Psychiatry Research 323 (2023) 115168

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

Establishing a training plan and estimating inter-rater reliability across the multi-site Texas childhood trauma research network

Jeffrey D. Shahidullah [a,*], James Custer [b], Oscar Widales-Benitez [a], Nazan Aksan [b], Carly Hatchell [a], D. Jeffrey Newport [a,c], Karen Dineen Wagner [d], Eric A. Storch [e], Cynthia Claassen [f], Amy Garrett [g], Irma T. Ugalde [h], Wade Weber [a], Charles B. Nemeroff [a], Paul J. Rathouz [b]

[a] Department of Psychiatry and Behavioral Sciences, Dell Medical School, The University of Texas at Austin, Austin, Texas, USA
[b] Department of Population Health, Dell Medical School, The University of Texas at Austin, Austin, Texas, USA
[c] Department of Women's Health, Dell Medical School, The University of Texas at Austin, Austin, Texas, USA
[d] Department of Psychiatry and Behavioral Sciences, University of Texas Medical Branch, Galveston, Texas, USA
[e] Department of Psychiatry and Behavioral Sciences, Baylor College of Medicine, Houston, Texas, USA
[f] John Peter Smith Hospital, Fort Worth, Texas, USA
[g] Department of Psychiatry, University of Texas Health Science Center San Antonio, San Antonio, Texas, USA
[h] Department of Emergency Medicine, McGovern Medical School at UTHealth Houston, Houston, Texas, USA

ARTICLE INFO

Keywords:
Psychiatry research
Reliability
Inter-rater reliability
Training
Measurement
Trauma

ABSTRACT

Objective: Minimal guidance is available in the literature to develop protocols for training non-clinician raters to administer semi-structured psychiatric interviews in large, multi-site studies. Previous work has not produced standardized methods for maintaining rater quality control or estimating interrater reliability (IRR) in such studies. Our objective is to describe the multi-site Texas Childhood Trauma Research Network (TX-CTRN) rater training protocol and activities used to maintain rater calibration and evaluate protocol effectiveness.

Methods: Rater training utilized synchronous and asynchronous didactic learning modules, and certification involved critique of videotaped mock scale administration. Certified raters attended monthly review meetings and completed ongoing scoring exercises for quality assurance purposes. Training protocol effectiveness was evaluated using individual measure and pooled estimated IRRs for three key study measures (TESI-C, CAPS-CA-5, MINI-KID [Major Depressive Episodes - MDE & Posttraumatic Stress Disorder – PTSD modules]). A random selection of video-recorded administrations of these measures was evaluated by three certified raters to estimate agreement statistics, with jackknife (on the videos) used for confidence interval estimation. *Kappa*, weighted *kappa* and intraclass correlations were calculated for study measure ratings.

Results: IRR agreement across all measures was strong (TESI-C median kappa 0.79, lower 95% CB 0.66; CAPS-CA-5 median weighted kappa 0.71 (0.62), MINI-MDE median kappa 0.71 (0.62), MINI-PTSD median kappa 0.91 (0.9). The combined estimated ICC was $\geq$0.86 (lower CBs $\geq$0.69).

Conclusions: The protocol developed by TX-CTRN may serve as a model for other multi-site studies that require comprehensive non-clinician rater training, quality assurance guidelines, and a system for assessing and estimating IRR.

1. Introduction

The assessment of psychological symptoms by research raters across multi-site research studies presents several methodological and measurement challenges. Ensuring consistency and replicability of measures is paramount, and requires fostering and demonstrating strong inter-rater reliability (IRR; i.e., comparability in scoring/measurement between different raters using the same instruments). Investigators in Texas have undertaken the challenging task of developing a large statewide multi-site pediatric trauma network. Development of the

* Corresponding author at: Department of Psychiatry and Behavioral Sciences, Dell Medical School, The University of Texas at Austin, 1601 Trinity St., Health Discovery Building, Austin, Texas, 78712.
E-mail address: jeff.shahidullah@austin.utexas.edu (J.D. Shahidullah).

https://doi.org/10.1016/j.psychres.2023.115168
Received 7 January 2023; Received in revised form 9 March 2023; Accepted 11 March 2023
Available online 12 March 2023

Texas Childhood Trauma Research Network (TX-CTRN) required thoughtful navigation of a number of challenges including training research raters (hereafter "raters") in the assessment protocol, guarding against rater drift, and estimating IRR, while doing so in a virtual context given challenges posed by the COVID-19 pandemic.

The TX-CTRN was established in 2020 as a multi-site collaboration to develop a state-wide registry of youth, ages 8–20, who had experienced a traumatic event. This registry facilitates the analysis of population health outcomes related to trajectories of mental health following trauma and supports development of predictive models of short- and long-term risks and resilience. The network uses a "hub-and-spoke" organization. The hub represents the anchor site where the research plan is developed and monitored, and training, outreach, and support is provided to the other sites. The spokes represent the 12 academic medical center sites across Texas where participants are recruited and data are collected. For all sites, recruitment is conducted at multiple settings, including hospitals, emergency departments, mental health inpatient and outpatient clinics, and primary care clinics.

Once informed consent from parents and/or legal guardians and assent from youth are obtained, baseline data are collected regarding trauma history, symptoms of PTSD, depression, and other psychiatric disorders, suicidal ideation and behavior, associated comorbidities, medical history, treatment history, service utilization, and social determinants of health. Follow-up assessments are then conducted at 1-month, 6-months, 12-months, 18-months, and 24-months, providing a rich portrayal of the trajectory of mental health outcomes and social supports. Data collection at these time points includes, among others, three rater-administered assessments: (1) Traumatic Events Screening Inventory – Child (TESI-C; Ford et al. 2000), (2) Clinician-Administered PTSD Scale for DSM-5, Child/Adolescent Version (CAPS-CA-5; Pynoos et al., 2015), and (3) MINI International Neuropsychiatric Interview for Children and Adolescents English Version 7.0.2 for DSM-5 (MINI-Kid; Sheehan et al., 2010), Major Depressive Episodes (MDE) and Post-traumatic Stress Disorder (PTSD) modules. Although this longitudinal design is excellent for identifying predictors and outcomes associated with trauma, it also comes with the challenge of establishing sufficiently robust IRR that is critical for ensuring the quality of the data.

There is limited guidance in the literature for establishing a training protocol or estimating and monitoring IRR that could be applied to the TX-CTRN (Rosen et al., 2008). Over 30 years ago, Castorr et al. (1990) raised concern over the lack of observational studies that reported IRR estimation in describing their methodology and acknowledged the lack of available information to guide researchers in these processes. Two decades later, also bemoaned the lack of IRR reporting, particularly in large multi-site studies. In a review of clinical trials of depressive disorders, Mulsant et al. (2002) found that only three multi-site studies reported IRR and the median number of total raters in these studies was five. This lack of inclusion of IRR reporting has remained consistent in psychological assessment studies in general, and trauma studies in particular.

Assessment of PTSD symptoms in youth presents additional challenges. The CAPS-5-CA is a semi-structured interview of both child and parent, which provides a measure of the severity of PTSD symptoms in youth. It is widely considered to be the gold-standard assessment tool with excellent psychometric properties (Weathers et al., 2018). However, excellent reliability presumes that raters are sufficiently trained to overcome the obstacles that make it difficult to assess PTSD in children. One such difficulty is that both parents and youth tend to under-report physical and sexual abuse, including up to 50% of incidents (Grasso et al., 2009; Grant et al., 2020). Additionally, assessing symptom frequency and intensity can be difficult as parental and child reports can be quite discrepant (Scheeringa et al., 2006). This can partly be explained as some of the symptoms of PTSD are not readily observable by parents such as an overgeneralized fear response, nightmares, and dissociation (Cohen and Scheeringa, 2009). Additionally, parental report of child symptoms has been shown to be more strongly associated with the parent's own reaction to the trauma than that of the child (Shemesh et al., 2005). While the child's report is sometimes a more accurate guide to symptom severity, it is limited by the trauma avoidance that is one of the hallmarks of the disorder (Cohen and Scheeringa, 2009). Finally, the wide range of trauma experiences that comprise the inclusion criteria for this study differs from many other trauma studies that target the assessment of trauma symptoms in the aftermath of specific types of traumatic events (e.g., war, gun violence, natural disasters).

In addition to the complexities associated with assessment of children, the background and longevity of the raters plays a role in IRR. While many studies of psychological assessment utilize clinician raters (e.g., Kobak et al., 2005), TX-CTRN predominantly (approximately 80%) used non-clinician (lay) raters. This constraint posed challenges for the reliable assessment of psychological phenomena in interviews using the MINI-KID, TESI-C, and CAPS-CA-5 – all scales in which clinical judgement is necessary. Typically, each of the twelve sites has between 2 and 4 raters at any one time with some staff turnover over the course of the study, requiring the expeditious training and certification of newly onboarded raters. Other raters, however, remain in the study for extended periods. For these raters, guarding against rater drift was an important consideration; processes were therefore established to provide iterative training and longitudinal performance monitoring.

In summary, ensuring consistency and replicability by fostering strong IRR is essential when conducting research using multiple raters, particularly multi-site research in which groups of raters are geographically dispersed, and may have varying degrees of rater experience and rater turnover over time. However, there is currently no gold standard method for training raters in multi-site psychological research using clinical interviews where clinical judgement is required of non-clinician raters. In addition, whereas there are psychometric reviews of the scales used in the study (e.g., Duncan et al., 2018; Ohan et al., 2002; Ribbe, 1996), there are no comprehensive guidelines for developing a system for assessing and reporting IRR in the extant literature. TX-CTRN represents an ambitious undertaking wherein a mixed team of (predominantly) non-clinician research raters were trained to adhere to a rigorous administration standard when conducting a psychological assessment battery across diverse and unique recruitment catchment areas throughout the state. The methods and approaches developed by TX-CTRN may serve as a model for other multi-site projects. Thus, this paper aims to: (1) describe the TX-CTRN rater training curriculum and the certification process for raters, (2) describe the method for empirically evaluating and estimating IRR, (3) present results of IRR estimation and formal statistical inferences (e.g., confidence intervals) at both the item-level and scale-level for rater-administered scales used in the study, and (4) describe a process for longitudinal monitoring for rater drift.

2. Method

2.1. Procedure

2.1.1. Establishing a training and certification plan for research raters

The Institutional Review Board of the University of Texas approved this study. The 12 participating sites were University of Texas Southwestern Medical Center, Baylor College of Medicine, University of Texas Rio Grande Valley, University of Texas Health Science Center at Tyler, University of Texas Health Science Center at Houston, Texas Tech University Health Science Center Lubbock, University of Texas at Austin Dell Medical School, University of Texas Health Science Center at San Antonio, University of North Texas Health Science Center, Texas A&M Health Science Center, Texas Tech University Health Science Center El Paso, University of Texas Medical Branch. At the outset of the study, all raters from the 12 sites attended a virtual orientation organized by the training team that outlined instructions for becoming certified as a rater in the study. Raters then completed a training and certification program consisting of synchronous and asynchronous didactic learning modules, didactic knowledge assessments, virtual standardized patient

encounters, and videotaped mock administrations. Training included both self-report survey scales and rater-administered scales.

2.1.2. Didactic learning modules and knowledge assessments

Training exercises included independent review of the scales used in the study and accompanying tip-sheets. Tip-sheets developed by the hub training team comprised an overview of each scale as well as additional suggestions for administration that were not included in scale instructions. These additional suggestions included considerations related to cultural and developmental factors such as age-related wording alterations and were created based on recommendations from community-based pediatric clinicians within the network who use these scales with children and adolescents in their practices.

Training on the three rater-administered scales (MINI-Kid, TESI-C, CAPS-CA-5) consisted of a live (virtual) training session and a pre-recorded, voice-over presentation with additional in-depth information on the scales and administration procedures. All training meetings were recorded and uploaded to a cloud-based file storage system and shared with site teams to allow the trainings to be viewed as needed when new raters joined the network. Additionally, raters viewed video recorded mock administrations of rater-administered scales completed by expert clinicians who were part of the Hub Training Team (JDS, CH, OW). Training procedures varied for each of the 3 scales. As an example, for the MINI-Kid, trainees completed an online training module produced by the developer of the instrument that included a didactic training on the scale and a mock MINI-Kid interview administration after which they were required to complete a passing score on a knowledge assessment quiz covering the content of the training.

2.1.3. Virtual standardized patient encounters

For the CAPS-5-CA, raters completed an online training module developed by the VA National Center for PTSD (U.S. Department of Veterans Affairs, National Center for PTSD, n.d). This training module included two parts: 1) didactics on trauma, PTSD diagnostic requirements per the DSM-5, and the administration of the CAPS-CA-5, and 2) a standardized virtual patient encounter with real time administration and scoring using voice recognition software. After completing the virtual patient encounter, raters were provided with a score report of their ratings and the correct ratings for each item. Raters were required to obtain a total symptom severity score within 5 points of the correct score to be certified on the CAPS-CA-5. Because this training used the adult version of the CAPS, other supplemental training specific to the child version of the CAPS was provided to raters.

2.1.4. Videotaped mock scale administrations

Raters also conducted mock administrations of the MINI-Kid and TESI-C that were video recorded and submitted along with the rater's scores for both scales to the Hub Training Team for review. The Hub Training Team provided a rating of *Excellent, Good, Fair*, or *Poor* on the following aspects of mock administration and rater scoring: *Fidelity to the Scale* (i.e., the rater states prompts verbatim, states prompts in order, follows all instructions), *Fluency with the Assessment* (i.e., the rater displays comfort with the wording of the scale, displays minimal errors in administration), *Consistency of Scoring* (i.e., the rater scores endorsed items in a similar way, the rater demonstrates an understanding of the scoring criteria), *Delivery* (i.e., personable delivery, the rater is attentive to participant's mood, demonstrates compassion and sensitivity), and *Efficiency* (i.e., the rater uses follow-up probes as needed, moves to the next item once necessary questions are answered, keeps participant on track [if and when applicable]).

Individualized feedback was provided to all raters and a passing score achieved if all aspects of administration and scoring were rated in the *Good* range or higher. If any aspect of the administration/scoring was rated below *Good*, then raters were required to meet with the Hub Training Team for more detailed performance feedback and training before being authorized to resubmit a new mock administration and rater score for that scale. Only after passing scores were obtained for both mock assessments and the CAPS-5 virtual patient encounter, were raters approved to begin assessment of subjects at their site.

2.1.5. Ongoing training and quality assurance monitoring

Once certified, raters attended monthly IRR training meetings and participated in IRR scoring exercises for quality assurance tracking. For these monthly training meetings, all raters were required to submit video recorded administrations of their MINI-Kid, TESI-C, and CAPS-CA-5 assessments with study participants (with consent) via upload to a HIPAA-compliant repository. The Hub Training Team then selected two to three of these videos per meeting and assigned all raters to watch and concurrently score the videos in a Research Electronic Data Capture (REDCap; Harris et al., 2009) survey that linked to a centralized inter-rater reliability monitoring database, which included no identifying PHI. Scoring breakdowns were sent to the raters and their respective site principal investigators (PIs) to illustrate how raters performed in these monthly scoring exercises and how their scores compared with other raters. Specific items with high rater scoring variability (i.e., "problem items") and/or overall scales with substantial discrepancies in scoring were identified as targets for review. These video clips were further analyzed during monthly IRR meetings, and targeted training was provided in which relevant administration and scoring guidelines were highlighted. These video files were erased using appropriate data deletion procedures at the conclusion of each meeting and were not used outside of the training and quality assurance tracking capacity.

The monthly training meetings also included ongoing didactics on topics including, but not limited to, secondary/vicarious trauma, rater self-care, cultural sensitivity, and safety risk assessment and reporting. Meetings included virtual breakout sessions where raters, in small groups, discussed study operation protocols that were successful or challenging at their site, any scoring or administration difficulties they had encountered, and scenarios where they would benefit from extra support. Hub-level "office hours" were held weekly to provide dedicated time to support all raters.

2.1.6. Training dashboard

The network created a central virtual repository (Training Dashboard) to store all training-related materials including scales, tip-sheets, scale-specific training videos, video recordings of all monthly training meetings, FAQ forum, and a question and answer submission form where raters could submit questions anonymously to the Hub Training Team.

2.1.7. Processes for remediation and support for select raters

Raters who demonstrated below expected levels of reliability in the monthly IRR exercises attended extra support sessions with the Hub Training Team for additional training and performance feedback activities. Among these specific activities were review and discussion of raters' video-recorded sessions with participants by the Hub Training Team and shadowing opportunities with more advanced raters. This remediation was successful in rater achievement of expected levels of reliability, and no raters had to be withdrawn due to performance.

2.1.8. Onboarding new raters

As new raters joined the study across the 12 sites, they were required to complete the same pre-certification training described previously and watch all recording training meeting videos stored on the Training Dashboard. Once certified, they are instructed to shadow other more experienced certified raters in the network and conduct at least two supervised administrations before conducting administrations independently.

2.2. Estimating inter-rater reliability

2.2.1. Video ascertainment

Raters video-recorded administrations of the MINI-Kid, TESI-C, and CAPS-CA-5 with TX-CTRN participants, after obtaining consent. Raters then uploaded these videos to the university's instantiation of Box.com, a HIPAA-compliant storage repository. All videos were assigned a unique code number by the Hub Training Team and then randomly selected for reliability analysis using a random number generator. Selected videos were reviewed by the Hub Training Team to ensure they were valid administrations (e.g., all items were administered) and that the audio/video quality was appropriate so that all items in the scale and subject responses could be clearly heard and understood. If a video had no symptom endorsement or there were audio/video glitches affecting one's ability to interpret a question or response, then another video was randomly selected in its place using the method described above.

Videos for reliability analysis were selected separately for each of the instruments analyzed. As the purpose was not to estimate prevalence or means of item endorsements or scales, but rather concordance among raters, selected videos were expected to have a positive signal for at least some of the items. Our aim was $n = 20$ to $n = 25$ videos per instrument. For each video, the interviewer (rater) who performed the in-person interview was noted.

Full videos of the TESI-C and CAPS-CA-5 were used. Video length ranged from 10 to 45 min for the TESI-C and 15 to 60 min for the CAPS-CA-5. Given the length of the MINI-Kid (17 modules for psychiatric diagnoses), the full administration was not used. Rather, the two most commonly endorsed modules in the study – Major Depressive Episodes (MDE) Module and the Posttraumatic Stress Disorder (PTSD) Module – were selected for inclusion in the IRR estimation process. Only videos with a positive endorsement on the screening items for each of these modules were selected for inclusion.

2.2.2. Statistical design

To simplify execution of the experimental design with regard to assignment of raters to videos, for each instrument (or MINI module), we selected raters from our active panel equal in number to the number of videos to be rated. We then used a random Latin square algorithm (via R packages *jmuOutlier* v.2.2 [Garren, 2019] or *magic* v1.6–0 [Hankin, 2005]) to assign the n raters to slots 1 through n for each of videos 1 through n. Because our design called for only 3 ratings per video, we retained only the first three columns of the Latin square. Finally, we ran a check to ensure that the rater was never assigned to rate the video in which they served as the interviewer; if so, we simply re-ran the Latin square algorithm; this constraint was easily satisfied in 4 or fewer iterations. The result is that each video was rated exactly three times, each rater provided exactly three ratings, and it was never the case that the same triplet of raters scored more than one video. It was also rare that any *pair* of raters appeared more than twice among the assignments. The strong balance of this design provided protection against any subset of raters being overly influential in estimation of agreement measures.

2.3. IRR scoring procedures

Raters were instructed to review their assigned videos and concurrently enter their scoring for that video via the REDCap survey.

2.3.1. Data analysis

We used Cohen's (1960) *kappa* as a measure of agreement for all items (which are binary) for the TESI-C, the MINI-Kid MDE, and the MINI-Kid PTSD. For the TESI-C, analysis also focused on the Event Type (experienced trauma vs. witnessed trauma) and Measurement Type (screening item vs. DSM-5 PTSD Criterion A item).

We used weighted *kappa* (Cohen, 1968) for the CAPS-CA-5 items as these are ordinal variables scored from 0 to 4. The following weights were pre-specified: Discordance where one rating is 0 and the other rating was >0: weight=0. Discordance where two ratings differ by one point, where both are >0: weight=0.50. Discordance where two ratings differ by two points, where both are >0: weight=0.25. All other discordances: weight=0. Finally, for the four CAPS-CA-5 summary scores, we used the intra-class correlation coefficient (ICC). This was computed assuming *de novo* independent raters for each video wherein "each target is rated by a different set of $k = 3$ judges, randomly selected from a larger population of judges" (Shrout and Fleiss, 1979). This model is not exactly true but given the large total number of raters (n>20) vs raters per video ($k = 3$), the approximation is very good, as others have found as well. We estimated the ICC using a one-way random effects model, which is the standard in this setting.

For each item's *kappa* and weighted *kappa* estimation, we generated a data set wherein each video triplet was expanded to 6 pseudo-pairs of ratings labeled "rater 1″ and "rater 2″. (Six is the number of unique ordered pairs generated from a set of 3.) We then estimated *kappa* using these pseudo data. This approach ensures exchangeability among all the raters and pairs of raters, analogous to the one-way random effects approach for the ICC; the estimates are statistically valid, although their nominal standard errors (SE) are not. For SE estimation, we jackknifed the n videos (i.e., by leaving out 1 video at a time and re-running the estimation algorithm; Efron and Tibshirani, 1993). We calculated jackknife SEs on the logit scale (for *kappa*, weighted *kappa*, and ICC), computed normal-theory 95% one-sided lower confidence bounds for these agreement statistics on the logit scale, and back-transformed to the natural scale for reporting results. Operating on the logit scale avoided irregularities arising from estimates occurring near the boundary value of one. For each of the four instruments, we also estimated the median and the 25th percentile of *kappa* (or weighted *kappa*, for the CAPS-CA-5) for all the items, along with jackknife confidence bounds for these parameters. We performed analysis in R, estimated *kappa* and weighted *kappa* using R package *psych* v.2.2.5 (Revelle, 2018), and estimated ICC using R package *irr* v.0.84.1 (Gamer et al., 2019).

3. Results

3.1. TESI-C

For TESI-C (Table 1), the data are summarized for each item as the number (count) of videos (out of 21) in which that item was endorsed 0, 1, 2, or 3 times by the three raters, along with the *kappa* and 95% one-sided lower confidence bound. Item specific *kappas* are generally strong, although confidence bounds are low when responses are concentrated in the "all zero" or "all three" columns. The estimated median *kappa* is 0.79 (95% lower confidence bound [LCB]: 0.66) and the estimated 25th percentile *kappa* is 0.71 (95% LCB: 0.60). Results are generally weaker for witnessed (W) than for experienced (E) event types; specifically, the estimated median *kappa* for all experienced events combined is 0.82 with LCB of 0.63 but 0.73 for witnessed events (LCB= 0.59). Similarly, the corresponding values for *kappa* at the 25th percentile is 0.75 (LCB=0.62) and 0.62 (LCB=0.42) for experienced and witnessed events respectively. A similar pattern can be seen for endorsing each traumatic event type versus endorsing DSM-5 PTSD Criteria A criteria for that event. The median *kappa* for the screening item across all event types is 0.87 with LCB of 0.82 and the criterion-A item is 0.72 with LCB=0.62. The corresponding values for *kappa* at the 25th percentile are 0.80 (LCB=0.63) and 0.63 (LCB=0.45).

3.2. CAPS-CA-5

Because CAPS-CA-5 items are ordinal, Table 2 reports the prevalence of each item scored >0, along with the mean score among the scores >0 (thus, that mean is always >1). Many items have high prevalence values in the IRR sample, with relatively few below 0.25. When prevalence is low, measures of agreement are generally weaker, leading to lower confidence bounds. Nonetheless, overall results for agreement are

Table 1
TESI-C Item inter-rater reliability estimation and kappa summaries.

Item Inter-Rater Reliability Estimation TESI-C Item	Event Type*	Measurement Type Endorsed Trauma Screening Item** Number of Raters Endorsing				*Kappa*	lB	Endorsed PTSD Criterion A Item** Number of Raters Endorsing				*Kappa*	lB
		0	**1**	**2**	**3**			**0**	**1**	**2**	**3**		
Experienced Accident	E	8	1	0	12	0.90	0.52	13	2	1	5	0.75	0.43
Natural Disaster	E	8	0	2	11	0.84	0.53	19	0	1	1	0.72	0.15
Hospitalization/Surgery	E	5	0	1	15	0.88	0.43	10	0	4	7	0.72	0.45
Separated from Family	E	11	1	2	5	0.75	0.45	14	3	1	3	0.61	0.26
Attacked	E	11	2	1	7	0.78	0.48	16	1	1	3	0.76	0.37
Threatened to be Attacked	E	18	0	0	3	0.94	0.86	18	1	0	2	0.79	0.20
Mugged	E	17	0	1	3	0.85	0.37	19	1	1	0	0.33	0.10
Kidnapped	E	18	0	0	3	0.94	0.86	18	1	0	2	0.79	0.20
Attacked by Animal	E	13	1	0	7	0.90	0.49	17	0	1	3	0.85	0.37
Sexual Assault	E	15	0	1	5	0.88	0.46	16	1	1	3	0.76	0.37
Bullying***	E	12	0	1	8	0.90	0.52	–	–	–	–	–	–
Cyberbulling***	E	20	0	0	1	0.87	0.16	–	–	–	–	–	–
Summary Median *Kappa*: Event type X measurement type	E					0.88	0.83					0.76	0.58
Summary 25th%ile *Kappa*: Event type X measurement type	E					0.85	0.73					0.72	0.61
Witnessed Accident	W	11	1	0	9	0.91	0.52	13	2	2	4	0.67	0.36
Knew Someone Severely Ill/Injured	W	5	1	0	15	0.88	0.46	10	1	4	6	0.65	0.39
Witnessed Physical Attack	W	14	0	0	7	0.97	0.95	17	0	1	3	0.85	0.37
Witnessed Verbal Attack	W	11	1	1	8	0.84	0.54	17	2	1	1	0.50	0.08
Someone Been in Jail and/or Prison	W	10	1	2	8	0.78	0.50	19	2	0	0	0.13	0.02
Saw People Fight Outside Home	W	14	0	2	5	0.82	0.50	15	2	1	3	0.68	0.32
Saw People Yell/Scream Outside	W	15	0	2	4	0.80	0.47	19	1	0	1	0.68	0.03
Media Exposed to Trauma	W	6	4	3	8	0.54	0.28	18	3	0	0	0.07	0.01
Summary Median *Kappa*: Event type X measurement type	W					0.83	0.66					0.66	0.48
Summary 25th%ile: Event type X measurement type	W					0.80	0.69					0.41	0.07
Other Trauma	–	15	2	0	4	0.78	0.39	19	0	0	2	0.92	0.73

Note: lB = lower bound;.
* *E* = Experienced event; *W* = Witnessed event;.
** 0 or 3 endorsements implies 3 pairwise agreements among the 3 reviewers; 1 or 2 endorsements implies 1 agreement and 2 disagreements among the three reviewers;.
*** Bullying and Cyberbullying items were not in the original TESI-C but were included for the purposes of this study.

strong, with median/25th percentile weighted *kappas* estimated to be 0.71/0.68 (95% LCB: 0.62/0.57). The ICC values for the subtotal scores are all >=0.86 (smallest LCB: 0.69).

3.3. MINI-Kid (MDE & PTSD)

Results for MINI-MDE and for MINI-PTSD (Table 3) are interpreted similarly to those for TESI-C; MINI results are strong, with median / 25th percentile *kappas* estimated to be 0.87/0.80 (95% LCB: 0.82/0.61) for MDD and 0.91/0.84 (95% LCB: 0.91/0.84) for PTSD. Several items deserve special mention: There are 9 items which all three raters either endorsed or did not endorse. Estimated *kappa* is nearly 1.0, and, owing to the lack of estimated sampling variability, the lower bound is typically above 0.96 as well. Two of those items, however, only had one case of "no endorsement", rending the lower bound inestimable. In addition, one item (reckless/destructive behavior) did not have any endorsements, so that information about agreement is non-existent.

4. Discussion

We demonstrate an approach to training and evaluating non-clinician raters and estimating inter-rater reliability (IRR) within a large multi-site longitudinal childhood trauma research study. As noted earlier, there is no gold standard approach to emulate in designing our training protocol to address the following key challenges: a) training non-clinician raters to conduct interviews requiring clinical judgment, b) across multiple sites c) within the constraints of the COVID-19 pandemic, and d) with pediatric populations exposed to a heterogenous array of traumas. Our training protocol included both a certification process and a process to guard against rater drift. The results of an embedded statistical IRR investigation showed very strong (Landis and Koch, 1977) IRR across the three rater-administered scales – TESI-C, MINI-Kid, and CAPS-CA-5. Because we are not aware of any prior studies that provided estimates of IRR among non-clinician raters for these specific instruments, our ability to compare the levels achieved in this study to others in the literature is limited. Nevertheless, the IRR values we obtained exceeded those reported for MINI-Kid (Sheehan et al., 2010). The IRR values we obtained for CAPS-CA-5 were comparable to those obtained in small sample studies (Weathers et al., 2001). These results suggest the training model we adopted is useful and can provide guidance to other large research networks aiming to accomplish similar objectives with naïve raters across multiple sites and with pediatric populations.

Despite our success in training raters and estimating IRR, there are several limitations to consider. First, it is likely that the TX-CTRN and our participants may have higher levels of trauma than the general population. Because our participants were recruited based on history of trauma, prevalence estimates for all trauma types are higher than national rates. However, we expect estimates of association, including (weighted) *kappa* and ICC to not be nearly as vulnerable to such selection bias. In any case, a population sample, without sample sizes that are larger by orders of magnitude, would have such low prevalence as to yield estimates of association with extremely wide confidence bounds. A second limitation regards the assessment setting. In an ideal world, IRR would be assessed via in-person interviews. However, for the CTRN study, in-person interviews were logistically prohibitive, and we have

Table 2
CAPS-5-CA item level statistics and *Kappa* summaries.

CAPS-5-CA Clusters/Items	Prev.	Item Mean**	*Kappa*	lB	Symptom Cluster ICC	Symptom Cluster 95% CI
Criterion B Cluster: Reexperiencing Symptoms						
B1. Intrusive Memories	0.56	1.76	0.71	0.53		
B2. Distressing Dreams	0.32	1.79	0.74	0.51		
B3. Dissociative Distress	0.29	1.32	0.71	0.47		
B4. Cued Psychological Distress	0.60	1.76	0.66	0.46		
B5. Cued Physiological Reaction	0.47	1.54	0.64	0.46		
Summary Cluster B Symptoms					0.86	.77–0.92
Criterion C Cluster: Avoidance Symptoms						
C1. Avoidance of Memories, Thoughts, Feelings	0.61	1.85	0.72	0.53		
C2. Avoidance of External Reminders	0.21	1.75	0.69	0.46		
Summary Cluster C Symptoms					0.89	.81–0.94
Criterion D Cluster: Changes in Mood & Cognition						
D1. Inability to Recall Import Aspects of Events	0.33	1.76	0.57	0.40		
D2. Exaggerated Negative Beliefs or Expectations	0.25	2.00	0.55	0.30		
D3. Distorted Cognitions Leading to Blame	0.43	1.59	0.71	0.51		
D4. Persistent Negative Emotional State	0.75	1.64	0.80	0.61		
D5. Diminished Interest or Participation in Activities	0.41	1.61	0.71	0.50		
D6. Detachment or Estrangement from Others	0.37	1.86	0.81	0.55		
D7. Persistent Inability to Experience Positive Emotions	0.33	2.32	0.72	0.52		
Summary Cluster D Symptoms					0.89	.69–0.97
Criterion E Cluster: Alterations in Arousal & Reactivity						
E1. Irritable Behavior & Angry Outburst	0.33	1.76	0.72	0.51		
E2. Reckless or Self-Destructive Behavior*	0.00	–	–	–		
E3. Hypervigilance	0.35	2.04	0.73	0.54		
E4. Exaggerated Startle Response	0.39	2.03	0.70	0.53		
E5. Problems with Concentration	0.36	1.89	0.79	0.56		
E6. Sleep Disturbance	0.39	2.21	0.57	0.39		
Summary Cluster E Symptoms					0.89	.74–0.96

Note: Prev = Prevalence lB = Lower bound.
* Kappa not estimable owing to zero response variability.
** Mean of positive responses.

found that video-monitoring is a close surrogate and provides an option for easy recording.

There are also a few limitations in the statistical design and analysis. First, the analyses do not take into account systematic differences in the raters. Second, an ideal design would be based on a completely novel set of raters for each video. While logistically not feasible, we have approximated that design with as many raters as we have videos and ensuring that the number of times any given pair of raters appears on a video rarely exceeds two. Technically, the fact that raters appear more than one time in the study introduces a small degree of within-rater correlation. However, the jackknife variance estimator, because it resamples videos, replicates the design even with repeated measures at the rater level. In addition, a more complex analysis would likely involve crossed random effects at both the rater and the video levels. Such an approach would be both computationally quite burdensome and difficult to interpret in terms of simple *kappa* and ICC statistics. Our analysis strikes a balance between the statistical ideal and that which is empirically feasible, valid and clear to interpret.

There are several future directions that can build-on and extend the efforts described in the present study. First, it will be beneficial for future efforts to carry out longitudinal (i.e., repeated serial) IRR monitoring and training to quantify and minimize "rater drift". While this present study did this in an informal way, more robust training and monitoring approaches would be needed. Second, future work can focus on replicating the training and exercises in Spanish and other languages. Third, future work can extend these methods to other measures (e.g., depression, anxiety symptoms scales) and to adult populations. Fourth, future efforts can focus on evaluating the impact of rater characteristics on IRR, including whether raters are clinicians versus non-clinicians, psychiatrists (MD) versus psychologist (PhD) versus Master's degree trained, and previous experience with research-based standard interviews (K-SADS, MINI, etc.) versus no experience. Identifying significant differences among these subgroups may help to establish guidelines for selecting raters for future studies. Finally, inclusion and analysis of those with medical comorbidities (i.e., in which there may be overlap between symptoms of the medical and psychiatric syndromes) will strengthen this kind of research.

Developing an evidence-based gold standard training protocol has potential benefits for many domains of psychiatric research, as it provides a standardized approach by which people are trained. This training protocol may be suitable for adaptation to other large multicenter studies outside of the childhood trauma realm, using naïve non-clinician and clinician raters alike. Given the dearth of mental health providers, establishing effective, reliable, and expeditious rater training protocols using virtual/remote processes will be a game changer within psychiatric research.

CRediT authorship contribution statement

Jeffrey D. Shahidullah: Conceptualization, Methodology, Data curation, Writing – original draft. **James Custer:** Methodology, Formal analysis, Writing – original draft. **Oscar Widales-Benitez:** Methodology, Data curation, Writing – original draft. **Nazan Aksan:** Conceptualization, Methodology, Formal analysis, Writing – original draft. **Carly Hatchell:** Methodology, Data curation, Writing – original draft. **D. Jeffrey Newport:** Conceptualization, Project administration, Methodology, Writing – original draft. **Karen Dineen Wagner:** Conceptualization, Project administration, Funding acquisition, Writing – original draft. **Eric A. Storch:** Writing – original draft. **Cynthia Claassen:** Writing – original draft. **Amy Garrett:** Writing – original draft. **Irma T. Ugalde:** Writing – original draft. **Wade Weber:** Methodology, Data curation, Writing – original draft. **Charles B. Nemeroff:** Project administration, Funding acquisition, Conceptualization, Methodology, Writing – original draft. **Paul J. Rathouz:** Conceptualization, Methodology, Data curation, Writing – original draft.

Table 3
MINI-Kid (MDE & PTSD Modules) item level statistics, distributional summaries and *Kappa* summary.

Module Description	Number of Raters Endorsing*					
	0	1	2	3	*Kappa*	lB**
MINI-Kid (MDE Current Episode)						
Low Mood	14	1	0	12	0.93	0.59
Anhedonia	14	1	0	12	0.93	0.59
Appetite Changes	14	1	2	10	0.83	0.58
Sleep Changes	16	1	1	9	0.87	0.60
Psychomotor Agitation/Retardation	21	1	0	5	0.89	0.45
Decreased Energy	14	0	0	13	0.98	0.98
Guilt	16	1	1	9	0.87	0.60
Concentration Problems	18	0	2	7	0.86	0.58
Suicidality	18	2	1	6	0.79	0.50
Impairment	13	0	0	14	0.98	0.98
Fulfilled Diagnostic Criteria	2	1	1	23	0.72	0.28
MINI-Kid (MDE Lifetime Episode)						
Low Mood	1	0	0	26	0.87	0.15
Anhedonia	1	0	0	26	0.87	0.15
Appetite Changes	2	2	0	23	0.74	0.36
Sleep Changes	4	1	0	22	0.88	0.44
Psychomotor Agitation/Retardation	7	3	3	14	0.67	0.42
Decreased Energy	3	0	2	22	0.76	0.32
Guilt	5	1	1	20	0.83	0.50
Concentration Problems	6	1	1	19	0.85	0.54
Suicidality	9	2	1	15	0.82	0.57
Impairment	1	1	1	24	0.61	0.12
Episodic	12	0	2	13	0.88	0.62
Fulfilled Diagnostic Criteria	2	1	1	23	0.72	0.28
Summary Median *Kappa*[a]					0.71	0.62
Summary 25th%ile *Kappa*[a]					0.68	0.57
MINI-Kid (PTSD)						
Intrusive Memories	3	0	1	17	0.84	0.31
Avoidance	4	0	0	17	0.95	0.90
Trouble Remembering	12	2	2	5	0.69	0.40
Negative Beliefs	7	0	0	14	0.97	0.95
Distorted Cognitions/Blame	7	1	2	11	0.78	0.48
Negative Emotional State	7	0	3	11	0.77	0.47
Diminished Interest	11	0	0	10	0.97	0.97
Detachment from Others	8	0	2	11	0.84	0.53
Diminished Pleasure	11	1	0	9	0.91	0.52
Irritability/Anger	11	0	1	9	0.91	0.52
Reckless/Destructive Behavior	21	0	0	0	–	–
Hypervigilance	10	0	1	10	0.91	0.53
Exaggerated Startle Response	9	0	0	12	0.97	0.96
Problems with Concentration	13	0	0	8	0.97	0.96
Sleep Disturbance	9	0	0	12	0.97	0.96
Symptom Onset	8	1	1	11	0.84	0.54
Impairment	11	0	1	9	0.91	0.52
Summary Median *Kappa*					0.91	0.91
Summary 25th%ile *Kappa*					0.84	0.81

Note:.

[a] The summary Kappa values are for the combined lifetime and current MDD data.

[*] 0 or 3 endorsements implies 3 pairwise agreements among the 3 reviewers; 1 or 2 endorsements implies 1 agreement and 2 disagreements among the three reviewers. **Items with perfect agreement generally have artificially high lower bounds owing to the lack of variability in agreement across jackknife samples.

Declaration of competing Interest

Disclosures: Drs. Shahidullah, Custer, Widales-Benitez, Aksan, Hatchell, Wagner, Garrett, Ugalde, Weber, have reported no biomedical financial interests or potential conflicts of interest.

Dr. Newport has received research support from Eli Lilly, Glaxo SmithKline (GSK), Janssen, the National Alliance for Research on Schizophrenia and Depression (NARSAD), the National Institutes of Health (NIH), Navitor, Sage Therapeutics, Takeda Pharmaceuticals, the Texas Health & Human Services Commission, and Wyeth. He has served on speakers' bureaus and/or received honoraria from Astra-Zeneca, Eli Lilly, GSK, Pfizer and Wyeth. He has served on advisory boards for GSK, Janssen, Merck, and Sage Therapeutics. He has served as a consultant to Sage Therapeutics. Neither he nor family members have ever held equity positions in biomedical or pharmaceutical corporations.

Dr. Eric Storch receives grant support from NIH, the Ream Foundation, Greater Houston Community Foundation, International OCD Foundation, and Texas Higher Education Coordinating Board. He receives book royalties from Elsevier, Springer, American Psychological Association, Jessica Kingsley, Oxford, and Lawrence Erlbaum. He holds stock in NView, where he serves on the clinical advisory board. He was a consultant for Levo Therapeutics, and is currently a consultant for Biohaven Pharmaceuticals and Brainsway. He co-founded and receives payment from Rethinking Behavioral Health, which is a consulting firm that provides support for implementing evidence-based psychological treatment strategies.

Dr. Nemeroff has received research support from National Institutes of Health (NIH). He has served as a consultant to AbbVie, ANeuroTech (division of Anima BV), Signant Health, Magstim, Inc., Intra-Cellular Therapies, Inc., EMA Wellness, Sage, Silo Pharma, Engrail Therapeutics, Pasithea Therapeutic Corp., EcoR1, GoodCap Pharmaceuticals, Inc., Senseye, Clexio, Ninnion Therapeutics, AncoraBio, SynapseBio, BioXcel Therapeutics. He is a stockholder with Seattle Genetics, Antares, Inc., Corcept Therapeutics Pharmaceuticals Company, EMA Wellness, Naki Health, Relmada Therapeutics. He has served on advisory boards for ANeuroTech (division of Anima BV), Brain and Behavior Research Foundation (BBRF), Anxiety and Depression Association of America (ADAA), Skyland Trail, Signant Health, Laureate Institute for Brain Research (LIBR), Inc., Heading Health, Pasithea Therapeutic Corp., Sage. He has served on the Board of Directors for Gratitude America, ADAA, Lucy Scientific Discovery, Inc. He holds the following patents: Method and devices for transdermal delivery of lithium (US 6375,990B1); Method of assessing antidepressant drug therapy via transport inhibition of monoamine neurotransmitters by ex vivo assay (US 7148,027B2).

Dr. Claassen has received support from the National Institutes of Health (NIH), the Borderline Research Foundation, Timberlawn Foundation and the Jordan Elizabeth Harris Foundation. She currently serves as a consultant to LivaNova PLC.

Dr. Rathouz receives research Suppor from the NIH; he serves on a data safety monitoring board for Sunovion Pharmaceuticals.

References

Castorr, A.H., Thompson, K.O., Ryan, J.W., Phillips, C.Y., Prescott, P.A., Soeken, K.L., 1990. The process of rater training for observational instruments: implications for interrater reliability. Res. Nurs. Health 13, 311–318. https://doi.org/10.1002/nur.4770130507.

Cohen, J.A., Scheeringa, M.S., 2009. Post-traumatic stress disorder diagnosis in children: challenges and promises. Dial. Clin. Neurosci. 11, 91–99. https://doi.org/10.31887/DCNS.2009.11.1/jacohen.

Cohen, J., 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol. Bull. 70, 213–220. https://doi.org/10.1037/h0026256.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46. https://doi.org/10.1177/001316446002000104.

Duncan, L., Georgiades, K., Wang, L., Van Lieshout, R.J., MacMillan, H.L., Ferro, M.A., Lipman, E.L., Szatmari, P., Bennett, K., Kata, A., Janus, M., Boyle, M.H., 2018. Psychometric evaluation of the mini international neuropsychiatric interview for children and adolescents (MINI-Kid). Psychol. Assess 30, 916–928. https://doi.org/10.1037/pas0000541.

Ford, J.D., Racusin, R., Ellis, C.G., Daviss, W.B., Reiser, J., Fleischer, A., Thomas, J., 2000. Child maltreatment, other trauma exposure, and Posttraumatic symptomotology among children with oppositional and attention deficit hyperactivity disorders. Child Maltreat. 5, 205–2017. https://doi.org/10.1177/1077559500005003001.

Efron, B., Tibshirani, R., 1993. An Introduction to the Bootstrap. Chapman and Hall, New York.

Gamer, M., Lemon, J., Fellows, I., Singh, P., 2019. irr: Various Coefficients of Interrater Reliability and Agreement. R Package Version 0.84.1 retrieved from. https://CRAN.R-project.org/package=irr.

Garren, S.T., 2019. jmuOutlier: Permutation Tests for Nonparametric Statistics. R Packageversion 2.2. https://CRAN.R-project.org/package=jmuOutlier.

Grant, B.R., O'Loughlin, K., Holbrook, H.M., Althoff, R.R., Kearney, C., Perepletchikova, F., Grasso, D.J., Hudziak, J.J., Kaufman, J., 2020. A multi-method and multi-informant approach to assessing post-traumatic stress disorder (PTSD) in children. Int. Rev. Psychiatry 32, 212–220. https://doi.org/10.1080/09540261.2019.1697212.

Grasso, D., Boonsiri, J., Lipschitz, D., Guyer, A., Houshyar, S., Douglas-Palumberi, H., Massey, J., Kaufman, J., 2009. Posttraumatic stress disorder: the missed diagnosis. Child Welfare 88, 157–176.

Hankin, R.K.S., 2005. Recreational Mathematics with R: Introducing the 'Magic' Package R News 5. (1).

Harris, P.A., Taylor, R., Thieklke, R., Payne, J., Gonzalez, N., Conde, J.G., 2009. Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. J. Biomed. Inform. 42, 377–381. https://doi.org/10.1016/j.jbi.2008.08.010.

Kobak, K.A., Lipsitz, J.D., Williams, J.B., Engelhardt, N., Bellew, K.M., 2005. A new approach to rater training and certification in a multicenter clinical trial. J. Clin. Psychopharmacol. 25, 407–412. https://doi.org/10.1097/01.jcp.0000177666.35016.a0.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

Mulsant, B.H., Kastango, K.B., Rosen, J., Stone, R.A., Mazumdar, S., Pollock, B.G., 2002. Interrater reliability in clinical trials of depressive disorders. Am. J. Psychiatry 159, 1598–1600. https://doi.org/10.1176/appi.ajp.159.9.1598.

Ohan, J.L., Myers, K., Collett, B.R., 2002. Ten-year review of rating scales. IV: scales assessing trauma and its effects. J. Am. Acad. Child Adolesc. Psychiatry 41, 1401–1422. https://doi.org/10.1097/00004583-200212000-00012.

Pynoos, R.S., Weathers, F.W., Steinberg, A.M., Marx, B.P., Layne, C.M., Kaloupek, D.G., Schnurr, P.P., Keane, T.M., Blake, D.D., Newman, E., Nader, K.O., Kriegler, J.A., 2015. Clinician-Administered PTSD Scale for DSM-5-Child/Adolescent Version. Scale available from the National Center for PTSD at. www.ptsd.va.gov.

Revelle, W., 2018. psych: Procedures for Personality and Psychological research, Version 1.8.12. Northwestern University, Evanston, Illinois, USA. Retrieved from. https://CRAN.R-project.org/package=psych.

Ribbe, D., 1996. Psychometric review of traumatic events screening inventory for children (TESI-C). In: Stamm, B.H. (Ed.), Measurement of Stress, Trauma, and Adaptation. Sidran, Lutherville, MD, pp. 386–387.

Rosen, J., Mulsant, B.H., Marino, P., Groening, C., Young, R.C., Fox, D., 2008. Web-based training and interrater reliability testing for scoring the Hamilton Depression Rating Scale. Psychiatry Res. 161, 126–130. https://doi.org/10.1016/j.psychres.2008.03.001.

Scheeringa, M.S., Wright, M., Hunt, J.P., Zeanah, C.H., 2006. Factors affecting the diagnosis and prediction of PTSD symptomatology in children and adolescents. Am. J. Psychiatry 163, 644–651. https://doi.org/10.1176/ajp.2006.163.4.644.

Sheehan, D.V., Sheehan, K.H., Shytle, D.R., Janavs, J., Bannon, Y., Rogers, J.E., Milo, K. M., Stock, S.L., Wilkinson, B., 2010. Reliability and validity of the Mini International Neuropsychiatric Interview for Children and Adolescents (MINI-KID). J. Clin. Psychiatry 71, 313–326. https://doi.org/10.4088/JCP.09m05305whi.

Shemesh, E., Newcorn, J.H., Rockmore, K., Shneider, B.L., Emre, S., Gelb, B.D., Rapaport, R., Noone, S.A., Annunziato, R., Schmeidler, J., Yehuda, R., 2005. Comparison of parent and child reports of emotional trauma symptoms in pediatric outpatient settings. Pediatrics 115, e582–e589. https://doi.org/10.1542/peds.2004-2201.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420–428. https://doi.org/10.1037//0033-2909.86.2.420.

U.S. Department of Veterans Affairs, National Center for PTSD. (n.d). Clinician-administered PTSD Scale for DSM-5 training curriculum. 2022 Retrieved from https://www.ptsd.va.gov/professional/continuing_ed/caps5_clinician_training.asp.

Weathers, F.W., Bovin, M.J., Lee, D.J., Sloan, D.M., Schnurr, P.P., Kaloupek, D.G., Keane, T.M., Marx, B.P., 2018. The Clinician-Administered PTSD Scale for DSM-5 (CAPS-5): development and initial psychometric evaluation in military veterans. Psychol. Assess 30, 383–395. https://doi.org/10.1037/pas0000486.

Weathers, F.W., Keane, T.M., Davidson, J.R., 2001. Clinician-administered PTSD scale: a review of the first ten years of research. Depress Anxiety 13, 132–156.