HHS Public Access

Author manuscript

Neuropsychology. Author manuscript; available in PMC 2024 May 01.

Published in final edited form as:

Neuropsychology. 2023 May ; 37(4): 398–408. doi:10.1037/neu0000823.

Harmonizing PTSD severity scales across instruments and sites

Eamonn Kennedy[1,2,3], **Emily L. Dennis**[1,2], **Hannah M. Lindsey**[1,2], **Terri deRoon-Cassini**[4], **Stefan Du Plessis**[5,6], **Negar Fani**[7], **Milissa L. Kaufman**[8,9], **Nastassja Koen**[5,6], **Christine L. Larson**[10], **Sarah Laskowitz**[11], **Lauren A. M. Lebois**[8,12], **Rajendra A. Morey**[11,13], **Mary R. Newsome**[14,15], **Cori Palermo**[8,12], **Nicholas J. Pastorek**[15,14], **Abigail Powers**[7], **Randall Scheibel**[15,14], **Soraya Seedat**[16], **Antonia Seligowski**[8,12], **Dan J. Stein**[5,6], **Jennifer Stevens**[7], **Delin Sun**[11,13], **Paul Thompson**[17], **Maya Troyanskaya**[14,15], **Sanne J. H. van Rooij**[7], **Amanda A. Watts**[11], **Carissa N. Tomas**[18], **Wright Williams**[15], **Frank G. Hillary**[19,20], **Mary Jo Pugh**[1,2,3], **Elisabeth A Wilde**[1,2,14], **David F. Tate**[1,2]

[1]Department of Neurology, University of Utah School of Medicine, Salt Lake City, UT.

[2]George E. Wahlen Veterans Affairs Medical Center, Salt Lake City, UT.

[3]Division of Epidemiology, University of Utah, Salt Lake City, UT.

[4]Department of Surgery, Division of Trauma & Acute Care Surgery and Comprehensive Injury Center, Medical College of Wisconsin, Milwaukee, WI.

[5]Dept of Psychiatry and Mental Health, University of Cape Town, Cape Town, South Africa.

[6]SAMRC Unit on Risk & Resilience in Mental Disorders, Dept of Psychiatry and Mental Health and Neuroscience Institute, University of Cape Town, Cape Town, South Africa.

[7]Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA.

[8]Department of Psychiatry, Harvard Medical School, Boston, MA.

[9]Division of Women's Mental Health, McLean Hospital, Belmont, MA.

[10]Department of Psychology, University of Wisconsin-Milwaukee, Milwaukee, WI.

[11]Brain Imaging and Analysis Center, Duke University, Durham, NC.

[12]Division of Depression and Anxiety Disorders, McLean Hospital, Belmont, MA.

[13]VISN 6 MIRECC, Durham VA, Durham, NC.

[14]H. Ben Taub Department of Physical Medicine and Rehabilitation, Baylor College of Medicine, Houston, TX.

[15]Michael E. DeBakey Veterans Affairs Medical Center, Houston, TX.

[16]SU/UCT MRC Unit on Risk and Resilience in Mental Disorders, Department of Psychiatry, Stellenbosch University, Stellenbosch, South Africa.

Corresponding author: eamonn.kennedy@utah.edu.

[17]Imaging Genetics Center, Stevens Neuroimaging & Informatics Institute, Keck School of Medicine of USC, Marina del Rey, CA.

[18]Department of Epidemiology and Comprehensive Injury Center, Medical College of Wisconsin, Milwaukee, WI.

[19]Department of Neurology, Hershey Medical Center, State College, PA.

[20]Department of Psychology, Penn State University, State College, PA.

Abstract

Objective: The variety of instruments used to assess post-traumatic stress disorder (PTSD) allows for flexibility, but also creates challenges for data synthesis. The objective of this work was to use a multi-site mega analysis to derive quantitative recommendations for equating scores across measures of PTSD severity.

Methods: Empirical Bayes harmonization and linear models were used to describe and mitigate site and covariate effects. Quadratic models for converting scores across PTSD assessments were constructed using bootstrapping and tested on hold out data.

Results: We aggregated 17 data sources and compiled an n=5,634 sample of individuals who were assessed for PTSD symptoms. We confirmed our hypothesis that harmonization and covariate adjustments would significantly improve inference of scores across instruments. Harmonization significantly reduced cross-dataset variance (28%, $p<0.001$), and models for converting scores across instruments were well fit (median $R^2=0.985$) with an average root mean squared error of 1.46 on sum scores.

Conclusions: These methods allow PTSD symptom severity to be placed on multiple scales and offer interesting empirical perspectives on the role of harmonization in the behavioral sciences.

Keywords

Harmonization; PTSD; Screening Instruments; Mega Analysis

Introduction

Large-scale data sharing initiatives offer opportunities to improve robustness by synthesizing multiple data sources (Thompson et al., 2020). However, in the behavioral sciences, differences in psychometric evaluation can confound the aggregation of data (Houtkoop et al., 2018; Towse, Ellis & Towse, 2021). For example, researchers and clinicians can select from a variety of instruments for measuring post-traumatic stress disorder (PTSD), which is a prevalent and burdensome mental health condition (Kessler et al., 2005; Norris & Hamblen, 2004). Instruments that assess PTSD symptom severity broadly classify into three groups: (1) Clinical interviews such as the Clinician-Administered PTSD Scale for Diagnostic and Statistical Manual of Mental Disorders-5 (5th ed.; DSM–5; American Psychiatric Association, 2013), abbreviated CAPS-5 (Weathers et al., 2013, 2018); (2) Brief self-assessments such as the Davidson Trauma Scale (DTS) which can briefly screen for provisional diagnosis (Davidson et al., 1997), and 3. Thematically specific severity scales which are designed to assess a particular group (Wilkins, Lang, & Norman, 2011).

This variety of assessments affords flexibility, but also creates challenges for data synthesis. Even within one assessment, test instructions and items are continually modified to match advancing diagnostic criteria. For example, the DSM-5 criteria for PTSD (American Psychiatric Association, 2013) are distinct from previous iterations (Brett, Spitzer, & Williams, 1988), and different factor solutions have been proposed (Shelby, Golden-Kreutz & Andersen, 2005).

Establishing standards for converting scores across PTSD symptom inventories could improve clinical and research consistency. However, it is challenging to isolate instrumental effects because severity scores depend on at least five factors: (1) Clinical features and presentation; (2) Intrinsic biological variables such as age; (3) Distinct procedures across studies and sites; (4) Instrumental variations, such as distinct question phrasings; and (5) Statistical error and randomness. To accurately convert scores, the instrumental component must be isolated from other sources of variation, but data from single sources is typically subject to specific biases since most studies recruit and sample for specific conditions or traits (Radua et al., 2020; Pugh et al., 2021). Therefore, a secondary mega analysis is a good solution for identifying and removing unwanted effects (Boedhoe et al., 2019).

We report a multi-site (n=17 datasets) mega study analysis of five common instruments used for PTSD assessment. We leveraged recent data harmonization algorithms (Pomponio et al., 2020) to remove site effects. The component of severity scores associated with instrumentation was isolated from covariate effects, and calculations for converting scores across measures were tested on hold out data (data not used during model construction). Our main hypothesis was that without corrections, percentage and percentile models (Kolen & Brennan, 2004) would be confounded, while performance would be significantly improved by the harmonization of data sources, covariate adjustments, and models tolerant of some nonlinearity across instruments.

Methods

Data sources

This secondary mega analysis draws from a range of military and civilian studies. We petitioned collaborators for item level data, drawing from the Psychiatric Genomics Consortium and the Enhancing NeuroImaging Genetics through Meta-Analysis consortium (PGC-ENIGMA) PTSD working group (Logue et al., 2018), the ENIGMA Brain Injury working group (Wilde et al., 2021), and the Long-term Impact of Military-relevant Brain Injury Consortium - Chronic Effects of Neurotrauma Consortium (LIMBIC-CENC) (Cifu et al., 2015). We obtained 17 datasets that performed different combinations of PTSD assessments. Data quality and consistency was confirmed during discussions among authors who performed the primary data collection. All assessments were conducted in English. The University of Utah provided overall IRB study approval and data use agreements for the following sources:

(1) DOD-ADNI: A Study of Brain Aging in Vietnam War Veterans (Weiner et al., 2017).

(2) iSCORE: The Imaging Support for the Study of Cognitive Rehabilitation (Tate et al., 2019).

(3) PT: The Personality Traits and brain matter aberrations as potential markers of mTBI and PTSD study.

(4) CE: The longitudinal study of Chronic Effects of TBI in Veterans and service members.

(5, 6) Blast I and Blast II: Blast I is an FMRI study of TBI associated with blast injury, and Blast II is a renewal of the initial study capturing similar data.

(7) NBS-DoD: Neural and Behavioral Sequelae of blast-related traumatic brain injury.

(8) SARChI: Stellenbosch's South African Research Chairs Initiative (Suliman et al., 2014).

(9) GTP: The Grady Trauma Project (Gillespie et al., 2009).

(10) iSTAR: Imaging study of trauma and resilience (Weis et al., 2021).

(11) TSS: The McLean Trauma Spectrum Study (Lebois et al., 2021).

(12) NEST: The McLean Neurocardiac Effects of Stress and Trauma Study.

(13) DCHS: The Drakenstein Child Health Study of prenatal mothers (Donald et al., 2018)

(14) VCTP: Neuroimaging meditation therapy in Veterans with comorbid mild TBI and PTSD.

(15) MC: A military Mission Connect study.

(16) MIRECC-DU: Mental Illness Research Education Clinical, Centers of Excellence, Duke University.

(17) MIRECC-D: Mental Illness Research Education Clinical, Centers of Excellence, Durham Veterans Affairs (Sun et al., 2020).

Additionally, data from LIMBIC-CENC (Cifu et al., 2015) was held out during model construction to enable demonstration and independent testing. CE, Blast I/II, and some of the NBS-DoD data were acquired using very similar inclusion/exclusion criteria, but collection was stratified over time.

Inclusion Criteria

Adults aged over 18 years who completed at least one assessment to a level of <20% missingness were included. PTSD severity screeners were obtained at first entry to care facilities or at the initiation of research studies. Any repeated measurements per person were excluded, alongside measurements after interventions. Since the number of non-symptom reporting cases is largely dependent on the inclusion criteria of studies, individuals who reported the lowest possible sum score (e.g., a total severity score of zero on PCL-5) were excluded (see Limitations).

Measures and Characteristics

PTSD inventories are typically designed to elicit one item level response per diagnostic criteria. This means 20 unique items were recorded for DSM-5 and 17 unique items were recorded for DSM-IV assessments. All instruments included overlapping items that facilitated harmonization.

The PTSD Checklist for DSM-5 (PCL-5) is a 20-item self-report measure that assesses the 20 DSM-5 symptoms of PTSD using a 5-point Likert scale ranging from 0 to 4 ("Not at all", to "Extremely"). The checklist asks the participant to consider the level of symptom severity over the last month. The PCL-5 severity score is the sum of all items scores and ranges from 0 to 80 (Weathers et al., 1993, Wortmann et al., 2016).

The PTSD Checklist for DSM-IV (PCL-C) is a 17-item self-report measure that assesses the severity of the 17 DSM-IV symptoms of PTSD using a 5-point Likert scale ranging from 1 to 5 ("Not at all", to "Extremely") as experienced over the last month. The PCL-C severity score is the sum of all items, which ranges from 17 (no symptoms) to 85 (Weathers et al., 1993).

The PTSD Checklist for DSM-IV – Military Version (PCL-M). Like the PCL-C, the PCL-M is a 17-item self-report that measures the severity of the 17 DSM-IV PTSD symptoms over the last month on a 5-point Likert scale ranging from 1 to 5 ("Not at all", to "Extremely"). The items of the PCL-M are the same as the PCL-C except that the PCL-M wording addresses a stressful military experience. A prior synthesis of the PCL-C and PCL-M demonstrated strong consistency, reliability, and convergence (Wilkins, Lang, & Norman, 2011), and in this work we denote both DSM-IV PCL screeners as ‘PCL’. Item level military/civilian differences are described in detail (see Results).

The Davidson Trauma Scale (DTS) is a 17-item self-report measure that assesses the frequency and severity of the 17 DSM-IV symptoms of PTSD. Each item of the DTS severity scale ranges from (0 = "not at all distressing" to 4 = "extremely distressing") and the responder is asked to consider symptoms within the last week. The total sum score ranges from 0 to 136. We considered only the severity scale of DTS which ranges from 0 to 68 (Davidson et al., 1997).

The modified PTSD Symptom Scale (mPSS) is a 17 item self-report measure that asks about how upsetting the 17 DSM-IV symptoms of PTSD severity have been within the last two weeks, with items rated on 4-point Likert scale consisting of 0 ("Not at all"), 1 (Once per week or less/a little bit/once in a while), 2 (Two to four times per week/somewhat/half the time) to 3 ("Five or more times per week/very much/ almost”). Sum severity scores range from 0 to 51 (Falsetti et al., 1993).

Covariates of age, sex/gender, site/study, and military/civilian status were included. PTSD-associated conditions such as substance use disorder, depression, and early-onset cognitive impairment (Kennedy et al., 2022) were not considered in the primary analysis as they were not recorded consistently across studies. Race/ethnicity characteristics were also not recorded consistently across studies. Military status was broadly defined, and included US

Veterans of the Vietnam war as well as Veterans of Operation Enduring Freedom/Operation Iraqi Freedom/Operation New Dawn (OEF/OIF/OND). While some studies recorded gender, others recorded biological sex, so these characteristics were aggregated into a single sex/gender variable.

Statistical analysis

Analysis was performed in Python 3. Kruskal-Wallis H-tests (omnibus) were used to test for significance differences across groups. Welch's t tests were used for post-hoc pairwise comparisons. Where multiple tests were performed, q statistics were calculated at a threshold of 0.05 to reduce false discovery rates. Absolute severity scores were converted to fraction/percentage severity scores by subtracting the minimum assessment score from each observed score and dividing by the full range of the instrument.

The aim of harmonization was to remove unwanted site effects while preserving instrumental effects for further analysis. If absolute scores were harmonized, this would remove both site effects and also the absolute differences between instruments (e.g. baseline offsets between scales). Therefore, we used the ComBat-GAM algorithm (Pomponio et al., 2020) to harmonize percentage scores over all datasets. This method explicitly protected covariates and instrumental effects. After harmonization, percentage severity scores were returned to absolute scores on their respective instruments with site effects removed. Harmonization efficacy was measured as the reduction in cross-dataset variance when comparing raw scores to their post-harmonized equivalents. Coefficients of determination (R^2) were used to calculate the deviation of data from models. An R^2 of 1 means a model perfectly fits the data.

Partitioning, training, and model description

Ordinary least squares (OLS) linear models were used to adjust for covariates and to convert scores across instruments. After removing covariate effects, the square of sum scores was used as an additional term during instrument conversion to capture potential nonlinearities across assessments. Comparing random subsamples of the data reduced bias associated with variations in clinical presentation, and allowed for a measurement of confidence on the inferred scores (Choi et al., 2014). The quantity of interest is a predicted line of model fit, so we measured confidence using root mean squared error (RMSE). Since the number of observations differed for each assessment, we elected to bootstrap using $N=\text{argmin}([n_A, n_B])/2$ samples in each model fit iteration, where n_A and n_B and the total observations of assessment A and B respectively. This means the training fraction had an upper bound of 50%. Summary results were determined by averaging over the coefficients recovered from all model iterations.

Transparency and openness

Raw data is available upon reasonable request pending study approval and data transfer agreements between all participating institutions. Codes used for analysis can be provided by the authors upon reasonable request.

Results

Data summary

The summary statistics of the 17 datasets (see Methods) are shown in Table 1. The total sample was n=6,771 but this reduced to n=5,634 after exclusions were applied. The median age was 36 years old with an interquartile range of 28 – 46 years, and 36.4% of all study participants were female. The data comprised seven civilian studies, nine military studies, and one mixed population study. The total counts per assessment were (1) PCL-5: n = 1,325, (2) PCL-C/M: n = 786, (3) DTS: n = 3,196, and (4) mPSS: n = 327.

Figure 1 provides an overview of symptom reporting across all datasets. Figure 1a plots the histogram of all percentage severity scores. Figure 1b shows percentage sum scores broken out as boxplots by military status and sex/gender. An omnibus test confirmed significant differences between the groups in Figure 1b. Female civilians reported the most severe symptoms overall ($p<0.001$). Figure 1c shows the mean severity for each item across all datasets set to the same scale (range: [0:4] as per PCL-5). Sleep disturbance, hypervigilance, and negative feelings were the most intensely reported symptoms. Military status was associated with elevated risk taking (Figure 1c - right). Civilian status was associated with increased report of upsetting reminders, and increased avoidance of upsetting reminders and feelings (Figure 1c - right).

Unadjusted conversion

We first consider the conversion from PCL-5 to PCL severity scores. PCL-5 sum scores range from [0,80], while PCL sum scores range from [17,85]. Therefore, the line where the percentage severity scores for the two instruments are the same is described by the equation

$$S_{PCL} = a_0 + rS_{PCL5}$$

where r is the ratio of scale ranges (85–17)/(80–0) = 0.85, and a_0 =17 is the intercept. This is the line of percentage equality across instruments. However, a_0 and r can differ from expectation in practice because inventories have unique characteristics and distinct items. To visualize the similarity of data sources, we performed a combinatoric analysis of the 17 datasets which generated 153 dataset pairs. Six of these pairs are shown in Figure 2 as ranked paired sum scores subsampled from each dataset. Figure 2a compares three dataset pairs that used the same instrument, whereas Figure 2b contrasts three dataset pairs that used different instruments.

The deviation from the line of percentage equality (gray lines) is a convolved measure of all the underlying differences between the two datasets. If all data were on this line, the two datasets would be identical and $R^2 = 1$. Overall, simple percentage conversions (gray lines) did not accurately predict raw scores across datasets in most cases, and percentile models also showed errors. These findings suggest corrections should be performed before linking across instruments.

Harmonization

The ComBat-GAM algorithm was implemented to correct for site-specific variations such as differences in severity-based inclusion criteria. Figure 3 shows all 153 R^2 comparison values grouped into boxplots by dataset. Intrinsic differences across datasets confound simple linear conversion as hypothesized (Figure 3a). After harmonization (Figure 3b) the average R^2 increased by 0.74, and the number of well-fit pairs (R^2>0.9) increased significantly from 29 to 37 (+28%, p<0.001).

Covariate adjustment

Table 2 shows the results for two blocked logistic regressions predicting symptom severity using (1) sociodemographic factors and (2) instrumentation. The models predicted post-harmonized severity scores binarized to above/below the clinical cutoff of each inventory (e.g., y=1 for scores >32/80 for PCL-5). The unharmonized model fits were poor, but improved after harmonization. After harmonization, military status showed much higher odds (OR: 2.74, CI95%: 2.38 – 3.16) of above-threshold severity, while age groups and sex showed broadly similar odds. Age, sex/gender, and population were adjusted out of the data using linear regression.

Converting scores across instruments

After harmonization and covariate correction, we again consider the conversion of PCL-5 to PCL scores. To derive an empirical relationship between the instruments, we aggregated two datasets; one containing all post-harmonized and covariate adjusted PCL-5 sum scores, and one containing all post-harmonized and covariate adjusted PCL sum scores.

As described in the methods, we selected N=argmin([n_a, n_b])/2 = 272 samples from each group at random and sorted and paired the scores. A regression was performed to estimate r, the ratio of scale ranges for PCL-5 to PCL (nominally 0.85), and a_0, the intercept (nominally 17). Over ten random subsampling iterations, the regression found a_0 = 19, and $r_{a/b}$ = 0.83 was the best fit with a mean R^2 of 0.965. The similarity of the coefficients and the high R^2 of this approach are encouraging, but nonlinearities could emerge for more distinct assessments. To account for this, a quadratic sum score term, $S_A{}^2$, was added to the conversion models. Table 3 shows the estimated parameters for different instrument pairs using the following model:

$$S_B = a_0 + \beta_1 S_A + \beta_2 S_A{}^2$$

where S_A and S_B are the sum scores of instruments A and B. This process was repeated for all ‘A→B’ conversions.

Worked example

A worked conversion example is as follows: A clinician wishes to infer a patient’s PCL-5 score using the patient’s DTS score. Since frequency scales were not considered in these models, the clinician calculates only the patient’s DTS severity score (S_{DTS} = 45 out of 68). The clinician consults Table 3 and finds the coefficients of DTS → PCL-5 are a_0 = −2.2, β_1

=1.44, and $\beta_2 = -0.0044$, with root mean squared error of e = 1.5. Using these parameters, S_{PCL5} is calculated as

$$S_{PCL5} = -2.2 + 1.44(45) - 0.0044(45^2) = 53.7 \pm 1.5$$

To demonstrate conversion on real data, we elected to hold out an independent sample of N=1,212 observations of the PCL-5 from LIMBIC-CENC (see Methods) not used in model training. The conversions from PCL-5 to other assessments are shown in Figure 4. Figure 4a plots the inferred severity scores as a function of the input PCL-5 severity scores. Figure 4b shows a histogram of the severity scores after conversion to other severity scales.

Discussion

'Harmonization' is often used to refer to data aggregation, but true data harmonization aims to minimize source and measure variations in ways that preserve meaning. New PTSD treatments continue to be assessed (Rauch et al, 2018), but persistent gaps in clinical/research consistency, and shifting trial admission standards make it challenging to generalize findings. Harmonization presents a promising solution to address these concerns, and it is interesting to consider what new insights may be gleaned from data where source and acquisition effects are mitigated.

In the process of generating models for converting across PTSD severity scales, we found that empirical Bayes harmonization methods can isolate variations induced by different settings and procedures. We also confirmed our hypothesis that harmonization and covariate adjustments would significantly improve conversion model performance. There are many points of distinction that could explain why harmonization improves performance. For example, research studies and clinical facilities typically perform assessments in different ways. However, an exhaustive list of all the subtle ways that sources may differ is not necessary in order for these effects to be empirically detected and removed in aggregate.

Similarly, we outlined a thorough description of all the ways that the instruments differed (see Methods), but ultimately, simple models effectively captured the instrumental variations without reference to their specific differences. Conceptually, these strategies draw from the observation that model explainability is distinct from predictive power (Kasirzadeh, 2021). Several facts suggest the instrumental components were well isolated. For example, simple percentage equivalence models fit well after corrections, but not before (Figure 2).

This work also provides some general insights into PTSD symptomology. If site, biological, and instrumental variations can be separated from individual symptom reports, then perhaps new intrinsic truths can be unearthed from previously confounded data. There are even hints of this possibility in this work, and while the influence of age and sex/gender variables were relatively unchanged by harmonization, military status exhibited a large increase in odds after accounting for site effects (from OR: 0.89 to OR: 2.74). These population specific differences were only apparent after harmonization, and similar approaches applied to problems in the behavioral sciences could help to identify hidden population effects.

Conclusion

We leveraged a multi-site mega analysis (n=17 datasets) to derive quantitative recommendations for the conversion of common PTSD severity scales. The data ensemble and the impact of site and covariates on severity scores were described. After isolating the instrumental component of severity scores, we produced accurate (median R^2=0.985) models for converting PTSD symptom severity scores, which were validated on data not used in model construction. This analysis suggests PTSD instrumentation has objective effects that can be isolated and removed, and these methods offer new empirical perspectives on the role of harmonization in the behavioral sciences.

Limitations

We did not consider quantitation of structured interviews and future work could explore crosswalks between structured clinical interviews and brief inventories. PTSD is a highly comorbid condition, and the data collection across sources did not facilitate consideration of a wider range of conditions. We also did not consider varying trauma exposure or repeated measures, and future work would benefit from the analysis of multiple measures per person. This limitation is mitigated in part by repeated subsampling drawn from all sites in aggregate when deriving models. This work exclusively considered English language assessment and lacked data on race/ethnicity. Future analysis of assessment languages and race/ethnicity would be beneficial.

Individuals who reported the lowest possible sum score (e.g., a total severity score of zero on PCL-5) were excluded in order to remove predictive biases associated with differing inclusion criteria across studies. Some studies exclusively recruited individuals with moderate/severe PTSD severity, while others were convenience samples with a large fraction of participants with no history of PTSD exposure, who scored zero on PTSD severity assessment. Converted scores below zero do not provide additional value, and should be truncated to zero after conversion.

Our conversion models were constructed on data after site and covariate effects were removed. This means the models are only recommended for within-study inference unless additional harmonization procedures are enacted. In accordance with APA guidelines (AERA, APA, & NCME, 2014), we caution against broad generalizations of the methods presented here to new data and samples, or to cases where there is limited exact measure overlap. The extent to which harmonization and adjustment remove unwanted effects is empirical, and some residual effects may have persisted. We intentionally did not force the intercepts of the models, and at the extremes the parameters can return a value outside of the inferred inventory range. Any values out of range (e.g. negative numbers) should be truncated to the maximum/minimum possible value.

Funding Acknowledgement:

This work was supported by the NIH National Institute of Neurological Disorders and Stroke, award #R61NS120249, and the LongTerm Impact of Military Relevant Brain Injury Consortium (LIMBIC) Award W81XWH18PH/TBIRPLIMBIC under Awards No. W81XWH1920067 and W81XWH1320095. Secondary data provided for the mega analysis were supported by: Award number R21MH112956, the Anonymous

Women's Health Fund, Kasparian Fund, Trauma Scholars Fund; The South African Medical Research Council for the “Shared Roots” Flagship Project, Grant no. MRC-RFA-IFSP-01-2013/SHARED ROOTS” through funding received from the South African National Treasury under its Economic Competitiveness and Support Package; Funding from the SAMRC Unit on Risk & Resilience in Mental Disorders;The Narsad Young Investigators award; The PA Health Research Grant SAP #4100077082 to Dr. Hillary; Award numbers: (NIH) K23MH125920; AHA 20CDA35310031; MH098212; MH071537; M01RR00039; UL1TR000454; HD071982; HD085850; U54 EB020403; R01 MH116147; R56 AG058854; P41 EB015922; R01 MH111671; F32MH109274; and K23AT009713. All contents and opinions expressed are sole responsibility of the authors and do not necessarily represent the official views of any funding sources.

References

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). doi:10.1176/appi.books.9780890425596

AERA, APA, & NCME (2014). Standards for Educational and Psychological Testing: National Council on Measurement in Education Washington DC: American Educational Research Association.

Boedhoe PSW, Heymans MW, Schmaal L, Abe Y, Alonso P, Ameis SH, … Twisk JWR (2019) An Empirical Comparison of Meta- and Mega-Analysis With Data From the ENIGMA Obsessive-Compulsive Disorder Working Group. Frontiers in Neuroinformatics,12, doi:10.3389/fninf.2018.00102

Brett EA, Spitzer RL, & Williams JB (1988). DSM-III-R criteria for posttraumatic stress disorder. The American Journal of Psychiatry, 145(10), 1232–1236. doi.org/10.1176/ajp.145.10.1232 [PubMed: 3421344]

Choi SW, Schalet B, Cook KF, & Cella D (2014). Establishing a common metric for depressive symptoms: linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. Psychological Assessment, 26(2), 513–527. 10.1037/a0035768 [PubMed: 24548149]

Cifu DX, Diaz-Arrastia R, Williams RL, Carne W, West SL, McDougal M, & Dixon K (2015). The VA/DoD Chronic Effects of Neurotrauma Consortium: An Overview at Year 1. Federal Practitioner : For the health care professionals of the VA, DoD, and PHS, 32(8), 44–48. PMID: [PubMed: 30766083]

Davidson JRT, Book SW, Colket LA, Tupler LA, Roth S, David D, … Feldman ME (1997). Assessment of a new self-rating scale for post-traumatic stress disorder. Psychological Medicine, 27(1), 153–160. doi:10.1017/S0033291796004229 [PubMed: 9122295]

Donald KA, Hoogenhout M, du Plooy CP, Wedderburn CJ, Nhapi RT, Barnett W, Hoffman N, Malcolm-Smith S, Zar HJ, & Stein DJ (2018). Drakenstein Child Health Study (DCHS): investigating determinants of early child development and cognition. BMJ paediatrics open, 2(1), e000282. doi:10.1136/bmjpo-2018-000282 [PubMed: 29942867]

Falsetti SA, Resnick HS, Resick PA, & Kilpatrick D (1993). The Modified PTSD Symptom Scale: A brief self-report measure of posttraumatic stress disorder. The Behavioral Therapist, 16, 161–162. Retrieved from http://psycnet.apa.org/record/2011-20330-001

Gillespie CF, Bradley B, Mercer K, Smith AK, Conneely K, Gapen M, … & Ressler KJ (2009). Trauma exposure and stress-related disorders in inner city primary care patients. General Hospital Psychiatry, 31(6), 505–514. doi:10.1016/j.genhosppsych.2009.05.003 [PubMed: 19892208]

Houtkoop BL, Chambers C, Macleod M, Bishop DVM, Nichols TE, Wagenmakers E-J (2018). Data Sharing in Psychology: A Survey on Barriers and Preconditions. Advances in Methods and Practices in Psychological Science, 70–85. doi:10.1177/2515245917751886

Kasirzadeh A (2021). Reasons, Values, Stakeholders: A Philosophical Framework for Explainable Artificial Intelligence. Computers and Society (cs.CY); arXiv:2103.00752

Kennedy E, Panahi S, Stewart IJ, Tate DF, Wilde EA, Kenney K, Werner JK, et al. Traumatic Brain Injury and Early Onset Dementia in Post 9–11 Veterans. Brain Injury 2022 Feb 5:1–8. doi: 10.1080/02699052.2022.2033846.

Kessler RC, Berglund P, Delmer O, Jin R, Merikangas KR, & Walters EE (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Archives of General Psychiatry, 62(6): 593–602. doi:10.1001/archpsyc.62.6.593 [PubMed: 15939837]

Kolen MJ, & Brennan RL. (2004) Test equating, scaling, and linking: Methods and practices Springer. ISBN : 978–1-4419–2304-2 doi: 10.1007/978-1-4757-4310-4.

Lebois L, Li M, Baker JT, Wolff JD, Wang D, Lambros AM, … Kaufman ML (2021). Large-Scale Functional Brain Network Architecture Changes Associated With Trauma-Related Dissociation. The American Journal of Psychiatry, 178(2), 165–173. doi:10.1176/appi.ajp.2020.19060647 [PubMed: 32972201]

Logue MW, van Rooij S, Dennis EL, Davis SL, Hayes JP, Stevens JS, … Morey RA (2018). Smaller Hippocampal Volume in Posttraumatic Stress Disorder: A Multisite ENIGMA-PGC Study: Subcortical Volumetry Results From Posttraumatic Stress Disorder Consortia. Biological Psychiatry, 83(3), 244–253. doi:10.1016/j.biopsych.2017.09.006 [PubMed: 29217296]

Norris Fran H. & Hamblen Jessica L. (2004). Standardized self-report measures of civilian trauma and PTSD. In Wilson JP, Keane TM & Martin T (Eds.), Assessing Psychological Trauma and PTSD (pp. 63–102). New York: Guilford Press. ISBN: 1–59385-035–2

Pomponio R, Erus G, Habes M, Doshi J, Srinivasan D, Mamourian E, … Davatzikos C (2020). Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. NeuroImage, 208, 116450. doi:10.1016/j.neuroimage.2019.116450 [PubMed: 31821869]

Pugh MJ, Kennedy E, Prager EM, Humpherys J, Dams-O'Connor K, Hack D, … Lumba-Brown A (2021). Phenotyping the Spectrum of Traumatic Brain Injury: A Review and Pathway to Standardization. Journal of Neurotrauma, doi:10.1089/neu.2021.0059.

Radua J, Vieta E, Shinohara R, Kochunov P, Quidé Y, Green MJ, … Pineda-Zapata J (2020). Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA, NeuroImage, 218, 116956. doi:10.1016/j.neuroimage.2020.116956. [PubMed: 32470572]

Rauch SAM, Kim HM, Powell C, Tuerk PW, Simon NM, Acierno R, … Hoge CW (2018). Efficacy of Prolonged Exposure Therapy, Sertraline Hydrochloride, and Their Combination Among Combat Veterans With Posttraumatic Stress Disorder: A Randomized Clinical Trial. JAMA Psychiatry, 76(2), 117–126. doi:10.1001/jamapsychiatry.2018.3412

Shelby RA, Golden-Kreutz DM, & Andersen BL (2005). Mismatch of posttraumatic stress disorder (PTSD) symptoms and DSM-IV symptom clusters in a cancer sample: exploratory factor analysis of the PTSD Checklist-Civilian Version. Journal of Traumatic Stress, 18(4), 347–357. doi:10.1002/jts.20033 [PubMed: 16281232]

Suliman S, Stein DJ, & Seedat S (2014). Clinical and neuropsychological predictors of posttraumatic stress disorder. Medicine, 93(22), e113. doi:10.1097/MD.0000000000000113 [PubMed: 25396328]

Sun D, Gold AL, Swanson CA, Haswell CC, Brown VM, Stjepanovic D, VA Mid-Atlantic MIRECC Workgroup, LaBar KS, & Morey RA (2020). Threat-induced anxiety during goal pursuit disrupts amygdala-prefrontal cortex connectivity in posttraumatic stress disorder. Translational Psychiatry, 10(1), 61. doi:10.1038/s41398-020-0739-4 [PubMed: 32066690]

Tate DF, Wade B, Velez CS, Drennon AM, Bolzenius JD, Cooper DB, … Bigler ED (2019). Subcortical shape and neuropsychological function among U.S. service members with mild traumatic brain injury. Brain Imaging and Behavior, 13(2), 377–388. doi:10.1007/s11682-018-9854-8 [PubMed: 29564659]

Thompson PM, Jahanshad N, Ching CRK et al. (2020). ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries. Translational Psychiatry, 10, 100. doi:10.1038/s41398-020-0705-1 [PubMed: 32198361]

Towse JN, Ellis DA & Towse AS (2021). Opening Pandora's Box: Peeking inside Psychology's data sharing practices, and seven recommendations for change. Behavioral Research, 53, 1455–1468. doi:10.3758/s13428-020-01486-1

Weathers F, Litz B, Herman D, Huska J, Keane T (1993). The PTSD checklist: reliability, validity, and diagnostic utility. In Annual Meeting of the International Society for Traumatic Stress Studies San Antonio, TX.

Weathers FW, Litz BT, Keane TM, Palmieri PA, Marx BP, & Schnurr PP (2013). The PTSD Checklist for DSM-5 (PCL-5). Retrieved from the National Center for PTSD at www.ptsd.va.gov.

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

Webb EK, Weis CN, Huggins AA, Fitzgerald JM, Bennett KP, Bird CM, … Larson CL (2021). Neural impact of neighborhood socioeconomic disadvantage in traumatically injured adults. Neurobiology of Stress, 15, 100385. doi:10.1016/j.ynstr.2021.100385 [PubMed: 34471656]

Weiner MW, Harvey D, Hayes J, Landau SM, Aisen PS, Petersen RC, … Department of Defense Alzheimer's Disease Neuroimaging Initiative (2017). Effects of traumatic brain injury and posttraumatic stress disorder on development of Alzheimer's disease in Vietnam Veterans using the Alzheimer's Disease Neuroimaging Initiative: Preliminary Report. Alzheimer's & Dementia, 3(2), 177–188. doi:10.1016/j.trci.2017.02.005

Weis CN, Webb EK, Damiano S, Larson CL, & deRoon-Cassini TA (2021). Scoring the Life Events Checklist: Comparison of three scoring methods Psychological Trauma: Theory, Research, Practice, and Policy. In press doi:10.1037/tra0001049

Wilde EA, Dennis EL, & Tate DF (2021). The ENIGMA Brain Injury working group: approach, challenges, and potential benefits. Brain Imaging and Behavior, 15(2), 465–474. doi:10.1007/s11682-021-00450-7 [PubMed: 33506440]

Wilkins KC, Lang AJ, & Norman SB (2011). Synthesis of the psychometric properties of the PTSD checklist (PCL) military, civilian, and specific versions. Depression and Anxiety, 28(7), 596–606. doi:10.1002/da.20837 [PubMed: 21681864]

Wortmann JH, Jordan AH, Weathers FW, Resick PA, Dondanville KA, Hall-Clark B, Foa EB, … Litz BT (2016). Psychometric analysis of the PTSD Checklist-5 (PCL-5) among treatment-seeking military service members. Psychological assessment, 28(11), 1392–1403. doi:10.1037/pas0000260 [PubMed: 26751087]

Key points:

Question:

The precise relationship between scores on different PTSD assessments remains unclear because it is hard to isolate the effects of instrumentation in practice.

Findings:

We found that individual data sources each come with distinct characteristics and biases that must be addressed before the relationship between different PTSD assessments can be observed and modeled independently.

Importance:

We propose methods that allow scores on different measures of PTSD symptom severity to compared, which may reduce burden on patients, participants, and researchers.

Next Steps:

Future work could use these ideas to situate the results of new studies within the larger body of historical literature.

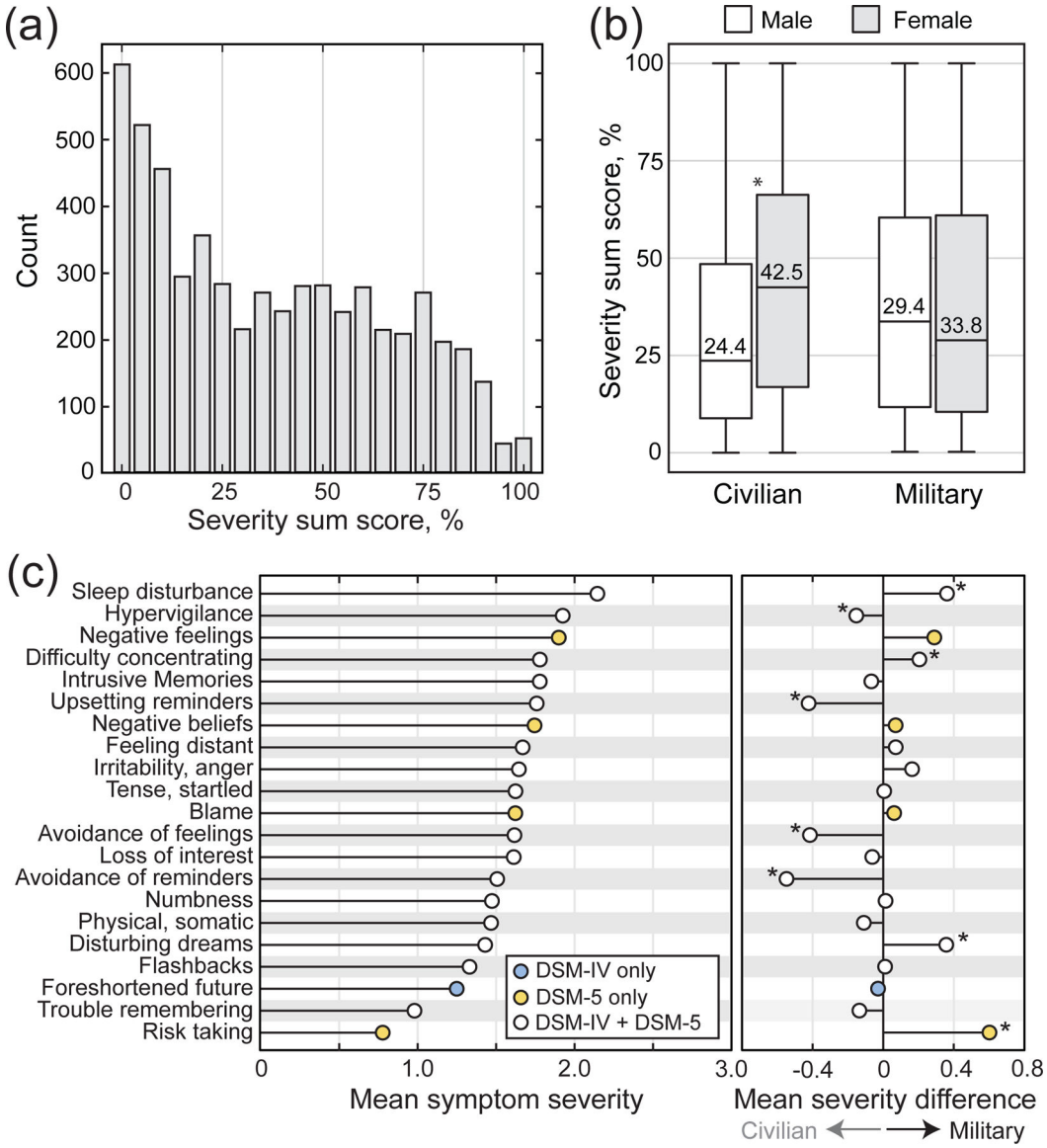

Figure 1: Summary of PTSD symptom reporting across all datasets.

(a) A histogram illustrates the distribution of all raw percentage severity sum scores. **(b)** Percentage severity sum scores are shown broken out as boxplots by military status and sex/gender. **(c)** A stem plot shows the mean severity reported for each item. **(c, right)** The average differences between military and civilian severity scores are shown for each item. (* indicates significance at $p<0.05$ after q correction for multiple comparisons).

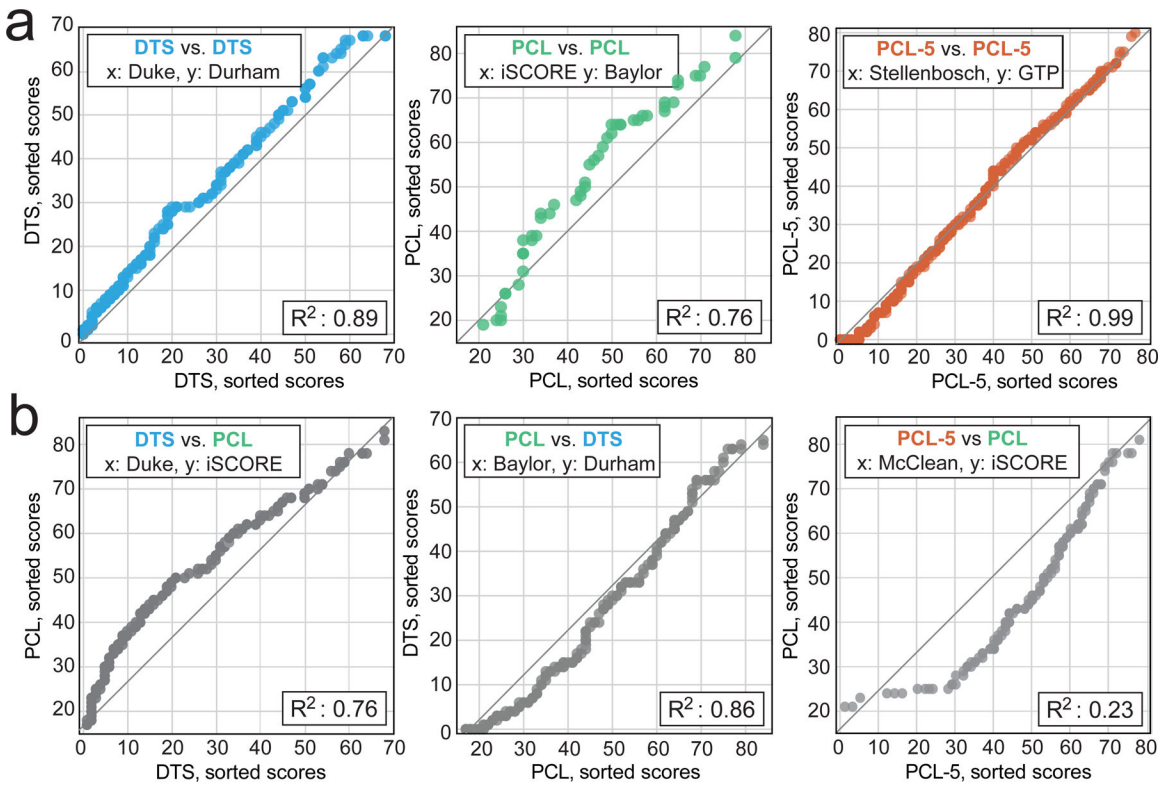

Figure 2: Visualizing raw severity scores across datasets and instruments.

(a) The sorted severity scores of datasets that used the same instrument are compared. The gray line indicates equality from the lowest to the highest possible score on each scale. The coefficients of determination (R^2, inset) measure deviation from the line. **(b)** Like **(a)** but comparing datasets that used different PTSD assessment instruments.

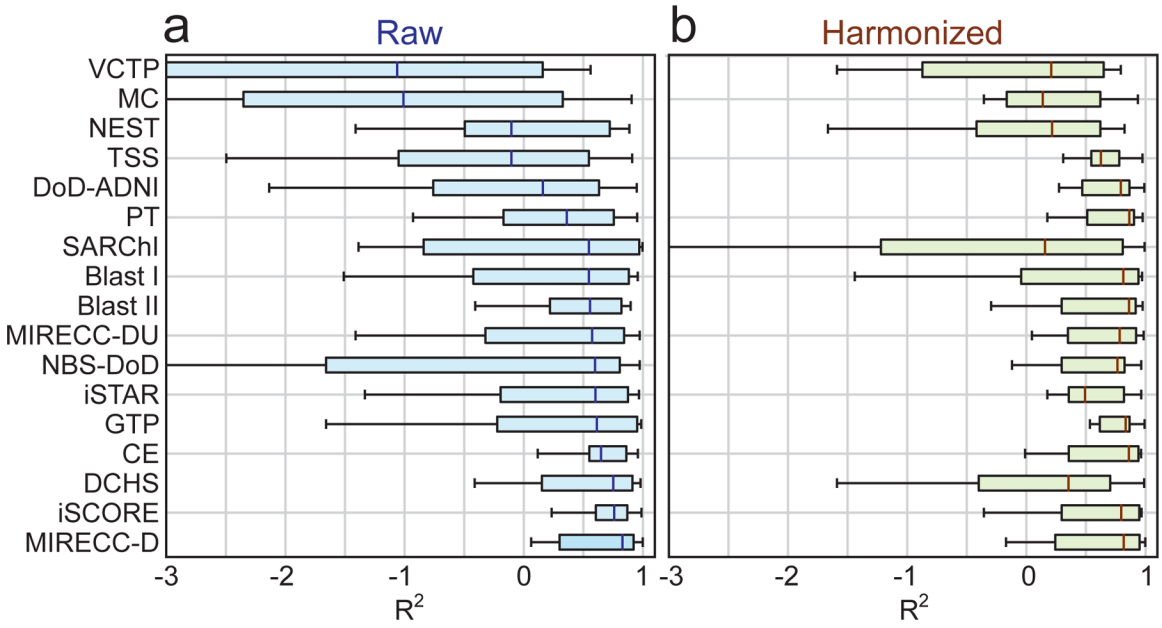

Figure 3: Comparison of pre/post harmonization fit quality.

(a) The distribution of coefficients of determination are shown as boxplots broken out by data source. **(b)** Like **(a)** but after Bayesian correction of site effects.

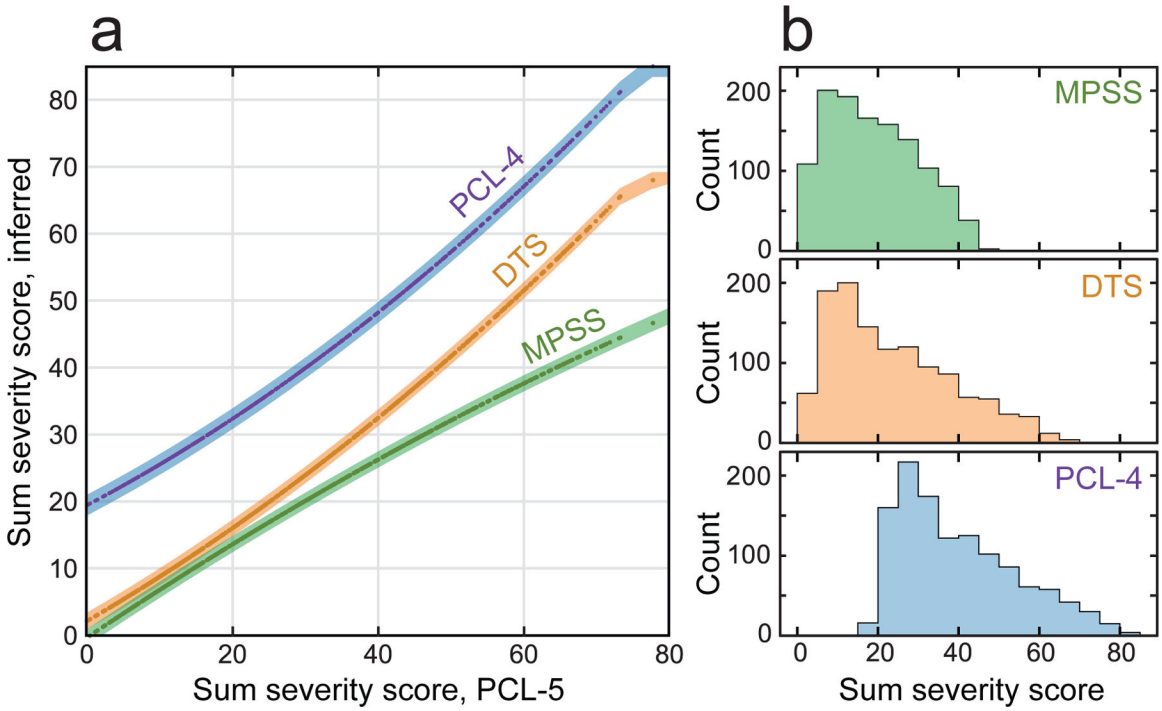

Figure 4: Converting severity scores.

Figure 4a plots model-inferred PTSD severity scores on different instruments as a function of real PCL-5 severity scores. The shaded areas indicate ±1 RMSE. Figure 4b shows the equivalent PCL-5 severity score distributions after conversion.

Table 1:

Summary statistics for the 17 datasets.

Statistics for each of the datasets are shown after exclusion criteria were applied (see Methods).

Dataset	Population	N	Female	Male	Median age	Measure	Mean score
VA-MIRECC-D	Military	2988	0.22	0.78	36	DTS	25.4
VA-MIRECC-DU	Military	208	0.19	0.81	39	DTS	20.8
DCHS	Civilian	327	1	0	26	MPSS	23.1
SARChI	Civilian	591	0.71	0.29	44	PCL-5	33
GTP	Civilian	335	0.86	0.14	44	PCL-5	32.4
iSTAR	Civilian	180	0.58	0.42	33	PCL-5	25.1
TSS	Civilian	135	0.83	0.17	35	PCL-5	49.1
NEST	Civilian	15	0.8	0.2	33	PCL-5	37.4
PT	Military	60	0.1	0.9	38	PCL-5	43.8
VCTP	Military	9	0.33	0.67	38	PCL-5	47.6
CE	Military	75	0.08	0.92	33	PCL-C	44.4
Blast II	Military	50	0.02	0.98	30	PCL-C	51.5
Blast I	Military	47	0.11	0.89	31	PCL-C	39.6
NBS-DoD	Mixed	31	0.29	0.71	29	PCL-C	36.1
DoD-ADNI	Military	242	0	1	69	PCL-M	10.8 *
iSCORE	Military	195	0.12	0.88	34	PCL-M	45.7
MC	Civilian	146	0.28	0.72	27	PCL-C	27.3

* indicates each item was recorded on a binary scale instead of a Likert scale, which results in an apparently low average sum score.

Table 2:
Results of blocked logistic regression predicting elevated PTSD symptom severity.

Raw scores (left) and post-harmonized scores (right) are shown for comparison.

OR (95% CI)	**Positive PTSD screen, Sum score>cutoff**	
Sociodemographic Characteristics		
	Raw $R^2 = 0.007$	*Harmonized* $R^2 = 0.091$
Age group (Ref: 18–29)		
30–39	1.18 (1.03 – 1.35) *	1.16 (1.0 – 1.34) *
40–49	1.14 (0.98 – 1.31)	1.01 (0.87 – 1.18)
50 and older	0.72 (0.60 – 0.85) *	0.7 (0.59 – 0.83) *
Sex/gender (Ref: Male)		
Female	1.16 (1.02 – 1.33)	1.06 (0.93 – 1.22)
Population (Ref: Civ.)		
Military	0.89 (0.58 – 0.77) *	2.74 (2.38 – 3.16) *
Assessment instrument		
	Raw $R^2 = 0.014$	*Harmonized* $R^2 = 0.058$
Inventory (Ref: PCL-5)		
DTS	0.72 (0.63 – 0.81) *	2.51 (2.2 – 2.86) *
PCL	0.58 (0.48 – 0.72) *	1.84 (1.5 – 2.25) *
mPSS	1.25 (0.98 – 1.6)	1.18 (0.93 – 1.5)

* indicates significance at $p<0.05$ after correction for multiple comparisons.

Table 3:
Model parameters for converting PTSD severity scales.

a_0: Offset, β_1: Slope, β_2: Quadratic coefficient, R^2: Coefficient of determination, and e: root mean squared error.

Conversion A→B	a_0	β_1	β_2	R^2	e (RSME)
PCL-5 → PCL-4	19.43	0.572	0.0037	0.982	1.54
PCL-4 → PCL-5	−33.5	1.82	−0.0065	0.983	1.65
PCL-5 → MPSS	−0.35	0.729	−0.0016	0.987	1.2
PCL-4 → DTS	−21.1	1.157	−0.0012	0.971	1.76
PCL-5 → DTS	2.128	0.63	0.00322	0.992	1.23
PCL-4 → MPSS	−22	1.252	−0.0053	0.976	1.42
DTS → PCL-5	−2.2	1.44	−0.0044	0.989	1.5
DTS → PCL-4	19.07	0.906	0.0008	0.971	1.73
DTS → MPSS	−3.14	1.088	−0.0058	0.992	1.06
MPSS → PCL-5	3.081	1.216	0.0077	0.986	1.54
MPSS → PCL-4	21.15	0.65	0.0158	0.975	1.66
MPSS → DTS	3.84	0.775	0.0135	0.993	1.18