Title

Estimating posttraumatic stress disorder severity in the presence of differential item functioning across populations, comorbidities, and interview measures: Introduction to Project Harmony

Journal

Journal of Traumatic Stress, 35(3)

ISSN

0894-9867

Authors

Morgan-López, Antonio A
Hien, Denise A
Saraiya, Tanya C
et al.

Publication Date

2022-06-01

DOI

10.1002/jts.22800

Peer reviewed

Published in final edited form as:

J Trauma Stress. 2022 June ; 35(3): 926–940. doi:10.1002/jts.22800.

Estimating posttraumatic stress disorder severity in the presence of differential item functioning across populations, comorbidities, and interview measures: Introduction to Project Harmony

Antonio A. Morgan-López[1], **Denise A. Hien**[2], **Tanya C. Saraiya**[3], **Lissette M. Saavedra**[1], **Sonya B. Norman**[4,5], **Therese K. Killeen**[3,6], **Tracy L. Simpson**[7], **Skye Fitzpatrick**[8], **Katherine L. Mills**[9], **Lesia M. Ruglass**[10], **Sudie E. Back**[3,6], **Teresa López-Castro**[10], **Consortium on Addiction, Stress and Trauma (CAST)**[1]

[1]RTI International, Research Triangle Park, North Carolina, USA

[2]Center for Alcohol Studies, Rutgers University–Piscataway, Piscataway, New Jersey, USA

[3]Department of Psychiatry and Behavioral Sciences, Medical University of South Carolina, Charleston, South Carolina, USA

[4]Veterans Affairs San Diego Healthcare System, San Diego, California, USA

[5]Department of Psychiatry, University of California–San Diego, San Diego, California, USA

[6]Ralph H. Johnson VA Medical Center, Charleston, South Carolina, USA

[7]Veterans Affairs Puget Sound Healthcare System, Seattle, Washington, USA

[8]Department of Psychology, York University, Toronto, Canada

[9]Sydney Medical SchoolUniversity of Sydney, Sydney, Australia

[10]Department of Psychology, City College of New York, New York, New York, USA

Abstract

Multiple factor analytic and item response theory studies have shown that items/symptoms vary in their relative clinical weights in structured interview measures for posttraumatic stress disorder (PTSD). Despite these findings, the use of total scores, which treat symptoms as though they are equally weighted, predominates in practice, with the consequence of undermining the precision of clinical decision-making. We conducted an integrative data analysis (IDA) study to harmonize PTSD structured interview data (i.e., recoding of items to a common symptom metric) from 25 studies (total $N = 2,568$). We aimed to identify (a) measurement noninvariance/differential

Correspondence: Antonio A. Morgan-López, RTI International, 3040 E. Cornwallis Road; Research Triangle Park, NC27709. amorganLopez@rti.org.

The Consortium on Addiction, Stress and Trauma (CAST) includes Steven Batki, Malcolm Battersby, Matthew Boden, Deborah Brief, Christy Capone, Kathleen Chard, Joan Cook, Annette Crisanti, Erica Eaton, Thomas Ehring, Paul Emmelkamp, Edna Foa, Linda Frisman, Moira Haller, Deborah Kaysen, Shannon Kehle-Forbes, Asa Magnusson, Meghan McDevitt-Murphy, Mark McGovern, Lisa Najavits, David Oslin, Jessica Peirce, Ismene Petrakis, M. Zachary Rosenthal, Michael Saladin, Claudia Sannibale, Rebecca Schacht, Ingo Schaefer, Jeremiah Schumm, Susan Sonne, Geraldine Tapia, Jessica Tripp, Debora Van Dam, Anka Vujanovic, and Caron Zlotnick.

item functioning (MNI/DIF) across multiple populations, psychiatric comorbidities, and interview measures simultaneously and (b) differences in inferences regarding underlying PTSD severity between scale scores estimated using moderated nonlinear factor analysis (MNLFA) and a total score analog model (TSA). Several predictors of MNI/DIF impacted effect size differences in underlying severity across scale scoring methods. Notably, we observed MNI/DIF substantial enough to bias inferences on underlying PTSD severity for two groups: African Americans and incarcerated women. The findings highlight two issues raised elsewhere in the PTSD psychometrics literature: (a) bias in characterizing underlying PTSD severity and individual-level treatment outcomes when the psychometric model underlying total scores fails to fit the data and (b) higher latent severity scores, on average, when using *DSM-5* (net of MNI/DIF) criteria, by which multiple factors (e.g., Criterion A discordance across *DSM* editions, changes to the number/type of symptom clusters, changes to the symptoms themselves) may have impacted severity scoring for some patients.

The field-wide standard method of characterizing the underlying severity of psychiatric disorders, including posttraumatic stress disorder (PTSD; e.g., King et al., 1998; Weathers et al., 2018), utilizes the total item or symptom score on a measure, where the quantitative values assigned to different ordered categories on Likert-scaled individual items are summed. The use of total scores is ubiquitous in PTSD research and clinical settings, where total scores from semistructured interviews or self-report measures are often used both as an outcome measure for differentiating average treatment arm differences in changes over time in randomized controlled trials and to maximally distinguish a probable PTSD diagnosis based on a screening cut-off score (e.g., Coffey et al., 2006).

Despite the use of total scores, researchers and clinicians have long recognized that total scores are problematic from both a psychometric perspective and, more importantly, a clinical perspective (Bauer & Curran, 2015; Campbell, 1960; McNeish & Wolf, 2020). From a psychometric perspective, a total score psychometric model is inconsistent with the considerable literature showing that PTSD symptoms have different clinical “weights” and can contribute differentially to the quantification of an underlying construct, either as weights as factor loadings in linear or nonlinear confirmatory factor analysis (CFA) models (Contractor et al., 2018; Lee et al., 2019; Savedra, Morgan-López, Hien, Back, et al., 2021) or as discrimination parameters from item response theory (IRT) models (Morgan-López, Killeen, et al., 2020; Silverstein et al., 2020). The parallel clinical analog to this psychometric perspective is to reflect the clinical view that some symptoms matter more than others (Bourne et al., 2013).

The total score problem as it relates to underlying PTSD severity was illustrated by Franklin et al. (2015), who noted that, for example, a total PTSD symptom count of six elicits 12,360 possible symptom combinations—although, in practice, there are considerably fewer observed combinations—but many of these combinations likely do not reflect equivalent underlying clinical severities (Morgan-López, Killeen, et al., 2020). Related concerns have been noted with regard to the accepted total score reduction on the Clinician-Administered PTSD Scale (CAPS) needed to indicate clinically significant individual-level change (e.g., 10- or 15-point decrease on the CAPS-IV; Back et al., 2012; Cook, 2006; Hien et al., 2015;

McGovern et al., 2015; Mills et al., 2012; Petrakis, 2006; Sannibale et al., 2013). The concern, as noted by Saavedra, Morgan-López, Hien, Back, et al. (2021), is that a 10-point decrease will not always be reflective of equivalent reductions in a patient's underlying severity if a different combination of symptoms contributes to the decrease across patients.

Indeed, despite recommendations to the contrary, total scores continue to be used, likely because of their practical utility and, perhaps, a field-wide acceptance of the use of assessment sum scores as long as the measure of internal consistency (i.e., Cronbach's alpha) is perceived to be "good enough" (Campbell, 1960; Sijtsma, 2009). Yet, the use of total scores constitutes a critical psychometric decision, namely an assumption of a de facto CFA that assumes that the factor loadings are equal across items; the analogous model in IRT is known as the one-parameter logistic model (Andrich, 1978; McNeish & Wolf, 2020). However, such a model, whether under CFA or IRT, is, in fact, testable and, in practice, has failed to fit most data on psychological constructs in general and PTSD symptoms in particular (He et al., 2014). This can have practical consequences, such as an unacceptably high proportion of individuals with high underlying levels of PTSD severity not receiving a diagnosis or, in contrast, individuals with lower underlying severity being diagnosed (Morgan-López, Killeen, et al., 2020).

An additional concern is that the measurement parameters that link symptoms to latent underlying PTSD severity can vary across time (i.e., in longitudinal settings), populations, psychiatric comorbidities (e.g., comorbid alcohol use disorders and substance use disorders [AUD/SUDs]), and measures. This phenomenon is described as *measurement noninvariance* (MNI) in factor analyses or *differential item functioning* (DIF) in IRT. For example, MNI/DIF on PTSD frequency or intensity items or symptoms has been observed between civilian and veteran samples (Jamison-Eddinger & McDevitt-Murphy, 2017) and across race/ethnicity (Hoyt & Yeater, 2010) to such an extent that failing to correct for MNI/DIF in scale scores can affect inferences and effect sizes for group differences in underlying PTSD severity (Ruglass et al., 2020). Similar MNI/DIF effects have been observed across gender within civilian (Chung & Breslau, 2008) samples but, generally, have not shown sufficient MNI/DIF to lead to different inferences when using total scores versus factor analysis or IRT (FA/IRT) scores that account for MNI/DIF (Frankfurt et al., 2016). Other factors that have been shown to significantly impact total PTSD severity scores but do not appear to have been specifically examined for item- or symptom-level MNI/DIF include differences between (a) civilians and incarcerated populations (Piper & Berle, 2019) and (b) variation across PTSD criteria outlined in the fourth edition, text revision, and fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (i.e., *DSM-IV-TR* and *DSM-5*, respectively; Hoge et al., 2014; Kaysen et al., 2019).

The assumption in psychiatric assessment that measurement is consistent across populations when, in fact, this may not be the case stands in stark contrast to other fields. For example, in educational testing, items with MNI/DIF are allowed to have parameters that vary to maximize the precision of latent scale score estimates and distributions so they are equivalent across populations and represent consistent measures of an underlying aptitude (e.g., math proficiency) even in the presence of MNI/DIF (Dorans, 2007; Kim & DeCarlo, 2016); this is a different issue from whether these scores are differentially predictive of

other outcomes (e.g., Marini et al., 2019). Hoge et al. (2014) illustrated the fundamental problem of how measurement imprecision impacts the clinical interpretation of comparisons between the *DSM-IV* and *DSM-5* versions of the PTSD Checklist (PCL), noting that a "high percentage of soldiers who met criteria by one (*DSM*) definition did not meet the other criteria. Clinicians need to consider how to manage discordant outcomes" (p. 269). However, the extent to which this discordance in outcomes is a function of a myriad of measurement differences across *DSM* editions (e.g., changes in the symptom criteria) and different interviews (e.g., differences how questions are worded or in the translation of symptom frequency and intensity) versus legitimate differences for some patients in the conceptualization of PTSD across editions (Hoge et al., 2014; Kaysen et al., 2019) remains an open question.

The systematic examination of MNI/DIF and its impact in estimating underlying PTSD severity across multiple predictors would simultaneously help to answer the question, "Which predictors contribute to measurement bias in estimating PTSD severity to a level where it would change estimates of "true" group differences?" Many single-dataset studies do not have a sufficient sample size nor sufficient variability across populations (e.g., demographic characteristics, comorbidities) to answer these questions. However, such studies can now be conducted within the integrative data analysis (IDA) framework, where data from multiple studies can be harmonized and combined into a single dataset, with specific considerations for MNI/DIF when the measures themselves may vary across studies (Hussong et al., 2020) in addition to MNI/DIF across populations.

To our knowledge, the present investigation was the first large-scale IDA study of individuals who present for treatment with comorbid PTSD and AUD/SUDs, with datasets contributed by 30 participating principal investigators who are part of the Consortium on Addiction, Stress, and Trauma (CAST; Hien et al., 2019). In this study, we use the moderated non-linear factor analysis (MNLFA) framework (Bauer, 2017; Bauer & Hussong, 2009), which, unlike multiple-group factor analysis or IRT, allows for multiple, simultaneous sets of categorical and/or continuous predictors of MNI/DIF. The present study stands as the most comprehensive study of potential measurement bias in assessing PTSD severity across multiple demographic characteristics, population types (i.e., civilian, veteran, incarcerated), psychiatric comorbidities (e.g., depression, AUD/SUDs), and structured interview types among patients who present for treatment. These analyses illustrate whether variation in symptom weights, population- or measure-specific MNI/DIF, or both, at least partially account for differences in underlying PTSD severity.

METHOD

Participants and procedure

Study participants (N = 2,658) were part of an integrated dataset of 25 studies that were shared with Project Harmony (Hien et al., 2019) by members of the CAST. The governing institutional review board (IRB) deemed the project to be "human subjects exempt" given the use of deidentified secondary data. For the current analysis, these 25 studies represent a subset of 42 studies in Project Harmony that included item-level data from a semistructured interview measure. Other studies that were part of a larger combined IDA (Curran et

al., 2008, 2020) and individual-patient meta-analyses were excluded from this analysis if they (a) only included self-report PTSD measures (e.g., PCL, PTSD Symptom Scale–Self-Report Version [PSS-SR], Impact of Event Scale), (b) only submitted total scores for PTSD outcome measures, (c) were only allowed to submit aggregated summary data (i.e., means, standard deviations, correlations) by their IRBs based on the original consent form, or (d) had not submitted their data for inclusion into Project Harmony in time to be harmonized as part of this analysis. Table 1 shows the studies that were included in the current analysis along with information on sample sizes, population type, semistructured interview type, and the within-study percentages regarding gender and full PTSD diagnosis.

Measures

PTSD symptoms—Assessment items were harmonized (i.e., recoded to a common item-level standard) to binary indicators of symptom presence following recommendations by Bauer and Hussong (2009) for harmonizing disparate items for the same construct. The resulting proportions for the harmonized symptoms are shown in Table 2. Most RCTs in this analysis did not report interrater reliabilities in the original articles, but, among those that did, no study reported interrater reliability below .70.

The CAPS-IV (Blake et al., 1995) was used in 21 of the 25 studies in this analysis. The CAPS-IV assesses the frequency and intensity of the 17 *DSM-IV* PTSD symptoms respondents have experienced in the past month; the measure is also used to determine PTSD diagnostic status and disorder severity. The CAPS-IV has three symptom cluster subscales: Reexperiencing, Avoidance/Numbing, and Hyperarousal. A *DSM-IV* diagnosis of PTSD requires the presence of a Criterion A traumatic event, at least one reexperiencing symptom, three avoidance/numbing symptoms, and two hyperarousal symptoms. For harmonizing CAPS-IV items to a common metric relative to the other interview measures (i.e., CAPS-5, PSS-Interview Version [PSSI-I] for *DSM-IV*), we employed the convention for converting frequency and intensity items to binary *DSM* symptoms based on a symptom frequency rating of at least once or twice in the previous month and a moderate or higher level of symptom intensity (Blake et al., 1990; Weathers et al., 1999).

CAPS-5—The CAPS-5 (Weathers et al., 2018) was used in two of the 25 studies in this analysis. For the 20 items that capture symptom severity based on *DSM-5* Criteria B–E, harmonization to a common metric was conducted based on the item-level rule of a severity score of 2 (*moderate*) or higher based on *DSM-5* PTSD criteria.

PSS-I—The PSS-I (Foa et al., 1993) was used in two of the 25 studies in this analysis. For the 17 items that capture symptom severity based on *DSM-IV* Criteria B–D, harmonization to a common metric was performed based on the item-level rule of a severity score of 2 (*2 to 4 times per week/somewhat*) or higher based on *DSM-IV* PTSD criteria.

Joint predictors of MNI/DIF and underlying PTSD severity

Predictor variables that were examined as both predictors of MNI/DIF across symptoms and latent underlying PTSD severity included the following demographic variables: age, gender, race/ethnicity, educational attainment, marital status, and population type (i.e., civilian,

veteran, currently incarcerated). Two dummy variable indicators were included for the measure (i.e., CAPS-IV, CAPS-5, PSS-I-IV), with CAPS-IV as the reference measure. Other psychiatric predictors that were commonly available across datasets included the number of days of alcohol use in the past month, any past-month cocaine use, any past-month opiate use, any past-month stimulant use, any past-month sedative use, concomitant psychiatric medications, and current depression diagnosis (e.g., Structured Clinical Interview for *DSM-5* diagnosis, severe depression as indicated by the Beck Depression Inventory). Descriptive statistics for predictors are shown in Table 3.

Data analysis

The initial set of analyses involved tests of unidimensionality, conducted using means-and-variance–adjusted weighted least squares (WLSMV) estimation for categorical indicators in *M*plus (Version 8; Muthén & Muthén, 1998–2017). A general single-factor model was fit in addition to a restricted single-factor model with equality constraints on factor loadings; the latter model was fit per the recommendations of McNeish and Wolf (2020) to formally test whether the model that is assumed when using PTSD total scores actually fits the data.

Next, a series of MNLFA models were fit to examine MNI/DIF in separate models for each PTSD symptom such that each set of predictors was assessed with regard to whether they contributed to significant MNI/DIF on each symptom above and beyond their effects on "true" latent underlying PTSD severity; it is this process of examining MNI/DIF across the three interview measures in particular (CAPS-IV, CAPS-5, PSS-I-IV) where the quality of the item harmonization process is assessed. The MNI/DIF parameters that were significant at $p \leq .05$ were then retained for a global MNLFA model. The parameters that remained significant in the global model were retained for the final MNLFA scale score estimation model. From the final MNLFA model, the MNI/DIF parameters that were significant are shown in Tables 4 and 5, and the predictors of underlying PTSD severity are described in the Results section. Predictors of underlying PTSD severity under the "total score" analog model are also presented in the Results section for comparison. *M*plus code is available in the Supplementary Materials, and additional detail on MNLFA modeling can be found in Bauer (2017) and Saavedra, Morgan-López, Hien, Back, et al. (2021).

RESULTS

Preliminary tests of model fit

The model for the conventional test of unidimensionality fit adequately, with the results showing that a single factor underlying the harmonized PTSD symptoms, comparative fit index (CFI) = .90, root mean square error of approximation (RMSEA) = .052, 95% CI [.048, .055], meeting the standard for essential unidimensionality (Millsap & Kwok, 2004). The total score analog (TSA) model, wherein factor loadings were constrained to equality, predictably failed in fitting the data, CFI = .786, RMSEA = .072, 95% CI [.068, .075], yet this is the psychometric model that underlies the use of symptom counts in the *DSM* (He et al., 2014; Morgan-López, Killeen, et al., 2020) and total scores in the majority of PTSD research.

MNLFA

Item parameters from the final MNLFA model are presented in Tables 4 and 5, with all predictors of MNI/DIF and PTSD severity centered so all comparisons are made against the sample means. Given the multitude of predictors of MNI/DIF examined, only three of the 21 cross-*DSM* PTSD symptoms showed no DIF across any predictor (i.e., "empirical anchor" symptoms): psychological cues, negative beliefs, and horror/shame/guilt. The emergence of psychological cues as an anchor is particularly important because it (a) is the only symptom that had the same measurement properties across the full sample, cutting across multiple populations, comorbidities, and interview measures, as negative beliefs and horror/shame/guilt were only relevant for the subsample with CAPS-5 data, and (b) had the largest loading/discrimination parameter and, thus, the largest clinical weight. Item information functions (IIF), the FA/IRT analog to item reliabilities, are shown in Figure 1 for the six symptoms with the highest IIF peaks, which were driven by the relative sizes of the factor loadings and are consistent with other IRT-based analyses of PTSD symptoms (e.g., Silverstein et al., 2020).

Threshold/difficulty MNI/DIF

Threshold/difficulty parameters refer to the level of PTSD severity required to reach a symptom endorsement likelihood of 50%. For the threshold/difficulty parameters (Table 4), key predictors that showed statistically significant MNI/DIF on at least four symptoms, compared to the overall sample average thresholds/difficulty parameters, included age, African American ancestry, being incarcerated, PSSI-I-IV assessment, and CAPS-5 assessment. The average threshold/difficulty DIF across symptoms, taken as the sum of the MNI/DIF parameters divided by the number of symptoms with MNI/DIF, that exceed a Cohen's d value of |0.20| included African American ancestry, $d = -0.34$; college education, $d = -0.54$; incarceration, $d = -0.37$; PSS-I-IV assessment, $d = -0.33$; and CAPS-5 assessment, $d = -0.30$.

Loading/discrimination MNI/DIF

For the factor loading/discrimination parameters (Table 5), key predictors that showed statistically significant MNI/DIF (i.e., variation in clinical weights) included gender, African American ancestry, incarceration status, and assessment using the CAPS-5. With regard to specific symptoms, men showed a significantly larger-than-average loading/discrimination parameter for foreshortened future. African American individuals showed significantly smaller-than-average loading/discrimination parameters for intrusive recollections, activity avoidance, and hypervigilance. Incarcerated women populations showed a significantly larger-than-average loading/discrimination parameter for nightmares. Participants who were assessed using the CAPS-5 showed a significantly smaller-than-average loading/discrimination parameter for thought avoidance and a significantly larger-than-average loading/discrimination parameter for diminished interest. The final PTSD severity scoring model under MNLFA, with DIF incorporated across all other items under a partial invariance model (Millsap & Kwok, 2004), fit significantly better than a base model with varying factor loadings across symptoms but no MNI/DIF, $\chi^2(41, N = 2{,}658) = 806.28$, $p < .001$.

Predictors of latent underlying severity: MNLFA model

Due to violations of the assumption of independence of observations (i.e., patients within a study reporting more similar underlying PTSD severity than patients across studies), corrections for study-level clustering were made to model standard errors as part of robust maximum likelihood estimation in *M*plus. After accounting for both MNI/DIF and study-level clustering, several factors emerged as predictors of higher levels of underlying PTSD severity that were either statistically significant, exceeded an absolute Cohen's *d* effect size of .20, or both. These predictors included younger age, $B = -0.006$, $SE = .002$, $z = -2.582$, $p = .013$, $d = 0.07$; being a veteran, $B = 0.155$, $SE = .067$, $z = 2.322$, $p = .024$, $d = 0.16$; being incarcerated, $B = 0.303$, $SE = .176$, $z = 1.715$, $p = .078$, $d = 0.30$; past-month opiate use, $B = 0.197$, $SE = .097$, $z = 2.030$, $p = .041$, $d = 0.20$; past-month sedative use, $B = 0.208$, $SE = .095$, $z = 2.183$, $p = .024$, $d = 0.21$; pretreatment psychiatric medication use, $B = 0.399$, $SE = .103$, $z = 3.868$, $p < .001$, $d = 0.40$; comorbid depression, $B = 0.577$, $SE = .058$, $z = 9.984$, $p < .001$, $d = 0.58$; and assessment using criteria from the *DSM-5* criteria versus the *DSM-IV*, $B = 0.488$, $SE = .113$, $z = 4.298$, $p < .001$, $d = 0.49$. Participants who were assessed using the PSS-I-IV demonstrated significantly lower levels of underlying PTSD severity, $B = -0.835$, $SE = .11$, $z = -4.878$, $p < .001$, $d = 0.84$, even after accounting for MNI/DIF across diagnostic systems (see Tables 4 and 5).

Predictors of latent underlying severity: TSA model

Although the TSA model demonstrated an extremely poor fit, we present a partial set of results that would have been analogous to the results observed without measurement modeling under MNLFA. Of the predictors of underlying severity that were significant or meaningful under MNLFA, several would have had roughly equivalent effect sizes under the TSA model (i.e., $(d_{MNLFA} - d_{TSA}) / d_{MNLFA} < |.20|$), including age, veteran status, past-month opiate use, past-month sedative use, pretreatment psychiatric medication use, and comorbid depression. Although participants who were assessed using *DSM-5* criteria still demonstrated significantly higher levels of underlying severity in the TSA model, $B = 0.294$, $SE = .089$, $z = 3.309$, $p = .001$, $d = 0.29$, these individuals showed a substantial decrease in *DSM-5* assessment effect size across measurement models (i.e., $d = 0.49$ vs. $d = 0.29$), suggesting that unmodeled *DSM-5* MNI/DIF in the TSA model would have had a substantial impact on exaggerating the differences in underlying PTSD severity.

Two predictors demonstrated different inferences altogether between the MNLFA and TSA models: African American ancestry and incarceration status. Under the TSA model, African American participants would have been estimated to have a significantly lower-than-average level of PTSD severity upon treatment entry, $B = -0.106$, $SE = .054$, $z = -1.983$, $p = .047$, $d = 0.11$; however, this difference was nonsignificant in the MNLFA model, $p = .45$, $d = -0.04$), suggesting this effect was observed in the TSA model due to unmodeled measurement artifacts stemming from (a) threshold/difficulty MNI/DIF on activity avoidance, inability to recall, foreshortened future and concentration problems (see Table 4), and (b) loading/discrimination MNI/DIF on intrusive recollections, activity avoidance, and hypervigilance (see Table 5; see also Ruglass et al., 2020). In addition, under the TSA model, incarcerated women participants showed nonsignificant differences from the average level of PTSD severity, $p = .513$, $d = 0.08$, that, under MNLFA, showed greater-

than-average PTSD severity before treatment entry, $d = 0.30$. This difference is attributable to threshold/difficulty DIF on six symptoms (see Table 4) as well as loading/discrimination MNI/DIF showing that nightmares had a substantially higher clinical weight for incarcerated women compared to the sample average under MNLFA, whereas this MNI/DIF is ignored under TSA (see Table 5).

DISCUSSION

The calculation of total scores for psychiatric outcome measures constitutes a critical, and potentially erroneous, psychometric decision that has ripple effects throughout the execution of an outcomes analysis, interpretation of findings, delineation of the underlying severity of clinical distress, and recommendations for treatment course. This initial integrative data analysis from Project Harmony attempted to consolidate work that has either explicitly examined MNI/DIF across several predictors (Engdahl et al., 2011; Hoyt & Yeater, 2010; Jamison-Eddinger & McDevitt-Murphy, 2017) or has raised the need to further examine MNI/DIF in other predictor domains (Grella et al., 2013; Hoge et al., 2014; Kaysen et al., 2019). MNI/DIF analysis, and the resulting scale scores from M*plus* that account for differences in the relative weighting of symptoms, can now be conducted with multiple predictors simultaneously under the MNLFA framework (Bauer, 2017) in ways that could not previously be done in multiple-group MNI/DIF analysis.

Most notably, our analyses revealed that MNI/DIF demonstrated the largest impacts among African American participants from both genders and incarcerated women, with results sufficient to change inferences regarding underlying PTSD severity in these populations. The variety of differences found across many *DSM* PTSD symptoms in MNI/DIF for African American individuals is notable but not new. Indeed, a previous multisite clinical trial analysis, which was the first analysis of its kind within the literature on PTSD/SUD clinical trials (Ruglass et al., 2020), reported similar measurement nonequivalence among African American women for six CAPS symptoms. The authors found that three symptoms in particular (i.e., inability to recall, foreshortened future, concentration problems) drove differences in underlying severity across scoring methods; adjustments for MNI/DIF under MNLFA would properly adjust the relative standing on underlying PTSD severity among African American individuals compared to the rest of the sample. Another important and understudied population is incarcerated individuals. In the present sample, we had access to two separate studies that included incarcerated women (Zlotnick et al., 2003, 2009). There are no previously examined studies of MNI/DIF for PTSD in this particular population; therefore, the present analysis points to the potential implications for measurement noninvariance. We note that although the differences in effect size compared to the rest of the sample differed by the method of scale score estimation, the incarcerated subsample only represented 2.8% of the total sample. Nevertheless, these findings suggest that it may be critical to disentangle the proportion of variation in underlying severity and treatment outcomes due to true differences in PTSD severity versus variation in outcomes due to unmodeled MNI/DIF, which has implications for mischaracterizing individuals' underlying levels of disorder severity (Ruglass et al., 2020) and, possibly, their diagnosis (Morgan-López, Killeen, et al., 2020).

Head-to-head comparisons of *DSM-IV* and *DSM-5* with regard to factor structure and diagnostic status are beginning to emerge (Hoge et al., 2014; Kaysen et al., 2019), and this study extends this literature by estimating the effects of interview type differences attributable to MNI/DIF versus underlying differences in severity. Our findings suggest the existence of both (a) MNI/DIF across versions of the *DSM* and (b) differences in underlying PTSD severity that are reduced, but still significantly different, after accounting for MNI/DIF in scale score estimation. The finding that CAPS-5 severity, even after accounting for partial MNI/DIF due to the measures used and other factors, is significantly higher than CAPS-IV severity is particularly notable and appears consistent with an increase in diagnostic rates under *DSM-5* in head-to-head comparisons with *DSM-IV* (Kaysen et al., 2019).

However, such general differences (e.g., changes in symptom criteria) between *DSM-IV* versus *DSM-5*, in addition to differences in measurement (e.g., wording differences, sequencing of prompts, handling of frequency and intensity of symptoms), cannot be disentangled from whether an index traumatic event would have met Criterion A under both *DSM-IV* and *DSM-5* versus meeting the criterion in one edition but not the other; previous research has indicated such discrepancies between the two editions of the *DSM* are present in an estimated 45% of patients (Hoge et al., 2014; Kaysen et al., 2019). This suggests that for any patient who is Criterion A– discrepant, legitimate differences in estimated PTSD severity should be expected under the different editions of the *DSM*, independent of measurement. However, in the present study, only 15% of participants had data on index trauma events submitted as part of Project Harmony's individual patient meta-analysis (Saavedra, Morgan-López, Hien, López-Castro, et al., 2021), limiting our ability to disentangle how much of the differences in underlying disorder severity for the *DSM-IV* versus the *DSM-5* were due to differences in the measures and other factors from cross-*DSM* stability or shifts in Criterion A. Investigators who have access data on participants' index traumatic events and similar secondary data for cross-*DSM* data integration, as well as those who have data for head-to-head single-dataset comparisons of *DSM-IV* and *DSM-5* symptom assessment among the same patients, are encouraged to look more closely at further disentangling the explanations for the differences in underlying severity across *DSM-IV* and *DSM-5*.

The present study directly builds upon the findings from multiple studies of MNI/DIF across multiple factors that could not previously be examined simultaneously within an MNI/DIF analysis (Caldas et al., 2020; Contractor et al., 2018; Engdahl et al., 2011; Jamison-Eddinger & McDevitt-Murphy, 2017; Ruglass et al., 2020), using a larger, more diverse sample that included multiple comorbidities (e.g., AUD/SUD, depression), women who were incarcerated, men, veterans, and civilians. Despite these advantages, there were some noted limitations. First, only two studies in this IDA consisted of women who were incarcerated (Zlotnick et al., 2003, 2009) which prevents our findings from generalizing to incarcerated men. Additional work in this area is paramount given that as recently as December 2020, 93.3% of incarcerated populations identified as male in the United States (Federal Bureau of Prisons, 2021), and 6.2% of these men were diagnosed with PTSD (Baranyi et al., 2018). Similarly, only two studies in our analysis utilized the CAPS-5 (Norman et al., 2018; Vujanovic et al., 2018), and only two studies used the PSS-I (Foa

et al., 2013; Schafer et al., 2019). Although these earlier studies examined factor structure among items, but not symptoms, and diagnostic rate differences (Hoge et al., 2014; Kaysen et al., 2019), the current study highlights the additional needs for understanding the impact of MNI/DIF on differences in underlying disorder severity as well as differences in change over time across measures integrated under IDA treatment studies (Hien et al., 2020; Saavedra, Morgan-López, Hien, López-Castro, et al., 2021).

Further, the present findings may not generalize to PTSD samples without AUD/SUD, and we encourage additional IDA studies across "pure" PTSD samples as well as other types of comorbid samples (e.g., PTSD/depression). However, this study could provide a set of guideposts for what may be expected. First, the general pattern of the measurement properties (e.g., relative rank and size of item thresholds and factor loadings) was consistent with other studies in noncomorbid samples (e.g., Silverstein et al., 2020). Second, if any particular substance would have impacted the measurement properties of the harmonized PTSD symptoms (i.e., substance-induced amplification), this may have been reflected in the MNI/DIF parameters for the alcohol and any other drug use (AOD) indicators. None of the substance use disorder indicators showed MNI/DIF, and only one symptom showed significant MNI/DIF on alcohol use (i.e., flashbacks), although there are no noncomorbid participants against which MNI/DIF and general underlying severity can be compared. PTSD/AOD comorbidity likely will have larger implications for examining treatment efficacy on PTSD outcomes, as it is well-known that comorbid samples typically demonstrate weaker treatment outcomes and are less likely to complete treatment (e.g., Hien et al., 2012); however, the comparative effectiveness of treatments for PTSD/AOD has been understudied and is an integral component of Project Harmony (Hien et al., 2019; Saavedra, Morgan-López, Hien, López-Castro, et al., 2021).

This study examined the impact of MNI/DIF in a large, diverse dataset of individuals with PTSD and co-occurring substance use disorders, leveraging contributions of data from 30 participating principal investigators in CAST. The present results demonstrate that some subpopulations are misrepresented by summed scores from semistructured interviews for PTSD, with large enough bias to change effect sizes for group differences in a number of cases as well as large enough bias to change inferences for both individuals with African American ancestry and incarcerated women. The cost of such imprecision is enormous, with misrepresentation in PTSD severity having a direct impact on incorrect clinical decision-making. Thus, the field must make a collective effort not just to complete factor analysis or IRT MNI/DIF studies on PTSD assessments but to use the resulting scale scores with patients in research and clinical practice. For example, it is likely possible to build mobile apps that can be used to score underlying PTSD severity where the FA/IRT item parameters (e.g., Tables 4 and 5) would be "under the hood" of the app, as is the case with computerized adaptive psychiatric scoring modules, such as the Patient-Reported Outcomes Measurement Information System (PROMIS; Cella et al., 2010) and CAT-MH (Gibbons et al., 2019). This would substantially contribute to improving the accuracy and precision of routine clinical assessment without burdening the end-user with complex scale scoring methodology by reflecting what Bourne et al.'s (2013) observation that some symptoms matter more than others (e.g., Figure 1). In the interim, the findings regarding CAPS total scores for African American individuals and incarcerated women should be interpreted with extreme caution.

Recent work has shown that using FA/IRT–estimated scale scores, in combination with methods to delineate clinical from nonclinical distributions (Jacobson & Truax, 1991), can reduce individual-level clinical decision-making errors as well as errors in judgment regarding clinically significant change, with research suggesting that such errors are made for an estimated 1 out of 4 patients when sum scores are used (Morgan-López, Killeen, et al., 2020; Saavedra, Morgan-López, Hien, Back, et al., 2021). The uptake of such practice may allow clinicians to better ensure that PTSD assessment accurately captures variation in the presentation of the disorder among different groups, which will have downstream effects on clinical decision-making and patient health, such as tailoring treatments to specific combinations of symptoms based on symptom weighting. The field's future collective work will need to focus on continuing to address such measurement discordance so there is accuracy and precision in assessment that is commensurate with what patients deserve.

References for IDA datasets are indicated with an *

OPEN PRACTICES STATEMENT

This study was preregistered with the International Prospective Register of Systematic Reviews (PROSPERO 2019 CRD42019146678). The study protocol was published as Saavedra et al., 2021, in *Contemporary Clinical Trials*). Individual-level data will be made available following NIAAA guidance whenever possible. Due to some confidentiality agreements in five of the original studies, a smaller harmonized and deidentified dataset will be made available for sharing at the completion of the current planned and funded aims.

Acknowledgments

The work presented in this manuscript was supported by grants from the National Institute on Drug Abuse (NIDA Clinical Trials Network Protocol 0015, Denise A. Hien, PI; T32DA007288, Jacqueline F. McGinty, PI) and the National Institute on Alcohol Abuse and Alcoholism (R01AA025853; Denise A. Hien and Antonio A. Morgan-López, MPIs). This study was registered with the International Database of Prospectively Registered Systematic Reviews (PROSPERO #2019 CRD42019146678).

REFERENCES

Andrich D (1978). A rating formulation for ordered response categories. Psychometrika, 43(4), 561–573. 10.1007/bf02293814

Asparouhov T, & Muthen B (2010) Bayesian analysis using Mplus: Technical Implementation [Technical report]. https://www.statmodel.com

*Back SE, Killeen T, Badour CL, Flanagan JC, Allan NP, Santa Ana E, … Brady KT (2019). Concurrent treatment of substance use disorders and PTSD using prolonged exposure: a randomized clinical trial in military veterans. Addictive Behaviors, 90, 369–377. [PubMed: 30529244]

*Back SE, Killeen T, Foa EB, Santa Ana EJ, Gros DF, & Brady KT (2012). Use of an integrated therapy with prolonged exposure to treat PTSD and comorbid alcohol dependence in an Iraq veteran. American Journal of Psychiatry, 169(7), 688–691. 10.1176/appi.ajp.20n.n091433 [PubMed: 22760188]

Back SE, McCauley JL, Korte KJ, Gros DF, Leavitt V, Gray KM, Hammer MB, DeSantis SM, Malcolm R, Brady KT, & Kalivas PW (2016). A double-blind, randomized, controlled pilot trial of N-acetylcysteine in veterans with posttraumatic stress disorder and substance use disorders. The Journal of Clinical Psychiatry, 77(11), 1439–1446. 10.4088/JCP.15m10239

Baranyi G, Cassidy M, Fazel S, Priebe S, & Mundt AP (2018). Prevalence of posttraumatic stress disorder in prisoners. Epidemiologic Reviews, 40(1), 134–145. 10.1093/epirev/mxx015 [PubMed: 29596582]

Bauer DJ (2017). A more general model for testing measurement invariance and differential item functioning. Psychological Methods, 22(3), 507–526. 10.1037/met0000077 [PubMed: 27266798]

Bauer D, & Curran P (2015). The discrepancy between measurement and modeling in longitudinal data analysis. In Harring JR, Stapleton LM, & Beretvas SN (Eds.), Advances in multi-level modeling for educational research: Addressing practical issues found in real-world applications (pp. 3–38). Information Age Publishing.

Bauer DJ, & Hussong AM (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. Psychological Methods, 14(4), 101–125. 10.1037/a0017642 [PubMed: 19485624]

*Boden MT, Kimerling R, Jacobs-Lentz J, Bowman D, Weaver C, Carney D, Walser R, & Trafton JA (2012). Seeking Safety treatment for male veterans with a substance use disorder and post-traumatic stress disorder symptomatology. Addiction, 107(3), 578–586. 10.1111/j.1360-0443.2011.03658.x [PubMed: 21923756]

Bourne C, Mackay CE, & Holmes EA (2013). The neural basis of flashback formation: the impact of viewing trauma. Psychological Medicine, 43(7), 1521–1532. 10.1017/S0033291712002358 [PubMed: 23171530]

Calhoun PS, Hertzberg JS, Kirby AC, Dennis MF, Hair LP, Dedert EA, & Beckham JC (2012). The effect of draft *DSM-V* criteria on posttraumatic stress disorder prevalence. Depression and Anxiety, 29(12), 1032–1042. 10.1002/da.2201 [PubMed: 23109002]

Campbell DT (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. American Psychologist, 15(8), 546–553. 10.1037/h0048255

Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, Amtmann D, Bode R, Byusse D, Choi S, Cook K, DeVellis R, DeWalt D, Fries JFF, Gershon R, Hahn EA, Lai J-S, Pilkonis P, Revicki D, … Hays R (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. Journal of Clinical Epidemiology, 63(11), 1179–1194. 10.1016/j.jclinepi.2010.04.011 [PubMed: 20685078]

Chung H, & Breslau N (2008). The latent structure of post-traumatic stress disorder: tests of invariance by gender and trauma type. Psychological Medicine, 38(4), 563–573. 10.1017/s0033291707002589 [PubMed: 18325132]

Coffey SF, Gudmundsdottir B, Beck JG, Palyo SA, & Miller L (2006). Screening for PTSD in motor vehicle accident survivors using the PSS-SR and IES. Journal of Traumatic Stress, 19(1), 119–128. 10.1002/jts.20106 [PubMed: 16568464]

Contractor AA, Caldas SV, Dolan M, Lagdon S, & Armour C (2018). PTSD's factor structure and measurement invariance across subgroups with differing count of trauma types. Psychiatry Research, 264, 76–84. 10.1016/j.psychres.2018.03.065 [PubMed: 29627700]

*Cook JM, Walser RD, Kane V, Ruzek JI, & Woody G (2006). Dissemination and feasibility of a cognitive-behavioral treatment for substance use disorders and posttraumatic stress disorder in the Veterans Administration. Journal of Psychoactive Drugs, 38(1), 89–92. 10.1080/02791072.2006.10399831 [PubMed: 16681179]

Dickerson DL, Spear S, Marinelli-Casey P, Rawson R, Li L, & Hser YI (2011). American Indians/Alaska Natives and substance abuse treatment outcomes: Positive signs and continuing challenges. Journal of Addictive Diseases, 30(1), 63–74. 10.1080/10550887.2010.531665 [PubMed: 21218312]

Dorans NJ (2007). Linking scores from multiple health outcome instruments. Quality of Life Research, 16(1), 85–94. 10.1007/s11136-006-9155-3 [PubMed: 17286198]

Elhai JD, Miller ME, Ford JD, Biehn TL, Palmieri PA, & Frueh BC (2012). Posttraumatic stress disorder in *DSM-5*: Estimates of prevalence and symptom structure in a nonclinical sample of college students. Journal of Anxiety Disorders, 26(1), 58–64. 10.1016/jjanxdis.2011.08.013 [PubMed: 21944437]

Foa E, Riggs DS, Dancu CV, Constance V, & Rothbaum BO (1993). Reliability and validity of a brief instrument for assessing posttraumatic stress disorder. Journal of Traumatic Stress, 6(4), 459–473. 10.1002/jts.2490060405

*Foa EB, Yusko DA, McLean CP, Suvak MK, Bux DA, Oslin D, O'Brien CP, Imms P, Riggs DS, & Volpicelli J (2013). Concurrent naltrexone and prolonged exposure therapy for patients with comorbid alcohol dependence and PTSD: A randomized clinical trial. JAMA, 310(5), 488–495. 10.1001/jama.2013.8268 [PubMed: 23925619]

Frankfurt SB, Armour C, Contractor AA, & Elhai JD (2016). Do gender and directness of trauma exposure moderate PTSD's latent structure? Psychiatry Research, 245, 365–370. 10.1016/j.psychres.2016.08.049 [PubMed: 27591411]

Franklin CL, Piazza V, Chelminski I, & Zimmerman M (2015). Defining subthreshold PTSD in the *DSM-IV* literature: A look toward *DSM-5*. The Journal of Nervous and Mental Disease, 203(8), 574–577. 10.1097/NMD.0000000000000332 [PubMed: 26133273]

Galatzer-Levy IR, & Bryant RA (2013). 636,120 ways to have posttraumatic stress disorder. Perspectives on Psychological Science, 8(6), 651–662. 10.1177/2F1745691613504115 [PubMed: 26173229]

Gibbons RD, & deGruy FV (2019). Without wasting a word: Extreme improvements in efficiency and accuracy using computerized adaptive testing for mental health disorders (CATMH). Current Psychiatry Reports, 21(8), 67. 10.1007/s11920-019-1053-9 [PubMed: 31264098]

*Haller M, Norman SB, Cummins K, Trim RS, Xu X, Cui R, Allard CB, Brown SA, & Tate SR (2016). Integrated cognitive behavioral therapy versus cognitive processing therapy for adults with depression, substance use disorder, and trauma. Journal of Substance Abuse Treatment, 62, 38–48. 10.1016/j.jsat.2015.n.005 [PubMed: 26718130]

He Q, Glas CA, & Veldkamp BP (2014). Assessing impact of differential symptom functioning on post-traumatic stress disorder (PTSD) diagnosis. International Journal of Methods in Psychiatric Research, 23(2), 131–141. 10.1002/mpr.1417 [PubMed: 24436035]

*Hien DA, Cohen LR, Miele GM, Litt LC, & Capstick C (2004). Promising treatments for women with comorbid PTSD and substance use disorders. American Journal of Psychiatry, 161(8), 1426–1432. 10.1176/appi.ajp.16L8.1426 [PubMed: 15285969]

*Hien DA, Levin FR, Ruglass LM, López-Castro T, Papini S, Hu M-C, Cohen LR, & Herron A (2015). Combining seeking safety with sertraline for PTSD and alcohol use disorders: A randomized controlled trial. Journal of Consulting and Clinical Psychology, 83(2), 359–369. 10.1037/a0038719 [PubMed: 25622199]

Hien DA, Morgan-López AA, Ruglass LM, Saavedra LM, Fitzpatrick S, Back SE, Killeen TK, & Norman SB (2019). Project Harmony: A systematic review and meta-analysis of individual patient data of behavioral and pharmacologic trials for comorbid posttraumatic stress, alcohol, and other drug use disorders. https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42019146678

*Hien DA, Wells EA, Jiang H, Suarez-Morales L, Campbell AN, Cohen LR, Miele GM, Killeen T, Brigham GS, Zhang Y, Hansen C, Hodgkins C, Hatch-Maillette M, Brown C, Kulaga A, Kristman-Valente A, Chu C, Sage R, Robinson JA, … Nunes EV (2009). Multisite randomized trial of behavioral interventions for women with co-occurring PTSD and substance use disorders. Journal of Consulting and Clinical Psychology, 77(4), 607–619. 10.1037/a0016227 [PubMed: 19634955]

Hoge CW, Riviere LA, Wilk JE, Herrell RK, & Weathers FW (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(4), 269–277. 10.1016/S2215-0366(14)70235-4 [PubMed: 26360860]

Hoyt T, & Yeater EA (2010). Comparison of posttraumatic stress disorder symptom structure models in Hispanic and White college students. Psychological Trauma: Theory, Research, Practice, and Policy, 2(1), 19–30. 10.1037/a0018745

Hussong AM, Cole VT, Curran PJ, Bauer DJ, & Gottfredson NC (2020). Integrative data analysis and the study of global health. In Chen X & Chen D-D (Eds.), Statistical methods for global health and epidemiology (pp. 121–158). Springer.

Jabrayilov R, Emons WHM, & Sijtsma K (2016). Comparison of classical test theory and item response theory in individual change assessment. Applied Psychological Measurement, 40(8), 559–572. 10.1177/2F0146621616664046 [PubMed: 29881070]

Jamison-Eddinger JR, & McDevitt-Murphy ME (2017). A confirmatory factor analysis of the PTSD Checklist 5 in veteran and college student samples. Psychiatry Research, 255, 219–224. 10.1016/j.psychres.2017.05.035 [PubMed: 28578182]

Kaysen D, Rhew IC, Bittinger J, Bedard-Gilligan M, Garberson LA, Hodge KA, Nguyen AJ, Logan DE, Dworkin ER & Lindgren KP (2019). Prevalence and factor structure of PTSD in *DSM-5* versus *DSM-IV* in a national sample of sexual minority women. Journal of Interpersonal Violence, 36(21–22), NP12388–NP12410. 10.1177/0886260519892960 [PubMed: 31833796]

Kim Y, & DeCarlo LT (2016). Evaluating equity at the local level using bootstrap tests. [Research Report 2016-4, Rep. No. 4]. College Board.

King DW, Leskin GA, King LA, & Weathers FW (1998). Confirmatory factor analysis of the Clinician-Administered PTSD Scale: Evidence for the dimensionality of posttraumatic stress disorder. Psychological Assessment, 10(2), 90–96. 10.1037/1040-3590.10.2.90

Lee DJ, Bovin MJ, Weathers FW, Palmieri PA, Schnurr PP, Sloan DM, Keane TM, & Marx BP (2019). Latent factor structure of *DSM–5* posttraumatic stress disorder: Evaluation of method variance and construct validity of novel symptom clusters. Psychological Assessment, 31(1), 46–58. 10.1037/pas0000642 [PubMed: 30113182]

Marini JP, Westrick PA, Young L, Ng H, Shmueli D, & Shaw EJ (2019). Differential validity and prediction of the SAT™: Examining first-year grades and retention to the second year. College Board.

*McDevitt-Murphy ME, Murphy JG, Williams JL, Monahan CJ, Bracken-Minor KL, & Fields JA (2014). Randomized controlled trial of two brief alcohol interventions for OEF/OIF veterans. Journal of Consulting and Clinical Psychology, 82(4), 562–568. 10.1037/a0036714 [PubMed: 24773573]

*McGovern MP, Lambert-Harris C, Alterman AI, Xie H, & Meier A (2011). A randomized controlled trial comparing integrated cognitive behavioral therapy versus individual addiction counseling for co-occurring substance use and posttraumatic stress disorders. Journal of Dual Diagnosis, 7(4), 207–227. 10.1080/15504263.2011.620425 [PubMed: 22383864]

*McGovern MP, Lambert-Harris C, Xie H, Meier A, McLeman B, & Saunders E (2015). A randomized controlled trial of treatments for co-occurring substance use disorders and post-traumatic stress disorder. Addiction, 110(7), 1194–1204. 10.1111/add.12943 [PubMed: 25846251]

McNeish D, & Wolf MG (2020). Thinking twice about sum scores. Behavior Research Methods, 52(6), 2287–2305. 10.3758/s13428-020-01398-0 [PubMed: 32323277]

*Mills KL, Teeson M, Back SE, Brady KT, Baker AL, Hopwood S, Sannibale C, Barrett EL, Merz S, Rosenfeld J, & Ewer PL (2012). Integrated exposure-based therapy for co-occurring post-traumatic stress disorder and substance dependence: A randomized controlled trial. Journal of the American Medical Association, 308(7), 690–699. 10.1001/jama.2012.9071 [PubMed: 22893166]

Millsap RE, & Kwok O-M (2004). Evaluating the impact of partial factorial invariance on selection in two populations. Psychological Methods, 9(1), 93–115. 10.1037/1082-989X.91.93 [PubMed: 15053721]

Morgan-López AA, Killeen TK, Saavedra LM, Hien DA, Fitzpatrick S, Ruglass LM, & Back SE (2020) Crossover between diagnostic and empirical categorizations of full and subthreshold PTSD. Journal of Affective Disorders, 274(4), 832–840. 10.1016/j.jad.2020.05.031 [PubMed: 32664022]

Muthén LK, & Muthén BO (1998–2017). Mplus user's guide (8th ed.). Muthén & Muthén.

*Myers US, Browne KC, & Norman SB (2015). Treatment engagement: Female survivors of intimate partner violence in treatment for PTSD and alcohol use disorder. Journal of Dual Diagnosis, 11(3–4), 238–247. 10.1080/15504263.2015.1113762 [PubMed: 26515712]

*Norman SB, Trim R, Haller M, Davis BC, Myers US, Colvonen PJ, Blanes E, Lyons R, Siegel EY, Angkaw AC, Norman GJ, & Mayes T (2019). Efficacy of integrated exposure therapy vs integrated coping skills therapy for comorbid posttraumatic stress disorder and alcohol use disorder: A randomized clinical trial. JAMA Psychiatry, 76(8), 791–799. 10.1001/jamapsychiatry.2019.063 [PubMed: 31017639]

*Petrakis IL, Desai N, Gueorguieva R, Arias A, O'Brien E, Jane JS, Sevarino K, Southwick S, & Ralevski E (2016). Prazosin for veterans with posttraumatic stress disorder and comorbid alcohol dependence: A clinical trial. Alcoholism: Clinical and Experimental Research, 40(1), 178–186. 10.1111/acer.12926 [PubMed: 26683790]

Petrakis IL, Poling J, Levinson C, Nich C, Carroll K, Ralevski E, Rounsaville B (2006). Naltrexone and disulfiram in patients with alcohol dependence and comorbid post-traumatic stress disorder. Biological Psychiatry, 60(7), 777–783. 10.1016/j.biopsych.2006.03.074 [PubMed: 17008146]

Piper A, & Berle D (2019). The association between trauma experienced during incarceration and PTSD outcomes: A systematic review and meta-analysis. The Journal of Forensic Psychiatry & Psychology, 30(5), 854–875. 10.1080/14789949.2019.1639788

Ruglass LM, Lopez-Castro T, Papini S, Killeen T, Back SE, & Hien DA (2017). Concurrent treatment with prolonged exposure for co-occurring full or subthreshold posttraumatic stress disorder and substance use disorders: A randomized clinical trial. Psychotherapy and Psychosomatics, 86(3), 150–161. 10.1159/000462977 [PubMed: 28490022]

*Ruglass LM, Morgan-López AA, Saavedra LM, Hien DA, Fitzpatrick S, Killeen TK, Back SE, & López-Castro T (2020) Measurement non-equivalence of the Clinician-Administered PTSD Scale by race/ethnicity: Implications for quantifying PTSD severity. Psychological Assessment, 32(11), 1015–1027. 10.1037/pas0000943 [PubMed: 32853005]

Saavedra LM, Morgan-López AA, Hien DA, Back SE, Killeen TK, Ruglass LM, & López-Castro T (2021). Putting the patient back in clinical significance: Using item response theory in estimating clinically significant change in treatment for PTSD and SUDs. Journal of Traumatic Stress, 34(2), 454–466. 10.1002/jts.22624 [PubMed: 33175470]

Saavedra LM, Morgan-López AA, Hien DA, López-Castro T, Ruglass LM, Back SE, Fitzpatrick S, Norman SB, Killeen TK, Ebrahimi CT, Hamblen J, & the Consortium on Addictions, Stress, and Trauma. (2021). Evaluating treatments for posttraumatic stress disorder, alcohol and other drug use disorders using meta-analysis of individual patient data: Design and methodology of a virtual clinical trial. Contemporary Clinical Trials, 107,106479. 10.1016/j.cct.2021.106479 [PubMed: 34157418]

*Saladin M. (n.d.) Trial comparing propranolol to placebo for treating PTSD and alcohol use disorders [Unpublished dataset].

*Sannibale C, Teesson M, Creamer Sitharthan T, Bryant RA, Sutherland K, Taylor K, Bostock-Matusko D, Visser P, & Peek-O'Leary M (2013). Randomized controlled trial of cognitive behaviour therapy for comorbid posttraumatic stress disorder and alcohol use disorders. Addiction, 108(8), 1397–1410. 10.1111/add.12167 [PubMed: 25328957]

Schacht RL, Brooner RK, King VL, Kidorf MS, & Peirce JM (2017). Incentivizing attendance to prolonged exposure for PTSD with opioid use disorder patients: A randomized controlled trial. Journal of Consulting and Clinical Psychology, 85(7), 689–701. 10.1037/ccp0000208 [PubMed: 28414485]

Schäfer I, Lotzin A, Hiller P, Sehner S, Driessen M, Hillemacher T, Schäfer M, Scherbaum N, Schneider B, & Grundmann J (2019). A multisite randomized controlled trial of Seeking Safety vs. Relapse Prevention Training for women with co-occurring posttraumatic stress disorder and substance use disorders. European Journal of Psychotraumatology, 10(1), Article 1577092. 10.1080/20008198.2019.1577092

Sijtsma K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika, 74(1), 107–120. [PubMed: 20037639]

Silverstein MW, Petri JM, Kramer LB, & Weathers FW (2020). An item response theory analysis of the PTSD checklist for *DSM-5*: Implications for *DSM-5* and *ICD-11*. Journal of Anxiety Disorders, 70, 102190. 10.1016/j.janxdis.2020.102190 [PubMed: 32106024]

*Sonne S. (n.d.) Trial comparing Paxil to placebo for treating PTSD and substance use disorders [Unpublished dataset].

*Vujanovic AA, Smith LJ, Green CE, Lane SD, & Schmitz JM (2018). Development of a novel, integrated cognitive-behavioral therapy for co-occurring posttraumatic stress and substance use disorders: A pilot randomized clinical trial. Contemporary Clinical Trials, 65, 123–129. 10.1016/j.cct.2017.12.013 [PubMed: 29287668]

Weathers FW, Bovin MJ, Lee DJ, Sloan DM, Schnurr PP, Kaloupek DG, Keane TM, & Marx BP (2018). The Clinician-Administered PTSD Scale for *DSM–5* (CAPS-5): Development and initial psychometric evaluation in military veterans. Psychological Assessment, 30(3), 383–395. 10.1037/pas0000486 [PubMed: 28493729]

Weathers FW, Ruscio AM, & Keane TM (1999). Psychometric properties of nine scoring rules for the Clinician-Administered Posttraumatic Stress Disorder Scale. Psychological Assessment, 11(2), 124–133. 10.1037/1040-3590.11.2.124

Wohlfarth TD, van den Brink W, Winkel FW, & ter Smitten M (2003). Screening for Posttraumatic Stress Disorder: An evaluation of two self-report scales among crime victims. Psychological Assessment, 15(1), 101–109. 10.1037/1040-3590.15.L101 [PubMed: 12674729]

*Zlotnick C, Najavits LM, Rohsenow DJ, & Johnson DM (2003). A cognitive-behavioral treatment for incarcerated women with substance abuse disorder and posttraumatic stress disorder: Findings from a pilot study. Journal of Substance Abuse Treatment, 25(2), 99–105. 10.1016/s0740-5472(03)00106-5 [PubMed: 14629992]

*Zlotnick C, Johnson J, & Najavits LM (2009). Randomized controlled pilot study of cognitive-behavioral therapy in a sample of incarcerated women with substance use disorder and PTSD. Behavior Therapy, 40(4), 325–336. 10.1016/j.beth.2008.09.004 [PubMed: 19892078]

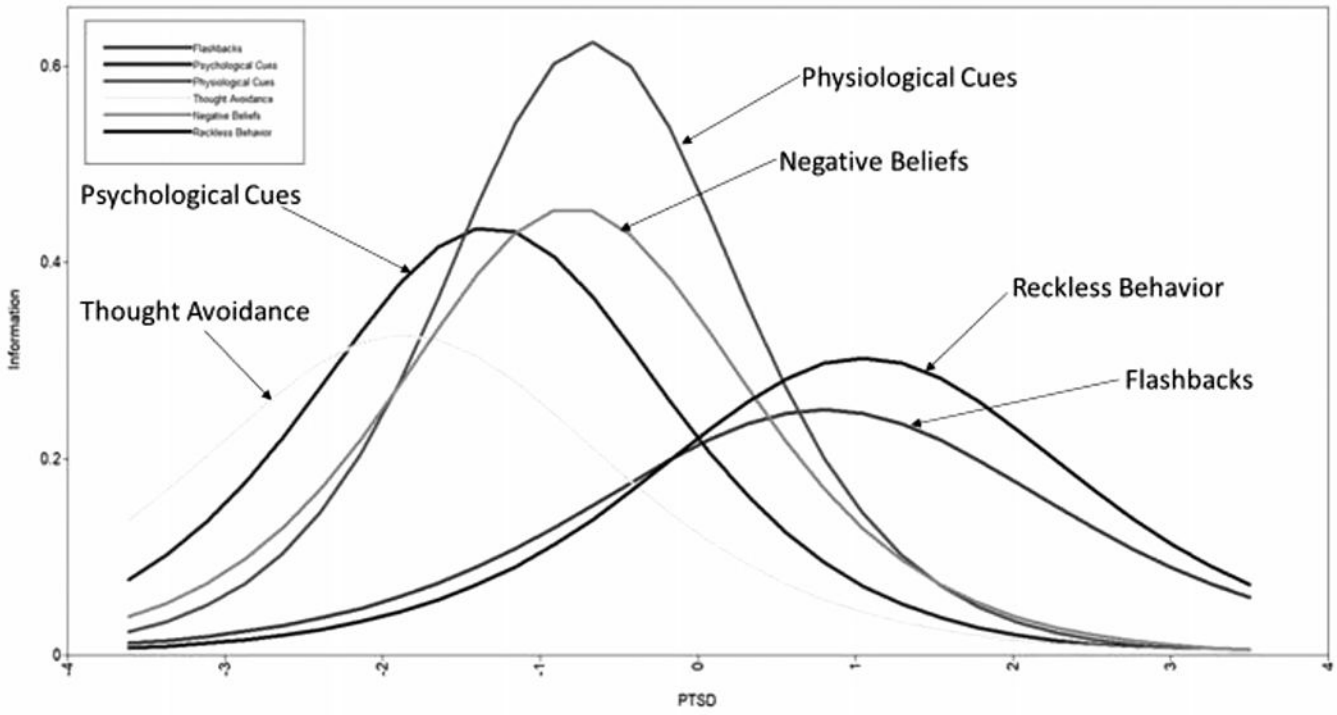

FIGURE 1.

Six highest peak Clinician-Administered Posttraumatic Stress Disorder (PTSD) Scale item information functions

J Trauma Stress. Author manuscript; available in PMC 2023 January 17.

TABLE 1

Study characteristics

First author (year)	*N*	Population type	Female (%)	Full PTSD diagnosis (%)	Measure
Back (2016)	27	Veteran	3.7	66.7	CAPS-IV
Back (2019)	81	Veteran	9.9	100.0	CAPS-IV
Boden (2012)	98	Veteran	0.0	86.7	CAPS-IV
Foa (2013)	165	Civilian/Veteran	34.5	72.1	PSS-I(-IV)
Hien (2004)	126	Civilian	100.0	82.5	CAPS-IV
Ruglass/Hien (2017)	95	Civilian	81.0	48.2	CAPS-IV
Hien (2009)	353	Civilian	100.0	80.1	CAPS-IV
Hien (2015)	113	Civilian	37.4	40.7	CAPS-IV
McDevitt-Murphy (2014)	68	Veteran	8.8	58.8	CAPS-IV
McGovern (2011)	53	Civilian	56.6	100.0	CAPS-IV
McGovern (2015)	284	Civilian	58.5	100.0	CAPS-IV
Mills (2012)	103	Civilian	62.1	95.2	CAPS-IV
Myers (2015)	41	Civilian	100.0	68.3	CAPS-IV
Norman (2019)	119	Veteran	10.1	86.6	CAPS-5
Haller/Norman (2016)	154	Veteran	11.7	64.3	CAPS-IV
Petrakis (2016)	96	Civilian	6.3	83.3	CAPS-IV
Petrakis (2020)	24	Veteran	8.4	29.1	CAPS-IV
Saladin (n.d.)	44	Civilian	50.0	93.1	CAPS-IV
Sannibale (2013)	62	Civilian	53.2	91.9	CAPS-IV
Schacht/Peirce (2017)	58	Civilian	79.3	100.0	CAPS-IV
Schafer (2019)	343	Civilian	100.0	63.3	PSS-I(-IV)
Sonne (n.d)	25	Civilian	32.0	92.0	CAPS-IV
Vujanovic (2018)	53	Civilian	50.9	73.6	CAPS-5
Zlotnick (2003)	24	Incarcerated	100.0	87.5	CAPS-IV
Zlotnick (2009)	49	Incarcerated	100.0	81.6	CAPS-IV

Note: *N* = 2,658. PTS*D* = posttraumatic stress disorder; *DSM* = *Diagnostic and Statistical Manual of Mental* Disorders; CAPS-IV = Clinician-Administered PTSD Scale for *DSM-IV*; CAPS-5 = Clinician-Administered PTSD Scale for *DSM-5*; PSS-I-IV = PTSD Symptom Scale–Interview for *DSM-IV*.

TABLE 2

Descriptive characteristics for cross-study harmonized symptoms

PTSD symptom (*DSM-IV* Criterion)	*N*	Proportion	*SD*
Intrusive recollections (B1)	2559	0.79	0.41
Nightmares (B2)	2548	0.60	0.49
Flashbacks (B3)	2551	0.36	0.48
Psychological cues (B4)	2550	0.79	0.41
Physiological cues (B5)	2549	0.68	0.47
Thought avoidance (C1)	2563	0.84	0.36
Activity avoidance (C2)	2556	0.64	0.48
Inability to recall (C3)	2558	0.42	0.49
Diminished interest (C4)	2561	0.64	0.48
Detachment (C5)	2564	0.77	0.42
Restricted affect (C6)	2560	0.74	0.44
Foreshortened future (C7)	2274	0.43	0.50
Sleep (D1)	2446	0.80	0.40
Irritability (D2)	2561	0.70	0.46
Concentration problems (D3)	2436	0.70	0.46
Hypervigilance (D4)	2559	0.74	0.44
Startle (D5)	2541	0.60	0.49
Negative beliefs[a]	172	0.81	0.40
Blame of self or others[a]	171	0.54	0.50
Horror/shame/guilt[a]	172	0.94	0.24
Reckless behavior[a]	172	0.37	0.49

Note: PTS*D* = posttraumatic stress disorder; *DSM* = *Diagnostic and Statistical Manual of Mental Disorders*.

[a]Symptom is specific to the fifth edition of the *DSM-5*.

J Trauma Stress. Author manuscript; available in PMC 2023 January 17.

TABLE 3

Descriptive statistics for predictors of measurement non-invariance/differential item functioning and underlying posttraumatic stress disorder (PTSD) severity

Predictor	***N***	***M***	***SD***
Age (years)	2587	40.22	11.43
	N	**Proportion**	***SD***
Male	2,616	0.41	0.49
Hispanic	2,623	0.07	0.26
White	2,623	0.64	0.48
Black	2,623	0.27	0.44
Other	2,623	0.04	0.19
Asian	2,623	0.01	0.09
High School or less	2,011	0.50	0.50
Some college	1,968	0.33	0.47
College degree	1,968	0.16	0.36
Married	2,225	0.18	0.39
Veteran	2,601	0.36	0.48
Civilian	2,601	0.62	0.49
Incarcerated	2,658	0.03	0.16
Depression diagnosis	2,273	0.57	0.51
CAPS-5	2,658	0.06	0.25
PSS-I-IV	2,658	0.19	0.39
Full PTSD Diagnosis	2,658	0.78	0.42
Past-month alcohol use (days)	2,496	9.96	10.87
Any past-month cocaine use	1,886	0.24	0.43
Any past-month opiate use	1,766	0.11	0.31
Any past-month stimulant use	1,744	0.10	0.30
Any past-month sedative use	1,522	0.10	0.30
Concomitant psychiatric medication	1,381	0.47	0.50

Note: CAPS-5 = Clinician-Administered PTSD Scale for *DSM-5*. PSS-I-IV = PTSD Symptom Scale–Interview for *DSM-IV*; *DSM = Diagnostic and Statistical Manual of Mental Disorders*.

J Trauma Stress. Author manuscript; available in PMC 2023 January 17.

TABLE 4

Final moderated nonlinear factor analysis threshold/difficulty parameter estimates

PTSD symptom	Threshold/difficulty	Age MNI/DIF[a]	Male MNI/DIF	Black MNI/DIF	College MNI/DIF	Veteran MNI/DIF	Incarcerated MNI/DIF	PSS-I-IV MNI/DIF	CAPS-5 MNI/DIF	Past-month alcohol MNI/DIF
Intrusive recollections	−1.67								2.06	
Nightmares	−0.46									
Flashbacks	0.78								−1.42	−0.03
Psychological cues	−1.80									
Physiological cues	−1.13	−0.01					0.82			
Thought avoidance	−2.18				−0.42		−0.44			
Activity avoidance	−0.68			0.38						
Inability to recall	0.39		−0.41	−0.84					−1.13	
Diminished interest	−0.63	0.02					−0.81			
Detachment	−1.52		0.62				1.43			
Restricted affect	−1.24							−0.38		
Foreshortened future	0.35	0.02	−0.52							
Sleep	−1.61						−1.23			
Irritability	−1.03	−0.02				0.52		−1.04	−0.71	
Concentration problems	−1.03			−0.39						
Hypervigilance	−1.20				−0.46		−1.64			
Startle	−0.47						−0.78	0.55		
Negative beliefs	−1.12									
Blame of self or others	−0.36				−0.74	−1.28				
Horror/shame/guilt	−2.63									
Reckless behavior	1.08	−0.04								

Note: MNI/DIF = measurement noninvariance/differential item functioning; *DSM = Diagnostic and Statistical Manual of Mental* Disorders; CAPS-5 = Clinician-Administered PTSD Scale for *DSM-5*; PSS-I-IV = PTSD Symptom Scale–Interview for *DSM-IV*.

[a]MNI/DIF estimates that were significant at $p < .05$ are shown and were included in the final scoring model.

J Trauma Stress. Author manuscript; available in PMC 2023 January 17.

TABLE 5

Final moderated nonlinear factor analysis (MNLFA) loading/discrimination parameter estimates

PTSD symptom	Loading/discrimination	Male MNI/DIF[a]	Black MNI/DIF[a]	Incarcerated MNI/DIF[a]	CAPS-5 MNI/DIF[a]
Intrusive recollections	1.07		−0.33		
Nightmares	0.93			0.41	
Flashbacks	1.00				
Psychological cues	1.33				
Physiological cues	1.58				
Thought avoidance	1.09				−0.88
Activity avoidance	0.88		−0.38		
Inability to recall	0.45				
Diminished interest	0.93				0.76
Detachment	1.09				
Restricted affect	0.97				
Foreshortened future	0.51	0.29			
Sleep	0.74				
Irritability	0.78				
Concentration problems	0.95				
Hypervigilance	0.76		−0.49		
Startle	1.02				
Negative beliefs	1.35				
Blame of self or others	0.92				
Horror/shame/guilt	1.02				
Reckless behavior	1.10				

Note: PTS*D* = posttraumatic stress disorder; MNI/DIF = measurement noninvariance/differential item functioning; CAPS-5 = Clinician-Administered PTSD Scale for *DSM-5*.

[a]Estimates that were significant at $p < .05$ are shown and were included in the final scoring model.