global mental health

OTHER

ORIGINAL RESEARCH PAPER

Development and cross-cultural testing of the International Depression Symptom Scale (IDSS): a measurement instrument designed to represent global presentations of depression

E. E. Haroz[1]*, J. Bass[1], C. Lee[2], S. S. Oo[3], K. Lin[4], B. Kohrt[5], L. Michalopolous[6], A. J. Nguyen[7] and P. Bolton[2]

[1] *Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, 624 N. Broadway, Baltimore, MD, USA*
[2] *Department of International Health, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., Baltimore, MD, USA*
[3] *Aung Clinic, Yangon, Myanmar*
[4] *Thu Kha Nwe Specialist Clinic, Yangon, Myanmar*
[5] *Duke University, Duke Global Health Institute & Department of Psychiatry and Behavioral Sciences, Durham, NC, USA*
[6] *School of Social Work, Columbia University, 1255 Amsterdam Avenue, New York, NY, USA*
[7] *University of Virginia Curry School of Education, Charlottesville, VA, USA*

Global Mental Health (2017), **4**, e17, page 1 of 13. doi:10.1017/gmh.2017.16

Background Self-report measurement instruments are commonly used to screen for mental health disorders in Low and Middle-Income Countries (LMIC). The Western origins of most depression instruments may constitute a bias when used globally. Western measures based on the DSM, do not fully capture the expression of depression globally. We developed a self-report scale design to address this limitation, the International Depression Symptom Scale-General version (IDSS-G), based on empirical evidence of the signs and symptoms of depression reported across cultures. This paper describes the rationale and process of its development and the results of an initial test among a non-Western population.

Methods We evaluated internal consistency reliability, test–retest reliability and inter-rater reliability of the IDSS-G in a sample $N = 147$ male and female attendees of primary health clinics in Yangon, Myanmar. For criterion validity, IDSS-G scores were compared with diagnosis by local psychiatrists using the Structured Clinical Interview for DSM (SCID). Construct validity was evaluated by investigating associations between the IDSS-G and the Patient Health Questionnaire (PHQ), impaired function, and suicidal ideation.

Results The IDSS-G showed high internal consistency reliability ($\alpha = 0.92$), test–retest reliability ($r = 0.87$), and inter-rater reliability ($ICC = 0.90$). Strong correlations between the IDSS-G and PHQ-9, functioning, and suicidal ideation supported construct validity. Criterion validity was supported for use of the IDSS-G to identify people with a SCID diagnosed depressive disorder (major depression/dysthymia). The IDSS-G also demonstrated incremental validity by predicting functional impairment beyond that predicted by the PHQ-9. Results suggest that the IDSS-G accurately assesses depression in this population. Future testing in other populations will follow.

Received 7 April 2017; Revised 21 June 2017; Accepted 29 June 2017

Key words: Assessment, depression, global mental health, validation.

Introduction

Task-sharing methods whereby non-specialists are trained to treat mental health disorders (Bolton *et al.*

* Address for correspondence: E. E. Haroz, Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, 624 N. Broadway, Baltimore, MD 21205, USA.
(Email: EHaroz1@jhu.edu)

2003; Patel *et al.* 2009), are viewed as a critical strategy to reduce the 'treatment gap' in global mental health. Treatment delivered by non-specialists has been shown to be effective for improving depression and other common mental health disorders (Bass *et al.* 2013; van Ginneken *et al.* 2013; Bolton *et al.* 2014).

A necessary first step in treating depression is identifying those individuals in need and who will likely benefit from treatment. As non-specialist workers do not have extensive training in recognizing the signs and symptoms of mental disorder, they typically rely on self-report instruments to conduct assessments. Many of these instruments were developed based on Western, and/or clinical populations (Hamilton, 1960; Beck *et al.* 1961; Zung *et al.* 1965; Radloff, 1977; Mulrow *et al.* 1995), and validity results have varied when using them in new settings and populations (Ali *et al.* 2016). Many of these instruments reflect DSM diagnostic criteria, a group of symptoms that appears to inadequately capture how depression is experienced around the world (Haroz *et al.* 2017). Some of these instruments have been successfully adapted (e.g. additional items, colloquial translations) and tested (Adewuya *et al.* 2006; Patel *et al.* 2008; Ghimire *et al.* 2013; Haroz *et al.* 2014). Other researchers have developed locally-specific screening instruments based on qualitative research in a particular context (Patel *et al.* 1997; Phan *et al.* 2004; Miller *et al.* 2006).

Both adapted and locally-specific instruments often perform well, but these approaches have limited generalizability (Ali *et al.* 2016). Moreover, instrument development and/or adaptation is a cumbersome process (Hollifield, 2002) that is not possible for many employers of non-specialist mental health care workers (e.g. non-governmental or community-based organizations) due to a lack of resources.

One option for addressing the limitations of existing instruments and processes of adaptation, is using an instrument that is less biased toward Western populations and more accurately reflects common ways of expressing depression around the world. An instrument that is based on empirical evidence of the commonalities in depression presentation globally should be more generalizable, and particularly useful for situations where local adaptation is not possible. We created such an instrument by empirically investigating symptoms associated with depression from a variety of populations around the world. The resulting instrument is the International Depression Symptom Scale that includes a General version (IDSS-G), which can be augmented with locally relevant symptoms (IDSS-L). In creating the IDSS we did not seek to create another depression instrument to diagnose DSM defined depression, but rather an instrument that better captures the experience of depression worldwide and reflects the constellation of symptoms associated with impaired functioning.

Testing the IDSS

We tested the reliability, validity and clinical utility of the IDSS in a community sample of adults in Yangon, Myanmar. To evaluate whether the IDSS performed better than a commonly used standard screening instrument (translated but not adapted), we compared the IDSS with the Patient Health Questionnaire-9 (PHQ-9) on ability to predict impaired functioning (i.e. *incremental validity*). This testing is the first of a series of tests in different parts of the world with the goal of developing a reliable and valid instrument to measure depression that better reflects global commonalities in symptoms across and within different cultural contexts.

Methods

Development of the IDSS

The first step in the development of the IDSS involved a systematic review of qualitative research to identify common depression symptoms across geographic regions, gender, and contexts (Haroz *et al*. 2017). The second step involved a quantitative analysis using Item Response Theory (IRT) of the 15 symptom HSCL-25 depression scale administered in eight distinct cultural settings (Haroz *et al*. 2016). The new instrument combined symptoms that were common across multiple regions identified during the literature review with the best performing symptoms from the quantitative analysis. An expert panel of researchers and practitioners from the fields of global mental health, anthropology, psychiatric epidemiology, and psychiatry reviewed the draft instrument and additional revisions were made based on their feedback, which included adding symptoms from the DSM-5. These experts included two psychiatrists with extensive experience in global mental health, the former chair of the Department of Mental Health at Johns Hopkins Bloomberg School of Public Health who is a sociologist by training and has worked on developing measures of depression, two medical anthropologists who have been heavily involved in global mental health work for over 20 years, and a clinical psychology professor who has done extensive work with displaced populations.

The IDSS is a modular instrument with 29 items in the global measure (IDSS-G) and additional items added, based on qualitative research, when used in different settings (IDDS local; IDSS-L). The 29 items on the IDSS-G and the development process that supported each symptom's inclusion are provided in Table 1.

https://doi.org/10.1017/gmh.2017.16 Published online by Cambridge University Press

Table 1. *Source of the supporting evidence for each symptom on the IDSS*

Item	Qualitative review[a]	IRT analysis[b]	DSM-5
D01 sad	◆	◆	◆
D02 no interest	◆	◆	◆
D03 crying	◆	◆	
D04 hopeless	◆	◆	
D05 lonely	◆	◆	
D06 social withdrawal	◆		
D07 tired/fatigue	◆	◆	◆
D08 weigh too little	◆	◆	◆
D09 weigh too much	◆	◆	◆
D10 increased appetite	◆	◆	◆
D11 sleep problems	◆	◆	◆
D12 feeling trapped		◆	
D13 worry	◆	◆	
D14 worthless	◆	◆	
D15 headaches	◆		
D16 stomachaches	◆		
D17 general aches and pains	◆		
D18 anger	◆		
D19 thinking too much	◆		
D20 confused	◆		◆
D21 heart weakness	◆		
D22 palpitations	◆		
D23 heavy heart	◆		
D24 heart pressure	◆		
D25 heart pain	◆		
D26 psychomotor			◆
D27 concentration			◆
D28 imp function			◆
D29 suicide			◆

[a] Haroz *et al.* (2017).
[b] Haroz *et al.* (2016).

Study procedures

All participants completed the assessment battery and were evaluated by a local psychiatrist. To assess test–retest and inter-rater reliability, $n = 54$ randomly selected participants were interviewed a second time. The same interviewer who administered the initial interview ($n = 24$) or a different interviewer ($n = 30$) conducted the re-interview.

Psychiatric evaluation.

Local psychiatrists conducted evaluations for each participant within 2–5 days of the initial assessment. Diagnoses from these interviews were treated as the primary criterion for validity analysis. The first $n = 40$ study participants were interviewed by psychiatrists in pairs (with independent ratings) in order to establish inter-rater reliability. The remainder of participants were interviewed by psychiatrists working individually.

Participants

Study participants were recruited from two medical clinics in Yangon, Myanmar. We purposively sampled participants from these clinics as there were reported high rates of psychiatric disorders (~30–40%), indicating a high likelihood we would include both participants with mental disorders, and some without. To be included in the study, participants had to be a clinic patient and over the age of 18. Exclusion criteria consisted of active psychosis or the presence of a major developmental delay. All participants provided informed verbal consent. The study was approved by the Johns Hopkins Internal Review Board (IRB #6011) and the Ethics Review Committee of the Department of Medical Research (Lower Myanmar).

Measures

The IDSS-G is a 29-item self-report measure. Participants were asked to indicate how often in the last 2 weeks they had experienced each symptom in the measure. Responses options ranged from 0 'none of the time' to 3 'almost all the time.'

PHQ-9 (Kroenke *et al.* 2001) is a nine-item self-report measure that asks participants how often in the past 2 weeks the symptom bothered him/her. Response options ranged from 0 'not at all' to 3 'nearly every day.' The PHQ-9 is a commonly used measure of depression and has been found to be valid in a variety of low-resource settings (Lotrakul *et al.* 2008; Marc *et al.* 2014; Zhong *et al.* 2014), although it had not been previously tested in Myanmar.

Structured Diagnostic Interview for DSM-IV (SCID; Spitzer *et al.* 1995) is a semi-structured interview designed for use by trained mental health professionals to facilitate diagnosing DSM Axis I disorders (American Psychiatric Association, 2000). For the current study only major depressive disorder (MDD), dysthymia and generalized anxiety disorder (GAD) were evaluated. Diagnosis of GAD was included due to its high co-morbidity with depression (Almeida *et al.* 2012) and overlap in diagnostic criteria (American Psychiatric Association, 2013).

A *Local measure of functional impairment* was previously developed and validated among Burmese refugees displaced in Thailand (Haroz *et al.* 2014). The measure includes tasks that men (16 items) and women (23 items) do to care for themselves, their families and their communities. Participants were asked how much difficulty he/she had in the last 4 weeks compared with other men/women of similar age.

Response options ranged from 0 ‘no difficulty’ to 4 ‘often cannot do.’

Functional impairment in this study was used as a validity criterion for two reasons. First, it is often the major domain that is adversely affected among people suffering from mental health symptoms and provides a way to assess the presence of symptoms and their impact on people’s daily lives. Second, assessing the association of symptoms with impaired functioning broadened our criteria beyond DSM diagnosis.

Translation

The assessment instruments were translated and back-translated by the local study coordinator and a local psychiatrist. Review of all translations took place as part of training the interviewers and psychiatrists. Each item was reviewed during each of the trainings. When minor problems with wording or phrasing arose, we discussed as a group, and settled on wording by consensus. No major problems with translations were identified during the trainings. In addition, a subset of participants ($n = 30$ men and $n = 30$ women) was asked to complete a cognitive interview to assess face validity and the comprehension of select items from the IDSS-G. For each symptom question, participants were asked: (1) *Please describe the meaning of this question in your own words*; (2) *Is there any part of this question you don’t understand or that does not make sense?*; (3) *Can you tell me what thought you had when deciding your answer choice*?; and (4) *Was this question easy or difficult to answer*? Sixteen items on the IDSS-G had been previously tested in a similar population (see Haroz *et al.* 2014) and were not included in the cognitive interview. The remaining 13 items from the IDSS-G were part of the cognitive interviews.

Interviewers and psychiatrists

Eight local interviewers administered the full assessment using mobile devices and facilitated the cognitive interviewing. Interviewers were people from the community with previous experience doing data collection. Interviewers were trained in study procedures, research ethics, and a safety protocol, during a 3-day training prior to data collection. The interviewers administered the IDSS verbally using tablets to record participants’ responses. This was done as literacy rates were inconsistent and we wanted a uniform interview procedure for all participants involved in the study.

Four local psychiatrists conducted the clinical interviews using the SCID. All psychiatrists had medical degrees obtained from medical schools in Myanmar. Three had been practicing for more than 5 years, while the fourth was finishing residency. All psychiatrists attended a 3-day SCID training prior to data collection. Each psychiatrist was given a handout with the DSM-IV criteria for the three study disorders (MDD, dysthymia, and/or GAD). They were instructed to use the SCID to inform their clinical judgment as to diagnosis.

Analysis

Average summary scores for the IDSS-G, PHQ-9 and functional impairment measures were generated. For the IDSS-G, two items were not included in summary scores: ‘difficulty doing your usual activities at home or work’ and ‘thoughts of wanting to kill yourself.’ These items were included in the instrument to assess severity and safety risk. All analyses were done using STATA-13 (StataCorp, 2013) and Mplus 7.3 (Muthén & Muthén, 1998–2012).

Reliability

We examined: (1) Exploratory Factor Analysis (EFA) with geomin rotation, (2) internal consistency reliability, and (3) test–retest and inter-rater reliability for the IDSS-G. The EFA examined factor loadings and item uniqueness. Cronbach’s alpha (α) (Cronbach, 1951) was used for internal consistency reliability. Pearson’s correlation coefficients (r) were calculated for test–retest reliability. Correlations of $|0.7|$ or above are considered very strong, $|0.4|$ to $|0.69|$ strong, $|0.3|$ to $|0.39|$ moderate, $|0.2|$ to $|0.29|$ weak, and less than $|0.2|$ are considered negligible (Cohen, 1988). Inter-rater reliability was assessed using intra-class correlation (ICC) by comparing scores from the first interview to scores on re-interview (done by a different interviewer). ICCs >0.75 are considered excellent; 0.40–0.75 fair to good; and <0.40 poor (Fleiss, 1986).

To establish the reliability of psychiatrist diagnosis, inter-rater reliability between pairs was calculated using a Kappa statistic. A Kappa of <0 indicates less than chance agreement; 0.01–0.20 slight agreement; 0.21–0.40 fair agreement; 0.41–0.60 moderate agreement; 0.61–0.80 substantial agreement; and 0.81–0.99 indicates almost perfect agreement (Viera & Garrett, 2005).

Validity

We examined construct, criterion, and incremental validity. Construct validity is defined as the degree to which a scale measures the theoretical construct that it was designed to measure and is correlated to other related constructs. Criterion validity is defined as the association of a scale to a criterion variable (i.e. psychiatric diagnosis and functional impairment) (Allen & Yen, 2002). Incremental validity refers to the ability

https://doi.org/10.1017/gmh.2017.16 Published online by Cambridge University Press

of a measure to increase predictive ability beyond another measure (Sackett & Lievens, 2008).

For construct validity, we use Pearson's correlation coefficients (r) and Spearman's correlation coefficients (ρ), to assess the strength of relationships between the IDSS-G and age, gender, functioning, PHQ-9, and the single functional impairment and suicidal ideation items. Based on evidence in the literature, we hypothesized that higher scores on the IDSS-G would be associated with increasing age (Jorm, 2000; Kessler *et al.* 2003; Bromet *et al.* 2011); female gender (Nolen-Hoeksema *et al.* 1999; Bromet *et al.* 2011); greater impairments in functioning (Ormel *et al.* 2008; Kessler & Bromet, 2013), and suicidal ideation (Nock *et al.* 2008). As both the IDSS-G and the PHQ-9 measure depression, we expected that scores on these measures would be highly correlated. Evidence for these associations would support construct validity.

For criterion validity, scores on the IDSS-G of participants diagnosed with a SCID disorder were compared with those without a disorder. This was followed by comparing any depressive disorder to no disorder. Criterion validity would be supported if IDSS-G scores were substantially and significantly higher among participants with any diagnosis and/or a depression disorder (depression/dysthymia) compared to those without a disorder. Determination of whether the difference of means between diagnostic categories was statistically significant was done using logistic regression.

Incremental validity

Incremental validity was assessed using a series of linear regression models in which variables were added stepwise to predict functional impairment. Model 1 examined the impact of age. In model 2, suicidal ideation was also added, followed by inclusion of the PHQ-9 in model 3, and addition of the IDSS-G in model 4. Incremental validity would be supported if scores on the IDSS-G significantly predicted functional impairment ($p < 0.05$), above and beyond the impact of age, suicidal ideation and scores on the PHQ-9, as measured by a statistically significant increase (F test) in the R^2 statistic when comparing model 4 with model 3 (Sackett & Lievens, 2008). We also examined the degree of collinearity between scores on the IDSS-G and PHQ-9 using a variance inflation factor (VIF). VIFs of 5 or greater are usually cause for concern (Craney & Surles, 2002) and indicate that variables are highly collinear.

Clinical utility

Receiver operating curves (ROC) were used to compare the area under the curve (AUC), for the IDSS-G and PHQ-9 across diagnostic comparisons. ROC curves plot the true positive rate (sensitivity) against the false positive rate (1-specificity). An AUC of 0.5 (50% sensitivity and 50% specificity) indicates that the test is of no diagnostic utility, while an AUC of 1.0 (100% sensitivity and 100% specificity) indicates perfect prediction of the criterion. AUC values of 0.50–0.70 indicate low accuracy; 0.70–0.90 moderate accuracy, and above 0.90 high accuracy (Fischer *et al.* 2003). An optimal cut-off point was generated for the IDSS-G based on maximizing sensitivity and specificity (Liu, 2012).

Results

Descriptive statistics

Overall $N = 151$ people were interviewed using the IDSS-G and associated measures; $n = 2$ refused to participate in the SCID evaluation and $n = 2$ had data that were mistakenly erased during uploading; leaving a final analytic sample of $n = 147$. Two-thirds of the participants were women ($n = 95$; 63.8%) and ages ranged from 18 to 81 with a mean age of 47.5.

Average scores on the IDSS-G ranged from 0 to 2.44 with a mean of 0.72 (s.d. = 0.49). Scores on all of the measures were positively skewed, indicating that most participants reported few symptoms and good functioning (Table 2). The positive skew across the sample was most likely a result of our sampling method, which aimed to have both participants who were well-functioning (i.e. no disorder), and participants who were less well.

Based on psychiatrist diagnosis with the SCID, $n = 31$ people met criteria for MDD, $n = 39$ people for dysthymia, and $n = 22$ for GAD (Table 2). Of those with

Table 2. *Mean scores and frequencies for scales on assessment battery and frequency of SCID diagnoses*

Measure	*N*	*M*	Range	s.d.	Skew
IDSS-G	147	0.72	0–2.44	0.49	1.07
PHQ-9	146	0.67	0–3	0.63	1.46
Functioning	147	0.61	0–2.43	0.60	1.08

SCID diagnosis ($N = 147$)	*N* (%)
Any disorder	71 (48.3)
Depression	31 (21.1)
Dysthymia	39 (26.5)
GAD	22 (15.0)
None of these disorders	63 (42.9)
Co-morbidity (two or more)	24 (16.3)

https://doi.org/10.1017/gmh.2017.16 Published online by Cambridge University Press

a comorbidity ($n = 24$), $n = 18$ had MDD and dysthymia; $n = 4$ had MDD and GAD; and $n = 2$ with dysthymia and GAD. One participant was diagnosed with all three disorders (Table 2). A little less than half of the total sample ($n = 63$; 42.9%) had none of the evaluated SCID disorders.

Cognitive interviewing

Most questions were easily understood, with the exception of 'feeling weakness in your heart' and 'feeling as though your heart was heavy' ($n = 15$ and 7 found it difficult to understand, respectively). Most of the items were believed to be related to mental health problems with the exception of the items representing somatic complaints. For example, the majority of people talked about 'stomach pain' being related to medical problems or eating spicy food. Only one person mentioned that stomach pain could come from stress. The item 'other bodily aches and pains' also overwhelmingly was reported to be related to physical and medical issues, with most respondents describing having this symptom after being sick or having a medical issue ($n = 29$), working too much ($n = 15$), or being caused by cold weather ($n = 14$). The meanings of the items 'feeling weakness in your heart,' 'heart palpitations' 'feeling pressure on your heart' and 'pain in your heart' were described as related to medical problems as well.

Reliability results

Factor analysis

We explored one- to five-factor solutions. The three-factor solution was selected as the most appropriate model based on loadings and what made theoretical sense. The majority of items (Table 3) loaded on the first factor, and include symptoms related to depressed mood, social isolation, and cognitive impairment. The items related to appetite and weight loaded on the second factor. The third factor included many of the somatic symptoms such as 'headaches' and all of the heart-related items. Four items do not appear to load on any of the factors and these include: 'tired/fatigue,' 'problems with sleep' and 'stomach aches,' and 'other aches and pains.'

Internal consistency reliability and item analysis

Cronbach's alpha was high for the IDSS-G ($\alpha = 0.92$). Analysis of item-level correlations supported dropping only one item, 'weighing too much,' as the item was negatively correlated with all other items. Alpha for the PHQ-9 was lower ($\alpha = 0.83$) and item analysis did not support the removal of any items.

Table 3. *Factor loadings for items on the IDSS-G*

	F1	F2	F3
D01 sad	0.713*	0.107	0.091
D02 no interest	0.688*	−0.001	−0.034
D03 crying	0.579*	0.287*	0.058
D04 hopeless	0.565*	−0.059	0.248*
D05 lonely	0.748*	0.006	−0.062
D06 social withdrawal	0.745*	0.096	−0.077
D07 tired/fatigue	0.282	0.325*	0.359*
D08 weigh too little	0.351	0.731*	−0.013
D09 weigh too much	0.041	−0.594*	0.167
D10 increased appetite	0.070	0.609*	0.275*
D11 sleep problems	0.276*	0.242*	0.278*
D12 trapped	0.903*	−0.019	−0.063
D13 worry	0.692*	−0.055	0.015
D14 worthless	0.565*	−0.001	0.165
D15 headaches	−0.023	0.238*	0.578*
D16 stomach_aches	−0.254	0.298*	0.344*
D17 other_aches	0.198	0.205	0.209
D18 anger	0.549*	−0.199*	0.109
D19 thinking too much	0.784*	−0.197	0.012
D20 confused	0.843*	−0.042	−0.063
D21 heart_weakness	0.067	0.154	0.543*
D22 palpitations	0.079	0.257*	0.600*
D23 heavy_heart	0.009	−0.059	0.910*
D24 heart_pressure	0.112	−0.014	0.861*
D25 heart_pain	−0.033	0.340*	0.550*
D26 psychomotor	0.608*	0.261*	−0.135
D27 concentration	0.605*	0.060	−0.086
D28 imp function	0.598*	0.084	0.144
D29 suicide	0.674*	0.345*	0.050

*$p < 0.05$.

Test–retest reliability

Re-interviews by the same interviewer were performed within 2–11 days of the initial administration of the IDSS-G (mean = 3.8 days; s.d. = 2.17). Visual inspection of the graph depicting the relationship between IDSS-G scores at the first interview and re-interviews indicated that a linear relationship fit the data well. The correlation between average scores on the first interview with average scores on the re-interview was $r = 0.87$, indicating a strong positive relationship and good test–retest reliability. The PHQ-9 also showed good test–retest reliability ($r = 0.88$).

Inter-rater reliability

On average, re-interviews with different interviewers were done 10.2 days (s.d. = 5.3; *range:* 2–19 days) after the initial administration of the IDSS-G. The average ICC across interviewers for the IDSS-G was *ICC* = 0.90 with a 95% CI of (0.79–0.95), indicating high inter-rater reliability. Inter-rater reliability was lower for the PHQ-9 ($ICC = 0.77$; 95% CI 0.53–0.89). Kappas between each pair of psychiatrists indicated substantial to almost perfect agreement for all diagnosis (range: $\kappa = 0.64$ for no diagnosis *v.* any diagnosis to $\kappa = 1.00$ for GAD *v.* no diagnosis), with the exception of the dysthymia rating in pair 1 for which only fair agreement was achieved ($\kappa = 0.38$).

Validity

Construct validity

Table 4 displays the polychoric correlation matrix for: (1) IDSS-G; (2) age; (3) gender; (3) functional impairment measure; (4) PHQ-9; (5) functional impairment item; and (6) suicidal ideation item. Construct validity was supported by a very strong correlation between the IDSS-G and the PHQ-9 ($r = 0.78$) and strong correlations between the IDSS-G and functional impairment scale ($r = 0.56$), and item ($\rho = 0.65$), and suicidal ideation item ($\rho = 0.65$).

Criterion validity

Average scores on the IDSS-G were higher among all disorder classifications (any disorder: mean = 0.87, S.D. = 0.47; depressive disorder: mean = 0.93, S.D. = 0.49; GAD: mean = 0.73, S.D. = 0.40) compared with participants classified as not having any of the SCID disorders (mean = 0.55, S.D. = 0.43). Logistic regressions indicated statistically significant differences between the mean score on the IDSS-G for participants classified as having any disorder and MDD/dysthymia compared with participants with none of these disorders. Results were similar for the PHQ-9 with higher average scores across disorder classifications compared with those classified as having no disorder (Fig. 1).

Incremental validity

Table 5 presents results from the incremental validity investigation. The final model (model 4) included all variables from model 3, as well as average scores on the IDSS-G. Thirty-four percent of the total variance in impaired functioning was explained by the variables in model 4 (additional 7% explained in model 4 compared with model 3). Results from model 4 indicated that after controlling for age, suicidal ideation and scores on the PHQ-9, every unit increase on the IDSS-G was associated with a 0.47 increase in impaired functioning. In model 4, both the IDSS and the PHQ-9 had VIFs of 2.8 and tolerances of 0.39. Moreover, after adding the IDSS-G, the PHQ-9 was no longer significantly associated with impaired functioning. The F test comparing models 3 and 4 indicated a statistically significant increase in R^2 when the IDSS-G was added to the model ($p = 0.001$), thus supporting incremental validity of the IDSS-G. Reversing the order the variables were added (i.e. adding the IDSS-G to model 3, then the PHQ-9 to model 4), resulted in no change in the variance explained between models 3 and 4, and the IDSS-G remained significant.

Sensitivity analysis. We performed a sensitivity analysis to check how stable our coefficients were in models 3 and 4 when using smaller samples. If collinearity is a problem with the IDSS and PHQ-9 score variables, we would expect increases in S.E. of the estimate, making it hard to reject the null hypothesis that there is no difference between the scores on each scale. Reducing our sample size would amplify this result. To test this, we randomly split our sample into equal groups of $n = 102$ each. We re-ran models 1–4 in these subsamples. In model 3 of the first sample, average scores on the PHQ-9 were significant ($b = 0.32$; $p = 0.014$, with an $R^2 = 0.30$). In model 4, the PHQ-9 became not significant and the IDSS was significant ($b = 0.48$; $p = 0.022$, with an $R^2 = 0.35$) with a 5% increase in R^2. In the second subsample, we found similar results: model 3

Table 4. *Correlations of IDSS-G and other measured variables*

	IDSS-G	Age	Gender	Functioning measure	PHQ-9	Function item	Suicide item
IDSS-G	1.00						
Age	−0.16	1.00					
Gender	0.17	−0.06	1.00				
Functioning measure	0.56*	−0.17*	−0.11	1.00			
PHQ-9	0.78*	−0.18*	0.06	0.50*	1.00		
Functioning item	0.65*	−0.16	−0.05	0.48*	0.62*	1.00	
Suicide item	0.65*	−0.40*	0.09	0.50*	0.56*	0.56*	1.00

*$p < 0.05$.

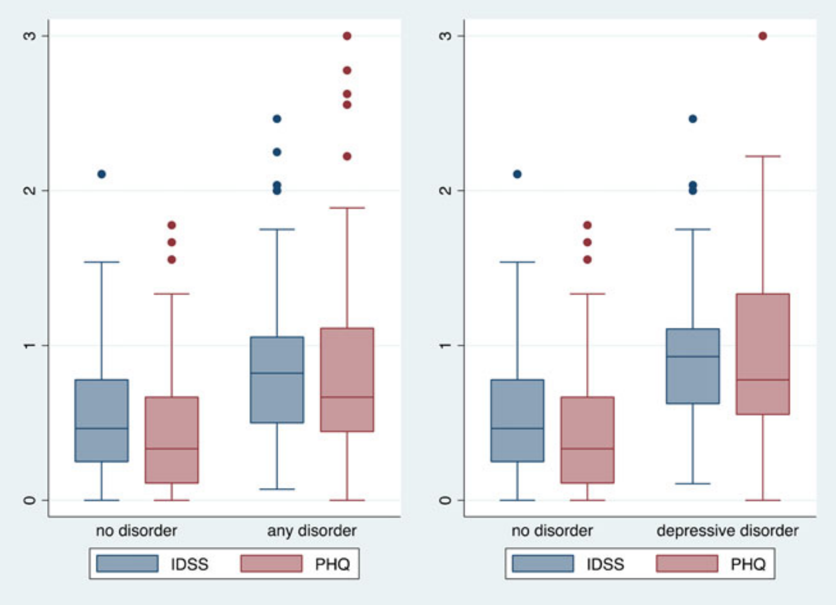

Fig. 1. Box plots of scores on the IDSS and PHQ-9 over SCID diagnoses.

indicated scores on the PHQ-9 were significant ($b = 0.34$; $p = 0.007$, with an $R^2 = 0.26$) and model 4 indicated that scores on the PHQ-9 were not significant and the IDSS was significant with a 5% increase in R^2 ($b = 0.44$; $p = 0.027$, with an $R^2 = 0.31$). Despite smaller sample sizes, parameter estimates and changes in R^2 remained consistent with our overall findings suggesting scores on the PHQ-9 and IDSS are not problematically collinear.

Table 5. *Effects of measured variables on impaired functioning presented as beta coefficients*

Model	β (S.E.)	t
Model 1		
Age	−0.008 (0.01)	−2.21*
Model 2		
Age	−0.004 (0.01)	−1.22
Suicidal ideation[a]	0.71 (0.15)	4.75**
Model 3		
Age	−0.003 (0.01)	−0.92
Suicidal ideation	0.25 (0.16)	1.56
PHQ-9	0.37 (0.08)	4.78**
Model 4		
Age	−0.003 (0.01)	−0.92
Suicidal ideation	0.22 (0.16)	1.40
PHQ-9	0.12 (0.11)	1.21
IDSS-G	0.47 (0.14)	3.35**

[a] For the purposes of the incremental validity testing, the item related to suicide ideation was dichotomized meaning that 0 = none of the time and 1 = some, most and almost all of the time.

*$p < 0.05$.

**$p < 0.001$.

Clinical utility

The IDSS-G had an AUC of 0.72 (95% CI 0.63–0.81) for the comparison on of any disorder to no disorder and an AUC of 0.75 (95% CI 0.66–0.83) when comparing depressive disorders (MDD/dysthymia) to no disorder (Fig. 2). The AUC results for the PHQ-9 were similar indicating moderate accuracy for both scales, across diagnostic comparisons. The PHQ-9 had an AUC of 0.74 (95% CI 0.65–0.82) for the comparison of any disorder to no disorder; and an AUC of 0.74 (95% CI 0.64–0.83) for the comparison of a depressive disorder (MDD/dysthymia) to no disorder (Fig. 2). We identified an optimal cut-off score for the IDSS-G of 0.56, which corresponded to a sensitivity and specificity of 73% and 67% for any *v.* no disorder, and 77% and 67% for a depressive disorder *v.* no disorder.

Discussion

The present study examined the reliability, validity, and clinical utility of the newly created IDSS-G, a self-report instrument developed based on an empirical investigation into the signs and symptoms of depression that occur in populations around the world. Reliability statistics for the IDSS-G were either equivalent or slightly higher than those of the commonly used PHQ-9. High correlation between the two self-report measures, as well as high correlation between

https://doi.org/10.1017/gmh.2017.16 Published online by Cambridge University Press

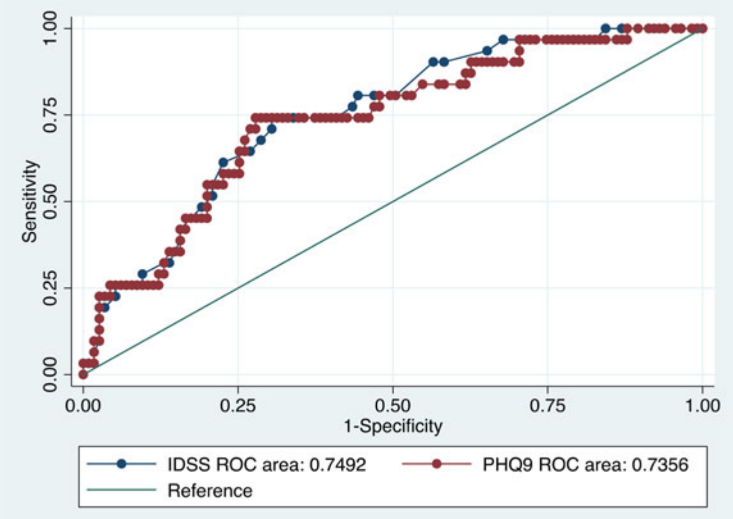

Fig. 2. ROCs for IDSS and PHQ-9 using a diagnosis of depressive disorder.

the IDSS-G and both functional impairment and suicidal ideation, support the instrument's construct validity. Criterion validity was demonstrated by significantly higher IDSS-G scores among individuals assessed by a psychiatrist as having a disorder than among those assessed as having no disorder. Both the IDSS-G and PHQ-9 showed low to moderate diagnostic utility for detecting DSM defined disorders; however, the IDSS-G predicted functional impairment slightly better than the PHQ-9 in this setting.

These results suggest that both the IDSS-G and PHQ-9 are both suitable screening instruments to detect DSM-defined depressive disorders in this context, particularly by non-specialist providers who lack the training to conduct extensive diagnostic interviews. Given that both the SCID and the PHQ-9 are derived from the DSM, it is not surprising that the PHQ-9 would perform well against SCID diagnoses. The IDSS-G, on the other hand, was developed based on cross-cultural descriptions of depression that capture symptoms beyond those included in the DSM. The ability of the IDSS-G to perform similarly well against SCID diagnoses is an important minimal standard supporting its use. However, that the IDSS-G slightly, but significantly, *outperformed* the PHQ-9 at predicting functional impairment – a major outcome of interest in mental health – is a key study finding demonstrating the important contribution of this new instrument over existing instruments. However, given the wide scale use of the PHQ-9 in studies around the world, this finding would need to be replicated in other settings and populations.

While, the IDSS-G slightly, but significantly, predicted higher levels of functional impairment compared with the PHQ-9 in this sample, suggesting that the Western DSM model of depression, as reflected by the PHQ-9, may be inadequate in this population. It may be that the IDSS-G includes elements of disorder that are even more closely related to functioning than the DSM diagnostic criteria or core symptoms of Western depression. This finding would need to be replicated in other settings to justify the use of the IDSS over a shorter and comparable instrument like the PHQ-9.

Local adaptation of existing instruments is typically a critical element of instrument testing in new settings for just this reason – existing models of depression are specific to Western presentations, and instruments based on these models, are likely to miss relevant local expressions of distress. Our findings suggest that some of these missing 'local' symptoms are, in fact, symptoms relevant to depression across multiple cultures that are simply not reflected in the DSM (Haroz *et al*. 2017). In this case, using an instrument developed based on global presentations of depression appears to be more likely to capture locally relevant impairment than a Western measure that reflects DSM diagnostic criteria such as the PHQ-9.

Because the IDSS-G was developed to be a global instrument, we did not conduct preliminary adaptation before testing it a new context. Despite studies showing that adapted Western-based instruments can be reliable and valid in other contexts (Bass *et al.* 2008; Haroz *et al.* 2014; Rasmussen *et al*. 2014; Ali *et al.* 2016), very few studies have actually evaluated the impact of adaptation on scale validity. Jayawickreme *et al*. (2012) conducted a study looking at the incremental validity of Western instruments that incorporated local idioms of distress. In this study, instruments that incorporated local idioms predicted functional impairment above and beyond simple translations of well-established Western measures. The authors stress the importance of doing brief

https://doi.org/10.1017/gmh.2017.16 Published online by Cambridge University Press

ethnographic work to inform scale adaptation (Applied Mental Health Research Group, 2013).

The development of the IDSS-G does not diminish the importance of locally relevant signs and symptoms of distress. When identified, these local indicators are important to include, as they represent common ways of expressing distress in each setting (Keys *et al.* 2012; Kohrt *et al.* 2014), may more saliently communicate illness, be less stigmatizing, and useful for measuring treatment success (Kohrt *et al.* 2014). The incremental validity of the un-adapted IDSS-G over the PHQ-9 suggests that it may be a better measure of depression in non-Western contexts, both as a starting point for local adaptation and when preliminary qualitative work and adaptation is not feasible.

Limitations

The study was conducted in a single site in Myanmar and involved a non-random sample in an urban setting, many of whom had medical illness. It is possible that the sampling strategy explains why we did not identify gender and age differences that would be expected based on the literature (Nolen-Hoeksema *et al*. 1999; Van de Velde *et al*. 2010; Ferrari *et al.* 2013). Likewise, the overrepresentation of participants with a medical illness may help to explain why the somatic items performed differently, even though these items are common globally (Haroz *et al*. 2017). Many of the symptoms included in the IDSS-G are based on English translations of depression symptoms found in qualitative research. However, despite efforts to find accurate translations of symptoms, direct translation often results in overlapping terms that do not necessarily fully capture the original meaning (Nichter, 2010). It is possible that some symptoms on the translated IDSS-G may not fully capture how distress is conveyed locally, pointing to the need for local adaptation when possible. Finally, it remains unclear as to whether the IDSS-G is diagnostically superior to a locally developed measure of depression that incorporates idioms of distress.

Conclusion

Overall, the findings show that the IDSS-G is a reliable and valid depression instrument in Yangon, Myanmar. Incremental validity found the IDSS-G to be a better able to predict impaired functioning than the PHQ-9. We speculate that this may be true for other non-Western populations. Further development and testing of the IDSS-G in multiple populations is necessary to determine whether this cross-culturally derived instrument is preferable to current standard instruments developed in the West.

Acknowledgements

We would like to thank our local research team and our local partners the Assistance Association for Political Prisoners for all their work on this project. The funding for this study was generously provided by the United States Agency for International Development Victims of Torture Fund (VOT): AID-DFD-A-00-08-00308 Dr Haroz was supported by a NIMH T32 Predoctoral training grant (no. MH014592-38) and postdoctoral training grant (no. MH103210) during the course of this research and writing of the manuscript.

Declaration of Interest

None.

Ethical Standards

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.' and 'The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional guides on the care and use of laboratory animals.

References

Adewuya AO, Ola BA, Afolabi OO (2006). Validity of the patient health questionnaire (PHQ-9) as a screening tool for depression amongst Nigerian university students. *Journal of Affective Disorders* **96**, 89–93. doi: 10.1016/j.jad.2006.05.021.

Ali G-C, Ryan G, De Silva MJ, Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B, Simon G, Abas M, Ali G, Nakimuli-Mpungu E, Chibanda D, Hotopf M, Chidgey J, Addington-Hall J, Ly KL, Brown L, Kroenke K, Lunda C, Breena A, Flishera A, Kakumab R, Corrigalla J, Joskaa J, Missinne S, Bracke P, Lindert J, von Ehrenstein O, Priebe S, Mielckd A, Brähler E, Bronstein I, Montgomery P, Vythilingum B, Field S, Kafaar Z, Baron E, Stein D, Sanders L, Abiodun O, Patel V, Kirmayer L, Adewuya A, Ola B, Dada A, Fasoto O, Bolton P, Bass J, Ryder R, Lammers M, Mukaba T, Bolton P, Tsai A, Scott J, Hung K, Zhu J, Matthews L, Psaros C, Reuland D, Cherrington A, Watkins G, Bradford D, Blanco R, Gaynes B, Wu C, Kelley L, Akena D, Joska J, Obuku E, Amos T, Musisi S, Stein D, Glasa A, Lijmerb J, Prinsc M, Bonseld G, Bossuyta P, Sheng L, Wang Y, Zhang L, Chen C, Yao G, Mumford D, Ayub M, Karim R, Izhar N, Asif A, Bavington J, Rumble S, Swartz L, Parry C, Zwarenstein M, Abiodun O, Adewuya A, Ola B, Afolabi O, Zhang Y, Liang W, Chen Z, Zhang H, Zhang J, Weng X, Khamseh M, Baradaran H, Javanbakht A, Mirghorbani M, Yadollahi Z, Malek M, Gelaye B, Williams M, Lemma S,

Deyessa N, Bahretibeb Y, Shibre T, Uwakwe R (2016). Validated screening tools for common mental disorders in low and middle income countries: a systematic review. *PLoS ONE* **11**, e0156939. doi: 10.1371/journal.pone.0156939.

Allen MJ, Yen WM (2002) *Introduction to Measurement Theory*. Waveland Press, Inc: Long Grove, IL.

Almeida OP, Draper B, Pirkis J, Snowdon J, Lautenschlager NT, Byrne G, Sim M, Stocks N, Kerse N, Flicker L, Pfaff JJ (2012). Anxiety, depression, and comorbid anxiety and depression: risk factors and outcome over two years. *International Psychogeriatrics* **24**, 1622–1632.

American Psychiatric Association (2000). *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR*.Author: Washington, DC: Washington, DC.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders*, 5th edn. American Psychiatric Publishing.

Applied Mental Health Research Group (2013). *Design, Implementation, Monitoring, and Evaluation of Mental Health and Psychosocial Assistance Programs for Trauma Survivors in low Resource Countries: a User's Manual for Researchers and Program Implementers*. Johns Hopkins University: Baltimore, MD.

Bass JK, Annan J, McIvor Murray S, Kaysen D, Griffiths S, Cetinoglu T, Wachter K, Murray LK, Bolton PA (2013). Controlled trial of psychotherapy for Congolese survivors of sexual violence. *New England Journal of Medicine* **368**, 2182–2191. doi: 10.1056/NEJMoa1211853.

Bass JK, Ryder RW, Lammers M-C, Mukaba TN, Bolton PA (2008). Post-partum depression in Kinshasa, Democratic Republic of Congo: validation of a concept using a mixed-methods cross-cultural approach. *Tropical Medicine and International Health: TM & IH* **13**, 1534–1542. doi: 10.1111/j.1365-3156.2008.02160.x.

Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J (1961). An inventory for measuring depression. *Archives of General Psychiatry* **4**, 561–571.

Bolton P, Bass J, Neugebauer R, Verdeli H, Clougherty KF, Wickramaratne P, Speelman L, Ndogoni L, Weissman M (2003). Group interpersonal psychotherapy for depression in Rural Uganda. *JAMA* **289**, 3117. doi: 10.1001/jama.289.23.3117.

Bolton P, Lee C, Haroz EE, Murray L, Dorsey S, Robinson C, Ugueto AM, Bass J (2014). A transdiagnostic community-based mental health treatment for comorbid disorders: development and outcomes of a randomized controlled trial among burmese refugees in Thailand. *PLoS Medicine* **11**, e1001757. doi: 10.1371/journal.pmed.1001757.

Bromet E, Andrade LH, Hwang I, Sampson NA, Alonso J, de Girolamo G, de Graaf R, Demyttenaere K, Hu C, Iwata N, Karam AN, Kaur J, Kostyuchenko S, Lépine J-P, Levinson D, Matschinger H, Mora MEM, Browne MO, Posada-Villa J, Viana MC, Williams DR, Kessler RC (2011). Cross-national epidemiology of DSM-IV major depressive episode. *BMC Medicine* **9**, 90. doi: 10.1186/1741-7015-9-90.

Cohen J (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates: Mahwah, NJ.

Craney TA, Surles JG (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering* **14**, 391–403.

Cronbach LJ (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* **16**, 297–334.

Ferrari AJ, Charlson FJ, Norman RE, Patten SB, Freedman G, Murray CJL, Vos T, Whiteford HA, Ferrari A, Somerville A, Baxter A, Norman R, Patten S, Ferrari A, Charlson F, Norman R, Flaxman A, Patten S, Charlson F, Ferrari A, Somerville A, Norman R, Patten S, Ustun T, Ayuso-Mateos J, Chatterji S, Mathers C, Murray C, Prince M, Patel V, Saxena S, Maj M, Maselko J, Murray C, Ezzati M, Flaxman A, Lim S, Lozano R, Murray C, Vos T, Lozano R, Naghavi M, Flaxman A, Lim S, Vos T, Flaxman A, Danaei G, Shibuya K, Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Salomon J, Vos T, Hogan D, Gagnon M, Naghavi M, Vos T, Flaxman A, Naghavi M, Lozano R, Michaud C, Whiteford H, Degenhardt L, Rehm J, Baxter A, Ferrari A, Moher D, Liberati A, Tetzlaff J, Altman D, The P, Kruijshaar M, Barendregt J, Vos T, de Graaf R, Spijker J, Moffitt T, Caspi A, Taylor A, Kokaua J, Milne B, Simon G, VonKorff M, Susser E, Shrout P, Baxter A, Page A, Whiteford H, Vos T, Haby M, Barendregt J, Kruijshaar M, Corry J, Barendregt J, Van Oortmarssen G, Vos T, Murray C, Baxter A, Charlson F, Somerville A, Whiteford H, Charlson F, Stapelberg N, Baxter A, Whiteford H, Li Z, Page A, Martin G, Taylor R, Eaton W, Kalaydjian A, Scharfstein D, Mezuk B, Ding Y, Fu T, Lee C, Gunnell D, Lee W, Cheng A, Charlson F, Steel Z, Degenhardt L, Chey T, Silove D, Jessop E, Murray C, Richards M, Newton J, Fenton L, McKnight P, Kashdan T, Sheehan K, Sheehan D, Sheehan D, Harnett-Sheehan K, Raj B, Kessler R, Berglund P, Demler O, Jin R, Koretz D (2013). Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. *PLoS Medicine* **10**, e1001547. doi: 10.1371/journal.pmed.1001547.

Fischer JE, Bachmann LM, Jaeschke R (2003). A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. *Intensive Care Medicine* **29**, 1043–1051.

Fleiss J (1986). *The Design and Analysis of Clinical Experiments*. John Wiley & Sons, Ltd: New York, NY.

Ghimire DJ, Chardoul S, Kessler RC, Axinn WG and Adhikari BP (2013). Modifying and validating the Composite International Diagnostic Interview (CIDI) for use in Nepal. *International Journal of Methods in Psychiatric Research* **22**, 71–81. doi: 10.1002/mpr.1375.

Hamilton M (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry* **23**, 56–62.

Haroz EE, Bass JK, Lee C, Murray LK, Robinson C, Bolton P (2014). Adaptation and testing of psychosocial assessment instruments for cross-cultural use: an example from the Thailand Burma border. *BMC Psychology* **2**, 31. doi: 10.1186/s40359-014-0031-6.

Haroz EE, Bolton P, Gross A, Chan KS, Michalopoulos L, Bass J (2016). Depression symptoms across cultures: an IRT analysis of standard depression symptoms using data from eight countries. *Social Psychiatry and Psychiatric Epidemiology* **51**, 981–991. doi: 10.1007/s00127-016-1218-3.

Haroz EE, Ritchey M, Bass JK, Kohrt BA, Augustinavicius J, Michalopoulos L, Burkey MD, Bolton P (2017). How is depression experienced around the world? A systematic review of qualitative literature. *Social Science and Medicine* **183**, 151–162. doi: 10.1016/j.socscimed.2016.12.030.

Hollifield M (2002). Accurate measurement in cultural psychiatry: will we pay the costs? *Transcultural Psychiatry* **39**, 419–421. doi: 10.1177/136346150203900401.

Jayawickreme N, Jayawickreme E, Atanasov P, Goonasekera MA, Foa EB (2012). Are culturally specific measures of trauma-related anxiety and depression needed? The case of Sri Lanka. *Psychological Assessment* **24**, 791–800. doi: 10.1037/a0027564.

Jorm AF (2000). Does old age reduce the risk of anxiety and depression? A review of epidemiological studies across the adult life span. *Psychological Medicine* **30**, 11–22.

Kessler RC, Berglund P, Demler O, Jin R, Koretz D, Merikangas KR, Rush AJ, Walters EE, Wang PS (2003). The epidemiology of major depressive disorder. *JAMA* **289**, 3095. doi: 10.1001/jama.289.23.3095.

Kessler RC, Bromet EJ (2013). The epidemiology of depression across cultures. *Annual Review of Public Health* **34**, 119–138. doi: 10.1146/annurev-publhealth-031912-114409.

Keys HM, Kaiser BN, Kohrt BA, Khoury NM, Brewster A-RT (2012). Idioms of distress, ethnopsychology, and the clinical encounter in Haiti's Central Plateau. *Social Science and Medicine* **75**, 555–564. doi: 10.1016/j.socscimed.2012.03.040.

Kohrt BA, Rasmussen A, Kaiser BN, Haroz EE, Maharjan SM, Mutamba BB, de Jong JT, Hinton DE (2014). Cultural concepts of distress and psychiatric disorders: literature review and research recommendations for global mental health epidemiology. *International Journal of Epidemiology* **43**, 365–406. doi: 10.1093/ije/dyt227.

Kroenke K, Spitzer RL, Williams JB (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine* **16**, 606–613. doi: 10.1046/j.1525-1497.2001.016009606.x.

Liu X (2012). Classification accuracy and cut pointselection. *Statistics in Medicine* **31**, 2676–2686. doi: 10.1002/sim.4509.

Lotrakul M, Sumrithe S, Saipanish R (2008). Reliability and validity of the thai version of the PHQ-9. *BMC Psychiatry* **8**, 46–53.

Marc LG, Henderson WR, Desrosiers A, Testa MA, Jean SE, Akom EE (2014). Reliability and validity of the haitian creole PHQ-9. *Journal of General Internal Medicine* **29**, 1679–1686.

Miller KE, Omidian P, Quraishy AS, Quraishy N, Nasiry MN, Nasiry S, Karyar NM, Yaqubi AA (2006). The Afghan Symptom Checklist: a culturally grounded approach to mental health assessment in a conflict zone. *American Journal of Orthopsychiatry* **76**, 423–433. doi: 10.1037/0002-9432.76.4.423.

Mulrow CD, Williams JW, Gerety MB, Ramirez G, Montiel OM, Kerber C (1995). Case-finding instruments for depression in primary care settings. *Annals of Internal Medicine* **122**, 913–921.

Muthén LK, Muthén BO (1998–2012). *Mplus User' s Guide*, 7th edn. Muthen & Muthen: Los Angeles, CA.

Nichter M (2010). Idioms of distress revisited. *Culture, Medicine, and Psychiatry* **34**, 401–416. doi: 10.1007/s11013-010-9179-6.

Nock MK, Borges G, Bromet EJ, Alonso J, Angermeyer M, Beautrais A, Bruffaerts R, Chiu WT, de Girolamo G, Gluzman S, de Graaf R, Gureje O, Haro JM, Huang Y, Karam E, Kessler RC, Lepine JP, Levinson D, Medina-Mora ME, Ono Y, Posada-Villa J, Williams D (2008). Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. *British Journal of Psychiatry* **192**, 98–105. doi: 10.1192/bjp.bp.107.040113.

Nolen-Hoeksema S, Larson J, Grayson C (1999). Explaining the gender difference in depressive symptoms. *Journal of Personality and Social Psychology* **77**, 1061–1072.

Ormel J, Petukhova M, Chatterji S, Aguilar-Gaxiola S, Alonso J, Angermeyer MC, Bromet EJ, Burger H, Demyttenaere K, de Girolamo G, Haro JM, Hwang I, Karam E, Kawakami N, Lepine JP, Medina-Mora ME, Posada-Villa J, Sampson N, Scott K, Ustun TB, Von Korff M, Williams DR, Zhang M, Kessler RC (2008). Disability and treatment of specific mental and physical disorders across the world. *British Journal of Psychiatry* **192**, 368–375. doi: 10.1192/bjp.bp.107.039107.

Patel V, Araya R, Chowdhary N, King M, Kirkwood B, Nayak S, Simon G, Weiss HA (2008). Detecting common mental disorders in primary care in India: a comparison of five screening questionnaires. *Psychological Medicine* **38**, 221–228. doi: 10.1017/S0033291707002334.

Patel V, Simon G, Chowdhary N, Kaaya S, Araya R (2009). Packages of care for depression in low- and middle-income countries. *PLoS Medicine* **6**, e1000159. doi: 10.1371/journal.pmed.1000159.

Patel V, Simunyu E, Gwanzura F, Lewis G, Mann A (1997). The shona symptom questionnaire: the development of an indigenous measure of common mental disorders in Harare. *Acta Psychiatrica Scandinavica* **95**, 469–475.

Phan T, Steel Z, Silove D (2004). An ethnographically derived measure of anxiety, depression and somatization: the Phan Vietnamese Psychiatric Scale. *Transcultural Psychiatry* **41**, 200–232. doi: 10.1177/1363461504043565.

Radloff LS (1977). The CES-D scale: a self-report depression scale for research in the general population. *Applied Psychological Measurement* **1**, 385–401. doi: 10.1177/014662167700100306.

Rasmussen A, Keatley E, Joscelyne A (2014). Posttraumatic stress in emergency settings outside North America and Europe: a review of the emic literature. *Social Science and Medicine (1982)* **109**, 44–54. doi: 10.1016/j.socscimed.2014.03.015.

Sackett PR, Lievens F (2008). Personnel selection. *Annual Review of Psychology*. **59**, 419–450. doi: 10.1146/annurev.psych.59.103006.093716.

Spitzer RL, Williams JB, Gibbon M, First MB (1995). *Structured Clinical Interview for DSM-IV (SCID)*. Biometrics Research: New York.

StataCorp (2013). *Stata Statistical Software: Release 13*. StataCorp LP: College Station, TX.

Van de Velde S, Bracke P, Levecque K (2010). Gender differences in depression in 23 European countries. Cross-national variation in the gender gap in depression. *Social Science and Medicine* **71**, 305–313. doi: 10.1016/j.socscimed.2010.03.035.

van Ginneken N, Tharyan P, Lewin S, Rao GN, Meera SM, Pian J, Chandrashekar S, Patel V (2013). Non-specialist health worker interventions for the care of mental, neurological and substance-abuse disorders in low- and middle-income countries. *Cochrane Database of Systematic Reviews* **11**, CD009149. doi: 10.1002/14651858.CD009149. pub2.

Viera AJ, Garrett JM (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine* **37**, 360–363.

Zhong Q, Gelaye B, Fann JR, Sanchez SE, Williams MA (2014). Cross-cultural validity of the spanish version of PHQ-9 among pregnant peruvian women: A rasch item response theory analysis. *Journal of Affective Disorders* **158**, 148–153.

Zung WWK, Grinker RR Sr., Overall JE, Friedman AS, Zunf WWK, Wessman AE, Hildreth HM, Hathaway S, McKinley C, Hutchinson JT, Smedberg D, Beck AT, Fleminger JJ, Groden B, Wechsler H (1965). A self-rating depression scale. *Archives of General Psychiatry* **12**, 63. doi: 10.1001/archpsyc.1965.01720310065008.

https://doi.org/10.1017/gmh.2017.16 Published online by Cambridge University Press