Published in final edited form as:
*Psychol Inj Law*. 2014 June 1; 7(2): 153–164. doi:10.1007/s12207-014-9188-9.

# The importance of establishing reliability and validity of assessment instruments for mental health problems: An example from Somali children and adolescents living in three refugee camps in Ethiopia

**Brian J. Hall, PhD**[1,2], **Eve Puffer, PhD**[3], **Laura K. Murray, PhD**[1], **Abdulkadir Ismael, BSc PHN**[4,5], **Judith K. Bass, PhD, MPH**[1], **Amanda Sim, MA**[4], and **Paul A. Bolton, MBBS, MPH**[1,6]

[1]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health

[2]Department of Psychology, The University of Macau

[3]Department of Psychology and Neuroscience, Global Health Institute, Duke University

[4]International Rescue Committee

[5]Jigjiga University & the Addis Continental Institute of Public Health

[6]Department of International Health, Johns Hopkins Bloomberg School of Public Health

## Abstract

Assessing mental health problems cross-culturally for children exposed to war and violence presents a number of unique challenges. One of the most important issues is the lack of validated symptom measures to assess these problems. The present study sought to evaluate the psychometric properties of two measures to assess mental health problems: the Achenbach Youth Self-Report and the Child Posttraumatic Stress Disorder Symptom Scale. We conducted a validity study in three refugee camps in Eastern Ethiopia in the outskirts of Jijiga, the capital of the Somali region. A total of 147 child and caregiver pairs were assessed, and scores obtained were submitted to rigorous psychometric evaluation. Excellent internal consistency reliability was obtained for symptom measures for children and their caregivers. Validation of study instruments based on local case definitions was obtained for the caregivers but not consistently for the children. Sensitivity and specificity of study measures were generally low, indicating that these scales would not perform adequately as screening instruments. Combined test-retest and inter-rater reliability was low for all scales. This study illustrates the need for validation and testing of existing measures cross-culturally. Methodological implications for future cross-cultural research studies in low- and middle-income countries are discussed.

### Keywords

PTSD; assessment; reliability; validity; children; adolescents; refugees

Corresponding Author: Brian J. Hall, Ph.D., Department of Psychology, The University of Macau, Av. Padre Tomas Pereira, Taipa, Macau, China, brianhall@umac.mo.

Armed conflict and political instability create humanitarian crises that lead to the displacement and relocation of populations worldwide. The United Nations High Commissioner for Refugees (UNCHR) reports that 7.2 million refugees live in protracted refugee situations for 'five or more years, without immediate prospects for durable solutions' (United Nations, 2012). This trend is likely to continue, as humanitarian consequences of armed conflicts occurring in Libya, and most recently and ongoing in Syria, have not yet been fully appreciated (Yan 2013). For displaced populations, mental health problems present an important public health concern. Identifying children and adolescents that suffer from these issues is critical for effective triage and provision of psychological interventions. In order to assess the mental health needs of populations at risk and to evaluate the effectiveness of treatment provision, accurate and valid measures of psychological distress and disorders are needed.

The circumstances of forced migration include exposure to violence and this is a leading risk factor for poor mental health outcomes (Reed et al. 2012). The mental health problems affecting children living in refugee camps include elevated symptoms of posttraumatic stress disorder (PTSD), depression and anxiety (Barenbaum et al. 2004; Sagi-Schwartz et al. 2008; Lustig et al. 2004). A qualitative study of Burmese refugees conducted in Ban Mai Soi refugee camp in Northwest Thailand demonstrated that children in the camps reported internalizing symptoms of depression, worry, sadness and loneliness. These symptoms related to chronic stressors in the camp setting, including violence, alcohol use, and child maltreatment (Meyer et al. 2013).

There are a number of important considerations in the assessment of mental health problems among refugee communities in low- and middle-income countries (LMIC). First, there may not be instruments that are validated for use in these contexts and within these populations (Hollifield et al. 2002; Mollica et al. 2004). Measures developed for use in one population, that are translated for use in another population, may not capture what it is intended to measure; that is, the scale may lack criterion validity (Bolton 2001). Second, the items included in western instruments may not be useful, appropriate, or relevant in another cultural context. Third, the local idioms of distress may or may not be consistent with western conceptualizations of mental health problems (Kohrt and Hruschka 2010; Rasmussen et al. 2011; Silove et al. 2008). Failure to assess local symptoms can result in missing key indicators of child well being. Finally, the psychometric properties (i.e., reliability and validity) of adapted mental health measures cannot be assumed based on the psychometrics of the original measures; they need to be locally evaluated. Using measures that have not been validated can lead to erroneous conclusions about the mental health of a community (Hobfoll et al. 2011).

Developing reliable and valid instruments, or validating existing tools, is particularly difficult when working in contexts with few mental health professionals where clinical diagnoses of mental health problems may not be possible. In LMIC, there are few trained mental health service providers that can assess children and provide diagnoses. In the country of Ethiopia, there are only .05 psychiatrists, .58 psychiatric nurses, and .02 psychologists (although not trained in clinical psychology) providing mental health treatment per 100, 000 people (Ethiopian Ministry of Health, 2012). Moreover, of the 40

available psychiatrists in the country, only 10 of them work outside the capital city of Addis Ababa. For populations living in remote areas, psychiatric care would be near impossible to access, and may not be familiar with the Somali refugee population. Thus, there is not capacity to diagnose disorders based on existing psychiatric nosology (i.e., Diagnostic and Statistical Manual of Mental Disorders or the International Classification of Disease). Further, and perhaps more importantly, the local relevance of that nosology may also be problematic or not fit the signs and symptoms that are expressed in the context (Phan and Silove 1999). Therefore, alternative methods for evaluating mental health conditions that incorporate locally relevant signs and symptoms is needed in order to establish criterion-related validity.

One field-based approach to instrument validation is to conduct a rapid ethnographic field study, identify local signs, symptoms and concepts of mental distress and illness, and utilize these local definitions to validate an instrument in the absence of a ‘gold standard’ diagnosis (Bolton 2001). These methods have been employed in numerous countries including in Kurdish northern Iraq (Bolton et al. 2013), Democratic Republic of Congo (Bass et al., 2008), northern (Betancourt et al. 2009) and southern Uganda (Bolton et al. 2004), and Zambia (Murray et al. 2011).

## The present study

Armed conflict has pervaded the Somali landscape for decades. In January 1991, the fall of the Mohammed Said Barre government was the impetus of mass migration into neighboring Ethiopia, Djibouti and Kenya. The Ethiopian government constructed 8 refugee camps in the Somali region of Ethiopia, and 7 of these closed following repatriation efforts in the early to mid-2000s.

Fighting escalated in Somalia after December 2006, and new refugees overwhelmed Kebribeyah Camp leading the UNHCR and the Association of Refugee and Returnee Affairs-Ethiopia to open a new camp in the town of Shedder and to reopen a camp in the town of Aw Barre. These camps are near the border of Ethiopia and Somaliland, roughly 60km-70km northeast from Jijiga. As of September 2013, the combined total population of these three camps was 41, 705 people, approximately 60% of whom are youth under the age of 18.

UNHCR conducted an assessment in 2008 and revealed circumstances of child abuse and exploitation. Many of the youth fled with relatives or neighbors, forcing some to be placed with foster families in which further abuse occurred. Children in camps reportedly experienced traumatic events such as men forcing them to fight and witnessing killings if they refused. In 2009, the International Rescue Committee (IRC), confirmed the challenges children face in the camps, with community members identifying lack of family stability and inadequate financial resources as two major risk factors that can contribute to violence and stress in the home. In addition to this 2009 assessment, the Women’s Refugee Commission conducted a qualitative assessment focusing on adolescent girls t in Aw Barre and Sheder in 2012. Findings are that girls are at extreme risk of sexual violence, exploitation and harmful traditional practices. They also face barriers in receiving services (Women’s Refugee

Commission, 2012). According to UNHCR statistics (25 Sep 2012), there were 9,016 children/adolescents at risk (i.e. child protection cases) and 1,117 unaccompanied or separated children across the three camps (UNHCR 2012).

The IRC offers gender-based violence and child protection programs in the camps (International Rescue Committee & UNHCR, 2009). In their comprehensive efforts to provide aid, the IRC initiated a project to assess the mental health needs of the children living in the camps with the goal of evaluating the efficacy of an evidence based mental health treatment. The current study describes the process of validating a mental health assessment instrument in Somali refugee camps and the results highlight the pressing need to undertake these investigations.

## Methods

The current study follows from a qualitative study that was conducted to identify the mental health problems among the Somali children living in Aw Barre, Kebribeyeh and Shedder refugee camps (Puffer et al. 2011). In brief, two qualitative methods were used (free listing and key informant interviews) to identify the mental health problems among children living in these camps. A synthesis of the qualitative data identified internalizing (e.g., sadness and guilt), externalizing (e.g., rule breaking and disruptive behaviors), and traumatic stress (e.g., reminders of traumatic events that occurred) symptoms among children. Specific symptoms and local terms for those symptoms were identified, and mental health assessment instruments that most closely matched these were chosen for adaptation and validation. Local signs and symptoms identified as important during the qualitative study not present on these instruments were included in the complete mental health assessment (termed qualitative items in this paper).

The Achenbach Child Behavior Checklist (CBCL)/Youth Self Report (YSR) (Achenbach and Rescorla 2001) and the Child Post Traumatic Stress Disorder Symptom Scale Interview format (CPSS-I) (Foa et al. 2001) were selected for adaptation as they were thought to best capture the presenting problems identified by the Somali community.

The CBCL/YSR contain two broad scales that measure internalizing and externalizing symptoms. Subsumed under the Internalizing scale are the Anxiety and Depression, Withdrawal and Depression, and Somatic Complaints subscales. The Externalizing scale is comprised of the Rule Breaking and Aggressive Behavior subscales. Additional scales included the Attention Problems, Social Problems, and Thought Problems scales. For the CBCL/YSR, respondents reported their frequency of each symptom in the past four weeks, with responses coded on a 3-point Likert scale ranging from ‘0’ (not true) to ‘2’ (very often true). The CBCL/YSR has evidenced excellent internal and test-retest reliabilities. According to the CBCL/YSR manual, Cronbach’s alphas and test-retest reliabilities for the CBCL Internalizing and Externalizing scale were all above .90. For the YSR, Cronbach’s alpha was .90 for the internalizing and externalizing scales, with test-retest being .80 and .89 respectively (Achenbach and Rescorla 2001).

The list of 15 potentially traumatic events on the CPSS-I was adapted to the local context. Ten of the original items were retained (e.g., *Being in a bad accident, Being hit, punched or kicked, Having your private sexual body parts touched*"). A total of five additional items were added to assess: seeing a dead body, having to flee their home country due to war, female genital cutting, an adult showing pornographic pictures, and an adult showing their private parts. All items were dichotomous, answered as either yes/no. The CPSS-I also includes 17 symptom items for PTSD occurring during the past two weeks with responses coded on 4-point Likert scale ranging from '0' (never) to '3' (all the time). The CPSS has evidenced excellent psychometric properties, with Cronbach's alpha and test-retest values exceeding .80 (Foa et al. 2001).

We used child and caregiver forms of the instrument with both forms inquiring about the child's symptoms. Although the CPSS-I does not have a parent report version, we used the adapted scale to assess parental report of children's symptoms in this study. The internal reliability and combined test-retest and internal reliability are displayed for the CBCL/YSR and CPSS-I in Table 4 and 5 respectively.

All study instruments were translated into Somali (by study author AK), integrating the exact Somali terms used by community members during the qualitative phase. A local language expert conducted backward translation (into English). As a group, study interviewers also participated in item-by-item discussion regarding the meaning, translation, and cultural appropriateness of each item in the local context (led by authors BH and EP). The complete assessment included 164 items: CBCL/YSR with 112 items, CPSS-I with 17 symptom items and 16 traumatic event items, and 19 qualitative items.

Two of the study authors (BJH & EP) conducted a 3-day training of twenty-one interviewers from Jijiga and 3 supervisors in structured interviewing techniques and research ethics. One of the study authors (AK) facilitated the training by providing all sequential live translations. All interviewers were native Somali speakers and a large number had also participated in the previous qualitative study so were familiar with the study topics. One supervisor in each camp oversaw the data collection procedures and BJH and EP monitored daily data collection. Study authors PB and LM provided field support and supervision via phone/ Skype throughout the conduct of the field study.

## Case identification

Criterion-related validity is the extent to which a measure agrees with another measure known to be valid, typically an existing gold standard. In our present study, criterion-related validity was explored based on whether scores on a scale thought to measure a certain concept (e.g., internalizing symptoms) were able to differentiate a group of children known to have that problem (i.e., "cases") from a group of children known to not have that problem. As in most LMIC, there was no locally established method of case identification of known good validity against which to test our new one – the study instrument. Without this established standard, a new one needed to be created.

We used a methodology of case identification based on local definitions of non-professional mental health service providers (see Applied Mental Health Research Group, DIME Module

2 for a complete description of these methods). IRC refugee camp social workers from each camp provided lists of children who they thought had one or more problems (i.e., internalizing, externalizing, traumatic stress), as well as children who were thought not to have any of these problems (i.e., 'non-cases'). They developed the lists using brief screening criteria that were derived from the qualitative data; these were the same as those used by the caregiver/child dyads as described below. These referral lists contained 129 children with internalizing symptoms, 115 with externalizing symptoms, 58 with trauma symptoms, and 210 non-case children. Some of these referral lists were overlapping, as some children were thought to have more than one of these problems.

Interviewers then went to Aw Barre, Kebribeyeh, and Sheder refugee camps in May of 2012 to locate the children and caregivers on the referral lists. This occurred within two weeks from case identification by camp social workers. As an additional method to confirm whether children really were likely to have the referral problems, interviewers asked the caregivers and the children themselves. Interviewers asked these child-caregiver pairs three screening questions. The following questions were asked of each child and caregiver for each problem:

- Internalizing: "We learned from the community about children who have problems of being alone, silent, thinking too much or disappointed. Do you think that any of these words describes you/your child?"
- Externalizing, "We learned from the community about children who are disobedient, bothersome or troublesome. Do you think that any of these words describes you/your child?"
- Traumatic Stress "We learned from the community about children who are fearful. Do you think that this describes you/your child?"

Responses to these questions were then used to establish whether there was agreement on the presence/absence of the problem(s) between the social worker referral, parent, and child. When there was concordance across all three people, the child was considered to have a "local diagnosis" of that problem category, meaning that the child was viewed to have that problem based on the perspective of the child and two people who knew them well. In the absence of a culturally-valid mental health diagnostic process, this local perspective provides a community-based form of "diagnosis" (Bolton 2001).

Following case identification, interviewers located concordant caregiver-child pairs and administered the full mental health questionnaire to both the children and their caregivers. A sub-sample of pairs was also selected for re-testing one week later to evaluate combined test-retest and inter-rater reliability. The sample was balanced by camp, age, and child sex. Different interviewers conducted the initial and re-test interviews. Verbal informed consent or assent was obtained for all participants prior to their participation.

## Statistical analyses

**Item endorsement and relevance—**The data were reviewed to identify individual symptoms that had a low frequency of affirmative responses. The purpose of this was to identify whether some items were less relevant in the local context. Items were considered

not relevant and candidates for deletion if greater than 90% of the study population did not respond to the item at all (response was either missing or refused) or responded with a non-affirmative response (i.e., a zero response indicating that they did not experience the symptom at all in the prior 2 or 4 weeks). Exception to this deletion criteria were items that measured sensitive topics or involve possible self- or other-harm, which were retained. These items often occur with low frequency, but when they do occur, they are clinically important.

**Scale reliability—**We evaluated reliability by calculating: (a) the Cronbach's alpha, which measures internal reliability; (b) the Cronbach's alpha if an item was deleted, which measures whether the Cronbach's alpha coefficient changes after the deletion; and (c) item-test correlations for each item. Item-test correlations measure how well each item on a scale is correlated with the other items and therefore how useful it is in measuring the underlying scale construct (e.g., externalizing symptoms). Decision to remove an item was based on an a priori specification of item-test correlation below .30 and large increases (>.10) in Cronbach's alpha when an item was removed. An increase in the Cronbach's alpha was a secondary consideration, as increases suggest increased confidence in the appropriateness of item removal.

We evaluated 19 items from the qualitative study (Puffer et al. 2011) to determine whether they were appropriate to add to the internalizing, externalizing or traumatic stress symptom scales. These items were considered based on their potential to improve the local relevance of these scales by broadening the symptoms of a concept to include locally occurring problems. Study authors (BJH, EP, LM, PB) independently rated each item for their inclusion on different scales based on item face validity and discussed discrepancies until consensus was reached. Final determination about whether the items were included and on which scales was based on author consensus and psychometric performance of the item (i.e., item-test correlation, change in Cronbach's alpha coefficient).

**Combined test-retest and inter-rater reliability—**Combined test-retest and inter-rater reliability was evaluated for each of the scales to determine whether the performance of the scales would remain consistent over time if it were to be used repeatedly with the same children and caregivers, with different interviewers (i.e., raters) and under similar testing conditions (e.g., the refugee camp setting). Test-retest reliability can be divided into a participant's true score plus measurement error and the natural change in the concept expected to occur in children or caregivers over time. Inter-rater reliability evaluates whether similar results would be found when two or more interviewers administer a test to the same child or caregiver. The combined inter-rater and test-retest correlation was computed using Pearson or Spearman correlations depending on the distribution of the scale. Pearson correlation is appropriate when two variables are linearly associated and Spearman correlation is appropriate when variables are associated non-linearly.

**Criterion-related validity—**Independent samples t-tests were conducted to compare children identified as having a specific problem (i.e., cases) with those identified as not having any of the specified mental health problems (i.e., non-cases). Criterion-related

validity was evaluated based on whether there were meaningful and statistically significant differences in means between the case and non-case children.

**Receiver operator characteristic curves (ROC)—**ROC curves were used to evaluate how well each assessment scale differentiated between cases and non-cases. The area under the curve (AUC) was used to indicate how well each scale was able to achieve this. We generated ROC curves separately for each scale. ROC analyses were then conducted to identify sensitivity and specificity values for each of the scales with the goal of establishing a cut-off score for each scale that could be used to identify children with clinically significant symptoms. We aimed to identify cut-off scores that had the highest sensitivity and specificity giving equal weight to both such that the scale would do as well at selecting children with clinically significant problems (i.e., sensitivity) as omitting children who do not have clinically significant problems (i.e., specificity). All analyses were conducted using STATA version 12.1 (StataCorp 2012).

## Results

### Participant screening

Of the 431 children listed on the referral sheets to be screened, 80 children were ineligible for reasons including being younger than 7 years old or older than 18, moving/resettlement, being unavailable when visited, or not having a caregiver. Nine child/caregiver pairs refused to participate, and 57 were not needed as the target recruitment numbers were reached before they were approached (see Table 1). Of the 288 children and caregivers who consented and were screened, 256 were concordant for either one or more of the specified mental health problems or for having none of these problems. Of this consented sample, 72 did not participate in the full assessment for reasons including refusals (2), being unavailable (2), and inability to verify the identity of the children (2); 66 were not approached because the target assessment sample size was reached. A final enrolled sample included 183 child and caregiver pairs. Of this number, 159 (87%) pairs were included in the analysis; others were excluded due to incomplete interviews and errors in data entry and/or participant identification.

The majority of the caregivers were the child's biological parents (72%). The average amount of time the caregivers reported living in the camps was 11 years (SD=8.60). See Table 1 for a summary of the participant characteristics and Table 2 for the summary of comorbidity occurring within this sample based on the case identification results. Comorbidity refers to when the child had more than one co-occurring problem. The most common comorbidity was between the internalizing and trauma groups – with 33.62% of the children having both these problems.

### Review of individual items

Although there were some items with low frequency of endorsement, none were deleted given the sensitive or critical nature of these items to children's safety and mental health. Examples of infrequent responses from the exposure items were *"Being in a bad accident," or "Being shown pornographic pictures,"* which were reported less than 10% of the time.

Some items from the Achenbach instrument were also infrequently reported, and these items included *"Drinking without approval", "I set fires," "I steal at home," "I think about sex too much,"* and *"Using drugs for nonmedical purposes."*

### Scale reliability

For the purpose of analysis, a total of 10 items were removed from the child YSR scales and 15 items were removed from the caregiver CBCL scales based on lower than acceptable item-test correlations. Examples include *"I am afraid I might think or do something bad," "I am too shy or timid." "I am too dependent on adults."* All subsequent analyses were performed without these items.

### Inclusion of qualitative items

With the exception of two items, all symptoms that were generated from the initial qualitative study but were not already represented in the standards measures functioned well on their respective scales and were retained. One item, *I think over something too much*, performed well on both internalizing and traumatic stress, so this item was retained on both scales. A list of these qualitative items and the scales on which they were included is displayed in Table 3.

### Internal consistency reliability

Each of the scales had adequate to high internal reliability. Cronbach's alpha for each of the child scales ranged from .76 to .96 and from .68 to .95 for the caregiver scales (see Table 4).

### Combined test-retest and inter-rater reliability

Initial results revealed low test-retest correlations (i.e., below .50) for many of the measures, indicating poor performance for both the child and the caregiver assessments. In order to identify the potential source of these low correlations, scatterplots were inspected for outliers: cases that performed very differently from others and were suspected to be highly influential in the analyses. Three caregivers were identified as reporting their child's symptoms very differently from the other caregivers in the sample on at least 8 of the scales and were removed from the analyses (ratings of symptoms were entirely unreliable with baseline being, for example, 20 at the initial assessment, and retest value of 0). Removal of these cases increased the combined test-retest and inter-rater reliability of the measures, though the coefficients remained low. Based on field reports, we suspected that the children who were 7 years old (n=10) may not have been able to answer the questions consistently and as a result, could be biasing the results. Therefore we conducted analyses again after removing a total of 13 child-caregiver pairs and this increased the reliability of the scales. After this adjustment, the test-retest correlations demonstrated that the caregiver CPSS ($r$ = .72), Internalizing scale ($r$ = .60), Social Problem scale ($r$ = .68) and Thought Problem scale ($r$ = .77) were performing well (i.e., scores above .60). However, this was true only for the caregiver scales. The child scales were still not adequate with only one scale combined test-retest and inter-rater reliability correlation exceeding .60 (thought problems). See Table 5.

### Criterion-related validity

Table 6 shows the means and standard deviations for each of the symptom groups comparing cases and non-cases. Results provided evidence for criterion-related validity for the Achenbach Internalizing Scale that included qualitative items and the CPSS-I. For the child report, the internalizing cases reported mean values of 21.32 (SD=18.02) versus 10.94 (SD=13.82) for the non-case group on the Achenbach Internalizing scale. The traumatic stress cases reported mean values of 20.49 (SD=15.55) versus 9.19 (SD=11.71) for the non-case group on the CPSS-I. No statistically significant difference was noted between the Externalizing cases and the non-cases for the Achenbach Externalizing scale.

For the caregiver cases, criterion-related validity was noted for the Achenbach Internalizing Scale, Achenbach Externalizing scale that included qualitative items, and the CPSS-I. The caregivers of Internalizing cases reported mean values of 20.01 (SD=11.60) versus 11.54 (SD=9.10) for the non-case group on the Achenbach Internalizing scale. The caregivers of Externalizing cases reported mean values of 14.33 (SD=12.93) versus 5.59 (SD=6.67) for the non-case group on the Achenbach Externalizing scale with qualitative items. Caregivers of traumatic stress cases reported mean values of 24.35 (SD=15.47) versus 10.29 (SD=11.95) for the non-cases on the CPSS-I.

### Receiver operator characteristics (ROC)

The area under the curve (AUC), sensitivity, specificity, and cut score values for the scales demonstrating criterion validity is displayed in Table 7. Most of these values were above .70 and the values for the 95% confidence intervals were typically above .50, indicating better than chance classification could be obtained for most scales. However, the identified cut-off scores for each measure were low, as were the sensitivity and specificity values.

## Discussion

The purpose of this paper is to illustrate the importance of psychometric evaluation of mental health measures cross-culturally. We describe the results of an instrument adaptation and psychometric evaluation process that is useful for this purpose in international contexts where indigenous mental health instruments have not been developed. This study attempted to validate instruments that would be useful for screening and symptom monitoring during mental health programming. Applying a rigorous methodology for instrument validation, a number of concerns were identified suggesting that the measures may not be appropriate for use as screening tools or to measure change over time. This further stresses the importance of validation efforts in cross-cultural work. Without conducting this study there would have been no way to assess the assumption of validity and reliability of the measures.

We obtained excellent results for internal consistency reliability for both the child and caregiver scales. This indicated that the items on the scales related to one another well. All but two of the locally derived items from the qualitative study were included on relevant scales. The Cronbach’s alpha values were increased when these items were included, which is expected as longer scales typically increase scale reliability. What is notable is that they did not decrease, suggesting that they were appropriately assigned to corresponding scales.

The inclusion of these items allowed us to broaden the traditional western symptom profiles to assess locally relevant and meaningful symptoms that might have been missed by using only items from the existing scales. This is of particular importance when working within communities that do not have a systematized mental health nosology for the arrangement of symptoms and syndromes (Bass et al., 2007). Indeed, one of the major strengths of approaching cross-cultural mental health assessment with this openness is that it can improve the ability of measures to accurately portray peoples' suffering within their 'local moral worlds' (Kleinman 1988).

The low correlations between assessments demonstrated weak support for combined test-retest and inter-rater reliability of the scales. We attempted to identify sub-groups of children (e.g., over the age of 7) and corrected for flawed reporting in order to eliminate noise from the results. However, correlations remained low between the two administrations. The child scales were particularly low, indicating that it would not be appropriate to rely on child self-report to assess changes in symptoms over the course of treatment. The caregiver test-retest correlation was also quite low for all measures except the CPSS-I. Given the high internal consistency of the measures, it may be possible that serial administrations by the same interviewer could mitigate one potential source of measurement error. The overall length of study instrument may have contributed to these results.

Caregiver instruments performed well for criterion related validity testing. The internalizing scale differentiated between internalizing cases and non-cases, as well as traumatic stress cases and non-cases. The externalizing scale differentiated between externalizing cases, internalizing cases, and traumatic stress cases. The CPSS-I discriminated between child traumatic stress, internalizing, and externalizing cases and non-cases. Children's scales performed similarly suggesting that child reported symptoms of internalizing and traumatic stress are consistent with their local diagnosis. The CPSS-I also provided evidence of discriminant validity since the mean value was not significantly different between the externalizing and non-case groups. Children who showed comorbidity for internalizing symptoms and traumatic stress also reported a greater number of traumatic exposures, which is consistent with a perspective of shared etiology of these symptoms, or may be related to the comorbidity in the current study.

Children with a local diagnosis of externalizing problems did not report symptoms consistent this diagnosis. Therefore, it is possible that this scale may not be valid in this population, either due to content or the ways in which the questions were asked. It is also possible that social desirability influenced their willingness to endorse specific signs and symptoms, even though they did endorse these types of problems during the screening phase. Previous studies suggest that children are better at reporting their internalizing rather than externalizing symptoms (Achenbach et al., 1987; Bass et al., 2013).

Results from ROC analyses indicated that caregiver scales were superior to child scales in terms of differentiating between cases and non-cases. This was most clearly true for externalizing cases. Despite this apparent superiority, statistical comparisons between AUC values showed that neither the caregiver trauma nor the internalizing scale were actually superior to the child-report for screening. For example, the AUC of .77 (caregiver) is

statistically equivalent to .73 (child) for the prediction of traumatic stress cases (results not shown).

Initial field-testing of the cut-scores identified in the ROC analysis was conducted with only the caregiver measures and yielded poor results. The cut-scores were unreasonably low, which resulted in frequent false positive screenings when scores were at or near to the cut score (e.g., 6 on externalizing). Since one of the most important uses of these measures was for screening for case identification, we set a higher cut score as has been done in previous studies by using a score that is equivalent to answering “sometimes” or higher on all of the items in a scale (e.g., Murray et al., 2011). This increased the cut score, provided additional confidence that a child was symptomatic, and still allowed for heterogeneity in the symptoms being reported by the children. We also used the child self-report questionnaires in addition to the caregiver since children are generally better at self-reporting internal states (internalizing problems) and worse at reporting on their behavioral problems (externalizing) than caregivers (Bass et al., 2013).

## Lessons learned

The results of the current study highlighted a number of context- and human resource-related challenges that may have influenced results. These are important to discuss, as they may apply to other similar measure validation efforts being conducted cross-culturally in low-resource or humanitarian settings. The first challenge was limited human resources, as is often the case in insecure contexts. Many interviewers did not have previous training or experience in structured interviewing techniques or in mental health or psychosocial assessment. Since administering the types of assessment measures evaluated in this study requires intensive training and practice for inexperienced interviewers, we found that additional training time would have been valuable, particularly for more hands-on practice. This would have minimized questions about whether the study results were truly due to the performance of the measure or to variable performance of the interviewers.

Although the rapid methodology is an asset in contexts that are in need of service provision, it also provides less flexibility and time for in-depth training, careful recruitment, and tracking a highly transient refugee sample. The logistical challenges also exacerbated these concerns. The camps were located 2–3 hours away from the main city, which meant long daily commutes for the study staff which contributed to fatigue and delays in data collection caused by vehicle malfunctions. Future studies conducted in similar challenging settings may need to include additional time for testing instruments.

Following the conclusion of the validity study, lessons continued to be learned that illuminated flaws in the measures, highlighting the need for flexibility during longer-term implementation of the assessment measures. Several issues emerged with the translations of study items and these items were re-translated. This highlights that concerns with translations can exist even following careful translation work with native speakers and the item-by-item checking that was done. The answer options for participants were also thought to be problematic. The response options for the Achenbach measure was changed from “Not True,” “Somewhat or Sometimes True” and “Very True or Often True” to “Not True,”

"Sometimes True" and "Often True," since the local team believed that this would be easier for children to follow.

The cultural appropriateness of some of the items was also revisited. When doing work with diverse communities, care needs to be taken in balancing the needs to assess sensitive topics (e.g., sexual abuse), while also respecting some boundaries related to these sensitive issues in some cultures. For example, in this study, the IRC ultimately decided to remove some items related to sex, including "*I think about sex too much*," and "*plays with own sex parts in public*."

The local team felt that asking such questions would cause offence in the religious and cultural context, and potentially jeopardize relationships with the local community, as most if not all participants were of Islamic faith. The study team disagreed about whether these items should be removed based on experience that interviewers and others in the community normally have greater concerns about such questions than interviewees, if the instrument is properly administered. The choice to remove an item based on religious and cultural considerations needs to be made judiciously and balanced against the overall study aims, participant welfare, and maintaining community/stakeholder engagement. In the present study, the items were removed out of respect for IRC's concerns.

Locating, identifying, and tracking study participants proved difficult in the refugee camp context, especially given that the study took place in three camps simultaneously. As is typical of refugee populations, participants moved quite a lot during the study (e.g., within the camp, the surrounding community, or out of the country due to resettlement). This made it difficult to conduct a study that requires immediate participation at multiple time-points. Related, problems relocating some of the children and establishing their identity as study participants was problematic for assessing combined test-retest and inter-rater reliability. It was difficult to locate our target number of 30 child and caregiver pairs for these analyses. This was compounded by problems encountered with the quality of the data obtained from some of these participants, necessitating the removal of some cases, which reduced the sample size. A small sample size may have influenced the reliability estimates. Further, the study population was referred based on whether they had one or more problems or were known not to have one of these problems. Therefore, the study population may not be fully representative of the refugee camp population.

There are technologies that can aid future studies. First, marking the residence with a QR or barcode code or other participant identifier in a stable place (e.g., under the bed) could allow quicker verification of whether a study participant lives in a certain home. Taking GPS coordinates of the home may also aid in this. Certainly, photographs taken of the caregivers and children would allow for easier identification of study participants, though explaining this would need to be very clear during informed consent procedures, and local consultation should be conducted first to uncover any safety, confidentiality and cultural concerns (e.g., taking photographs of females is considered taboo in many Islamic societies). We also used paper and pencil methods for the data collection. This led to significant delays in evaluating the quality of the data and understanding how the measure was performing in the field. Implementing electronic data collection procedures could allow for real-time and rapid

analysis in the field to evaluate problems when course corrections could more easily be made. This study was conducted in a humanitarian setting and as such, the adaption of technologies should be carefully assessed for safety or ethical concerns given the politically sensitive nature of refugee camp settings.

The approach to assessing test-retest and inter-rater reliability simultaneously was problematic. We were not able to decipher whether the interviewers were inconsistent in their interviewing of the children and caregivers or if the measure itself was poorly performing or a combination. Future studies could implement a design that utilizes the same interviewers to evaluate test-retest, and a separate group of different interviewers to assesses the inter-rater reliability. A comparison of the results for both of these methods could provide better insights into the functioning of the instrument.

### Conclusion

The present study illustrates the need to undertake the task of rigorous psychometric evaluation of mental health measures. Although the instruments chosen in this study have excellent psychometric properties in diverse settings, the performance of these instruments cannot be assumed to be equivalent in all contexts. The use of non-validated instruments can lead to conclusions about mental health burden and response to treatment intervention that may be inaccurate. In instances such as the one described in this study, alternative qualitative assessment of intervention impact may be needed, while the measures are used cautiously, in light of the limitations uncovered.

## Acknowledgments

The present research was made possible by a grant from the Bill and Melinda Gates Foundation: Changing Lives: A Learning Initiative to End Violence Against Women and Children in Emergencies. Dr. Hall was supported by the National Institute of Mental Health T32 in Psychiatric Epidemiology T32MH014592-35 and through the Fogarty Global Health Fellows Program Consortium comprised of the University of North Carolina, John Hopkins Bloomberg School of Public Health, Morehouse and Tulane (1R25TW009340-01). We thank the dedicated study interviewers, Sarah Katherine Baird, IRC Community Well-Being Initiative Manager, Aden Abdi Hiss, IRC Transportation Officer, Asyia Abdulahi Elabe, IRC Caring for Child Survivors Officer, Tensay Tefera IRC CYPD-Child, Youth and protection Department Manager, Shewaye, IRC CYPD coordinator, Ayalew, IRC Child Protection Manager, Anjuli Shivshanker, IRC Research and Evaluation Officer, for her assistance with data collection, and the children and caregivers who participated in this study.

## References

Achenbach, TM.; Rescorla, LA. Manual for the ASEBA School-Age Forms & Profiles. Burlington, VT: 2001.

Applied mental health research group. DIME module 2. 2013 http://www.jhsph.edu/research/centers-and-institutes/center-for-refugee-and-disaster-response/response_service/AMHR/dime/VOT_DIME_MODULE2_FINAL.pdf2013.

Barenbaum J, Ruchkin V, Schwab-Stone M. The psychosocial aspects of children exposed to war: practice and policy initiatives. Journal of Child Psychology and Psychiatry. 2004; 45(1):41–62. [PubMed: 14959802]

Bass JK, Ayash C, Betancourt TS, Haroz EE, Verdeli H, Neugebauer R, et al. Mental health problems of displaced war-affected adolescents in northern uganda: Patterns of agreement between self and caregiver assessment. Journal of Child and Family Studies. 2013

Bass JK, Ryder RW, Lammers MC, Mukaba TN, Bolton P. Post-partum depression in Kinshasa, Democratic Republic of Congo: validation of a concept using a mixed-methods cross-cultural approach. Tropical Medicine International Health. 2008; 13(12):1534–1542. [PubMed: 18983279]

Betancourt TS, Bass J, Borisova I, Neugebauer R, Speelman L, Onyango G, et al. Assessing local instrument reliability and validity: A field-based example from northern Uganda. Social Psychiatry and Psychiatric Epidemiology. 2009; 44(8):685–692. [PubMed: 19165403]

Bolton P. Cross-cultural validity and reliability testing of a standard psychiatric assessment instrument without a gold standard. Journal of nervous and mental disease. 2001; 189(4):238–242. [PubMed: 11339319]

Bolton P, Michalopoulos L, Ahmed MA, Murray LK, Bass J. The mental health and psychosocial problems of survivors of torture and genocide in Kurdistan, Northern Iraq: a brief qualitative study. Torture. 2013; 23(1):1–14. [PubMed: 23831815]

Bolton P, Wilk CM, Ndogoni L. Assessment of depression prevalence in rural Uganda using symptom and function criteria. Social Psychiatry and Psychiatric Epidemiology. 2004; 39(6):442–447. [PubMed: 15205728]

Federal Democratic Republic of Ethiopa Ministry of Health. National Mental Health Strategy. 2012/13 – 2015/6

Foa EB, Johnson KM, Feeny NC, Treadwell KRH. The Child PTSD Symptom Scale: A preliminary examination of its psychometric properties. Journal of Clinical Child Psychology. 2001; 30(3): 376–384. [PubMed: 11501254]

Hobfoll SE, Canetti D, Hall BJ, Brom D, Palmieri PA, Johnson RJ, et al. Are community studies of psychological trauma's impact accurate? A study among Jews and Palestinians. [Research Support, N.I.H., Extramural]. Psychological Assessment. 2011; 23(3):599–605. [PubMed: 21381832]

Hollifield M, Warner TD, Lian N, Krakow B, Jenkins JH, Kesler J, et al. Measuring trauma and health status in refugees: A critical review. JAMA: Journal of the American Medical Association. 2002; 288(5):611–621.

The International Rescue Committee & UNHCR. Assessment Report on the Situation of Children in Kebribeyeh and Sheder Refugee Camps. 2009

Kleinman, A. Rethinking psychiatry: from cultural category to personal experience. New York City: Free Press; 1988.

Kohrt BA, Hruschka DJ. Nepali concepts of psychological trauma: The role of idioms of distress, ethnopsychology and ethnophysiology in alleviating suffering and preventing stigma. Culture, Medicine and Psychiatry. 2010; 34(2):322–352.

Lustig SL, Kia-Keating M, Knight WG, Geltman P, Ellis H, Kinzie JD, et al. Review of child and adolescent refugee mental health. Journal of the American Academy of Child & Adolescent Psychiatry. 2004; 43(1):24–36. [PubMed: 14691358]

Meyer S, Murray LK, Puffer ES, Larsen J, Bolton P. The nature and impact of chronic stressors on refugee children in Ban Mai Nai Soi camp, Thailand. Global Public Health. 2013; 8(9):1027–1047. [PubMed: 23886374]

Mollica RF, Cardozo BL, Osofsky HJ, Raphael B, Ager A, Salama P. Mental health in complex emergencies. The Lancet. 2004; 364(9450):2058–2067.

Murray LK, Bass J, Chomba E, Imasiku M, Thea D, Semrau K, et al. Validation of the UCLA Child Post Traumatic Stress Disorder-Reaction Index in Zambia. International Journal of Mental Health Systems. 2011; 5

Phan T, Silove D. An overview of indigenous descriptions of mental phenomena and the range of traditional healing practices amongst the Vietnamese. Transcultural Psychiatry. 1999; 36(1):79–94.

Puffer ES, Larsen J, Murray LK, Meyer S, Bolton P. Qualitative Findings: Community perspectives on children’s mental health in Somali refugee camps in Ethiopia. Report prepared for the International Rescue Committee and Bill and Melinda Gates Foundation. 2011

Rasmussen A, Katoni B, Keller AS, Wilkinson J. Posttraumatic idioms of distress among Darfur refugees: Hozun and Majnun. Transcultural Psychiatry. 2011; 48(4):392–415. [PubMed: 21911508]

Reed RV, Fazel M, Jones L, Panter-Brick C, Stein A. Mental health of displaced and refugee children resettled in low-income and middle-income countries: Risk and protective factors. The Lancet. 2012; 379(9812):250–265.

Sagi-Schwartz A, Seginer R, Abdeen Z. Chronic exposure to catastrophic war experiences and political violence: Links to the well-being of children and their families: Introduction to the special issue. International Journal of Behavioral Development. 2008; 32(4):257–259.

Silove D, Bateman CR, Brooks RT, Fonseca CAZ, Steel Z, Rodger J, et al. Estimating clinically relevant mental disorders in a rural and an urban setting in postconflict Timor Leste. Archives of General Psychiatry. 2008; 65(10):1205–1212. [PubMed: 18838637]

StataCorp. STATA Version 12.1. College Station, TX: 2012.

UNHCR. Jijiga population statistics (as of 25 September, 2012). 2012

United Nations. Report of the United Nations High Commissioner for Refugees. New York: United Nations; 2012.

Women's Refugee Commission. In Search of Safety and Solutions: Somali Refugee Adolescent Girls at Sheder and Aw Barre Camps, Ethiopia. 2012

Yan H. State of Syria: Exodus reaches 1 million; seesaw battles rage on. 2013 http://www.cnn.com/2013/03/06/world/meast/syria-civil-war/index.html.

**Table 1**

Sample characteristics

| | Child | Caregiver |
|---|---|---|
| Age M (SD) | 11.02 (2.90) | 39.16 (9.96) |
| Sex (% female) | 58.0 | 88.0 |
| Education | | |
| None | 5.0 | 65.0 |
| Some or graduated primary | 76.0 | 20.0 |
| Middle school | 13.0 | 6.0 |
| Some high school | 7.0 | 4.0 |
| High school graduate or higher | 0 | 6.0 |

## Table 2

Comorbidity among problem categories

| | Yes | No | % within category |
|---|---|---|---|
| Trauma only | 9 | 83 | 9.78[1] |
| Internalizing only | 6 | 79 | 7.05[1] |
| Externalizing only | 10 | 52 | 16.12[1] |
| Internalizing and trauma | 39 | 77 | 33.62[2] |
| Externalizing and trauma | 12 | 98 | 10.34[2] |
| Internalizing and externalizing | 17 | 99 | 14.65[2] |
| Internalizing, externalizing, and trauma | 23 | 93 | 19.82[2] |

*Note*.

[1]Percentages refer to proportion within categories (i.e., percent of people with only trauma among those being referred for trauma problems).

[2]Percentages refer to proportion of children with comorbid problems of the total sample of 116 children with problems.

**Table 3**

Qualitative items and scale inclusion decisions

| Qualitative item | Scale on which the item was included |
|---|---|
| I shout and yell | Externalizing |
| I despise the people | Externalizing |
| I don’t like myself | Internalizing |
| I am impatient | Externalizing |
| I faint | Internalizing |
| I think over something too much | Traumatic stress and internalizing |
| I am demoralized | Internalizing |
| I feel hopeless | Internalizing |
| I feel futureless | Internalizing |
| I hate this life / refuse this life | Internalizing |
| I feel pressure and overloaded | Internalizing |
| I have dropped out of school | Externalizing |
| I am surprised and shocked | Traumatic stress |
| I hold grudged | Internalizing |
| My mind has declined | Internalizing |
| Bad behavior (since trauma) | Traumatic stress |
| I am less active than I used to be | Internalizing |
| I do not take care of physical hygiene | Item was deleted |
| I am afraid of adults | Item was deleted |

**Table 4**

Internal consistency reliability for child and caregiver report

| Scale | Child | Caregiver |
|---|---|---|
| Child Posttraumatic Stress Scale-I | .94 | .94 |
| Child Posttraumatic Stress Scale-I with qualitative items | .95 | .95 |
| Achenbach: Internalizing Scale | .95 | .92 |
| Achenbach: Internalizing Scale with qualitative items | .96 | .95 |
| Achenbach: Anxiety and Depression Subscale | .91 | .88 |
| Achenbach: Withdrawal and Depression Subscale | .76 | .68 |
| Achenbach: Somatic Complaints Subscale | .87 | .83 |
| Achenbach: Externalizing Scale | .92 | .93 |
| Achenbach: Externalizing Scale with qualitative items | .93 | .94 |
| Achenbach: Rule Breaking Subscale | .78 | .84 |
| Achenbach: Aggressive Behavior Subscale | .89 | .91 |
| Achenbach: Attention Problems Scale | .82 | .81 |
| Achenbach: Social Problems Scale | .85 | .76 |
| Achenbach: Thought Problems Scale | .87 | .75 |

*Note*. Cronbach's Alpha refers to the internal consistency reliability for the scales with items removed in the final analytic sample of 147 children and caregiver pairs.

CPSS-I = Child PTSD Symptom Scale-Interview.

## Table 5

Combined test-retest and inter-rater reliability for child and caregiver measures

| | Child | | Caregiver | |
|---|---|---|---|---|
| **Scale** | **Pearson test-retest correlation** | **Spearman test-retest correlation** | **Pearson test-retest correlation** | **Spearman test-retest correlation** |
| Child Posttraumatic Stress Scale-I | - | .47* | .69*** | - |
| Child Posttraumatic Stress Scale-I with qualitative items | - | .45* | .72*** | - |
| Achenbach: Internalizing Scale | .38 | - | .60** | - |
| Achenbach: Internalizing Scale with qualitative items | .48* | - | - | .49* |
| Achenbach: Anxiety and Depression Subscale | - | .28 | .61** | - |
| Achenbach: Withdrawal and Depression Subscale | .49* | - | .59** | - |
| Achenbach: Somatic Complaints Subscale | .25 | - | - | .38 |
| Achenbach: Externalizing Scale | - | .34 | - | .54** |
| Achenbach: Externalizing Scale with qualitative items | - | .35 | - | .55** |
| Achenbach: Rule Breaking Subscale | .25 | - | .58** | - |
| Achenbach: Aggressive Behavior Subscale | - | .35 | - | .59** |
| Achenbach: Attention Problems Scale | .20 | - | .58** | - |
| Achenbach: Social Problems Scale | - | .53** | .68*** | - |
| Achenbach: Thought Problems Scale | - | .72*** | - | .77*** |

*Note*.

* <.05;

** <.01;

*** <.001.

CPSS-I = Child PTSD Symptom Scale-Interview.

## Table 6

Criterion-related validity: Results of independent samples t-tests comparing children referred with problems to those referred with no problems.

| **Child scales** | **Internalizing Cases (n = 79)** | **Externalizing Cases (n = 57)** | **Trauma Cases (n = 74)** | **Non-cases (n = 38)** |
|---|---|---|---|---|
| Trauma exposure | 3.79 (2.98)** | 2.92 (2.54) | 4.24 (2.91)*** | 2.18 (2.38) |
| CPSS-I with qualitative items | 18.26 (15.22)** | 13.28 (13.81) | 20.49 (15.55)*** | 9.19 (11.71) |
| Achenbach: Internalizing with qualitative items | 21.32 (18.02)** | 16.94 (15.95) | 22.65 (18.19)** | 10.94 (13.82) |
| Achenbach: Anxiety/depression | 5.66 (5.67)** | 4.20 (4.83)* | 5.82 (5.72)** | 2.16 (3.76) |
| Achenbach: Withdrawal/depression | 4.41 (3.34)* | 3.64 (2.90) | 4.86 (3.13)** | 3.13 (2.13) |
| Achenbach: Somatic Complaints | 5.62 (4.60)** | 4.91 (4.47) | 5.88 (4.81)** | 3.33 (2.92) |
| Achenbach: Externalizing with qualitative items | 9.30 (10.27)* | 8.08 (8.60) | 9.96 (10.44)* | 5.23 (8.69) |
| Achenbach: Rule Breaking | 2.87 (3.85) | 2.14 (2.17) | 2.99 (3.76) | 1.70 (3.49) |
| Achenbach: Aggressive Behavior | 5.74 (6.06)* | 5.35 (5.68) | 6.22 (6.36)* | 3.14 (4.90) |
| Achenbach: Attention Problems | 4.93 (4.18)** | 4.12 (3.68)* | 5.21 (4.08)** | 2.47 (3.26) |
| Achenbach: Social Problems | 4.30 (4.16)** | 3.18 (3.56) | 4.52 (4.20)** | 2.26 (3.37) |
| Achenbach: Thought Problems | 4.34 (5.05)** | 3.09 (4.12)* | 5.09 (3.95)** | 1.56 (2.60) |
| **Caregiver scales** | **Internalizing** | **Externalizing** | **Trauma** | **Normal** |
| Trauma exposure | 4.72 (3.29)** | 3.73 (3.29) | 5.41 (3.41)*** | 2.89 (3.18) |
| CPSS-I with qualitative items | 22.14 (15.65)*** | 16.52 (14.50)* | 24.35 (15.47)*** | 10.29 (11.95) |
| Achenbach: Internalizing | 20.01 (11.60)*** | 17.15 (11.57)* | 21.18 (11.49)*** | 11.54 (9.10) |
| Achenbach: Anxiety/depression | 8.87 (5.80)*** | 7.56 (5.63)** | 9.28 (5.51)*** | 4.73 (4.20) |
| Achenbach: Withdrawal/depression | 4.69 (2.75)*** | 3.91 (2.75)* | 4.80 (2.85)*** | 2.75 (2.46) |
| Achenbach: Somatic Complaints | 6.40 (4.25)** | 5.64 (4.38) | 7.05 (4.55)** | 3.99 (4.13) |
| Achenbach: Externalizing with qualitative items | 14.12 (12.59)*** | 14.33 (12.93)*** | 15.32 (13.10)*** | 5.59 (6.67) |
| Achenbach: Rule breaking | 3.61 (4.22)* | 3.86 (3.88)*** | 4.22 (4.52)*** | 1.39 (2.26) |
| Achenbach: Aggressive behavior | 9.28 (7.89)*** | 9.11 (8.22)*** | 9.77 (7.94)*** | 3.79 (3.98) |
| Achenbach: Attention Problems | 5.77 (4.13)*** | 4.93 (3.64)** | 6.25 (5.34)*** | 2.92 (3.05) |
| Achenbach: Social problems | 5.53 (3.88)** | 4.28 (3.46) | 6.11 (3.68)*** | 3.54 (3.15) |
| Achenbach: Thought problems | 3.15 (2.98)*** | 2.43 (2.78)* | 3.46 (3.18)*** | 1.30 (1.44) |

*Note*.

* <.05;

** <.01;

*** <.001.

CPSS-I = Child PTSD Symptom Scale-Interview.

**Table 7**

Receiver operating characteristic (ROC) results for area under the curve

| | Internalizing Cases | | | | Externalizing Cases | | | | Trauma Cases | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **AUC (se), [CI]** | **Cut Score** | **SENS** | **SPEC** | **AUC (se), [CI]** | **Cut Score** | **SENS** | **SPEC** | **AUC (se), [CI]** | **Cut Score** | **SENS** | **SPEC** |
| Child scales | | | | | | | | | | | | |
| Achenbach: Internalizing | .70 (.05), [.60–.80] | 11 | 68.35 | 63.16 | .63 (.06), [.52–.75] | 10 | 59.65 | 60.53 | .72 (.05), [.62–.82] | 11 | 67.57 | 63.16 |
| CPSS-I | .69 (.05), [.59–.79] | 8 | 64.56 | 65.79 | .60 (.06), [.48–.71] | 6 | 54.39 | 57.89 | .73 (.05), [.63–.83] | 13 | 64.86 | 81.58 |
| Caregiver Scales | | | | | | | | | | | | |
| Achenbach: Internalizing | .73 (.04), [.63–.82] | 14 | 65.82 | 65.79 | .66 (.06), [.54–.77] | 14 | 57.89 | 65.79 | .76 (.05), [.66–.85] | 15 | 71.62 | 68.42 |
| Achenbach: Externalizing | .72 (.05), [.63–.81] | 6 | 65.82 | 68.42 | .70 (.05), [.60–.81] | 10 | 59.65 | 60.53 | .75 (.05), [.66–.85] | 6 | 68.92 | 68.42 |
| CPSS-I | .73 (.05), [.63–.82] | 12 | 69.62 | 68.42 | .64 (.06), [.52–.75] | 9 | 59.65 | 60.53 | .74 (.04), [.68–.87] | 14 | 71.62 | 71.05 |

*Note.* All scales except for the caregiver internalizing scale include qualitative items.

CPSS-I = Child PTSD Symptom Scale-Interview. AUC = Area under the curve. SENS = Sensitivity. SPEC = Specificity.