from the self-report/clinical assessment "gold standard" measure on scale score and effect size distributions. Further examination of the individual scores showed that, at the lower range of PTSD severity, uncalibrated CAPS scale scores were higher than uncalibrated MPSS-SR scale scores while the reverse was true for scores at the higher range of PTSD severity.

Although the use of raw scores to estimate treatment efficacy from self-report and clinical interview measures is a common practice in reporting findings from treatment outcome studies (which often conflict), the results shown here illustrate that the use of scores from IRT calibration of symptoms from self-report and interview measures will more accurately reflect true treatment effect sizes—assuming that a multi-informant measure can be treated as "more true" than each single, uncalibrated measure alone. Further, IRT calibration procedures could eliminate the need to make measure-specific inferences regarding treatment effect sizes, which has been the norm in and outside of treatment contexts. If concurrency between patient and clinicians scale scores can be properly established using IRT calibration methodology, there may be opportunities for introducing multiple types of efficiencies into data collection and PTSD severity scoring. For example, self-report measures could be readily translatable into remote, digital assessments (such as ecological momentary assessment) that are now being used for symptom monitoring, with symptom weights calibrated to whatever gold standard is appropriate (i.e., CAPS, combined CAPS/self-report measure). Other forms of efficiency could be introduced across timepoints where self-report and clinical interview measures of PTSD are typically collected together (e.g., baseline, end-of-treatment, follow-ups), such as planned missingness designs (Graham, Taylor, Olchowski, & Cumsille, 2006) where randomly selected patients would only get the self-report measure, but the joint item parameters from both the self-report and the interview could be calibrated for the subset with both measures and applied to the subset with just the

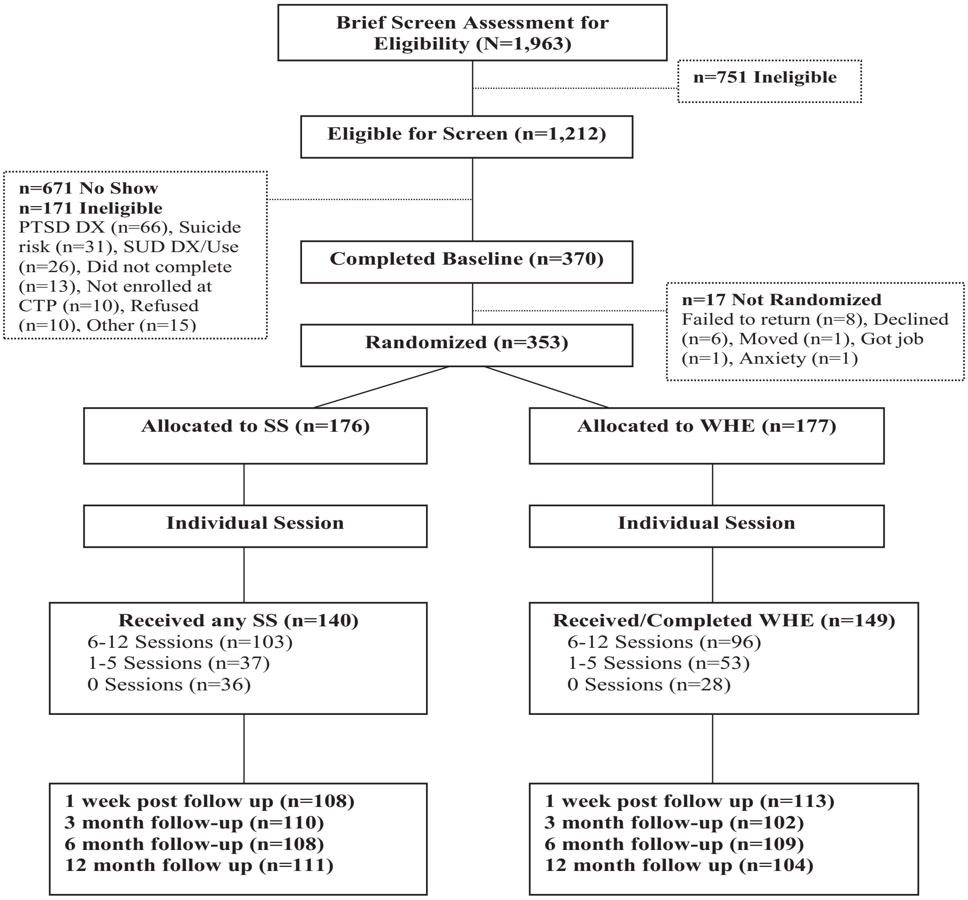
Brief Screen Assessment for Eligibility (N=1,963)
n=751 Ineligible
Eligible for Screen (n=1,212)
n=671 No Show
n=171 Ineligible
PTSD DX (n=66), Suicide risk (n=31), SUD DX/Use (n=26), Did not complete (n=13), Not enrolled at CTP (n=10), Refused (n=10), Other (n=15)
Completed Baseline (n=370)
n=17 Not Randomized
Failed to return (n=8), Declined (n=6), Moved (n=1), Got job (n=1), Anxiety (n=1)
Randomized (n=353)
Allocated to SS (n=176)
Allocated to WHE (n=177)
Individual Session
Individual Session
Received any SS (n=140)
6-12 Sessions (n=103)
1-5 Sessions (n=37)
0 Sessions (n=36)
Received/Completed WHE (n=149)
6-12 Sessions (n=96)
1-5 Sessions (n=53)
0 Sessions (n=28)
1 week post follow up (n=108)
3 month follow-up (n=110)
6 month follow-up (n=108)
12 month follow up (n=111)
1 week post follow up (n=113)
3 month follow-up (n=102)
6 month follow-up (n=109)
12 month follow up (n=104)

Figure 4. Women and trauma study CONSORT.

self-report, similar to what we did here with the PSS calibration. Given advances in comorbidity treatment research aiming to study dynamic changes over the "in-treatment" phase, where examining lagged effects of one set of symptom changes over another (such as PTSD changes impacting substance use) is used for illuminating the real time impact of particular treatments, the use of the self-report calibrated in combination with the clinical interview may be a critical advance.

The specific implications for IRT calibration on treatment outcomes analysis, which largely deals with differences in distributions across treatment arms, have to be made in a much more nuanced fashion when addressing the implications for IRT calibration for individual scores because, as Hanson et al. (2001) note, calibration of individual scores may not be as close for some individuals, particularly at the points of the scale where the joint reliabilities of both sets of scores are less reliable (Lee et al., 2012). While the literature has shown that the strength of effect sizes from treatment trials using the PSS/PCL and CAPS varies within trial (e.g., Blanchard et al., 2003; Hien et al., 2009; Monson et al., 2004, 2008; Price et al., 2015), the current findings actually show that, without calibration, certain patients may score higher on one measure or the other depending on their underlying PTSD severity. Under IRT calibration, the confidence interval for the IRT scale score difference between the CAPS and MPSS-SR contained 0 throughout the range of interest for GS severity scores in tests for local first-order equity. For applications that are geared more toward scoring of individuals from self-report measures, such as assessment and diagnostic contexts (see, e.g., Coffey, Gudmundsdottir, Beck, Palyo, & Miller, 2006; Gudmundsdottir & Beck, 2004; Mouthaan, Sijbrandij, Reitsma, Gersons, & Olff, 2014; Rash, Coffey, Baschnagel, Drobes, & Saladin, 2008; Ruggiero, Del Ben, Scotti, & Rabalais, 2003), greater consideration may need to be made regarding a) the points at which scores from the interview and the self-report are maximally reliable (Lee et al., 2012) and b) the determination of the clinical cut score at which a diagnosis is or is not made (Jacobson & Truax, 1991; Nunes Baptista et al., 2017). The ideal would be if both of these coincide (Chiesi et al., 2017), and, though this study was not a study on calibration for assessment and diagnosis per se, the approaches used in this study may offer direction to those interested in such contexts for improved assessment that goes above and beyond the utility of uncalibrated sum scores.

We wish to note several limitations of this study. This analysis was limited to the baseline and immediate posttreatment follow-up timepoints, which is notable for two reasons. As an initial demonstration of IRT calibration, we restricted the analysis to the two timepoints (pre and post for both the CAPS and MPSS-SR) where primary outcome effect sizes are typically evaluated, whereas in the dataset, additional in- and posttreatment timepoints are available. In addition to not incorporating the full complement of available data for this demonstration, it is again noted that just as self-report measures may be subject to patient bias, clinical interviews may be equally subject to assessor bias. This bias due to joint perspectives on the same construct may place a ceiling on the amount of symptom level concordance that could ever be observed between patient and clinician, particularly when considering variation in how symptom endorsement is coded from frequency and severity items in self-report and clinical interview measures (Blake et al., 1995; Weathers et al., 1999). This limits the extent to which any application of linking methodology could ever reach the high standards for score equating, but the patient/clinician measurement context seems to fit conditions for score calibration: ". . . instruments of unequal reliability that are nonetheless designed to measure the same construct" where "a content framework (e.g., self-report and interview content validity match to *DSM* criteria) is used to ensure that the construct being measured is the same from one instrument to another" (Dorans, 2007, p. 87).

That said, this methodology does not necessarily resolve the issue regarding which reporter has greater value in reflecting a construct of interest, as there are situations where differences in reporters' impressions may truly reflect different contexts (e.g., self-reported vs. partner-reported PTSD or a child at home vs. in school; De Los Reyes & Kazdin, 2009; Monson et al., 2004) and thus need to reflect different "gold standards" across contexts. For example, in treatment outcomes contexts such as in the current paper, a standard that reflects an "equal" contribution of the clinical impressions of both patient and clinician of PTSD symptomatology may be warranted, whereas for screening and diagnosis, the clinician's impression would likely be the optimal gold standard. Thus, for clinical assessment, instead of calibrating the clinician and patient measure against a combined measure as we have done here, it may make more sense clinically to calibrate the patient report directly against the clinical interview. Nevertheless, the calibration of separate self-report and interview measures for estimation of virtually identical treatment effects and effect sizes will allow treatment researchers to reflect the treatment effect on PTSD as a construct and not effect sizes that are specific to the PTSD measure.

Another limitation is the use of data from the *DSM–IV* conceptualization of PTSD. While many disorders remained unchanged regarding symptoms and criteria under *DSM–5*, substantial changes were made to PTSD under *DSM–5*, including a) the separation of PTSD from other anxiety disorders; b) the addition of a new criterion D (negative alterations in cognitions and mood), which includes symptoms such as persistent negative emotional states and distorted cognitions regarding the cause and consequences of the traumatic event; c) the addition of a symptom for self-destructive behaviors to the hyperarousal cluster; and d) also the removal of the requirement that the response involved fear, helplessness, or horror. However, in studies that have compared symptom endorsement and diagnosis rates, the comparison of symptoms that remain common between *DSM–IV* and *DSM–5* within the same patients have a concordance rate equivalent to or better than what would be have been expected for test–retest reliability with either system, and the three new symptoms do not appear to contribute much in additional clinical utility above and beyond the 17 symptoms that overlap across *DSM–IV* and *DSM–5* (see, e.g., Hoge, Riviere, Wilk, Herrell, & Weathers, 2014). Thus, while the analyses presented in this article lean more toward emphasizing the utility of IRT calibration, the specific clinical implications of this work for researchers studying current conceptualizations of PTSD under *DSM–5* are likely not compromised by the use of *DSM–IV* symptoms.

Despite these limitations, this study makes a number of notable contributions to the literature. In particular, this work represents the transfer of IRT calibration methodology that has a long-standing history and use in educational testing (e.g., Lord, 1980; Morris, 1982) for the improvement of accuracy in PTSD measurement but may have implications for other areas of intervention research. This work also illustrates ways in which it may be possible to reduce clinician time burden by utilizing repeated self-report measures that could be calibrated against the clinical interview. This could be done in a manner that would generate scale scores from the self-report measure as though it came from the clinical interview for possible integration into digital assessments even in real time and may lead to other efficiencies in data collection. We hope that this work would spur researchers across the spectrum of treatment and clinical foci to examine the utility of these approaches for capturing treatment effects on the construct of interest, irrespective of the reporter, rather than being restricted to the estimation of treatment effects and effect sizes that are specific to any single measure.

References

American Psychological Association. (2018). *Clinical practice guideline for the treatment of posttraumatic stress disorder*. Retrieved from https://www.apa.org/ptsd-guideline/assessment/

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573. http://dx.doi.org/10.1007/BF02293814

Back, S. E., Killeen, T., Badour, C. L., Flanagan, J. C., Allan, N. P., Ana, E. S., . . . Brady, K. T. (2019). Concurrent treatment of substance use disorders and PTSD using prolonged exposure: A randomized clinical trial in military veterans. *Addictive Behaviors, 90,* 369–377. http://dx.doi.org/10.1016/j.addbeh.2018.11.032

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods, 14,* 101–125. http://dx.doi.org/10.1037/a0015583

Beaujean, A. A., & Osterlind, S. J. (2008). Using item response theory to assess the Flynn effect in the National Longitudinal Study of Youth 79 Children and Young Adults data. *Intelligence, 36,* 455–463. http://dx.doi.org/10.1016/j.intell.2007.10.004

Blake, D. D., Weathers, F. W., Nagy, L. M., Kaloupek, D. G., Gusman, F. D., Charney, D. S., & Keane, T. M. (1995). The development of a clinician-administered PTSD scale. *Journal of Traumatic Stress, 8,* 75–90. http://dx.doi.org/10.1002/jts.2490080106

Blanchard, E. B., Hickling, E. J., Malta, L. S., Jaccard, J., Devineni, T., Veazey, C. H., & Galovski, T. E. (2003). Prediction of response to psychological treatment among motor vehicle accident survivors with PTSD. *Behavior Therapy, 34,* 351–363. http://dx.doi.org/10.1016/S0005-7894(03)80005-9

Boos, D. D., & Brownie, C. (1989). Bootstrap methods for testing homogeneity of variances. *Technometrics, 31,* 69–82. http://dx.doi.org/10.1080/00401706.1989.10488477

Brewin, C. R., Lanius, R. A., Novac, A., Schnyder, U., & Galea, S. (2009). Reformulating PTSD for *DSM-V*: Life after criterion A. *Journal of Traumatic Stress, 22,* 366–373. http://dx.doi.org/10.1002/jts.20443

Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures.

Clinical Therapeutics, 36, 648–662. http://dx.doi.org/10.1016/j.clinthera.2014.04.006

Chen, F., Huang, X., & MacGregor, D. (2009). *Equating or linking: Basic concepts and a case study*. Accessed 17 December 2018. Retrieved from http://online.fliphtml5.com/xrgx/bfuj/#p=1

Chen, W. H., Revicki, D. A., Lai, J. S., Cook, K. F., & Amtmann, D. (2009). Linking pain items from two studies onto a common scale using item response theory. *Journal of Pain and Symptom Management, 38,* 615–628. http://dx.doi.org/10.1016/j.jpainsymman.2008.11.016

Chiesi, F., Primi, C., Pigliautile, M., Ercolani, S., Della Staffa, M. C., Longo, A., . . . Mecocci, P. (2017). The local reliability of the 15-item version of the Geriatric Depression Scale: An item response theory (IRT) study. *Journal of Psychosomatic Research, 96,* 84–88. http://dx.doi.org/10.1016/j.jpsychores.2017.03.013

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20,* 19–27. http://dx.doi.org/10.1111/j.1745-3992.2001.tb00072.x

Cizek, G. J., & Wollack, J. A. (Eds.). (2017). *Handbook of quantitative methods for detecting cheating on tests*. New York, NY: Routledge.

Coffey, S. F., Gudmundsdottir, B., Beck, J. G., Palyo, S. A., & Miller, L. (2006). Screening for PTSD in motor vehicle accident survivors using the PSS-SR and IES. *Journal of Traumatic Stress, 19,* 119–128. http://dx.doi.org/10.1002/jts.20106

Curran, P. J., Hussong, A. M., Cai, L., Huang, W., Chassin, L., Sher, K. J., & Zucker, R. A. (2008). Pooling data from multiple longitudinal studies: The role of item response theory in integrative data analysis. *Developmental Psychology, 44,* 365–380. http://dx.doi.org/10.1037/0012-1649.44.2.365

De Los Reyes, A., & Kazdin, A. E. (2009). Identifying evidence-based interventions for children and adolescents using the range of possible changes model: A meta-analytic illustration. *Behavior Modification, 33,* 583–617. http://dx.doi.org/10.1177/0145445509343203

Dorans, N. J. (2007). Linking scores from multiple health outcome instruments. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 16*(S1), 85–94. http://dx.doi.org/10.1007/s11136-006-9155-3

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation,* 16(S1), 5–18. http://dx.doi.org/10.1007/s11136-007-9198-0

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

Engelhard, I. M., van den Hout, M. A., Weerts, J., Arntz, A., Hox, J. J., & McNally, R. J. (2007). Deployment-related stress and trauma in Dutch soldiers returning from Iraq. *The British Journal of Psychiatry, 191,* 140–145. http://dx.doi.org/10.1192/bjp.bp.106.034884

Falsetti, S. A., Resnick, H. S., Resick, P. A., & Kilpatrick, D. G. (1993). The modified PTSD symptom scale: A brief self-report measure of posttraumatic stress disorder. *The Behavior Therapist, 16,* 161–162.

Feingold, A. (2009). Effect sizes for growth-modeling analysis for controlled clinical trials in the same metric as for classical analysis. *Psychological Methods, 14,* 43–53. http://dx.doi.org/10.1037/a0014699

Feingold, A. (2015). Confidence interval estimation for standardized effect sizes in multilevel and latent growth modeling. *Journal of Consulting and Clinical Psychology, 83,* 157–168. http://dx.doi.org/10.1037/a0037721

Foa, E. B., Riggs, D. S., Dancu, C. V., & Rothbaum, B. O. (1993). Reliability and validity of a brief instrument for assessing posttraumatic stress disorder. *Journal of Traumatic Stress, 6,* 459–473. http://dx.doi.org/10.1002/jts.2490060405

Gerardi, M., Rothbaum, B. O., Ressler, K., Heekin, M., & Rizzo, A. (2008). Virtual reality exposure therapy using a virtual Iraq: Case report. *Journal of Traumatic Stress, 21,* 209–213. http://dx.doi.org/10.1002/jts.20331

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods, 11,* 323–343. http://dx.doi.org/10.1037/1082-989X.11.4.323

Gudmundsdottir, B., & Beck, J. G. (2004). Understanding the pattern of PTSD symptomatology: A comparison of between versus within-group approaches. *Behaviour Research and Therapy, 42,* 1367–1375. http://dx.doi.org/10.1016/j.brat.2003.09.005

Hanson, B. A., Harris, D. J., Pommerich, M., Sconing, J. A., & Yi, Q. (2001). *Suggestions for the evaluation and use of concordance results*. Iowa City, IA: ACT.

Hien, D. A., Morgan-Lopez, A. A., Campbell, A. N. C., Saavedra, L. M., Wu, E., Cohen, L., . . . Nunes, E. V. (2012). Attendance and substance use outcomes for the Seeking Safety program: Sometimes less is more. *Journal of Consulting and Clinical Psychology, 80,* 29–42. http://dx.doi.org/10.1037/a0026361

Hien, D. A., Wells, E. A., Jiang, H., Suarez-Morales, L., Campbell, A. N. C., Cohen, L. R., . . . Nunes, E. V. (2009). Multisite randomized trial of behavioral interventions for women with co-occurring PTSD and substance use disorders. *Journal of Consulting and Clinical Psychology, 77,* 607–619. http://dx.doi.org/10.1037/a0016227

Hoge, C. W., Riviere, L. A., Wilk, J. E., Herrell, R. K., & Weathers, F. W. (2014). The prevalence of post-traumatic stress disorder (PTSD) in U.S. combat soldiers: A head-to-head comparison of *DSM–5* versus *DSM–IV–TR* symptom criteria with the PTSD checklist. *The Lancet Psychiatry, 1,* 269–277. http://dx.doi.org/10.1016/S2215-0366(14)70235-4

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement, 40,* 559–572. http://dx.doi.org/10.1177/0146621616664046

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12–19. http://dx.doi.org/10.1037/0022-006X.59.1.12

Kim, Y., & DeCarlo, L. T. (2016). *Evaluating equity at the local level using bootstrap tests*. (Research Report 2016–4, Rep. No. 4). New York, NY: The College Board.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking methods and practices*. New York City, NY: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4757-4310-4

Lee, E., Lee, W.-C., & Brennan, R. L. (2012). *Exploring equity properties in equating using AP examinations*. (Research Report 2012–4). New York, NY: The College Board.

Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 349–364). Hillsdale, NJ: Erlbaum.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lunney, C. A., Schnurr, P. P., & Cook, J. M. (2014). Comparison of clinician- and self-assessments of posttraumatic stress symptoms in older versus younger veterans. *Journal of Traumatic Stress, 27,* 144–151. http://dx.doi.org/10.1002/jts.21908

Miller, S., Pagan, D., & Tross, S. (1998). *Women's health education: Peer activism for female partners of injection drug users*. Unpublished treatment manuscript, Columbia University, New York.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: ETS.

Monson, C. M., Gradus, J. L., Young-Xu, Y., Schnurr, P. P., Price, J. L., & Schumm, J. A. (2008). Change in posttraumatic stress disorder symptoms: Do clinicians and patients agree? *Psychological Assessment, 20,* 131–138. http://dx.doi.org/10.1037/1040-3590.20.2.131

Monson, C. M., Schnurr, P. P., Stevens, S. P., & Guthrie, K. A. (2004). Cognitive-Behavioral couple's treatment for posttraumatic stress disorder: Initial findings. *Journal of Traumatic Stress, 17,* 341–344. http://dx.doi.org/10.1023/B:JOTS.0000038483.69570.5b

Morgan-Lopez, A. A., Saavedra, L. M., Hien, D. A., Campbell, A. N., Wu, E., & Ruglass, L. (2013). Synergy between seeking safety and twelve-step affiliation on substance use outcomes for women. *Journal of Substance Abuse Treatment, 45,* 179–189. http://dx.doi.org/10.1016/j.jsat.2013.01.015

Morgan-Lopez, A. A., Saavedra, L. M., Hien, D. A., Campbell, A. N., Wu, E., Ruglass, L., . . . Bainter, S. C. (2014). Indirect effects of 12-session seeking safety on substance use outcomes: Overall and attendance class-specific effects. *The American Journal on Addictions, 23,* 218–225. http://dx.doi.org/10.1111/j.1521-0391.2014.12100.x

Morris, G. M. (1982). On the foundations of test equating. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 169–191). New York, NY: Academic Press.

Mouthaan, J., Sijbrandij, M., Reitsma, J. B., Gersons, B. P. R., & Olff, M. (2014). Comparing screening instruments to predict posttraumatic stress disorder. *PLoS ONE 9*(5), e97183.

Muthén, L. K. and Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Author.

Najavits, L. M. (2002). *Seeking safety: A treatment manual for PTSD and substance abuse*. New York, NY: Guilford Press.

Nunes Baptista, M., Primi, R., de Francisco Carvalho, L., Gomes Oliveira, J., & Elhai, J. D. (2017). Constructing a common scale between tests of depression: The use of item response theory for transferring of norms from the BDI to EBADEP-A. *Universitas Psychologica, 16,* 256–266.

Palmieri, P. A., Weathers, F. W., Difede, J., & King, D. W. (2007). Confirmatory factor analysis of the PTSD Checklist and the Clinician-Administered PTSD Scale in disaster workers exposed to the World Trade Center Ground Zero. *Journal of Abnormal Psychology, 116,* 329–341. http://dx.doi.org/10.1037/0021-843X.116.2.329

Petrakis, I. L., & Simpson, T. L. (2017). Posttraumatic stress disorder and alcohol use disorder: A critical review of pharmacologic treatments. *Alcoholism: Clinical and Experimental Research, 41,* 226–237. http://dx.doi.org/10.1111/acer.13297

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63,* 539–569. http://dx.doi.org/10.1146/annurev-psych-120710-100452

Price, M., Maples, J. L., Jovanovic, T., Norrholm, S. D., Heekin, M., & Rothbaum, B. O. (2015). An investigation of outcome expectancies as a predictor of treatment response for combat veterans with PTSD: Comparison of clinician, self-report, and biological measures. *Depression and Anxiety, 32,* 392–399. http://dx.doi.org/10.1002/da.22354

Ramaswamy, S., Driscoll, D., Reist, C., Smith, L. M., Albers, L. J., Rose, J., . . . Hollifield, M. (2017). A double-blind, placebo-controlled randomized trial of vilazodone in the treatment of posttraumatic stress disorder and comorbid depression. *The Primary Care Companion for CNS Disorders, 19*(4), 17m02138. http://dx.doi.org/10.4088/PCC.17m02138

Rash, C. J., Coffey, S. F., Baschnagel, J. S., Drobes, D. J., & Saladin, M. E. (2008). Psychometric properties of the IES-R in traumatized substance dependent individuals with and without PTSD. *Addictive Behaviors, 33,* 1039–1047. http://dx.doi.org/10.1016/j.addbeh.2008.04.006

Roberts, N. P., Roberts, P. A., Jones, N., & Bisson, J. I. (2015). Psychological interventions for post-traumatic stress disorder and comorbid substance use disorder: A systematic review and meta-analysis. *Clinical Psychology Review, 38,* 25–38. http://dx.doi.org/10.1016/j.cpr.2015.02.007

Ruggiero, K. J., Del Ben, K., Scotti, J. R., & Rabalais, A. E. (2003). Psychometric properties of the PTSD Checklist-Civilian Version. *Journal of Traumatic Stress, 16,* 495–502. http://dx.doi.org/10.1023/A:1025714729117

Ruglass, L. M., Hien, D. A., Hu, M. C., Campbell, A. N., Caldeira, N. A., Miele, G. M., & Chang, D. F. (2014). Racial/ethnic match and treatment outcomes for women with PTSD and substance use disorders receiving community-based treatment. *Community Mental Health Journal, 50,* 811–822. http://dx.doi.org/10.1007/s10597-014-9732-9

Saavedra, L. M., Morgan-López, A. A., Hien, D. A., Killeen, T. K., Fitzpatrick, S., Ruglass, L. M., . . . López-Castro, T. (2019, November). *Putting the patient back in clinical significance: Using item response theory in estimating clinically significant change in treatment for PTSD and SUDs*. Poster presented at the Annual Meeting of the Association for Behavioral and Cognitive Therapies, Atlanta, GA.

Schnurr, P. P., Friedman, M. J., Engel, C. C., Foa, E. B., Shea, M. T., Chow, B. K., . . . Bernardy, N. (2007). Cognitive behavioral therapy for posttraumatic stress disorder in women: A randomized controlled trial. *Journal of the American Medical Association, 297,* 820–830. http://dx.doi.org/10.1001/jama.297.8.820

Schumm, J. A., Monson, C. M., O'Farrell, T. J., Gustin, N. G., & Chard, K. M. (2015). Couple treatment for alcohol use disorder and posttraumatic stress disorder: Pilot results from U.S. military veterans and their partners. *Journal of Traumatic Stress, 28,* 247–252. http://dx.doi.org/10.1002/jts.22007

Sijbrandij, M., Reitsma, J. B., Roberts, N. P., Engelhard, I. M., Olff, M., Sonneveld, L. P., & Bisson, J. I. (2013). *Self-report screening instruments for post-traumatic stress disorder (PTSD) in survivors of traumatic experiences (protocol)*. Retrieved from http://dare.ubvu.vu.nl/bitstream/handle/1871/43985/Sijbrandij?sequence=15August2019

Torchalla, I., Nosen, L., Rostam, H., & Allen, P. (2012). Integrated treatment programs for individuals with concurrent substance use disorders and trauma experiences: A systematic review and meta-analysis. *Journal of Substance Abuse Treatment, 42,* 65–77. http://dx.doi.org/10.1016/j.jsat.2011.09.001

Weathers, F. W., Litz, B. T., Huska, J. A., & Keane, T. M. (1994). *PTSD checklist—Military version*. Boston, MA: Nation Center for PTSD, Behavioral Sciences Division.

Weathers, F. W., Litz, B. T., Keane, T. M., Palmieri, P. A., Marx, B. P., & Schnurr, P. P. (2013). *The PTSD checklist for DSM–5 (PCL-5)*. Retrieved from www.ptsd.va.gov

Weathers, F. W., Ruscio, A. M., & Keane, T. M. (1999). Psychometric properties of nine scoring rules for the Clinician-Administered Posttraumatic Stress Disorder Scale. *Psychological Assessment, 11,* 124–133. http://dx.doi.org/10.1037/1040-3590.11.2.124

Weiss, D. S., & Marmar, C. R. (1997). The impact of event scale-Revised. In J. P. Wilson & T. M. Keane (Eds.), *Assessing psychological trauma and PTSD* (pp. 399–411). New York, NY: Guilford Press.

Wicklin, R. (2013). *Simulating data with SAS*. Cary, NC: SAS Institute Inc.

Witkiewitz, K., Hallgren, K. A., O'Sickey, A. J., Roos, C. R., & Maisto, S. A. (2016). Reproducibility and differential item functioning of the alcohol dependence syndrome construct across four alcohol treatment studies: An integrative data analysis. *Drug and Alcohol Dependence, 158,* 86–93. http://dx.doi.org/10.1016/j.drugalcdep.2015.11.001

Wolf, R. (2014). *Assessing the impact of characteristics of the test, common-items, and examinees on the preservation of equity properties in mixed-format test equating*. (Unpublished doctoral dissertation), University of Pittsburgh, Pittsburgh, PA.

Yu, C. H., & Popp, S. E. O. (2005). Test equating by common items and common subjects: Concepts and applications. *Practical Assessment, Research & Evaluation, 10,* 1–19.

Received April 16, 2019
Revision received October 17, 2019
Accepted October 23, 2019 ■