Published in final edited form as:
J Appl Meas. 2006 ; 7(1): 16–38.

Using Rasch Analysis to Test the Cross-Cultural Item Equivalence of the Harvard Trauma Questionnaire and the Hopkins Symptom Checklist Across Vietnamese and Cambodian Immigrant Mothers

Yoonsun Choi,
University of Chicago

Amy Mericle, and
University of Chicago

Tracy W. Harachi
University of Washington

Abstract

A major challenge in conducting assessments in ethnically and culturally diverse populations, especially using translated instruments, is the possibility that measures developed for a given construct in one particular group may not be assessing the same construct in other groups. Using a Rasch analysis, this study examined the item equivalence of two psychiatric measures, the Harvard Trauma Questionnaire (HTQ), measuring traumatic experience, and the Hopkins Symptom Checklist (HSCL), assessing depression symptoms across Vietnamese- and Cambodian American mothers, using data from the Cross-Cultural Families (CCF) Project. The majority of items were equivalent across the two groups, particularly on the HTQ. However, some items were endorsed differently by the two groups, and thus are not equivalent, suggesting Cambodian and Vietnamese immigrants may manifest certain aspects of trauma and depression differently. Implications of these similarities and differences for practice and the use of IRT in this arena are discussed.

Comparisons of psychiatric disorders across cultural groups implicitly assume that measures assessing disorders are equivalent across groups (Byrne and Campbell, 1999). A major challenge in examining the similarities and differences in rates and patterns of psychiatric disorders arises with the possibility that measures developed for a given construct in one particular group may not be assessing the same construct in other groups as a result of conceptual or metric differences (Good and Kleinman, 1985; Hughes, Seidman, and Williams, 1995; Tran, Ngo, and Conway, 2003). Validity can be affected as a result of translation difficulties, irrelevancy of item contents and/or inappropriate norm scores when using a measure in a culture or language different from its originally intended audience (Custers, Hoijtink, van der Net, and Helders, 2000). Without invariance, it is unclear whether the observed differences across groups are due to true differences in rates and patterns of psychological phenomenon or to different patterns of responses to the items of the measure (Cheung and Rensvold, 2002). Studies concur that cultures may differ in how they express specific symptoms of psychopathology; for example, depression is often expressed in psychosomatic complaints rather than mood changes in many Asian cultures

Requests for reprints should be sent to Yoonsun Choi, The School of Social Service Administration, University of Chicago, 969 East 60th Street, Chicago, IL 60637, USA, yoonsun@uchicago.edu.

(Mollica et al., 1992; Mollica, Wyshak, de Marneffe, Khuon, and Lavelle, 1987). Thus, an examination of cross-cultural equivalence is a prerequisite for accurate and meaningful comparisons across diverse groups (Hui and Triandis, 1985).

Southeast Asian immigrants tend to report high levels of depression and numerous traumatic experiences (Kroll et al., 1989). Many of these psychiatric symptoms are chronic and co-occur with other psychiatric disorders, are exacerbated by resettlement, and constitute a major impediment to their adjustment. Among Southeast Asian mental health clients, Kinzie and his colleagues (1990) report a mere 6% recovery from the PTSD symptoms after 10 to 15 years, evidencing the chronicity of the disorder. The impact of numerous traumatic experiences and high rates of chronic psychiatric disorders is multifold, resulting in consequences related to unemployment and poor health (Uba and Chung, 1991) as well as difficult family relations and diminished psychosocial well-being of children (Tran and Ferullo, 1997). Parental psychopathology is an important risk factor for child mental health (Lahey, Miller, Gordon, and Riley, 1999). Parents with such difficulties are more likely to struggle with their own adjustment, may be unable to provide guidance and adequate parenting to their children, and hence may put their children at risk (Ascher, 1985; Carpio, 1981; Tobin and Friedman, 1984; Tran and Ferullo, 1997). Thus, it is important to better understand psychiatric disorders among Southeast Asian immigrants.

This study examines the item equivalence of two psychiatric scales, the Harvard Trauma Questionnaire (HTQ; Mollica, Wyshak, and Lavelle, 1987), measuring traumatic experience, and the Hopkins Symptom Checklist (HSCL, Derogatis, Lipman, Rickels, Uhlenhuth, and Covi, 1974), which assesses depression symptoms, across a non-clinical sample of Vietnamese- and Cambodian immigrant mothers. Several options are available to examine the internal measurement structure across groups, including inter-item reliability coefficients, confirmatory factor analysis (CFA), or item response theory (IRT) models. We tested the applicability of these scales for item equivalence using the Rasch model (Rasch, 1960/1980), one family of IRT models. The Rasch model is superior to additive techniques typically employed to analyze rating scale data because measures produced by Rasch analysis are linear, sample- and test-free, and robust to missing data (Bode and Wright, 1999; Fox and Jones, 1998; Wright and Masters, 1982). Rasch analysis provides statistics that help to determine the fit of the data to the Rasch model (Wright, Linacre, Gustafen, and Martin-Lof, 1994), the reliability with which the items separate individual respondents (Smith, 2001; Wright, 1996, 1998), and whether the rating scale response categories are being used by respondents in the intended manner (Linacre, 2002). It can also be used to test item equivalence by assessing whether scale items function differently across subpopulations of interest controlling for level of the latent trait (e.g., psychopathology) (Smith, 1994). Further, Rasch models can compare scales across groups regardless of the distribution of the latent trait among the groups (Fischer and Molenaar, 1997; Wright and Tennant, 1996). Thus, Rasch analysis is appropriate to examine psychiatric scales with non-normal distributions.

Psychopathology Among Southeast Asian Immigrants

The prevalence of chronic psychiatric disorders is high in refugees, prisoners of war, and concentration camp survivors, especially among those who have experienced severe trauma (Kroll et al., 1989). Many Southeast Asian immigrants and refugees report serious multiple traumatic premigration experiences including war, torture, death, violent sexual abuse, and starvation, and the prevalence of serious psychiatric disorders is high among these groups (Carpio, 1981; Kinzie, Fredrickson, Ben, Fleck, and Karls, 1984; Mollica, Wyshak, and Lavelle, 1987; Ngo, Tran, Gibbons, and Oliver, 2001; Tobin and Friedman, 1984). Depression and posttraumatic stress disorder (PTSD) are the most common psychiatric problems for these groups (Kinzie et al., 1990; Kinzie et al., 1984; Kroll et al., 1989;

Mollica, Wyshak, and Lavelle, 1987; Ngo et al., 2001; Tran and Ferullo, 1997; Uba and Chung, 1991) and comorbidity of these disorders is high (Kinzie et al., 1984; Mollica, Wyshak, and Lavelle, 1987; Ngo et al., 2001). Cambodian refugees are at a heightened risk because they experienced more trauma and torture, and report higher levels of psychiatric disorders and psychosocial distress than any other Southeast Asian group (Kinzie et al., 1990; Mollica, Wyshak, and Lavelle, 1987; Ngo et al., 2001; Uba and Chung, 1991).

Studies of psychiatric disorders among Southeast Asians frequently use existing measures that were originally developed for populations other than Southeast Asians. Using such measures without investigating item equivalence may result in misdiagnosis and inaccurate estimates of the prevalence of disorders (Kroll et al., 1989). For instance, a number of studies have found that the expression of depression and the manner in which individuals manifest symptoms differ between Southeast Asians and their western counterparts (Kinzie et al., 1982; Kroll et al., 1989; Mollica et al., 1992). It is also suggested that there are culture-bound syndromes for PTSD (Mollica et al., 1992).

Several efforts have been made to remedy these problems. For example, Fawzi and his colleagues tested whether the PTSD symptoms defined by the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV; American Psychiatric Association, 1994) generalize to Vietnamese refugees (Fawzi et al., 1997). Kinzie and his colleagues (1982) developed a depression scale that contains culturally consistent items describing various depressive symptoms for Vietnamese psychiatric patients. Mollica and colleagues (1992) developed a PTSD scale, the Harvard Trauma Questionnaire (HTQ), based on their clinical experiences serving various Southeast Asian psychiatric patients and tested the validity and reliability of the scale with their patients. They also translated and tested the Southeast Asian language versions of the Hopkins Symptoms Checklist (HSCL; Derogatis et al., 1974) which was not originally developed for Southeast Asians, and tested it for use with these groups.

Although the HTQ and HSCL scales have been used in several studies of Southeast Asian immigrants and refuges, these measures have rarely been tested for cross-cultural equivalence across different Southeast Asian ethnic groups. Southeast Asian immigrants are culturally diverse. Each group has its own history, language, family structure, parenting values and styles, and level of urbanization and westernization in the country of origin (Kim and Chun, 1993; Tran et al., 2003). Given these differences, it is incorrect to assume that measures to assess psychosocial functioning are equally valid and reliable across various Southeast Asian ethnic groups. Although a number of similarities are also noted among these groups, one cannot assume that cultural differences may not exist across Southeast Asian ethnic groups which affect the expression of psychiatric disorders and/or correlates of the disorders for each group (Boehnlein et al., 1995; Tran et al., 2003). Thus, an examination of cross-cultural equivalence of these measures is a critical step.

Cross-Cultural Equivalence of Measurement

Hui and Triandis (1985) organized the concept of cross-cultural equivalence into several categories, including conceptual, functional, construct operationalization, item, and scalar. Conceptual equivalence suggests that a construct has a similar meaning in different cultures and functional equivalence indicates similar precursors, consequents, correlates, and goals. These two equivalences are the first requirements for cross-cultural comparisons, and are related to underlying theories of a measure (Hui and Triandis, 1985). Equivalence in construct operationalization indicates that a construct is operationalized in the same manner in different cultures, and is akin to conceptual/func tional equivalence. Item and scalar equivalence concern psychometric properties. Item equivalence, a more concrete and micro level of equivalence, presupposes conceptual, functional and operationalization equivalence, and indicates that a construct can be measured with the same instrument across groups (Hui

and Triandis, 1985). In other words, each item carries the same meaning across cultures. Item equivalence enables meaningful numerical comparisons between cultures or other groups (Hui and Triandis, 1985; Reise, Widaman, and Pugh, 1993). Lastly, scalar equivalence suggests a particular score on a measure rep resents the same degree, intensity, or magnitude of the construct across groups. This state of equivalence is the most difficult to achieve (Hui and Triandis, 1985), but is required particularly for diagnostic tools.

There are several options available to examine the categories of cross-cultural equivalence. Conceptual and construct operationalization equivalence can be established in the development of measures by incorporating procedures like ethnographic interviews with potential respondents, review of measures by bilingual expert committees and adopting a rigorous dual translation procedure (i.e., translation and back-translation) (Hui and Triandis, 1985; Tran et al., 2003). Functional equivalence can be tested by examining the similarities and differences of the relationships between a measure and its correlates across groups. In order to examine item equivalence, some studies use inter-item reliability analyses. However, studies have shown that inter-item reliability does not guarantee equivalence of measures (Byrne and Campbell, 1999; Choi and Harachi, 2002). Choi and Harachi (2002) demonstrated that an instrument with a similar level of inter-item reliability in two different cultural groups may show significant differences in magnitude and direction of the factor loadings from confirmatory factor analyses (CFA). CFA and item response theory (IRT) are two alternative and more sophisticated methods to examine item and scalar equivalence. IRT, in particular, is widely used to test measures in large-scale achievement testing programs, but is underutilized in testing measurement invariance. CFA examines the similarities and differences in factor structures among groups. This maximum likelihood (ML) factor analysis uses a variance-co-variance matrix to investigate differences in parameters of the measurement model such as factor loadings, factor variance and covariance, and error variance and covariance. CFA provides goodness-of-fit indices that are readily comparable across groups including the Likelihood Ratio Test (differences in chi-square between models), comparative fit index (CFI) (Bentler, 1990), and root mean square error approximation (RMSEA) (Brown, Lamborn, and Steinberg, 1993). ML estimation requires data to be multi-variate normal which may be problematic with some measures, for example, measures of psychiatric disorders which tend to have non-normal distributions.

Because Rasch measures can be estimated using joint maximum likelihood (JML) techniques (Fischer and Molenaar, 1997; Wright and Masters, 1982), one can compare subgroups of respon dents regardless of the raw score distributions. As such, Rasch analysis is more appropriate to examine cross-cultural equivalency for psychiatric measures likely to be non-normal. IRT models also posit more stringent sets of measurement invariance constraints because they account for the item difficulties, which are ignored in CFA (Reise et al., 1993). Further, examining the properties of a measure for different subgroups of the sample can indicate differential fit to the Rasch model, and examining item equivalence or differential item functioning (DIF)[1] can determine whether the item is endorsed similarly across different groups (Gerber et al., 2002). Taken together, various methods help investigate possible differences in how constructs are conceptualized in different cultures and whether items are equivalent.

More studies are needed to better understand the extent to which traumatic experiences affect psychosocial adaptation of adults, family relations, and their children's well-being among Southeast Asian immigrants. However, these estimates must be derived from culturally appropriate measures. This study seeks to contribute to this very important area of

[1]The DIF function of the Rasch model is similar to examining the invariance of factor loadings across groups in multiple-group CFA analyses.

research by investigating two psychiatric measures, the HTQ and the HSCL, to determine whether these measures have cross-cultural item equivalence across two Southeast Asian ethnic groups, specifically, Vietnamese and Cambodians. This study is the first attempt to use a sophisticated method, the Rasch model, to examine measurement invariance of these measures across Vietnamese and Cambodian immigrant mothers.

Methods

Overview of Project and Sample Selection

The Cross Cultural Families (CCF) Project is a longitudinal study to follow a panel of Cambodian (*n*=164) and Vietnamese (*n*=163) adolescents (P.I. Tracy W. Harachi, MH59777). The primary aims are to investigate the occurrence of problem behaviors and the developmental trajectories of immigrant children, and the relationships between risk and protective factors and different outcome behaviors. Potential respondents were identified through locator information obtained from an urban Pacific Northwest school district. A random sample of Vietnamese and Cambodian families with a child enrolled between third and sixth grades in the school district were contacted. With an overall consent rate of approximately 85%, data collection for the longitudinal study includes annual maternal interviews, adolescent interviews, teachers' report of student behaviors, and school and police records. This paper used only maternal data that were collected in 2001.

Sample Description

Three hundred eighteen mothers comprise the analysis sample. The average age of the Cambodian mothers was 42.8 years and 43.3 years for Vietnamese mothers. On average, the Cambodian mothers arrived in the U.S. 16 years ago and 11 years ago for the Vietnamese mothers. Seventy-two percent of mothers in both groups reported receiving public assistance, food stamps, or qualifying for the free/reduced lunch program. In terms of education, 81.7% of the Cambodian mothers had less than a high school education in contrast to 66.3% of the Vietnamese; 14.5% of the Cambodians and 17.2% of the Vietnamese had graduated from high school, and 3.8% of the Cambodians and 16.5% of the Vietnamese had some college or higher education either in their country of origin or the in U.S. Twenty-four percent of the Vietnamese live in single parent households in contrast to 48% of the Cambodians. Hence there were differences between the Vietnamese and Cambodian mothers with respect to age, highest level of education (completed in the native country and in the U.S.), year of arrival to the U.S, and the proportion of single parent households.

Measures

The Indochinese Psychiatric Clinic (IPC) developed the Harvard Trauma Questionnaire (HTQ) based on their 10 years of clinical experiences with Southeast Asian psychiatric patients and previous research (Mollica et al., 1992). The HTQ is a self-report scale consisting of three sections. The first section contains 17 items describing a range of traumatic experiences. The second section has open-ended questions asking respondents for qualitative descriptions of the most traumatic events during their refugee experiences. The third section includes 30 symptom items; 16 of them were from the DSM-III-R criteria for PTSD and an additional 14 items were generated by the IPC to capture symptoms related with the traumatic events specific to Indochinese populations. Linguistic equivalence for each item was established in Khmer, Lao, and Vietnamese, following guidelines for cross-cultural instrument development (Mollica et al., 1992). Psychometric properties of the scale, including validity and reliability, were acceptable based on responses from a combined group of the IPC Southeast Asian patients based on alpha reliability, sensitivity and specificity tests (Mollica et al., 1992). The CCF project only utilized the third section (items

are listed in Appendix A). Response options ranged from (1) "Not at all" to (4) "Extremely." An average raw score of 2.5 and above on the HTQ (30 symptom items) indicates PTSD (Mollica et al., 1992).

The Hopkins Symptoms Checklist (HSCL) was originally developed by Parloff, Kelman, and Frank in 1954 and was further refined by Derogatis et al. (1974) into a self-report inventory of five dimensions that include somatization, obsessive-compulsive, interpersonal sensitivity, anxiety, and depression. The scale has been widely used for both clinical and nonclinical samples, and has shown good reliability, factor invariance, and validity in numerous studies (Derogatis et al., 1974). Mollica and his coleagues developed Southeast Asian versions of the depression sub-scale of the HSCL by establishing linguistic equivalence that included dual translations in Vietnamese, Khmer, and Laotian. In addition, they examined the psychometric properties of the HSCL-25 among a group of Southeast Asian patients that included Vietnamese, Cambodians, and Laotians (Mollica, Wyshak, de Marneffe et al., 1987). Because there were insufficient samples for each group, their psychometric examination was based on the combination of the three groups. The CCF project adopted the Vietnamese and Khmer versions of the HSCL-25 that Mollica and his colleagues developed. Response options ranged from (1) "Not at all" to (4) "Extremely." An average raw score of 1.75 and above on the HSCL indicates clinical depression (Derogatis et al., 1974).

Both instruments have been translated into many different languages and used with diverse racial and ethnic groups. The Southeast Asian versions of the two instruments have also been used in a number of studies. However, studies to date have tested and used the scales often with a combined group of various Southeast Asians ethnic groups without testing the cross-cultural equivalence for the target groups (Mollica et al., 1992; Mollica, Wyshak, de Marneffe et al., 1987). This study seeks to determine whether the instruments have item equivalency across two South east Asian ethnic groups, Vietnamese and Cambodians.

Analysis Strategy

Data analyses were conducted using a computer program called WINSTEPS (Linacre, 2005b). Our strategy entailed three steps. First, we used Rasch rating scale analysis (Andrich, 1978; Wright and Masters, 1982) to examine the appropriateness of the HTQ and HSCL for use across the Vietnamese and Cambodian respondents. Rasch rating scale analysis allows assessment of the validity and reliability of the measures and examination of whether respondents are using the rating scale categories in the intended manner. These analyses were conducted with full samples to construct common and psychometri-cally sound metrics for comparing the groups to one another in subsequent analyses. We then conducted differential item functioning (DIF) analyses to assess item equivalence across the Vietnamese and Cambodian subgroups.[2] Finally, in order to determine whether observed differences between subgroups could be attributable to DIF and whether non-biased measures could be constructed from the HTQ and the HSCL items, we examined the properties of measures constructed only with the core set of items that did not exhibit DIF.

Step 1: Rating scale analysis of the HTQ and HSCL data

The Rasch model is a log odds model that uses the principles of conjoint additivity and inverse probability to produce objective measures of person ability and item difficulty (Wright and Stone, 1979). When applied to the analysis of rating scale data, the model

[2]Unlike separate calibration techniques, DIF analyses in WINSTEPS allow one to isolate the item difficulty parameter for each item and test for differences between subgroups because it anchors the person and the step measures from those calibrated on the full sample (Linacre, 2005a).

specifies that the log odds of a respondent choosing any given category on an item is an additive function of respondent ability, item difficulty, and step difficulty of the rating scale response categories (Andrich, 1988; Wright and Masters, 1982). The item difficulty is expressed in logits. It theoretically ranges from +/− infinity but typically ranges from −5 to 5 when the mean item difficulty is set at 0. In this study, we rescaled our measures so that the mean item difficulty was anchored at 50 and a shift in 10 units up or down the measure equaled a shift in one logit. The result of this transformation is a measure that ranges from roughly 0 to 100, depending on the upper level of the latent trait.

We used several statistics to assess validity of the HTQ and HSCL. First, we examined the infit mean square (INFIT MNSQ) and outfit mean square (OUTFIT MNSQ) statistics (Smith, 2001; Wright and Stone, 1979). These statistics compare and test the fit of the observed data to the values of expected by the Rasch model (Smith, 2001; Smith, 2000; Wright and Stone, 1979). Although a variety of ranges have been suggested to indicate adequate fit, we considered items to “fit” if their MNSQ falls within the range of 0.6 to 1.4 as suggested by Wright and his colleagues (1994). However, items with fit statistics within a range of 0.5 to 1.5 may still be considered productive for measurement (Linacre, 2005a).

To further ensure detection of misfit and to aid in decisions about what should be done with misfitting items, we also examine the standardized MNSQ fit statistics (ZSTD) and other statistics such as the point-measure (PTMEA) correlation and the item discrimination. The ZSTD fit statistics are standardized to an approximate unit-normal distribution with a mean of 0 and standard deviation of 1. The +2.0 value is often used as an indication of misfit and has been found to be a more sensitive indicator of misfit across simulations with varying sample sizes (Smith et al., 1998). The PTMEA correlation ranges from −1 to 1 and negative values indicate items that are improperly scored (Linacre, 2005a) or interpreted in the opposite manner than intended. The discrimination value in WINSTEPS is the estimated item discrimination. Values less than 1 indicate under-discrimination, which indicates weak differentiation from one level to the next (Linacre, 2005a).

To ensure that the response categories of the HTQ and the HSCL are being used by respondents in the intended manner, we followed the guidelines offered by Linacre (1997; 2002). We first examined category usage for infrequently and irregularly used response options. We then examined the average measure of item difficulty and respondent ability and the step calibration for each response category to ensure that these values advance monotonically with each advance in response options. We finally examined OUT FIT MNSQ and Coherence of each response option. The OUTFIT MNSQ of response options is the average of the OUTFIT MNSQs associated with the responses in each category (Linacre, 2005a). An OUTFIT MNSQ statistic that is smaller than 2.0 signals expected category usage. The Coherence statistics compare observed and expected category usage. We considered the categories to be coherent if at least half (50%) of the expected responses are actually observed in each response category.

We then generated person summary statistics for subgroups of the Vietnamese and Cambodian respondents using the PSELECT specification command in WINSTEPS. Differences between subgroups of the Vietnamese and Cambodian respondents were assessed by comparing average person measures, average ZSTD statistics, and separation reliability statistics. The Rasch reliability statistic is conceptually similar to other measures of reliability such as the KR20 and Cronbach's alpha (Linacre, 2005a). In our analyses, we considered a separation reliability to be ideal if it is greater than .90 and adequate if it is falls between .80 and .90.

Finally, we created maps of items and respondents to graphically display how the items are arranged from most easy to endorse at the bottom to most difficult to endorse at the top along the latent traits measured. For these scales, "hard items" are symptoms which respondents are less likely to endorse at higher levels. The subjects who do endorse higher levels of these symptoms have more of the latent trait (e.g., they are more likely to be experiencing symptoms of severe trauma or depression).[3] These maps are also useful ways to see the average level of the latent trait for each subgroup and to see how different subgroups of respondents are distributed along the latent trait.

Step 2: Differential Item Functioning (DIF) analyses

To determine whether observed differences between the Vietnamese and Cambodian subgroups could be explained by non-equivalence of the items, we conducted differential item functioning (DIF) analyses. In traditional item response theory (IRT) terms, DIF refers to the situation in which an item displays different properties for different groups after controlling for the abilities of these groups (Angoff, 1993). Differences in item difficulties from one subgroup to another can be tested for significance by dividing the difference by the joint standard errors of the items. The resulting *t* statistic can be interpreted to detect items biased against particular subgroups of respondents (Du, 1995). According to Smith (1994) who researched the question of sample size for the detection of DIF with the separate calibrations technique, a sample size of 500 gives reasonable power to detect bias of items in 0.2 logit range and that of 100 in 0.4 logit range. Using this as a guide and based on the number of Vietnamese and Cambodians in our sample, we considered items to display significant DIF if the difference in item difficulty between the two respondent groups was greater than 4 (0.4 logits) with a *t* statistic greater than +/− 1.96.

Step 3: Rating scale analysis of HTQ and HSCL core items

To achieve measurement equivalence, some researchers suggest discarding items with DIF to reconstruct measures (Lange, Thalbourne, Houran, and Lester, 2002). In the final stage of the analysis, we omitted items displaying statistically significant DIF to determine whether valid, reliable, and culturally equivalent measures could be constructed with the remaining items and to determine whether observed differences between subgroups were real or an artifact of item bias. To do so, we recalibrated the remaining "core" items, reassessed the items and response options for fit, and compared average person statistics between the Vietnamese and Cambodian subgroups on the new measures constructed with the core items.

Extreme responses

Subjects with extreme responses are dropped from WINSTEPS analyses because they do not provide information regarding the progress of the individuals along a continuum of the latent trait. The data set contained 318 observations. One Vietnamese mother did not respond to any of the HTQ or HSCL items (leaving 317 valid observations). On the HTQ, there were a total of 43 extreme responses; 31 Vietnamese mothers responded "Not at all" to every item, and 12 Cambodian mothers responded "Not at all" to every item. On the HSCL, there were a total of 60 extreme responses; 41 Vietnamese mothers responded "Not at all" to every item, while18 Cambodian mothers responded "Not at all" to every item and one Cambodian mother responded "Extremely" to every item. These respondents were dropped in the subsequent analyses. This resulted in the calibration of HTQ estimates from 129 Vietnamese and 145 Cambodian respondents (*N*=274). The calibration of the HSCL estimates was based

[3]There is no precisely ordered hierarchy of symptoms in HTQ and HSCL. In other words, mild symptoms are not required for severe symptoms to occur. However, there is a general hierarchy of symptoms in which some items represent more severe symptoms than others. For example, feeling sad is more common and milder than having thoughts of suicide.

on 119 Vietnamese and 138 Cambodian respondents (*N*=257). A higher number of Vietnamese samples were excluded from the estimates because they were more likely to endorse “not at all” to every item than Cambodian samples.

Results

Rating Scale Analysis of the HTQ and HSCL Data

We first examined the properties of the HTQ and HSCL items with the full sample of respondents. The results for the HTQ items are shown in Table 1. Items are arranged by degree of item difficulty, from most difficult to endorse at the top (63.76), to least difficult to endorse at the bottom (41.48). Items such as having trouble sleeping and having difficulty concentrating were easier to endorse, indicating lower symptom severity, while items like feeling others are hostile towards you and feeling split into two different people were more difficult to endorse, indicating higher symptom severity. Item fit statistics, indicated by the INFIT MNSQ and OUTFIT MNSQ, showed that all of the items in this scale had adequately fit the Rasch model. The INFIT MNSQ statistics for all the items were in the 0.6 to 1.4 range. The OUTFIT MNSQ statistics for all the items were also in the same range except for the item labeled as “Split” (0.57). However, this item had a positive and fairly strong PTMEA. Other items (unable to feel emotions, feeling irritable, avoiding activities, and trouble sleeping) had ZSTDs greater than 2.0, but these items also had acceptable discrimination and fairly strong PTMEAs. In addition, removing these items was detrimental to the properties of the measures. Thus, all items were kept for the subsequent analyses.

Table 2 displays the item fit statistics for the HSCL items. Like the HTQ items, the HSCL items are listed from most difficult to endorse at the top to least difficult to endorse at the bottom. Items indicating lower levels of depression included having trouble sleeping, worrying too much about things, and feeling low on energy. At the other end of the spectrum was the item of suicidal thoughts. All of the items in this scale had INFIT MNSQ and OUTFIT MNSQ statistics in the 0.6 to 1.4 range. As with the HTQ, some items (suicidal thoughts, loss of sexual interest, crying, and difficulty sleeping) had ZSTDs greater than 2.0, but these items also had acceptable discrimination and fairly strong PTMEA correlations

We also examined the properties of the response options for the HTQ and the HSCL scales (Table 3). For the most part, the response options were used in a manner that supported the construction of the measures. For example, the average measure of item difficulty and respondent ability increased with each increase in the response categories from “not at all” to “extremely” (−29.18 to 6.82 for HTQ and −31.77 to 13.69 for HSCL). Additionally, OUTFIT MNSQs for each category of both scales were below 2.0, indicating expected category usage (Linacre, 1997; Linacre, 2002). For both scales, however, a smaller than expected proportion of the respondents used the category “Quite a bit,” as indicated by the step calibration and the coherence statistic. The step calibration did not increase in both scales between the category “quite a bit” to the category “extremely,” meaning that there was no meaningful step between them. In addition, the coherence statistics for the category “quite a bit” were less than 50% in both scales, confirming the under-usage. Despite the irregular usage of this category, when we combined the “quite a bit” category with the “extremely” category, the fit of the data to the Rasch model did not substantially improve. Thus, we opted to maintain the measures constructed with the original four response options.

Table 4 shows person-summary statistics for the full sample as well as for Vietnamese and Cambodian subgroups. The average level of trauma, as indicated by the HTQ measure, was higher for the Cambodian subgroup (36.15 for Cambodians vs. 25.02 for Vietnamese). The average fit of the respondents, indicated by average INFIT MNSQ for the subgroups, was

adequate. The separation reliability of the HTQ measure was higher for the Cambodian than for the Vietnamese respondents (0.93 vs. 0.86). Similar patterns were found for respondents with respect to the HSCL: the Cambodian respondents reported higher level of depression (34.17 vs. 27.38), the average INFIT MNSQs were adequate for both groups, and the scale worked more reliably for the Cambodian than for Vietnamese re spondents (0.90 vs. 0.83).

Maps of items and respondents illustrate how the items were arranged along the latent traits of trauma and depression. Figures 1A through 1C depict how the combined sample, the Vietnamese respondents, and the Cambodian respondents were distributed along with the HTQ items. Figures 2A through 2C depict the distribution for the HSCL items. Similar patterns emerged from both scales. First, the mean level of trauma and depression among respondents was much lower than the average level of item difficulty (a full two logits for the HTQ). Cambodian mothers had higher means on both scales than Vietnamese mothers. Many of the respondents from both groups were clustered at the bottom indicating no or low symptomatology. Cambodian respondents, however, were distributed a bit more evenly along the measures, creating greater separation among respondents and higher reliability estimates.

DIF Analyses

The results of the differential item functioning (DIF) analyses for the HTQ items are presented in Table 5. A total of 10 of the 30 items displayed DIF. The level of difficulty and accompanying standard error is listed by item for both subgroups of respondents. The DIF contrast represents the difference between the item difficulty measures for each group. The statistical significance of this contrast is represented by the item's *t* statistics. Items are arranged by the magnitude of their *t* statistics from high to low. Given identical levels of trauma, items with a positive DIF contrast were more difficult to endorse for Cambodians and, as such, were less likely to be endorsed by them. A total of five items displayed significant DIF of this nature. The items that were significantly more difficult to endorse for Cambodians were: having trouble sleeping, feeling hostile, being unable to feel emotions, feeling on guard, and feeling irritable. Items with a negative DIF contrast are those that were significantly more difficult to endorse for the Vietnamese respondents. The items that were significantly more difficult for the Vietnamese respondents were: feeling helpless, feeling ashamed, feeling like you are going crazy, having a hard time daily performing tasks, and feeling split off from yourself. The magnitude of the DIF contrast for the HTQ items is depicted in Figure 3.

Table 6 lists the results for the HSCL for both groups of respondents including the level of item difficulty and accompanying standard error for each item. Nine of the 15 items on the HSCL showed significant DIF. The items that the Cambodian respondents were less likely to endorse were crying easily, experiencing poor appetite, having difficulty sleeping, and experiencing a loss in sexual interest or pleasure. The Vietnamese respondents were less likely to endorse items such as feeling lonely, hopeless, worrying too much, feeling trapped, and feeling sad. The magnitude of the DIF contrast for the HSCL items is depicted in Figure 4.

Rating Scale Analysis of Core Items

In this final step, we reanalyzed the 20 items from the HTQ and the 6 items from the HSCL that did not exhibit significant DIF to assess the psychometric properties of the measures constructed only from the non-biased items. Overall, removing the items with significant DIF reduced the number of non-extreme cases available to construct the HTQ (*N*=260) and the HSCL (*N*=230) measures. As Table 7 and Table 8 display, all items continued to contribute productively to the construction of the measures. The functioning of the rating

scale response categories for the measures did not appreciably change; the category “quite a bit” was still underused (Table 9). As Table 10 shows, removing the items with significant DIF did not adversely affect the separation reliability of the HTQ. However, removing the items with significant DIF adversely affected the separation reliability of the HSCL for both subgroups of respondents (reliability dropped to .52 for Vietnamese respondents and .77 for Cambodian respondents). The lower separation reliability of the HSCL is likely due to the fact that there was a smaller pool of the items with which to separate the respondents, rendering it a less than psychometrically sound measure of depression.

Discussion

This study examined the cross-cultural item equivalence of two psychiatric scales, the Harvard Trauma Questionnaire and the Hopkins Symptom Checklist, across two Southeast Asian ethnic groups, Vietnamese and Cambodian immigrant mothers. The findings show that the validity and reliability of both scales were adequate for both subgroups, as indicated by the adequate fit of all the items and the good separation reliability. However, as seen in the maps of items and respondents, most of the respondents were clustered at the lower end of the measures, indicating low or no symptoms, while the majority of the items were concentrated at the higher level. This is not unusual because the scales were developed as diagnostic tools. In other words, the items were designed to screen clinical populations who are more likely to endorse many of the items and/or higher level of symptoms and our non-clinical samples would be less likely to endorse items or higher level.[4]

The reliability and the validity of the two psychiatric scales were in the adequate range for both groups, although the reliability of both scales was higher for the Cambodian group, likely as a result of the larger variation among the Cambodian group. Our community samples of Cambodian mothers reported greater symptoms of PTSD and depression than Vietnamese mothers. This is consistent with findings of previous studies which used mostly clinic samples, that Cambodian immigrants and refugees experienced higher levels of trauma and thus are more likely to report higher levels of psychiatric disorders than other Southeast Asian groups.

The findings also showed that the response options of each scale were used in the intended manner, except that one of the options, “quite a bit,” was relatively underused. However, revising the response options to a 3-point scale did not substantially improve the fit, so we retained the scale with 4 options. In a study by Kinzie and his colleagues, their Vietnamese informants noted that the internal contrasts expressed in Likert scale type responses are not as self-evident for them as for white middle-class Americans, thus they adopted in their study a 3-degree continuum rather than the 5-point Likert scale common to many instruments (Kinzie et al., 1982). Although a 4-option scale worked acceptably in our study, our findings and Kinzie et al.'s study suggest that we may consider using less than 4-point Likert scale type responses for the examined constructs in future studies with Southeast Asians.

Examination of item equivalence showed that many items were equivalent across the two groups of Southeast Asian immigrants. However, there were also some items that were not equivalent. Approximately a third of the HTQ items displayed DIF and over half of the HSCL items displayed DIF. With respect to symptoms of trauma, it appeared that Vietnamese were less likely to endorse feeling hopeless and having difficulty in functioning (e.g., having a hard time daily performing tasks) and more likely to endorse symptoms

[4]Our sample may include some clinic clients, but the results suggest that when the sample is randomly selected from community, the sample includes non-clinical samples, thus, showing less symptoms than samples specifically derived from clinic.

related to the inability to feel emotions or feeling irritable, feeling hostile, and feeling on guard. Cambodians showed an opposite pattern. On the depression scale, the items that the Vietnamese respondents were less likely to endorse pertained to sadness and emotional distress (e.g., feeling lonely, worrying too Vietnamese respondents (e.g., poor appetite, difficulty sleeping, and loss in sexual interest). This finding is also consistent with what Kinzie and his colleagues (1982) found. In fact, they added somatic symptoms and behavioral changes due to depression to their depression scale because somatic complaints were common among Vietnamese clients (Kinzie et al., 1982). This was not true of Cambodian respondents. The differences of items found in this study may further confirm that cultures may differ in their manifestation of depressive symptoms (Kinzie et al., 1982; Kroll et al., 1989; Mollica et al., 1992). However, un like gender differences found in depression, in which the significant differences in factor structure implies variant underlying constructs of depression across genders (Butler and Nolen-Hoeksema, 1994; Nolen-Hoeksema, 1987; Williamson, 1987), Vietnamese and Cambodian immigrant mothers showed similarities in factor structure, suggesting invariant conceptual equivalence. Given the similarity of the factor structures across groups, one may proceed to examine the correlations of these constructs with others of interest, but caution should be given to analyses which examine means across groups.

In order to examine the extent to which traumatic experiences and resulting psychiatric disorders influence psychosocial adaptation of individuals and their families, sound cross-culturally equivalent measurements are required. Our study demonstrated a process to examine cross-cultural item equivalence of measures with the Rasch model that can be utilized in other studies. Few studies undertake this process, yet the results of this study illustrate the need to attend to these sorts of measurement issues. Based on our findings, caution is advised in relying solely on the cut-points of the scales given by the authors to determine the diagnosis of disorders among different subgroups of Southeast Asian immigrants. These results underscore the challenge in advancing our understanding of the psychiatric symptomatology among immigrant groups and the need to assess the appropriateness of measures across different populations.

Acknowledgments

This investigation was supported by a grant from the National Institute of Mental Health and the National Institute of Child and Human Development (MH59777-03) and a grant from the Louise R. Bowler Junior Faculty Research Fund at the University of Chicago.

Appendix

HTQ questions

I would like to ask you some questions about your past history and how you are feeling now

The following are things that people sometimes feel after experiencing hurtful or terrifying events in their lives. Please carefully decide how much these things bothered you in the past week. [variable name in tables/figures]

(Would you say…)

1. Recurring thoughts or memories of the most hurtful or terrifying events [thoughts]
2. Feeling as though the hurtful or terrifying event is happening again [ruminate]
3. Recurrent nightmares [nightmares]

4. Feeling detached or withdrawn from people [detached]
5. Unable to feel emotions [unable]
6. Feeling jumpy or easily startled [jumpy]
7. Difficulty concentrating [concentrating]
8. Trouble sleeping [sleeping]
9. Feeling on guard [guard]
10. Feeling irritable or having outburst of anger [irritable]
11. Avoiding activities that remind you of the traumatic or hurtful event [a_avoid]
12. Inability to remember parts of the most traumatic or hurtful events [remember]
13. Less interest in daily activities [interest]
14. Feeling as if you don't have a future [future]
15. Avoiding thoughts or feeling associated with the traumatic or hurtful events [t_avoid]
16. Sudden emotional or physical reaction when reminded of the most hurtful or traumatic events [reaction]
17. Feeling that people do not understand what happened to you [understand]
18. Difficulty performing work or daily tasks [perform]
19. Blaming yourself for things that have happened [blame]
20. Feeling guilty for having survived [guilt]
21. Feeling hopelessness [hopeless]
22. Feeling ashamed of the hurtful or traumatic events that have happened to you [ashamed]
23. Spending time thinking about why these things happened to you [happening]
24. Feeling as if you are going crazy [crazy]
25. Feeling that you are the only one who suffered these events [alone]
26. Feeling others are hostile toward you [hostile]
27. Feeling that you have no one to rely on [rely]
28. Finding out or being told by other people that you have done something that you cannot remember [told]
29. Feeling as if you are split into two people and one of you is watching what the other is doing [split]
30. Feeling someone you trusted betrayed you [betray]

HSCL questions

[CONTINUED FROM SECTION ABOVE] Please carefully decide how much these things bothered you in the past week. [variable name in tables/figures]

1. Feeling low in energy, slowed down [energy]
2. Blaming yourself for things [blame]

3. Crying easily [crying]
4. Loss of sexual interest or pleasure [sexual]
5. Poor appetite [appetite]
6. Difficulty falling asleep, staying asleep [sleep]
7. Feeling hopeless about the future [hopeless]
8. Feeling sad [sad]
9. Feeling lonely [lonely]
10. Thoughts of ending your life [suicide]
11. Feeling of being trapped or caught [trapped]
12. Worrying too much about things [worry]
13. Feeling no interest in things [interest]
14. Feeling everything is an effort [effort]
15. Feeling of worthlessness [worthless]

References

American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th. Washington, DC: American Psychiatric Association; 1994.

Andrich D. A rating formulation for ordered response categories. Psychometrika. 1978; 43:561–573.

Andrich, D. Rasch models for measurement. Thousand Oaks, CA: Sage; 1988.

Angoff, WH. Perspectives on differential item functioning methodology. In: Holland, PW.; Wainer, H., editors. Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum; 1993. p. 3-23.

Ascher C. The social and psychological adjustment of Southeast Asian refugees. The Urban Review. 1985; 17:147–152.

Bentler PJ. Fit indexes, Lagrange multipliers, constraint changes and incomplete data in structural models. Multivariate Behavioral Research. 1990; 25:163–172.

Bode, RK.; Wright, BD. Rasch measurement in higher education. In: Smart, JC.; Tierney, WG., editors. Higher education: Handbook of theory and research. Vol. XIV. New York: Agathon Press; 1999. p. 287-316.

Boehnlein JK, Tran HD, Riley C, Vu KC, Tan S, Leung PK. A comparative study of family functioning among Vietnamese and Cambodian refugees. The Journal of Nervous and Mental Disease. 1995; 183:768–773. [PubMed: 8522939]

Bond, TG.; Fox, CM. Applying the Rasch model: Fundamental measurement in the human sciences. Mahwah, NJ: Lawrence Erlbaum Associates; 2001.

Brown BB, Lamborn SD, Steinberg L. Parenting practices and peer group affiliation in adolescence. Child Development. 1993; 64:467–482. [PubMed: 8477629]

Butler LD, Nolen-Hoeksema S. Gender differences in responses to depressed mood in a college sample. Sex Roles. 1994; 30:331–346.

Byrne BM, Campbell TL. Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure: A look beneath the surfact. Journal of Cross Cultural Psychology. 1999; 30:555–574.

Byrne BM, Shavelson RJ, Muthen B. Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychological Bulletin. 1989; 105:456–466.

Carpio B. The adolescent immigrant. Canadian Nurse. 1981; 77:27–31. [PubMed: 6907034]

Cheung GW, Rensvold RB. Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling. 2002; 9:233–255.

Choi Y, Harachi TW. The cross-cultural equivalence of the Suinn-Lew Asian Self-Identity Acculturation Scale among Vietnamese and Cambodian Americans. The Journal of Social Work Research and Evaluation. 2002; 3:5–17.

Custers JWH, Hoijtink H, van der Net J, Helders PJM. Cultural differences in functional status assessment: Analyses of person fit according to the Rasch model. Quality of Life Research. 2000; 9:571–578. [PubMed: 11190011]

Derogatis LR, Lipman RS, Rickels K, Uhlenhuth EH, Covi L. The Hopkins Symptom Checklist (HSCL): A self-report symptom inventory. Behavioral Science. 1974; 19:1–15. [PubMed: 4808738]

Du Y. When to adjust for differential item functioning. Rasch Measurement Transactions. 1995; 9:414.

Fawzi MCS, Pham T, Lin L, Nguyen TV, Ngo D, et al. The validity of post-traumatic stress disorder among Vietnamese refugees. Journal of Traumatic Stress. 1997; 10:101–108. [PubMed: 9018680]

Fischer, GH.; Molenaar, IW. Rasch Models: Foundations, recent developments, and applications. New York: Springer Verlag; 1997.

Fox CM, Jones JA. Uses of Rasch Modeling in Counseling Psychology Research. Journal of Counseling Psychology. 1998; 45:30–45.

Gerber B, Smith EV, Girotti M, Pelaez L, Lawless K, et al. Using Rasch measurement to investigate the cross-form equivalence and clinical utility of Spanish and English versions of a Diabetes Questionnaire: A pilot study. Journal of Applied Measurement. 2002; 3:243–271. [PubMed: 12147912]

Good, BJ.; Kleinman, AM. Culture and anxiety: cross-cultural evidence for the patterning of anxiety disorders. In: Tuma, AH.; Maser, JD., editors. Anxiety and the anxiety disorders. Hillsdale, NJ: Lawrence Erlbaum Associates; 1985. p. 297-323.

Hughes D, Seidman E, Williams N. Cultural phenomena and the research enterprise: toward a culturally anchored methodology. American Journal of Community Psychology. 1995; 21:687–703. [PubMed: 8085565]

Hui CH, Triandis HC. Measurement in cross-cultural psychology: a review and comparison of strategies. Journal of Cross Cultural Psychology. 1985; 16:131–152.

Kim LS, Chun CA. Ethnic differences in psychiatric diagnosis among Asian American adolescents. The Journal of Nervous and Mental Disease. 1993; 181:612–617. [PubMed: 8409959]

Kinzie JD, Boehnlein JK, Leung PK, Moore LJ, Riley C, et al. The prevalence of posttraumatic stress disorder and its clinical significance among Southeast Asian refugees. American Journal of Psychiatry. 1990; 147:913–917. [PubMed: 2356877]

Kinzie JD, Fredrickson RH, Ben R, Fleck J, Karls W. Posttraumatic stress disorder among survivors of Cambodian concentration camps. American Journal of Psychiatry. 1984; 141:645–650. [PubMed: 6711684]

Kinzie JD, Manson SM, Vinh DT, Tolan NT, Anh B, Pho TN. Development and validation of a Vietnamese-Language Depression Rating Scale. American Journal of Psychiatry. 1982; 139:1276–1281. [PubMed: 7124979]

Kroll J, Habenicht M, Mackenzie T, Yang M, Chan S, et al. Depression and posttraumatic stress disorder in Southeast Asian Refugees. American Journal of Psychiatry. 1989; 146:1592–1597. [PubMed: 2589553]

Lahey, BB.; Miller, TL.; Gordon, RA.; Riley, AW. Developmental epidemiology of the disruptive behavior disorders. In: Quay, HC.; Hogan, AE., editors. Handbook of disruptive behavior disorders. New York: Kluwer Academic/Plenum Publishers; 1999. p. 23-48.

Lange R, Thalbourne M, Houran J, Lester D. Depressive response sets due to gender and culture-based differential item functioning. Personality and Individual Differences. 2002; 33:937–954.

Linacre, JM. Guidelines for rating scales. MESA Research Note #2; Paper presented at the Midwest Objective Measurement Seminar; Chicago, IL: 1997 June.

Linacre JM. Optimizing rating scale category effectiveness. Journal of Applied Measurement. 2002; 3:85–106. [PubMed: 11997586]

Linacre, JM. A user's guide to WINSTEPS. Chicago: Winsteps.com; 2005a.

Linacre, JM. WINSTEPS Rasch measurement [Computer program]. Chicago: Winsteps.com; 2005b.

Mollica R, Caspi-Yavin Y, Bollini P, Truong T, Tor S, et al. Harvard Trauma Questionnaire: Validating a cross-cultural instrument for measuring torture, trauma, and post-traumatic stress disorder in Indochinese refugees. The Journal of Nervous and Mental Disease. 1992; 180:111–116. [PubMed: 1737972]

Mollica RF, Wyshak G, de Marneffe D, Khuon F, Lavelle J. Indochinese versions of the Hopkins Symptom Checklist-25: A screening instrument for the psychiatric care of refugees. American Journal of Psychiatry. 1987; 144:497–500. [PubMed: 3565621]

Mollica RF, Wyshak G, Lavelle J. The psychosocial impact of war trauma and torture on Southeast Asian refugees. American Journal of Psychiatry. 1987; 144:1567–1572. [PubMed: 3688280]

Ngo D, Tran TV, Gibbons JL, Oliver J. Acculturation, premigration traumatic experiences, and depression among Vietnamese Americans. Journal of Human Behavior in the Social Environment. 2001; 3:225–242.

Nolen-Hoeksema S. Sex differences in unipolar depression: Evidence and theory. Psychological Bulletin. 1987; 101:259–282. [PubMed: 3562707]

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danish Institute for Educational Research; 1960. Expanded edition, 1980. Chicago: University of Chicago Press

Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychological Bulletin. 1993; 114:552–566. [PubMed: 8272470]

Santor D, Ramsay JO. Progress in the technology of measurement: Applications of item response models. Psychological Assessment. 1998; 10:345–359.

Smith EV Jr. Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. Journal of Applied Measurement. 2001; 2:281–311. [PubMed: 12011511]

Smith RM. Detecting item bias in the Rasch rating scale model. Educational and Psychological Measurement. 1994; 54:886–896.

Smith RM. Fit analysis in latent trait measurement models. Journal of Applied Measurement. 2000; 1:199–218. [PubMed: 12029178]

Smith RM, Schumacker RE, Bush MJ. Using item mean squares to evaluate fit to the Rasch model. Journal of Outcome Measurement. 1998; 2:66–78. [PubMed: 9661732]

Tobin JJ, Friedman J. Intercultural and developmental stresses confronting Southeast Asian refugee adolescents. Journal of Operational Psychiatry. 1984; 15:39–45.

Tran T, Ferullo DL. Indochinese mental health in North American: measures, status and treatments. Journal of Sociology and Social Welfare. 1997; 24:3–20.

Tran TV, Ngo D, Conway K. A cross-cultural measure of depressive symptoms among Vietnamese Americans. Social Work Research. 2003; 27:56–64.

Uba L, Chung RCY. The relationship between trauma and financial and physical well being among Cambodians in the United States. The Journal of General Psychology. 1991; 118:215–225. [PubMed: 1757781]

Williamson MT. Sex differences in depression symptoms among adult family medicine patients. The Journal of Family Practice. 1987; 25:591–594. [PubMed: 3681221]

Wright BD. Reliability and separation. Rasch Measurement Transactions. 1996; 9:472.

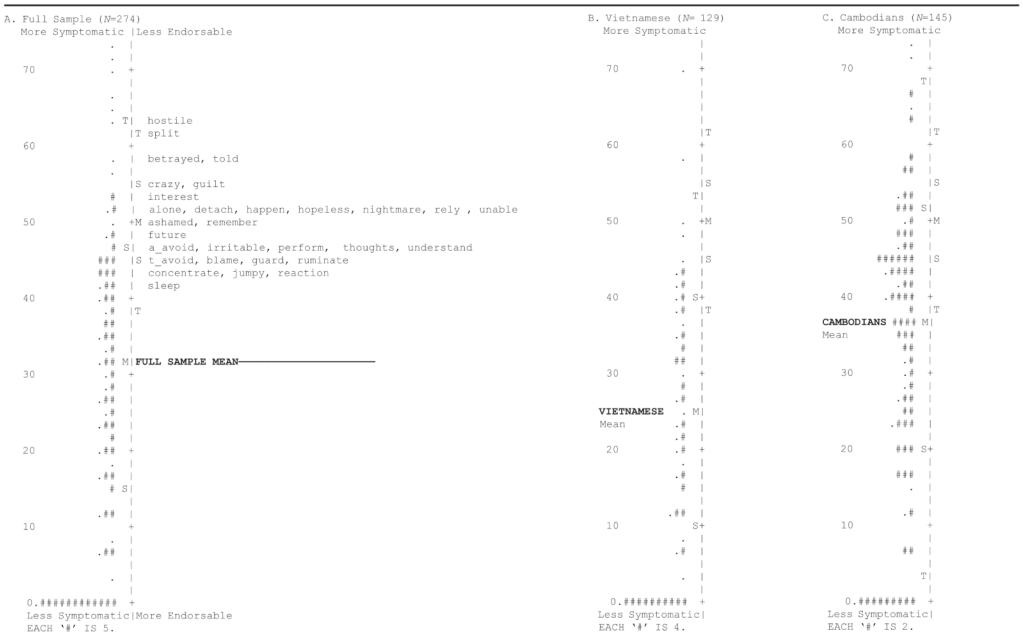
Wright BD. Interpreting reliabilities. Rasch Measurement Transactions. 1998; 11:602.

Wright BD, Linacre JM, Gustafen JE, Martin-Lof P. Reasonable mean-square fit values. Rasch Measurement Transactions. 1994; 8:370.

Wright, BD.; Masters, GN. Rating scale analysis. Chicago: MESA Press; 1982.

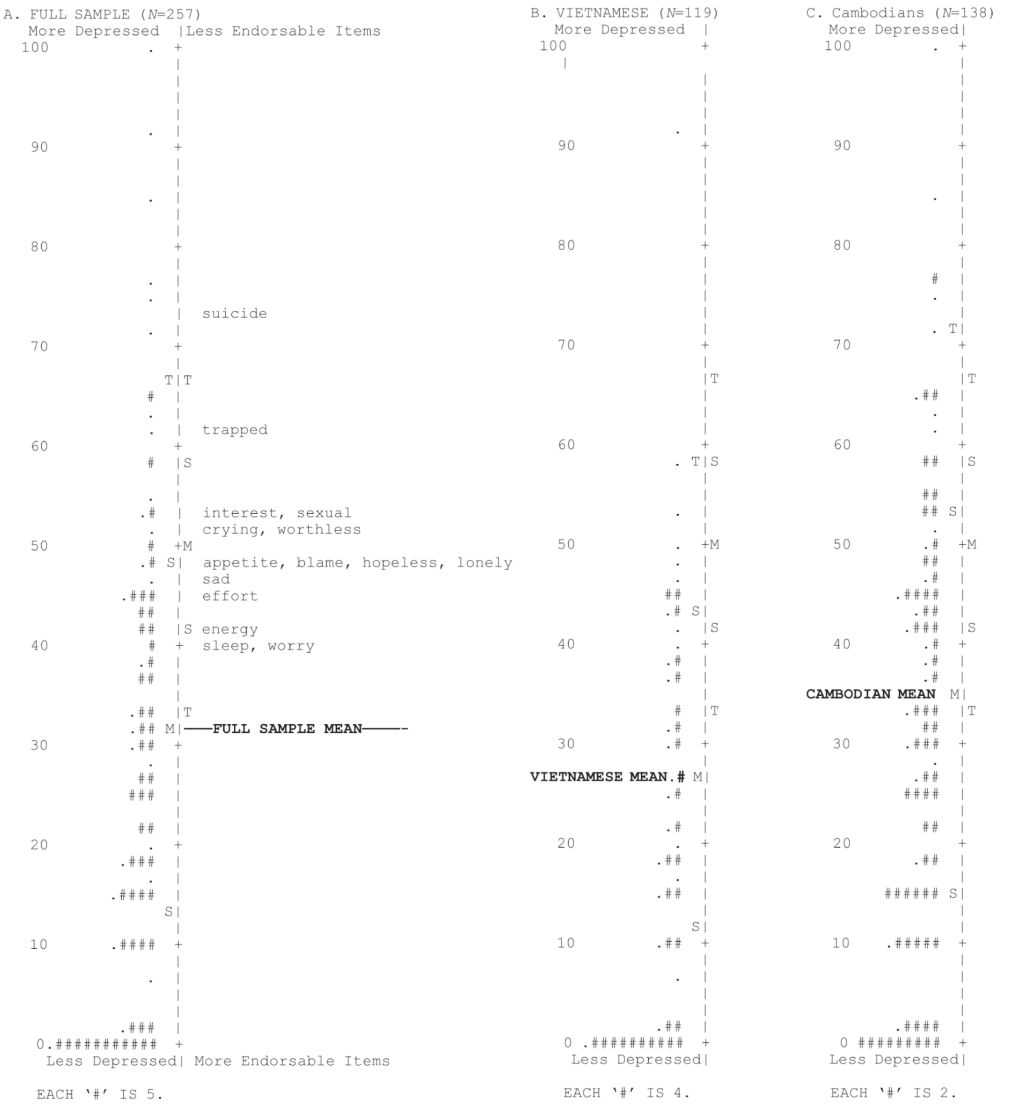
Wright, BD.; Stone, MH. Best test design. Chicago: MESA Press; 1979.

Wright BD, Tennant A. Sample size again. Rasch Measurement Transactions. 1996; 9:468.



Note: The measure was scaled such that the mean item difficulty is 50 and 10 units is equal to 1 logit.

Figure 1. ***Figures 1A-1C*****. Distribution of Subjects and HTQ Items**



Note: The measure was scaled such that the mean item difficulty is 50 and 10 units is equal to 1 logit.

Figure 2. ***Figures 2A-2C*****. Distribution of Subjects and HSCL Items**

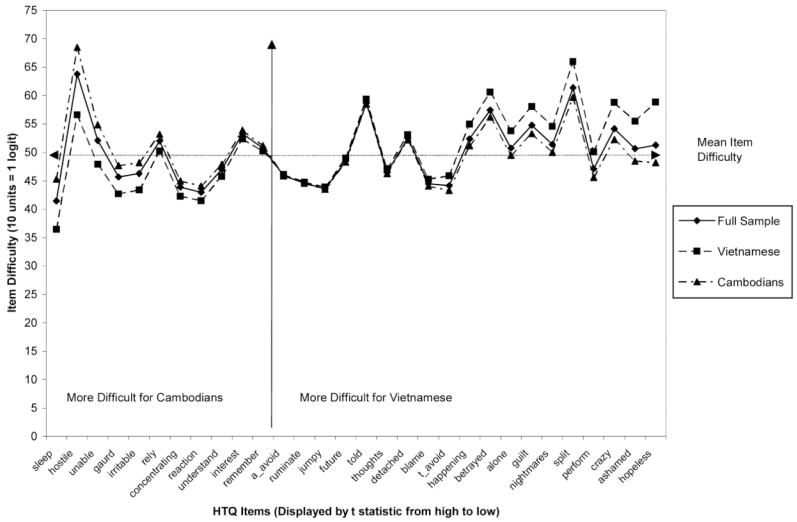


Figure 3. HTQ Item DIF For Cambodians and Vietnamese

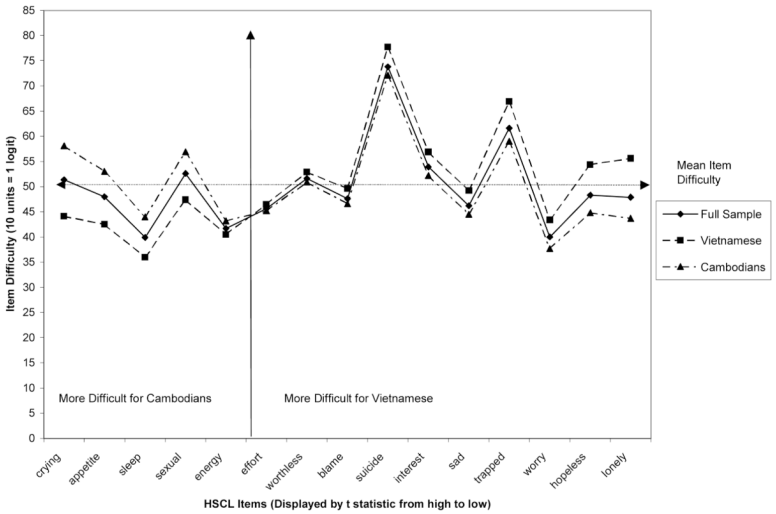


Figure 4. HSCL Item DIF For Cambodians and Vietnamese

Table 1

Item Fit Statistics for the HTQ Measure (N=274)

				INFIT		OUTFIT			
Item	**Raw Score**	**Item Difficulty**	***SE***	**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**	**PTMEA Correlation**	**Discrimination**
hostile	350	63.76	1.40	1.18	1.39	1.16	0.65	0.47	0.89
split	356	61.40	1.34	0.92	−0.60	0.57	−2.06	0.56	1.20
told	372	58.71	1.26	1.20	1.69	0.90	−0.44	0.54	0.97
betrayed	383	57.51	1.22	1.13	1.12	0.86	−0.70	0.57	1.03
guilt	403	54.78	1.15	1.00	0.05	0.70	−1.86	0.62	1.10
crazy	409	54.25	1.14	1.01	0.15	0.70	−1.88	0.63	1.18
interest	417	53.33	1.11	1.14	1.23	1.04	0.27	0.58	0.92
detached	425	52.49	1.09	1.21	1.83	1.08	0.53	0.58	0.90
happening	426	52.37	1.09	1.05	0.48	0.87	−0.82	0.62	1.05
unable	428	52.14	1.09	1.27	2.35	1.37	2.18	0.55	0.78
rely	428	52.00	1.08	1.16	1.48	1.13	0.84	0.59	0.88
nightmares	431	51.41	1.07	0.99	−0.04	1.03	0.22	0.62	0.95
hopeless	429	51.33	1.08	1.11	1.01	0.80	−1.33	0.64	1.13
alone	438	50.84	1.06	1.05	0.46	0.82	−1.28	0.65	1.10
remember	438	50.80	1.06	0.89	−1.09	0.89	−0.70	0.65	1.09
ashamed	435	50.68	1.06	1.00	0.00	0.72	−2.03	0.66	1.19
future	459	48.51	1.01	1.13	1.27	1.35	2.37	0.62	0.89
perform	473	47.07	0.99	0.75	−2.67	0.79	−1.68	0.70	1.13
understand	476	47.02	0.99	0.96	−0.42	0.99	−0.03	0.67	0.98
thoughts	481	46.54	0.98	0.85	−1.55	1.02	0.23	0.68	1.03
irritable	483	46.25	0.97	1.12	1.22	1.32	2.38	0.62	0.81
a_avoid	488	45.88	0.97	0.91	−0.95	0.76	−2.13	0.71	1.26
guard	489	45.65	0.96	1.23	2.21	1.23	1.77	0.64	0.80
ruminate	501	44.58	0.95	0.84	−1.75	0.82	−1.62	0.72	1.21
blame	499	44.52	0.95	0.70	−3.45	0.83	−1.48	0.73	1.21
t_avoid	504	44.18	0.94	0.98	−0.18	0.87	−1.08	0.71	1.11
concentrating	509	43.89	0.94	0.83	−1.91	0.85	−1.36	0.72	1.18

				INFIT		OUTFIT			
Item	**Raw Score**	**Item Difficulty**	***SE***	**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**	**PTMEA Correlation**	**Discrimination**
jumpy	511	43.59	0.93	1.16	1.66	1.11	0.95	0.67	0.88
reaction	518	43.03	0.93	0.98	−0.13	1.07	0.68	0.69	0.85
sleep	536	41.48	0.91	1.25	2.63	1.68	5.27	0.63	0.45

Note: Items are listed in decreasing order of difficulty to endorse (symptom severity) from high to low.

SE=Standard Error.

Table 2

Item Fit Statistics for the HSCL Depression Measure (N=257)

				INFIT		OUTFIT			
Item	**Raw Score**	**Item Difficulty**	***SE***	**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**	**PTMEA Correlation**	**Discrimination**
suicide	307	73.78	1.78	1.40	2.56	0.68	−0.88	0.51	0.98
trapped	358	61.59	1.37	1.22	1.79	0.80	−1.01	0.64	1.03
interest	402	53.87	1.21	0.93	−0.66	1.02	0.19	0.70	1.04
sexual	406	52.57	1.19	1.36	2.97	1.24	1.67	0.67	0.79
worthless	418	51.56	1.16	1.06	0.60	0.97	−0.20	0.71	1.00
crying	420	51.40	1.16	1.36	3.04	1.24	1.73	0.67	0.68
hopeless	444	48.31	1.10	0.92	−0.76	0.72	−2.55	0.76	1.23
appetite	448	47.98	1.10	1.05	0.54	1.12	1.03	0.71	0.91
lonely	450	47.95	1.09	1.00	0.04	0.85	−1.30	0.74	1.15
blame	449	47.64	1.09	1.00	0.08	0.92	−0.68	0.74	1.01
sad	463	46.23	1.07	0.73	−2.91	0.69	−3.03	0.80	1.32
effort	468	45.57	1.06	0.78	−2.33	0.85	−1.31	0.77	1.18
energy	504	41.72	1.00	0.98	−0.21	1.10	0.96	0.75	0.93
worry	525	39.97	0.98	0.86	−1.53	1.16	1.43	0.80	1.08
sleep	521	39.88	0.98	1.16	1.68	1.35	2.92	0.74	0.73

Note: Items are listed in decreasing order of difficulty to endorse (symptom severity) from high to low.

SE=Standard Error.

Table 3

Response Scale Structure for HTQ and HSCL Scales (N=317)

Scale	Category	%	Average Measure	Expected Measure	OUTFIT MNSQ	Step Calibration	Coherence: M→C
HTQ							
	1=Not at all	53	−29.18	−28.90	1.00		81%
	2=A little	34	−11.35	−11.80	.92	−15.24	57%
	3=Quite a bit	7	−0.58	−1.33	.97	8.99	31%
	4=Extremely	5	6.82	8.47	1.19	6.25	68%
HSCL							
	1=Not at all	48	−31.77	−30.40	.98		80%
	2=A little	37	−12.22	−12.60	.86	−19.10	92%
	3=Quite a bit	7	2.54	.89	1.01	11.31	28%
	4=Extremely	7	13.69	15.53	1.37	7.79	73%

Note: M→C refers to the percentage of responses expected to be in the category by the measure that are actually observed in the category.

Table 4

Person Summary Statistics for the HTQ Measure (30 Items and 4 Categories) and HSCL Measure (15 Items, 4 Categories)

Measur	e Sample	*N*	Average Measure	Average ZSTD (SD)		Adj. *SD*	RMSE	Separation	Reliability
				INFIT	OUFIT				
HTQ									
	Vietnamese	129	25.02	0.00 (1.4)	0.00 (1.3)	13.52	5.34	2.53	.86
	Cambodians	145	36.15	1.80 (1.8)	1.80 (1.8)	15.52	4.14	3.75	.93
	Full Sample	274	30.91	0.00 (1.6)	−0.10 (1.6)	15.63	4.75	3.29	.92
HSCL									
	Vietnamese	119	27.38	−0.20 (1.2)	−0.20 (1.1)	13.99	6.25	2.24	.83
	Cambodians	138	34.17	0.10 (1.4)	0.10 (1.3)	17.80	5.88	3.03	.90
	Full Sample	257	31.03	0.00 (1.3)	0.00 (1.2)	16.50	6.06	2.73	.88

Note: RMSE refers to the real root mean square error.

Table 5

HTQ Item DIF for Cambodians and Vietnamese (N=274)

Respondent Group	Item Difficulty	*SE*	Respondent Group	Item Difficulty	*SE*	DIF Contrast	Joint S.E.	*t*	*df*	Item Name
Cambodians	45.27	1.21	Vietnamese	36.36	1.32	8.91	1.79	4.97	270	**sleep**
Cambodians	68.51	1.91	Vietnamese	56.51	2.04	12.00	2.80	4.29	272	**hostile**
Cambodians	54.89	1.41	Vietnamese	47.80	1.66	7.10	2.18	3.26	272	**unable**
Cambodians	47.69	1.25	Vietnamese	42.56	1.48	5.13	1.94	2.65	271	**guard**
Cambodians	48.16	1.26	Vietnamese	43.32	1.50	4.83	1.96	2.47	271	**irritable**
Cambodians	53.15	1.37	Vietnamese	50.07	1.75	3.08	2.22	1.39	271	rely
Cambodians	44.97	1.21	Vietnamese	42.23	1.46	2.75	1.90	1.45	271	concentrating
Cambodians	44.12	1.20	Vietnamese	41.41	1.44	2.72	1.87	1.45	271	reaction
Cambodians	47.84	1.25	Vietnamese	45.71	1.58	2.14	2.02	1.06	272	understand
Cambodians	53.91	1.39	Vietnamese	52.35	1.85	1.56	2.31	0.67	271	interest
Cambodians	51.22	1.33	Vietnamese	50.11	1.75	1.11	2.20	0.50	271	remember
Cambodians	45.86	1.22	Vietnamese	45.96	1.59	−0.10	2.00	−0.05	272	a_avoid
Cambodians	44.54	1.20	Vietnamese	44.67	1.55	−0.13	1.96	−0.07	271	ruminate
Cambodians	43.48	1.19	Vietnamese	43.78	1.51	−0.29	1.92	−0.15	270	jumpy
Cambodians	48.32	1.26	Vietnamese	48.89	1.70	−0.57	2.12	−0.27	270	future
Cambodians	58.46	1.53	Vietnamese	59.34	2.22	−0.88	2.70	−0.33	267	told
Cambodians	46.31	1.23	Vietnamese	46.99	1.63	−0.68	2.04	−0.34	272	thoughts
Cambodians	52.23	1.35	Vietnamese	53.07	1.88	−0.84	2.31	−0.36	272	detached
Cambodians	44.11	1.19	Vietnamese	45.24	1.58	−1.13	1.98	−0.57	269	blame
Cambodians	43.26	1.18	Vietnamese	45.80	1.59	−2.53	1.98	−1.28	270	t_avoid
Cambodians	51.16	1.32	Vietnamese	54.91	1.96	−3.75	2.37	−1.58	272	happening
Cambodians	56.18	1.46	Vietnamese	60.58	2.28	−4.41	2.71	−1.63	269	betrayed
Cambodians	49.46	1.29	Vietnamese	53.69	1.91	−4.23	2.30	−1.84	271	alone
Cambodians	53.34	1.37	Vietnamese	58.01	2.14	−4.68	2.55	−1.84	269	guilt
Cambodians	49.96	1.30	Vietnamese	54.50	1.97	−4.55	2.36	−1.93	270	nightmares
Cambodians	59.72	1.56	Vietnamese	65.96	2.76	−6.24	3.16	−1.97	267	**split**
Cambodians	45.56	1.21	Vietnamese	50.00	1.75	−4.44	2.13	−2.08	270	**perform**
Cambodians	52.25	1.35	Vietnamese	58.72	2.16	−6.47	2.55	−2.54	271	**crazy**
Cambodians	48.46	1.27	Vietnamese	55.48	2.01	−7.02	2.38	−2.95	268	**ashamed**

Respondent Group	Item Difficulty	*SE*	Respondent Group	Item Difficulty	*SE*	DIF Contrast	Joint S.E.	*t*	*df*	Item Name
Cambodians	48.19	1.27	Vietnamese	58.83	2.19	−10.64	2.53	−4.20	268	**hopeless**

Note: Items in bolded text display significant DIF. Bolded items at the top are significantly more difficult for Cambodians while bolded items at the bottom are significant more difficult for Vietnamese.

Table 6

HSCL Item DIF for Cambodians and Vietnamese (N=257)

Respondent Group	Item Difficulty	*SE*	Respondent Group	Item Difficulty	*SE*	DIF Contrast	Joint *SE*	*t*	*df*	Item Name
Cambodians	58.03	1.65	Vietnamese	43.81	1.57	14.22	2.28	6.25	253	**crying**
Cambodians	52.98	1.53	Vietnamese	42.18	1.52	10.80	2.16	5.00	254	**appetite**
Cambodians	43.87	1.38	Vietnamese	35.46	1.39	8.41	1.95	4.31	252	**sleep**
Cambodians	56.89	1.64	Vietnamese	47.19	1.69	9.70	2.35	4.13	250	**sexual**
Cambodians	43.05	1.36	Vietnamese	40.13	1.47	2.92	2.00	1.46	253	energy
Cambodians	45.09	1.39	Vietnamese	46.26	1.65	−1.17	2.16	−0.54	253	effort
Cambodians	50.83	1.49	Vietnamese	52.75	1.88	−1.92	2.40	−0.80	252	worthless
Cambodians	46.50	1.41	Vietnamese	49.35	1.76	−2.85	2.25	−1.26	253	blame
Cambodians	72.10	2.12	Vietnamese	77.63	3.38	−5.54	3.99	−1.39	255	suicide
Cambodians	52.18	1.51	Vietnamese	56.82	2.05	−4.64	2.55	−1.82	253	interest
Cambodians	44.36	1.38	Vietnamese	49.00	1.73	−4.63	2.22	−2.09	254	**sad**
Cambodians	58.96	1.67	Vietnamese	66.80	2.52	−7.84	3.03	−2.59	255	**trapped**
Cambodians	37.58	1.30	Vietnamese	43.12	1.55	−5.54	2.02	−2.74	255	**worry**
Cambodians	44.74	1.38	Vietnamese	54.28	1.94	−9.54	2.38	−4.01	253	**hopeless**
Cambodians	43.58	1.36	Vietnamese	55.50	1.98	−11.92	2.40	−4.96	255	**lonely**

Note: Items in bolded text display significant DIF. Bolded items at the top are significantly more difficult for Cambodians while bolded items at the bottom are significant more difficult for Vietnamese.

J Appl Meas. Author manuscript; available in PMC 2012 March 2.

Table 7

Rasch Fit Statistics for the HTQ Core Items (N=260)

				INFIT		OUTFIT			
Item	**Raw Score**	**Item Difficulty**	***SE***	**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**	**PTMEA Correlation**	**Discrimination**
told	359	59.75	1.28	1.26	2.10	0.99	0.03	0.55	0.92
betrayed	369	58.53	1.24	1.20	1.71	0.93	−0.34	0.57	0.97
guilt	389	55.69	1.17	1.09	0.83	0.83	−1.09	0.62	1.01
interest	403	54.19	1.13	1.18	1.56	1.09	0.66	0.59	0.86
detached	411	53.31	1.11	1.23	1.96	1.14	0.96	0.59	0.87
happening	412	53.19	1.11	1.07	0.67	0.91	−0.58	0.64	1.02
rely	414	52.81	1.10	1.18	1.63	1.15	1.04	0.60	0.86
nightmares	417	52.17	1.10	1.01	0.08	1.22	1.55	0.62	0.91
alone	424	51.60	1.08	1.11	1.03	0.88	−0.85	0.65	1.05
remember	424	51.55	1.08	0.92	−0.68	0.88	−0.89	0.67	1.07
future	446	49.18	1.03	1.18	1.70	1.36	2.68	0.63	0.85
understand	462	47.63	1.00	0.91	−0.91	1.01	0.11	0.69	0.99
thoughts	467	47.13	1.00	0.81	−1.98	0.94	−0.47	0.71	1.07
a_avoid	474	46.45	0.98	0.89	−1.15	0.76	−2.28	0.73	1.27
ruminate	487	45.11	0.96	0.84	−1.67	0.81	−1.78	0.74	1.21
blame	485	45.04	0.97	0.73	−3.10	0.84	−1.49	0.74	1.18
t_avoid	491	44.69	0.96	0.91	−0.89	0.81	−1.87	0.74	1.18
concentrating	496	44.39	0.95	0.84	−1.70	0.94	−0.54	0.73	1.15
jumpy	498	44.09	0.95	1.19	1.97	1.25	2.16	0.67	0.80
reaction	504	43.49	0.94	0.97	−0.26	1.06	0.57	0.71	0.85

Note: Items are listed in decreasing order of difficulty to endorse (symptom severity) from high to low.

SE=Standard Error.

Table 8

Rasch Fit Statistics for the HSCL Core Items (N=230)

				INFIT		OUTFIT			
Item	**Raw Score**	**Item Difficulty**	***SE***	**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**	**PTMEA Correlation**	**Discrimination**
suicide	277	75.52	1.92	1.50	3.11	0.92	−0.12	0.58	0.85
interest	373	52.17	1.33	1.01	0.14	1.02	0.25	0.76	0.96
worthless	390	49.37	1.29	1.06	0.61	0.96	−0.33	0.80	1.03
blame	419	44.33	1.23	1.01	0.16	0.94	−0.56	0.82	1.07
effort	439	41.81	1.20	0.80	−2.04	0.80	−1.97	0.84	1.23
energy	475	36.81	1.14	1.14	1.33	1.28	2.32	0.79	0.86

Note: Items are listed in decreasing order of difficulty to endorse (symptom severity) from high to low.

SE=Standard Error.

Table 9

Response Scale Structure for HTQ and HSCL Core Items (N=317)

Scale	Category	%	Average Measure	Expected Measure	OUTFIT MNSQ	Step Calibration	Coherence: M→C
HTQ							
	1=Not at all	49	−27.86	−27.60	.99		80%
	2=A little	37	−11.13	−11.30	.89	−16.16	58%
	3=Quite a bit 4=Extremely	8 6	.02 7.52	−.90 8.75	1.02 1.24	9.42 6.74	31% 70%
HSCL							
	1=Not at all	45	−37.78	−37.20	.95		78%
	2=A little	42	−14.83	−15.40	.84	−25.30	68%
	3=Quite a bit	7	4.95	2.23	1.04	12.20	31%
	4=Extremely	6	17.82	21.29	1.66	13.10	71%

Note: M→C refers to the percentage of responses expected to be in the category by the measure that are actually observed in the category.

J Appl Meas. Author manuscript; available in PMC 2012 March 2.

Table 10

Person Summary Statistics for the HTQ Core Measure (20 Items and 4 Categories) and HSCL Core Measure (6 Items, 4 Categories)

				Average ZSTD (SD)					
Measure	**Sample**	***N***	**Average Measure**	**INFIT**	**OUFIT**	**Adj. SD**	**RMSE**	**Separation**	**Reliability**
HTQ									
	Vietnamese	117	26.79	0.00 (1.3)	0.00 (1.3)	12.98	5.93	2.19	.83
	Cambodians	143	37.34	−0.10 (1.6)	−0.10 (1.5)	14.90	4.69	3.18	.91
	Full Sample	260	32.59	−0.10 (1.4)	−0.10 (1.4)	15.01	5.28	2.84	.89
HSCL									
	Vietnamese	103	22.55	−0.3 0 (0.8)	−0.10 (0.8)	9.77	9.36	1.04	.52
	Cambodians	127	32.39	0.20 (1.2)	0.30 (1.1)	17.68	9.70	1.82	.77
	Full Sample	230	27.98	0.00 (1.0)	0.00 (1.0)	15.47	9.55	1.62	.72