Published in final edited form as:

Soc Psychiatry Psychiatr Epidemiol. 2016 July ; 51(7): 981–991. doi:10.1007/s00127-016-1218-3.

Depression symptoms across cultures: an IRT analysis of standard depression symptoms using data from eight countries

E. E. Haroz[1], **P. Bolton**[2], **A. Gross**[3], **K. S. Chan**[4], **L. Michalopoulos**[5], and **J. Bass**[1]

[1]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, 624 N. Broadway, Baltimore, MD 21205, USA

[2]Department of International Health, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Baltimore, MD 21205, USA

[3]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Baltimore, MD 21205, USA

[4]Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Baltimore, MD 21205, USA

[5]School of Social Work, Columbia University, 1255 Amsterdam Avenue, New York, NY 10027, USA

Abstract

Purpose—Prevalence estimates of depression vary between countries, possibly due to differential functioning of items between settings. This study compared the performance of the widely used Hopkins symptom checklist 15-item depression scale (HSCL-15) across multiple settings using item response theory analyses. Data came from adult populations in the low and middle income countries (LMIC) of Colombia, Indonesia, Kurdistan Iraq, Rwanda, Iraq, Thailand (Burmese refugees), and Uganda ($N = 4732$).

Methods—Item parameters based on a graded response model were compared across LMIC settings. Differential item functioning (DIF) by setting was evaluated using multiple indicators multiple causes (MIMIC) models.

Results—Most items performed well across settings except items related to suicidal ideation and “loss of sexual interest or pleasure,” which had low discrimination parameters (suicide: $a = 0.31$ in Thailand to $a = 2.49$ in Indonesia; sexual interest: $a = 0.74$ in Rwanda to $a = 1.26$ in one region of Kurdistan). Most items showed some degree of DIF, but DIF only impacted aggregate scale-level scores in Indonesia.

Conclusions—Thirteen of the 15 HSCL depression items performed well across diverse settings, with most items showing a strong relationship to the underlying trait of depression. The results support the cross-cultural applicability of most of these depression symptoms across LMIC

Correspondence to: E. E. Haroz.

Electronic supplementary material The online version of this article (doi:10.1007/s00127-016-1218-3) contains supplementary material, which is available to authorized users.

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

settings. DIF impacted aggregate depression scores in one setting illustrating a possible source of measurement invariance in prevalence estimates.

Keywords

Depression; Global mental health; Psychometrics; Measurement invariance; Item response theory

Introduction

Major depression is a major contributor to the global burden of disease [1]. However, prevalence estimates vary across settings, despite use of standardized protocols and assessment instruments [2]. Twelve-month prevalence rate estimates of major depressive disorder range from 2.2 % in Japan to 10.4 % in Brazil [2]. In contexts of war and displacement, Steel et al. [3] found depression prevalence rates among adults range between 3 and 85.5 %.

While variation may reflect true differences in prevalence, measurement error as a result of methodological factors such as differences in item performance across settings may also contribute to heterogeneity. The degree to which this error is relevant impacts policy decisions, program planning/evaluation, and service provision, particularly in low-resource settings [4, 5].

One potential source of measurement error is variation in performance of standardized instruments between populations [5, 6]. Most measurement instruments that have been used in global mental health research were originally developed for use with Western, clinical populations. Their applicability to other populations has not been well explored using quantitative analytic methods.

Item response theory (IRT) is one method that can be used to address the question of cross-population utility. IRT is a type of latent variable analysis that models the probability of a given response as a function of a respondent's underlying level of a latent trait. For example, IRT assesses the probability of endorsing each individual symptom on a depression instrument based on a respondent's level (or severity) of depression. Severity of depression is measured by a summary score of the items (i.e., symptoms) administered. IRT methods allow for a better understanding of how a depression instrument performs in different populations by providing information on item characteristics and identifying potential item-level bias across populations (i.e., differential item functioning; DIF) [7].

IRT has been used to investigate the performance of depression measures by sex [8], race/ethnicity [9, 10], administration methods [11], and language [12]. Few studies have used IRT across socio-cultural settings. Nuevo et al. [13] found that items on the Beck depression inventory (BDI) performed differently across five different European countries. Canel-Çınarbaş and colleagues [14] found similar findings comparing populations from Turkey and the United States. We could not find any studies comparing the performance of a depression measure in multiple low- and middle income countries (LMIC).

The present study examined the performance of the Hopkins symptom checklist 15-item depression scale (HSCL-15) [15], a widely used measure of depression in LMIC. The specific aim of the study was to assess the psychometric properties of the HSCL-15 for the measurement of depression across different language versions to determine the extent of cross-cultural variation in item response. We examined item parameters within each study population (referred to as "setting" herein) and evaluate these items for differential item functioning (DIF) by setting. Data are from adults (over 16 years old) in eight different LMIC: Colombia, Indonesia, Kurdistan Iraq (separated into two linguistically distinct settings: Dohuk and Erbil/Sulaymaniyah), Rwanda, Iraq, Thailand (Burmese refugees), and Uganda. The methods used in this study provide an example of methods that could be useful for similar investigations of other instruments of depression and/or other disorders cross-culturally. Results from this study will inform our understanding of how symptoms of depression present and vary across settings and whether cross-cultural variation in item response may be an important issue when explaining variation in findings between populations.

Methods

Data

Data come from: (1) displaced Afro-Colombians; (2) a conflict-affected population in Aceh Indonesia; torture- and trauma-affected populations in (3) the Dohuk region and (4) the Erbil/Sulaymaniyah regions of Kurdistan Iraq; (5) a genocide-affected population from Rwanda, (6) a torture-and trauma-affected population in Iraq; (7) torture- and trauma-affected Burmese refugees living in Thailand; and (8) adults affected by high death rates (due to HIV) in Uganda. The combined data represent individual-level depression data for N = 4732 participants (Table 1). Secondary data analysis was approved by the Johns Hopkins University IRB (IRB#4721).

The samples from Colombia, Indonesia, and Thailand were collected as part of studies to test the psychometrics of adapted instruments and as screening for randomized control trials (RCTs) of psychotherapeutic interventions [16–18]. The Kurdistan data (Dohuk and Erbil/Sulaymaniya) are from a clinic-based monitoring system established for a RCT of psychotherapeutic interventions [19]. The Rwanda data come from a population-based survey in rural communities [20]. The Iraq data are from a psychometric study of an instrument to measure psychological distress among victims of torture [21]. The data from Uganda come from a clustered-based random survey in southwest Uganda [22].

Measurement instrument

The HSCL-15 was adapted from the original 58-item Hopkins symptom checklist (HSCL), which was created to reflect symptom profiles of outpatient populations in the United States [23]. The 15-item depression sub-scale was created for use in family practice settings in the U.S. [24] and then validated for use with refugees from Southeast Asia [25]. It has been one of the most commonly used measures of depression in global mental health research [15].

In the current analysis, respondents were asked how much each item (i.e., symptom) bothered him/her in the past weeks. Responses ranged from 0 “not at all” to 3 “extremely.” The timeframe for the presence of symptoms varied by setting: 2 weeks in Indonesia, Kurdistan Iraq, Rwanda and Iraq; 4 weeks in Colombia and Thailand; and 1 week in Uganda. Prior to use, the HSCL-15 was adapted and validated in each setting [16, 18–22]. Adaptation consisted of using qualitative data to add additional local items. Qualitative data were also used to aid in translation. For translation, qualitative data are used as a key source for translating key concepts on the instrument, in words and phrases that local people actually use. Translators in each setting were instructed to translate all signs, symptoms and phrases on the instrument using phrases gathered during the qualitative research. If there were concepts not mentioned in the qualitative data, then translators would use their judgment to choose the appropriate terms. Following initial translation, the HSCL-15 was back translated in each setting. All items on the HSCL-15 were retained in most settings, with the exception of Indonesia where two items, “loss of sexual interest/pleasure” and “worry,” were not administered. In all settings the HSCL-15 was administered as part of a longer instrument aimed at assessing other domains of mental and behavioral health.

Analysis

Exploratory analysis examined basic descriptive statistics, distribution of item responses, average scores, and internal consistency reliabilities (Cronbach’s alpha; α) [26] across (N= 4732) and within settings. Factor analyses were conducted using the full dataset (N= 4732). Principal components analysis (PCA) with polychoric correlation matrix and exploratory factor analysis (EFA) were used to explore the dimensionality of the data. Confirmatory factor analysis (CFA) tested model fit of the proposed dimensional model. The dataset was randomly split into a development sample for the EFA and a validation sample for the CFA. Mean and variance-adjusted weighted least squares estimator (WLSMV) with Geomin rotated standardized factor loadings in Mplus 7.3 [27] were used for all factor analyses. Fit of confirmatory models was evaluated using the comparative fit index (CFI), the Tucker-Lewis index (TLI), and the root mean square error of approximation (RMSEA). RMSEA values lower than 0.06 and TLI/CFI values above 0.95 are indicative of good model fit [28].

To test for measurement differences across settings, the data were analyzed comparing each setting to all other data in the dataset (e.g., Colombia vs. all others). This was done because we did not want to specify one comparison group given that we were interested to see how the HSCL-15 and its items performed in each setting compared to all other LMIC settings for which we had data. Our exact measurement model was specified as: ($y_{is} = \mu_i + \lambda_{is}F_s + e_{is}$) where the observed response for item i for person s is equal to the item intercept μ_i plus the person’s latent trait F (i.e., depression) weighted by the factor loading λ plus error e.

We tested configural, metric, and scalar invariance across setting comparisons. Configural invariance tests if the same set of factors is present and indicates if the factor structure of the measure (i.e., HSCL-15) is similar across settings. Metric invariance tests if factor loadings are the same across settings and indicates whether the items are correlated with similar magnitudes to the underlying latent trait (i.e., depression) across settings. Scalar invariance is more restrictive than metric invariance and tests if item thresholds and factor loadings are the same, and reflects whether there are systematic differences in the way individuals from different settings respond to the items. Fit of each invariance model was evaluated using global fit indices as described above (Hu and Bentler 1998).

For the IRT analysis, item discrimination (*a*) and item location (*b*) parameters were estimated using the graded response model (GRM) [29]. The GRM was selected as the most appropriate model given the HSCL-15′s ordered response categories and the option to estimate different discrimination parameters across settings. For each item, one discrimination and three locations parameters were estimated. The GRM is specified as: $(P_{ik}(\theta) = P^{*}_{ik}(\theta) - P^{*}_{i,k-1}(\theta))$ where $(P^{*}_{ik}(\theta) = \frac{exp(Da_{i}(\theta - b_{ik}))}{1 + exp(Da_{i}(\theta - k))})$ and $P^{*}_{ik}(\theta)$ is the probability of a randomly chosen examinee with depression of θ endorsing the response category k or above on an item i and D is a scaling constant.

Parameters were estimated using the whole dataset and then for each setting separately. Discrimination parameters (*a*) are analogous to factor loadings and indicate how strongly an item is correlated to the underlying latent trait and how well it discriminates between people with different levels of the latent trait. Generally, item discrimination values of 0.01–0.34 are considered very low; 0.35–0.64 low; 0.65–1.34 moderate; 1.35–1.69 high; and 1.70 and above, very high [30].

Location parameters (*b*) are defined as the level of the underlying latent trait (θ) where the probability of endorsing the item with a particular response category is 0.50. Based on the GRM, three item location parameters (b_1, b_2, b_3) were estimated which correspond to the four possible response options on the HSCL-15. The first location parameter (b_1) represents the level of the underlying latent trait where the probability of endorsing the item with a "0" instead of a "1," "2," or "3" is 0.50. The second location parameter (b_2) is for the response of <2 and the third location parameter (b_3) for the response of <3.

In addition, measurement precision (or "information") was estimated for each item independently and for the scale as a whole (Test Information Curve; TIC). Information represents the certainty with which an item or scale measures the underlying latent trait (θ) and can vary as a function of the level of θ. All IRT models were implemented using IRTPRO [31].

Differential Item Functioning (DIF) occurs when respondents with the same level of latent trait (i.e., depression) have different probabilities of endorsing an item based on some other characteristic (i.e., setting). There are two types of DIF: non-uniform DIF, or DIF in the discrimination parameters; and uniform DIF, or DIF in the location parameters. Non-uniform DIF is analogous to effect modification and represents an interaction between the level of the latent trait, group membership and the item response. Uniform DIF is analogous to confounding or when the differences in responses to items can be found at all levels of the latent trait [32].

Non-Uniform and uniform DIF by setting were evaluated by comparing data from one setting (comparison group) to all other settings combined in the dataset (reference group) using MIMIC (multiple indicator, multiple causes) models with a WLSMV estimator in

Mplus v7.3 [27]. A Bonferroni correction was used to adjust the level of significance to account for multiple comparisons. DIF was indicated to be present if there was a statistically significant ($p < 0.001$) difference between the parameters in the comparison group (one setting) compared to the reference group (all other settings). For all DIF analyses, gender was included as an exogenous variable to control for its impact on item response.

DIF was evaluated for a total of eight comparison groups (representing each study setting) with eight different reference groups (e.g., reference group 1 included data from all settings but without the Colombia sample; reference group 2 included data from all settings but without the Indonesian sample, etc.). During calibration, the scale for IRT parameters are based group mean = 0, and 1 standard deviation = 1 of the reference group. Given the variation in the composition of reference croups, we compared the latent means and standard deviations of scores on the HSCL-15 for the different references groups. If the resulting means and standard deviations of the reference groups were relatively similar, then IRT parameter estimates for each setting can be reasonably compared. For example, if the means and standard deviations for each reference group were similar than the magnitude of DIF for an item in Colombia can be compared to the magnitude of DIF for that same item in Indonesia.

To investigate the impact of DIF, differences in latent mean scores for depression were examined comparing models that accounted for DIF to models that did not account for DIF. If the difference in latent mean score for depression between the two models was statistically significant, then DIF was considered to have a salient impact on scale-level scores.

Results

Exploratory analysis

Most participants were women (62.1 %) and between the ages of 25–44 years (46.5 %) (Table 2). Age distributions within each setting were comparable across settings and was unrelated to total HSCL-15 score. Item response distributions indicated that 0 “not at all” or 1 “A little bit” were the most frequent responses. Average scores on the HSCL-15 ranged from $\mu = 0.61$ in Rwanda to $\mu = 1.47$ in Dohuk. Across settings the internal consistency reliability (α) for the HSCL-15 was good ($\alpha = 0.87$; range: 0.79–0.93).

Factor analysis

Principal components analysis (PCA) and associated scree plot indicated one predominant factor (eigenvalue = 6.6; next highest eigenvalue = 1.1). Factor loadings from the one-factor EFA ranged from $\lambda = 0.50$ for “Loss of sexual interest/pleasure” to $\lambda = 0.79$ for “Feeling sad.” The CFA of a one-factor model yielded good model fit indices (RMSEA = 0.05; CFI = 0.95; TLI = 0.94).

Confirmatory between group comparison of the HSCL factor structure

Configural measurement invariance was largely supported in all settings, demonstrating that a 1-factor structure fits the data in all settings (see supplemental material table S1). RMSEA values for all configural models ranged from 0.64 in Dohuk to 0.75 in Colombia, Thailand,

and Uganda. CFI values for configural models ranged from 0.94 in Thailand to 0.965 in Indonesia, while TLI models ranged from 0.931 in Thailand to 0.958 in Indonesia were all above 0.90. Metric and scalar invariance showed less adequate fit across settings. For all setting comparisons and across all levels of measurement models, Chi-squared difference tests were significant indicating that only configural invariance was supported across settings.

Factor loadings for each item by setting comparison (supplemental material table S2), showed the lowest factor loadings for the item "loss of sexual interest or pleasure." Relatively large differences in factor loadings between settings was observed for the items "thoughts of death/suicide" in Thailand ($\lambda = 0.20$) compared to all other settings ($\lambda = 0.64$) and "crying too much" in Dohuk ($\lambda = 0.21$) compared to all other settings ($\lambda = 0.66$).

Item response theory model

Model fit indices for the GRM in the combined dataset indicated acceptable fit (RMSEA = 0.07; CFI = 0.95; TLI = 0.94). Model fit statistics in each setting indicated relatively good fit and ranged from RMSEA = 0.08, CFI = 0.96; TLI = 0.95 in Thailand to RMSEA = 0.04, CFI = 0.98; TLI = 0.98 in Uganda. The GRM model did not fit the data well in Indonesia (RMSEA = 0.11; CFI = 0.86; TLI = 0.86). However, after modifying the model by allowing "problems with appetite" and "problems with sleep," "crying" and "feeling sad," and "low energy" and "feeling everything is an effort" to correlate with each other, model fit improved (RMSEA = 0.06; CFI = 0.96; TLI = 0.95).

The IRT analysis in the combined dataset ($N = 4732$) indicated most items performed well. Discrimination parameters ranged from $a = 0.97$ ("lack of interest or pleasure in sex") to $a = 2.09$ ("feel sad") (Table 5). Location parameters (for b_1) ranged from $b_1 = -1.02$ ("feel sad") to $b_1 = 1.38$ ("thoughts of death/suicide") (Table 4). Measurement precision was best for people in the $\theta = 0.5$ to $\theta = 1.0$ range, or for those with depression slightly above the average level of depression across settings.

In the setting-specific IRT analyses, discrimination parameters ranged from $a = 0.31$ for "thoughts of death/suicide" in Thailand to $a = 3.10$ for "no interest" in Erbil/Sulaymaniyah. The item "lack of interest or pleasure in sex" had low or moderate discrimination parameters across settings; $a = 0.74$ in Rwanda to $a = 1.26$ in Erbil/Sulaymaniyah. Low discrimination parameters were also observed for "feeling everything is an effort" in Colombia ($a = 0.88$) and the items "loss of interest" ($a = 0.74$) and "self-blame" ($a = 0.98$) in Rwanda (Table 3).

Across settings "feeling sad" was most commonly endorsed by individuals with low levels of depression ($b_1 = -1.02$; $b_2 = -0.10$; $b_3 = 0.94$), while "thoughts of death or suicide" was commonly endorsed by people with higher levels of depression ($b_1 = 1.38$; $b_2 = 2.33$; $b_3 = 3.58$). Dohuk had the lowest item location parameter for "crying" ($b = -3.67$ for a response of <1) meaning that individuals in Dohuk had to have relatively low levels of depression to endorse this item. In Thailand "thoughts of death/suicide" was "difficult" ($b = 2.74$ response of <1) meaning that individuals had to have relatively high levels of depression to endorse this item (Table 4).

Differential item functioning (DIF)

DIF detection was done by comparing scale parameters in one setting (comparison group) to scale parameters in all other settings (reference group) combined (e.g., Colombia vs. all others). Mean of latent depression scores did not vary widely across reference groups (Range: 1.00 for the total sample without the Colombia data to 1.14 for the total sample without the Thailand data). Standard deviations of the mean latent depression scores ranged from SD = 0.59 for the total sample without the Colombia data to SD = 0.68 for the total sample without the Southern Iraq data. These results allow interpretation of DIF between comparison groups.

Non-uniform DIF was detected in all settings for at least one item. In Thailand, eight items showed non-uniform DIF. For example, the item "low energy" is more closely related to depression and is better at discriminating between levels of depression in Thailand compared to all other settings, but "thoughts of death/suicide" seems less related to depression in Thailand compared to all other settings. The items "loss of sexual interest" and "changes in appetite" were free of non-uniform DIF across all setting comparisons indicating these two items are similarly related to depression regardless of the setting (Table 5).

The most uniform DIF was observed for Indonesia compared to all other settings. Five items showed uniform DIF indicating that these items were most informative at different levels of underlying depression in Indonesia compared to all other settings. All items in Rwanda and Southern Iraq were free of uniform DIF meaning that items had the same location parameters when comparing these settings to all other settings (Table 4).

Impact of DIF—Salient impact of DIF (both uniform and non-uniform DIF) on aggregate latent mean scores of depression was observed for Indonesia. Without accounting for item-level DIF, participants in Indonesia had, on average, 0.07 units less depression compared to people in all other settings. Once DIF was accounted for participants in Indonesia had, on average, 0.28 units less depression compared to people in other settings. This difference of 0.21 suggests that by not accounting for DIF, average scores of depression in the Indonesian sample were overestimated. DIF did not have a significant impact on latent mean values of depression in Columbia, Dohuk, Rwanda, Thailand, Southern Iraq, Uganda and Erbil/Sulaymaniyah (Table 5).

Discussion

The aim of this study was to evaluate the psychometric properties of the HSCL-15 across different language versions to determine the extent of cross-cultural variation in item response. Our goal was to examine whether cross-cultural variation in item response may be important issue in explaining variation in findings between populations. Our findings suggest that the HSCL-15 items generally perform well and are therefore applicable and un-biased across multiple populations.

Factor analyses indicated that a unidimensional model was appropriate across settings. Most items had high discrimination parameters indicating a strong relationship of these items to depression regardless of the setting. These include: "feeling hopeless," "feeling sad," "feeling low in energy or slowed down," "problems with sleep," "feeling trapped," "worrying too much," and "feeling worthless." Location parameters indicate that most items were most informative for measuring slightly higher than average levels of depression. However, "feeling sad" and "worry" were most informative for lower than average levels of depression, and "thoughts of death/suicide" and "feeling worthless" were more informative for higher than average levels of depression.

Across settings "loss of sexual interest or pleasure" had relatively low discrimination parameters ($a = 0.74$ to $a = 1.26$). In many non-Western populations topics related to sex are not discussed openly. In response to cultural norms researchers have modified [16] or considered modifying [33] measures, despite the robust evidence that reporting this symptom is related to depression in high-income settings [34–36]. However, the evidence from the current study demonstrated that this item is not strongly related to depression in these settings, perhaps due to unwillingness to provide frank responses.

In setting-specific analyses, the lowest discrimination parameter was observed for "thoughts of killing oneself/suicide" ($a = 0.31$) in Thailand. Frequencies of item response categories were comparable across settings meaning that the low discrimination parameter in Thailand was not due to infrequent endorsement of the item. Instead, it appears that this item may not be a good indicator for depression in Thailand. Perhaps, in this context, suicidal ideation is not being driven by depression. Recent findings have shown that while depression is a risk factor for suicide ideation in high-income countries, impulse control disorders are more strongly associated with thoughts of death and suicide in many LMIC [37].

Uniform DIF findings indicated that setting often confounds the relationship of item response to depression. Setting may affect the symptoms that are indicative of mild or severe depression. For example, an individual from Indonesia with mild depression might indicate experiencing "low energy and fatigue," but this symptom may be more representative of severe depression in all other settings.

While DIF was found for all items, the impact of this DIF on overall scale scores was variable. There was no observed impact on latent mean depression levels in Colombia, Dohuk, Rwanda, Thailand, Southern Iraq, Uganda and Erbil/Sulaymaniyah. The only impact of DIF was for Indonesia, where depression level is likely to be overestimated if DIF is not accounted for.

The impact of DIF on aggregate scores may be one source of measurement error contributing to heterogeneity in prevalence. A recent review by Kessler and Bromet [38] suggested that measurement factors do not play a significant role in cross-national variability of depression. However, this review did not consider possible DIF. Despite the use of non-epidemiologic samples evidence from the current study suggests that in some cases, item-level response bias has a significant impact on aggregate estimates of depression. Future epidemiologic studies should investigate the impact of DIF on variability in prevalence rates.

Limitations

All of the data used in this analysis came from trauma-affected, non-representative populations, limiting the generalizability of the findings. While the data were based on validated scales and collected using similar methodology, it is possible that differences in symptom recall timeframes (i.e., 1–4 weeks), recruitment strategies, or other study procedures may have led to variability in responses and, subsequently, item parameters. As our main aim was to explore DIF across a wide range of settings to determine which items performed well across populations, we limited our investigation of DIF to only examine study setting while controlling for gender. We did not explore other variables that could be responsible for the observed DIF, such as age, education level, or other unmeasured variables. Future studies should evaluate DIF related to these and other potential sources of response bias. Finally, translation and slight changes in meaning of items across settings may have affected the DIF analysis.

Conclusions

The depression items from the HSCL performed well across diverse settings, with most showing a strong relationship to the underlying trait of depression. Overall, items were most informative for people with higher than average levels of depression. Some items performed poorly across settings including “loss of sexual interest and pleasure” and “thoughts of killing oneself/suicide.” Almost all items showed DIF; however, the impact of this DIF was salient in only two settings. This was the first study to examine the performance of depression related measurement items across multiple settings in LMIC. The methods used in this investigation illustrate the richness of information provided by IRT for scale development and/or refinement. Results suggest that the majority of the depression symptoms included in this analysis are applicable and un-biased across multiple populations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Whiteford HA, et al. The global burden of mental, neurological and substance use disorders: an analysis from the global burden of disease study 2010. PLoS One. 2015; 10(2):e0116820. [PubMed: 25658103]
2. Moussavi S, et al. Depression, chronic diseases, and decrements in health: results from the World Health Surveys. Lancet. 2007; 370(9590):851–858. [PubMed: 17826170]
3. Steel Z, et al. Association of torture and other potentially traumatic events with mental health outcomes among populations exposed to mass conflict and displacement: a systematic review and meta-analysis. JAMA. 2009; 302(5):537–549. [PubMed: 19654388]
4. Kohrt BA, et al. Validation of cross-cultural child mental health and psychosocial research instruments: adapting the depression self-rating scale and child PTSD symptom scale in Nepal. BMC Psychiatry. 2011; 11:127–144. [PubMed: 21816045]
5. Wessells MG. Do no harm: toward contextually appropriate psychosocial support in international emergencies. Am Psychol. 2009; 64(8):842–854. [PubMed: 19899908]
6. Bass JK, Bolton PA, Murray LK. Do not forget culture when studying mental health. Lancet. 2007; 370(9591):918–919. [PubMed: 17869621]

7. Hambleton, RK., Waminathan, H., Rogers, HJ. Fundamentals of item response theory. Vol. 1. Sage Publications Inc; California: 1991.
8. Bares C, et al. Differential item functioning due to gender between depression and anxiety items among Chilean adolescents. Int J Soc Psychiatry. 2011; 58(4):386–392. [PubMed: 21628359]
9. Hambrick JP, et al. Cross-ethnic measurement equivalence of measures of depression, social anxiety, and worry. Assessment. 2010; 17(2):155–171. [PubMed: 19915199]
10. Kim G, Chiriboga DA, Jang Y. Cultural equivalence in depressive symptoms in older white, black, and Mexican–American adults. J Am Geriatr Soc. 2009; 57(5):790–796. [PubMed: 19484834]
11. Bjorner JB, et al. Difference in method of administration did not significantly impact item response: an IRT-based analysis from the patient-reported outcomes measurement information system (PROMIS) initiative. Qual Life Res. 2014; 23(1):217–227. [PubMed: 23877585]
12. Paz SH, et al. Evaluation of the patient-reported outcomes information system (PROMIS$^{®}$) Spanish-language physical functioning items. Qual Life Res. 2013; 22(7):1819–1830. [PubMed: 23124505]
13. Nuevo R, et al. Cross-cultural equivalence of the Beck depression inventory: a five-country analysis from the ODIN study. J Affect Disord. 2009; 114(1):156–162. [PubMed: 18684511]
14. Canel-Çınarbaş D, Cui Y, Lauridsen E. Cross-cultural validation of the Beck depression inventory–II across US and Turkish samples. Meas Eval Couns Dev. 2011; 44(2):77–91.
15. Mollica, RF., et al. Measuring trauma, measuring torture. Harvard University; Cambridge: 2004.
16. Bass J, et al. A controlled trial of problem-solving counseling for war-affected adults in Aceh, Indonesia. Soc Psychiatry Psychiatr Epidemiol. 2012; 47(2):279–291. [PubMed: 21246186]
17. Bolton P, et al. a transdiagnostic community-based mental health treatment for comorbid disorders: development and outcomes of a randomized controlled trial among Burmese refugees in Thailand. PLoS Med. 2014; 11(11):e1001757. [PubMed: 25386945]
18. Haroz EE, et al. Adaptation and testing of psychosocial assessment instruments for cross-cultural use: an example from the Thailand Burma border. BMC Psychol. 2014; 2(1):31–40. [PubMed: 25685351]
19. Bolton P, et al. A randomized controlled trial of mental health interventions for survivors of systematic violence in Kurdistan, Northern Iraq. BMC Psychiatry. 2014; 14(1):360–375. [PubMed: 25551436]
20. Bolton P, Neugebauer R, Ndogoni L. Prevalence of depression in rural Rwanda based on symptom and functional criteria. J Nerv Ment Dis. 2002; 190(9):631–637. [PubMed: 12357098]
21. Weiss, W., Bolton, P. Assessment of torture survivors in Southern Iraq: development and testing of a locally-adapted assessment instrument. United States Agency for International Development; Washington, DC: 2010.
22. Bolton P, Wilk CM, Ndogoni L. Assessment of depression prevalence in rural Uganda using symptom and function criteria. Soc Psychiatry Psychiatr Epidemiol. 2004; 39(6):442–447. [PubMed: 15205728]
23. Derogatis LR, et al. The Hopkins Symptom Checklist (HSCL): a self-report symptom inventory. Behav Sci. 1974; 19(1):1–15. [PubMed: 4808738]
24. Hesbacher PT, et al. Psychiatric illness in family practice. J Clin Psychiatry. 1980; 41:6–10.
25. Khuon F, Lavelle J. Indochinese versions of the Hopkins symptom checklist-25: a screening instrument for the psychiatric care of refugees. Am J Psychiatry. 1987; 144(4):497–500. [PubMed: 3565621]
26. Bland JM, Altman DG. Cronbach’s alpha. Br Med J. 1997; 314(7080):572. [PubMed: 9055718]
27. Muthén, LK., Muthén, BO. Mplus User’s Guide. The comprehensive modelling program for applied researchers: user’s guide. 7. Muthén & Muthén; Los Angeles: 2012.
28. Hu L-T, Bentler PM. Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification. Psychol Methods. 1998; 3(4):424–453.
29. Samejima, F. Graded response model. Handbook of modern item response theory. Springer; New YorK: 1997. p. 85-100.
30. Baker, FB. The basics of item response theory. ERIC Clearing House on Assessment and Education; Washington, DC: 2001.

31. Cai, L., Du Toit, SHC., Thissen, D. IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling. Scientific Software International; Chicago: 2011.

32. Crane PK, Belle GV, Larson EB. Test bias in a cognitive test: differential item functioning in the CASI. Stat Med. 2004; 23(2):241–256. [PubMed: 14716726]

33. Kojima M, et al. Cross-cultural validation of the Beck depression inventory-II in Japan. Psychiatry Res. 2002; 110(3):291–299. [PubMed: 12127479]

34. Fabre LF, Smith LC. The effect of major depression on sexual function in women. J Sex Med. 2012; 9(1):231–239. [PubMed: 21883948]

35. Kennedy SH, Rizvi S. Sexual dysfunction, depression, and the impact of antidepressants. J Clin Psychopharmacol. 2009; 29(2):157–164. [PubMed: 19512977]

36. Kitamura T, et al. Factor structure of the Zung self-rating depression scale in first-year university students in Japan. Psychiatry Res. 2004; 128(3):281–287. [PubMed: 15541786]

37. Nock MK, et al. Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. Br J Psychiatry. 2008; 192(2):98–105. [PubMed: 18245022]

38. Kessler RC, Bromet EJ. The epidemiology of depression across cultures. Annu Rev Public Health. 2013; 34:119–138. [PubMed: 23514317]

Table 1

Description of data included in analysis

Study setting	Type of study	*N*
Colombia	Screening, validity	1263
Kurdistan Iraq		
Dohuk	Clinical monitoring	294
Erbil/Sulaymaniyah	Clinical monitoring	680
Indonesia	Screening, validity	588
Iraq	Validity	149
Rwanda	Epidemiologic study	368
Thailand	Screening, validity	803
Uganda	Epidemiologic study	587
Total		4732

Table 2

Sample characteristics (*N* = 4732)

	***N* (%)**
Gender	
Male	1778 (37.6)
Female	2939 (62.1)
Missing	15 (0.3)
Age	
16–24	834 (17.6)
25–44	2200 (46.5)
45–66	1353 (28.6)
67–79	247 (5.2)
80+	75 (1.6)
Missing	18 (0.4)

Table 3

Item discrimination parameters (a) and standard errors overall ($N = 4732$) and by setting

	Overall	Colombia	Indonesia[a]	Dohuk	Rwanda	Thailand	S. Iraq	Uganda	Erbil/ Sulaymaniyah
Hopeless	1.75 (0.05)	1.79 (0.58)	1.92 (0.26)	1.94 (0.39)	1.72 (0.26)	1.93 (0.19)	2.74 (0.44)*	1.44 (0.18)	2.19 (0.31)
Crying	1.50 (0.05)	1.56 (0.49)	1.26 (0.17)	0.69 (0.28)*	1.47 (0.25)	1.38 (0.15)*	1.43 (0.30)	1.34 (0.17)	1.33 (0.21)*
Sad	2.09 (0.06)	2.26 (0.66)	1.71 (0.22)	2.29 (0.57)	2.83 (0.39)	2.79 (0.29)*	2.98 (0.46)	1.86 (0.23)	2.68 (0.30)
Lonely	1.89 (0.06)	1.98 (0.57)	2.22 (0.29)	2.01 (0.46)	2.24 (0.31)	2.14 (0.21)	2.81 (0.42)	1.58 (0.19)	1.76 (0.24)*
Lost sex	0.97 (0.04)	1.01 (0.28)		0.91 (0.33)	0.74 (0.16)	1.17 (0.15)	1.00 (0.25)	0.85 (0.13)	1.26 (0.16)
No interest	1.44 (0.05)	0.98 (0.27)*	1.69 (0.21)	1.76 (0.46)	0.76 (0.18)	1.70 (0.17)	2.70 (0.42)*	1.67 (0.19)	3.10 (0.33)
Low energy	1.28 (0.04)	1.32 (0.35)*	1.41 (0.20)	1.70 (0.46)	1.55 (0.21)	2.01 (0.20)*	2.15 (0.36)	1.36 (0.16)	1.46 (0.20)
Appetite	1.05 (0.04)	1.09 (0.32)	1.23 (0.18)	1.41 (0.53)	1.60 (0.22)	1.05 (0.13)	1.44 (0.30)	1.11 (0.16)	1.33 (0.19)
Sleep	1.29 (0.04)	1.30 (0.39)*	1.24 (0.19)	1.67 (0.37)	1.66 (0.23)	2.23 (0.22)*	1.91 (0.35)	1.46 (0.18)	1.17 (0.19)*
Suicide	1.12 (0.04)	1.41 (0.41)	2.49 (0.47)	1.98 (0.32)	2.27 (0.52)	0.31 (0.10)*	1.59 (0.35)	1.03 (0.22)	1.60 (0.24)
Trapped	1.14 (0.04)	1.46 (0.43)	1.52 (0.21)	2.08 (0.33)	2.68 (0.39)*	2.01 (0.19)*	2.49 (0.38)*	1.26 (0.16)	1.83 (0.27)
Worry	1.57 (0.05)	1.21 (0.38)*		1.62 (0.53)*	1.35 (0.21)	2.83 (0.26)*	1.70 (0.29)	1.52 (0.18)	1.52 (0.22)*
Self-blame	1.22 (0.04)	1.10 (0.33)*	1.29 (0.19)	1.48 (0.33)	0.98 (0.19)	1.66 (0.16)	1.77 (0.32)	0.63 (0.12)*	1.80 (0.25)
Effort	1.47 (0.04)	0.88 (0.27)*	1.75 (0.24)	1.67 (0.42)	2.08 (0.27)	2.03 (0.19)*	2.48 (0.42)*	1.88 (0.20)	1.90 (0.24)
Worthlessness	1.47 (0.05)	1.48 (0.43)	2.84 (0.29)*	1.28 (0.31)	2.43 (0.32)	1.63 (0.17)	2.65 (0.40)*	1.58 (0.19)	1.89 (0.26)

Item discrimination parameters estimated using maximum likelihood

[*] Identifies statistically significant non-uniform DIF ($p < 0.001$) when comparing the setting in each column to all other settings in the combined dataset

[a] Items left blank were not included in the Indonesia dataset

Table 4

Item location parameters (b_1, b_2, b_3) and standard errors overall ($N = 4732$) and by settinga

Threshold	Overall	Colombia	Indonesia[a]	Dohuk	Rwanda	Thailand	S. Iraq	Uganda	Erbil/ Sulaymaniyah
D1.Hopeless, b_1	−0.14 (0.02)	−0.79 (0.39)	0.43 (0.09)	−1.30 (0.31)	−0.53 (0.12)	−0.90 (0.11)*	−0.62 (0.12)	−0.39 (0.11)	−0.76 (0.11)
D1.Hopeless, b_2	0.54 (0.03)	−0.15 (0.19)	0.79 (0.11)	0.07 (0.14)	0.24 (0.14)	−0.20 (0.08)*	−0.09 (0.10)	0.14 (0.11)	−0.03 (0.07)
D1.Hopeless, b_3	1.45 (0.04)	0.59 (0.08)	1.32 (0.17)	1.18 (0.16)	1.56 (0.31)	0.85 (0.10)	0.72 (0.12)	0.85 (0.16)	1.10 (0.16)
D2.Crying, b_1	−0.39 (0.03)	−1.10 (0.48)	−0.56 (0.11)	−3.67 (1.42)*	0.07 (0.14)	−0.94 (0.12)	−0.85 (0.17)	−0.39 (0.10)	−1.40 (0.20)
D2.Crying, b_2	0.38 (0.03)	−0.40 (0.27)	0.11 (0.10)	−0.57 (0.26)*	0.86 (0.23)	0.13 (0.08)	0.23 (0.16)	0.13 (0.11)	−0.36 (0.09)
D2.Crying, b_3	1.42 (0.04)	0.46 (0.06)	0.88 (0.16)	2.55 (1.06)	2.47 (0.54)	1.70 (0.21)	0.97 (0.25)	0.87 (0.16)	1.35 (0.24)
D3.Sad, b_1	−1.02 (0.03)	−1.62 (0.63)	−1.68 (0.21)	−1.10 (0.33)	−1.41 (0.16)	−1.37 (0.14)	−1.02 (0.14)	−1.34 (0.16)	−1.38 (0.16)
D3.Sad, b_2	−0.10 (0.02)	−0.71 (0.36)	−0.95 (0.14)*	−0.03 (0.17)	−0.45 (0.09)	−0.52 (0.08)	0.09 (0.09)	−0.55 (0.10)	−0.59 (0.09)
D3.Sad, b_3	0.94 (0.03)	0.07 (0.14)	0.02 (0.09)*	1.23 (0.17)	0.86 (0.15)	0.74 (0.08)	0.84 (0.13)	0.20 (0.09)	0.67 (0.09)
D4.Lonely, b_1	−0.31 (0.02)	−0.95 (0.44)	−0.27 (0.08)	−1.10 (0.33)	−0.74 (0.11)	−0.65 (0.10)	−0.89 (0.13)	−1.00 (0.14)	−0.94 (0.13)
D4.Lonely, b_2	0.38 (0.02)	−0.31 (0.25)	0.12 (0.07)	−0.03 (0.17)	−0.16 (0.10)	−0.00 (0.08)	0.21 (0.10)	−0.32 (0.09)	−0.16 (0.08)
D4.Lonely, b_3	1.34 (0.04)	0.43 (0.06)	0.70 (0.10)	1.23 (0.17)	1.19 (0.22)	0.91 (0.11)	1.16 (0.17)	0.70 (0.12)	1.29 (0.19)
D5.Lost sex, b_1	−0.02 (0.04)	−0.51 (0.32)	–	−1.01 (0.31)	−0.82 (0.17)	0.15 (0.11)	−1.02 (0.26)	−0.59 (0.14)	−0.75 (0.12)
D5.Lost sex, b_2	0.97 (0.05)	0.42 (0.11)	–	0.55 (0.30)	0.42 (0.3)	1.08 (0.19)	0.55 (0.28)	0.15 (0.16)	0.06 (0.09)
D5.Lost sex, b_3	1.97 (0.08)	0.98 (0.15)	–	2.41 (0.90)	2.04 (0.61)	2.59 (0.38)	1.72 (0.50)	1.11 (0.26)	1.22 (0.17)
D6.No interest, b_1	−0.08 (0.03)	−0.41 (0.28)	0.31 (0.09)	−1.63 (0.28)	0.49 (0.34)	−0.89 (0.11)	−1.11 (0.16)	−0.22 (0.10)	−1.04 (0.13)
D6.No interest, b_2	0.85 (0.03)	0.78 (0.12)	0.88 (0.14)	−0.02 (0.16)	2.57 (0.80)	0.04 (0.08)	0.09 (0.10)	0.49 (0.11)	−0.24 (0.07)
D6.No interest, b_3	2.19 (0.06)	2.01 (0.43)	1.93 (0.26)	1.73 (−0.42)	5.54 (1.64)	1.74 (0.20)	1.13 (0.18)	1.53 (0.20)	0.83 (0.11)
D7.Low energy, b_1	−1.07 (0.04)	−1.43 (0.56)	−2.02 (0.26)*	−2.17 (0.45)	−1.83 (0.21)	−0.83 (0.11)	−1.53 (0.21)	−1.24 (0.16)	−2.11 (0.27)
D7.Low energy, b_2	0.14 (0.03)	−0.24 (0.23)	−1.13 (0.16)*	−0.26 (0.13)	−0.72 (0.13)	0.06 (0.08)	−0.08 (0.11)	−0.53 (0.11)	−0.51 (0.10)
D7.Low energy, b_3	1.56 (0.05)	0.98 (0.15)	−0.14 (0.10)*	1.29 (0.38)	0.89 (0.21)	1.35 (0.15)	1.01 (0.19)	0.56 (0.13)	1.40 (0.21)*
D8.Appetite, b_1	−0.58 (0.04)	−0.84 (0.41)	−1.84 (0.24)*	−1.73 (0.47)	−1.08 (0.15)	−1.24 (0.15)	−1.31 (0.22)	0.23 (0.13)*	−1.09 (0.15)
D8.Appetite, b_2	0.68 (0.04)	0.22 (0.13)	−1.09 (0.16)	0.21 (0.20)	−0.03 (0.13)	0.62 (0.14)	0.14 (0.15)	0.95 (0.19)	0.09 (0.09)
D8.Appetite, b_3	2.22 (0.08)	1.39 (0.28)	0.42 (0.13)	1.54 (0.58)	1.43 (0.28)	2.84 (0.39)	1.64 (0.37)	1.89 (0.31)	1.83 (0.27)*
D9.Sleep, b_1	−0.63 (0.03)	−0.87 (0.41)	−2.01 (0.26)*	−1.63 (0.29)	−1.17 (0.15)	−0.80 (0.10)	−1.55 (0.23)	−0.48 (0.11)	−1.57 (0.22)

Threshold	Overall	Colombia	Indonesia[a]	Dohuk	Rwanda	Thailand	S. Iraq	Uganda	Erbil/ Sulaymaniyah
D9.Sleep, b_2	0.14 (0.03)	−0.22 (0.22)	−1.30 (0.18) *	−0.18 (0.15)	−0.33 (0.13)	−0.42 (0.08)	−0.30 (0.12)	0.05 (0.10)	−0.44 (0.10)
D9.Sleep, b_3	1.40 (0.05)	0.73 (0.11)	−0.02 (0.10) *	1.07 (0.32)	1.43 (0.28)	0.86 (0.10)	0.77 (0.18)	0.73 (0.14)	1.24 (0.22)
D10.Suicide, b_1	1.38 (0.05)	0.87 (0.14)	1.46 (0.21)	−0.68 (0.20) *	0.85 (0.23)	2.74 (1.08)	0.83 (0.20)	2.36 (0.53)	0.37 (0.10)
D10.Suicide, b_2	2.33 (0.09)	1.45 (0.30)	1.75 (0.25)	0.39 (0.18) *	1.12 (0.29)	9.12 (3.06)	1.73 (0.38)	2.79 (0.61)	1.12 (0.18)
D10.Suicide, b_3	3.58 (0.15)	2.22 (0.51)	2.14 (0.31)	1.80 (0.30)	2.30 (0.64)	13.38 (4.45)	2.57 (0.61)	3.28 (0.72)	2.32 (0.34)
D11.Trapped, b_1	−0.11 (0.03)	0.16 (0.12) *	−0.23 (0.09)	−0.87 (0.20)	−0.44 (0.10)	−1.62 (0.16) *	−0.60 (0.12)	−0.94 (0.13)	−0.52 (0.10)
D11.Trapped, b_2	0.96 (0.04)	0.81 (0.13) *	0.45 (0.11)	0.29 (0.18)	0.07 (0.09)	−0.32 (0.08) *	0.28 (0.11)	−0.23 (0.10)	0.29 (0.09)
D11.Trapped, b_3	2.35 (0.08)	1.60 (0.36)	1.06 (0.17)	1.41 (0.25)	1.27 (0.23)	1.19 (0.13)	1.20 (0.19)	0.72 (0.15) *	1.61 (0.23)
D12.Worry, b_1	−1.05 (0.04)	−2.30 (0.83)	–	−1.53 (0.33)	−0.87 (0.14)	−1.22 (0.13)	−1.56 (0.21)	−1.64 (0.19)	−1.59 (0.21)
D12.Worry, b_2	−0.14 (0.03)	−1.21 (0.50)	–	−0.19 (0.15)	0.19 (0.16)	−0.59 (0.09) *	−0.36 (0.12)	−0.66 (0.11)	−0.66 (0.11)
D12.Worry, b_3	1.07 (0.04)	−0.06 (0.16) *	–	1.86 (0.47)	2.29 (0.48)	0.54 (0.07)	0.69 (0.17)	0.42 (0.11)	1.04 (0.18)
D13.Self-blame, b_1	−0.11 (0.03)	−0.50 (0.30)	0.44 (0.12)	−2.04 (0.37)	−0.04 (0.20)	−1.04 (0.12) *	−1.70 (0.24)	0.95 (0.31)	−1.00 (0.13)
D13.Self-blame, b_2	0.93 (0.04)	0.45 (0.09)	1.13 (0.20)	0.07 (0.19)	1.76 (0.49)	−0.04 (0.08)	−0.25 (0.12)	2.62 (0.61)	−0.03 (0.08)
D13.Self-blame, b_3	2.40 (0.08)	1.70 (0.40)	2.33 (0.37)	1.86 (0.47)	3.27 (0.81)	1.49 (0.17)	0.79 (0.19)	4.11 (0.91)	1.49 (0.22)
D14.Effort, b_1	−0.81 (0.03)	−2.69 (0.94) *	−1.33 (0.17)	−1.71 (0.31)	−1.23 (0.15)	−0.62 (0.09)	−0.85 (0.14)	−1.14 (0.14)	−1.25 (0.16)
D14.Effort, b_2	0.09 (0.03)	−1.34 (0.53) *	−0.72 (0.11) *	−0.17 (0.16)	−0.36 (0.11)	0.01 (0.08)	0.14 (0.11)	−0.47 (0.10)	−0.24 (0.08)
D14.Effort, b_3	1.23 (0.04)	0.14 (0.13) *	0.12 (0.09) *	1.45 (0.40)	1.14 (0.22)	1.29 (0.14)	1.17 (0.20)	0.39 (0.10)	1.20 (0.18) *
D15.Worthless, b_1	0.38 (0.03)	0.32 (0.08)	0.25 (0.07)	−1.18 (0.24)	−0.43 (0.11)	−0.20 (0.08)	−0.57 (0.11)	−0.43 (0.10)	−0.48 (0.09)
D15.Worthless, b_2	1.16 (0.04)	0.96 (0.17)	0.68 (0.09)	0.32 (0.24)	0.13 (0.11)	0.63 (0.11)	0.04 (0.10)	0.25 (0.10)	0.29 (0.09)
D15.Worthless, b_3	2.21 (0.07)	1.64 (0.37)	1.19 (0.14)	1.85 (0.52)	1.23 (0.22)	1.85 (0.22)	0.79 (0.15)	1.12 (0.16)	1.59 (0.22)

Item difficulty parameters estimated using maximum likelihood; b_1 = difficulty parameter for an item response of <1; b_2 = difficulty parameter for an item response of <2; b_3 = difficulty parameter for an item response of < 3

[*] Identifies statistically significant uniform DIF ($p < 0.001$) comparing the setting in each column to all other settings in the combined dataset

[a] Items left blank were not included in the Indonesia dataset

Table 5

Difference in latent meant scores of depression accounting for/not accounting for DIF

	Difference in means β (SE)
Colombia	
Accounting for DIF	0.22 (0.02)
Not accounting for DIF	0.22 (0.02)
Difference	0.00
Indonesia	
Accounting for DIF	−0.28 (0.02)
Not accounting for DIF	−0.07 (0.02)
Difference	0.21 *
Dohuk	
Accounting for DIF	0.17 (0.02)
Not accounting for DIF	0.18 (0.02)
Difference	0.01
Rwanda	
Accounting for DIF	−0.26 (0.02)
Not accounting for DIF	−0.27 (0.02)
Difference	0.01
Thailand	
Accounting for DIF	−0.14 (0.02)
Not accounting for DIF	−0.17 (0.02)
Difference	0.03
S. Iraq	
Accounting for DIF	0.01 (0.02)
Not accounting for DIF	0.01 (0.02)
Difference	0.00
Uganda	
Accounting for DIF	−0.12 (0.01)
Not accounting for DIF	−0.14 (0.01)
Difference	0.02
Erbil/Sulaymaniyah	
Accounting for DIF	0.13 (0.02)
Not accounting for DIF	0.13 (0.02)
Difference	0.00

[*] Difference is statistically significant $p < 0.05$