Measuring antenatal depressive symptoms across the world: A validation and cross-country invariance analysis of the Patient Health Questionnaire – 9 (PHQ-9) in eight diverse low resource settings

AUTHORS

Aja Louise Murray, Chad Lance Hemady, Do Phuc Huyen, Michael Dunne, Sarah Foley, Joseph Osafo, Siham Sikander, Bernadette Madrid, Adriana Baban

Diana Taut, Catherine Ward, Asvini Fernando, Vo Van Thang, Manuel EisnerC, laire Hughes, Pasco Fearon, Sara Valdebenito, Mark TomlinsonS, usan Walker

Abstract

Measures that produce valid and reliable antenatal depressive symptom scores in low resource country contexts are important for efforts to illuminate risk factors, outcomes, and effective interventions in these contexts. Establishing the psychometric comparability of scores across countries also facilitates analyses of similarities and differences across contexts. To date, however, few studies have evaluated the psychometric properties and comparability of the most widely used antenatal depressive symptom measures across diverse cultural, political and social contexts. To address this gap, we used data from the *Evidence for Better Lives Study – Foundational Research (EBLS-FR)* to examine the internal consistency reliability, nomological network validity, and cross-country measurement invariance of the 9-item version of the Patient Health Questionnaire (PHQ-9) in antenatal samples across eight low-resource contexts. We found that the PHQ-9 scores had good internal consistency across all eight countries. Correlations between PHQ-9 scores and constructs conceptually associated with depression were generally consistent, with a few exceptions. In measurement invariance analyses, only partial metric invariance held and only across four of the countries. Our results suggest that the PHQ-9 yields internally consistent scores when administered in culturally diverse antenatal populations; however, the meaning of the scores may vary. Thus, interpretation of PHQ-9 scores should consider local meanings of symptoms of depression to ensure that context-specific conceptualisations and manifestations of antenatal depressive symptoms are adequately reflected.

Keywords: antenatal depressive symptoms; patient health questionnaire; global mental health; measurement invariance

Public Significance Statement

Antenatal depression is a significant global problem; however, there has been limited research into symptom measures in a diversity of low resource contexts. In this study, common measure of antenatal depressive symptoms (PHQ-9) gave scores with good reliability across eight different

countries and nine languages; however, scores cannot always be compared across settings. We provide recommendations for making symptom measures more relevant within and comparable across diverse country contexts.

Clinically significant depressive symptoms are estimated to affect approximately 15-20% of pregnant women, with evidence of higher rates in low- and middle-income countries (LMICs) than in high income countries (Biaggi et al., 2016; Fisher et al., 2012; Mahendran et al., 2019). A recent review focusing on LMICs suggested that antenatal depression is associated with poorer birth outcomes such as preterm birth and low birthweight (Fekadu Dadi et al., 2020). Given the potential adverse impacts of antenatal depression, the availability of measures that can successfully identify symptoms and track their progression during pregnancy offers considerable benefits. Measures that yield comparable scores across different countries are particularly valuable from a global health perspective, as they facilitate cross-national investigations of the prevalence, risk factors, and outcomes of antenatal depressive symptoms (e.g., Fisher et al., 2012).

The Patient Health Questionnaire (PHQ) is a Diagnostic and Statistical Manual of Mental Disorders (DSM) criteria-based tool that measures the symptoms of depression and other mental disorders (Spitzer et al., 2000). The nine item version (PHQ-9) is often favoured for screening for depressive symptoms as it is argued to achieve brevity and simplicity whilst still capturing a diverse set of depressive symptoms (Carroll et al., 2020). Though not specifically designed to screen for depressive symptoms during pregnancy, psychometric studies have supported its utility for estimating depressive symptom levels in antenatal populations (Barthel et al., 2015; Bindt et al., 2012; Flynn et al., 2011; Sidebottom et al., 2012; Zhong et al., 2014). However, despite having been translated into over 70 languages and dialects, the extent to which the PHQ-9 produces scores that are cross-culturally valid and comparable is virtually unknown (Carroll et al., 2020). In particular, there is limited research examining the validity of the PHQ-9 scores in low resource country contexts (Ali et al., 2016).

Mental health measures developed in one country or culture do not necessarily function in the same way in otherss (Kohrt et al., 2011, 2016); indeed, there is ample evidence for cultural influences on the conceptualisation and manifestation of mental health issues (Kohrt et al., 2014). Acknowledging this, the DSM 5 states that ‘all forms of distress are locally shaped’ (American Psychiatric Association, 2013). For example, Ali et al. (2016) note that in sub-Saharan Africa, depression is more likely to be reported in terms of somatic symptoms and local idioms (i.e., culture-

specific concepts of distress) than via symptoms commonly included in depression inventories developed in the English language in Western contexts. A similar idea has been applied to the contrast between Chinese and Western contexts, where the latter is more likely to 'psychologise' their depression and the former to 'somatisize' (Ryder et al., 2008).

Cross-cultural differences in stigma may also influence reporting of depressive symptoms. A recent large-scale study found that Vietnamese women were at high risk of antenatal depression (24.5%;Hue et al., 2020) and typically present to health centres with psychosomatic complaints, a finding which has been speculated to reflect stigma regarding mental health difficulties. In contrast, a Sri Lankan survey showed that depression is seen as having biological roots and help-seeking is less stigmatised (Ediriweera et al., 2012). Similar considerations may apply to specific depressive symptoms such as loss of sexual interest or pleasure, which may be taboo topics in some settings and thus not openly discussed. These items can be poor markers of mental health in such contexts (Dere et al., 2015)

Further, for more practical reasons, indicators that are relevant in one context may be poor markers of mental health in another. For example, loss of appetite is likely to be a poor marker for depression in countries where there are high levels of gastrointestinal infections or other physical problems impacting appetite (Kohrt et al., 2016).

These kinds of contextual differences have significant implications for mental health measurement, as the majority of measure development and validation work has relied on an 'etic' approach (Ali et al., 2016). In an etic approach, measures are developed in one context, most commonly a Western Educated Industrialised Rich and Democratic (WEIRD; Henrich et al., 2010) context, and translated and adapted for use in other contexts, rather than being developed from the starting point of local concepts (Berry, 1969). Accordingly, when instruments developed and adapted/translated in this way are applied and investigated in low resource country contexts it is common for measurement differences to be observed. Indeed, research has suggested that screens for post-natal depression in LMICs can miss cases when they neglect common culturally-specific manifestations of, or ways of describing, distress in these populations (Ali et al., 2016). A recent systematic review of the functioning of the Edinburgh Postnatal Depression Scale (EPDS) in women

in low- and lower-middle income countries suggested that its psychometric properties with respect to measuring perinatal depression are poorer than in high-income countries (Shrestha et al., 2016).

Measurement variations resulting from cross-country differences in the understanding and manifestation of mental health symptoms mean that measures developed and validated in WEIRD contexts may not yield valid or comparable scores in other contexts (e.g., see Kohrt et al., 2011). This makes it difficult to compare rates, risk factors for, and outcomes of antenatal depressive symptoms across different cultures. In turn, this lack of conceptual equivalence complicates assessments of the cross-cultural generalisability of findings, the identification of the settings where investment is most needed, and the targeting and tailoring of interventions to meet local needs.

Analyses of cross-country measurement invariance can provide insights into the extent to which respondents in different countries share similar conceptions of depression and obtain similar scores on depressive symptom assessments when experiencing comparable levels of underlying severity (Stevanovic et al., 2015; Svetina et al., 2020). Measurement invariance analysis can be conducted within a confirmatory factor analysis (CFA) framework, which models individual items as indicators of underlying latent depression factors (Svetina et al., 2020). Factor loadings capture the strength of relation between the latent depression factors and the observed items, while item intercepts capture the 'severity' of a symptom. Different levels of measurement invariance support different types of cross-country comparisons. Configural invariance is when the same items load on the same factors across counties and is a minimum level of invariance. Metric invariance is identified when the factor loadings of a confirmatory factor analysis (CFA) model for a set of items are equal across different countries. Metric invariance allows variances and covariances involving the latent depression construct(s) to be compared across countries. This would, for example, support a comparison of the strength of relations between depression and candidate risk factors across countries. Scalar invariance is when the intercepts (or thresholds) of the depression items are equal across countries. Achieving scalar invariance allows levels of the latent depression construct(s) to be compared across countries within the CFA model. Measurement non-invariance does not necessarily mean that countries cannot be validly compared, but indicates that the non-invariance should be modelled in a 'partial invariance' model in order to avoid biasing these comparisons (Pokropek et al.,

2019). Further, violations of measurement invariance can provide insights into cultural differences that can inform more global understandings of mental health. Given the need for an improved understanding of the extent to which antenatal depressive symptoms can be measured with reliability and validity, and comparably across countries, we conducted a psychometric validation and measurement invariance analysis of the PHQ-9 across eight culturally diverse middle-income countries.

Method

Overview

The PHQ-9 was translated into nine languages and its factorial validity, reliability, nomological network correlations, and cross-country invariance were evaluated in eight middle-income countries. The study was not pre-registered.

Participants

Data came from the Evidence for Better Lives Study - Foundational Research (EBLS-FR) dataset (Valdebenito et al., 2020). In EBLS-FR, data were collected from 1,206 pregnant women in their third trimester of pregnancy from eight sites. Sites were selected on the basis of being medium-sized cities/conurbations in an LMIC and to sample a wide range of cultural, political, and social contexts: Valenzuela, Philippines; Hue, Vietnam; Ragama, Sri Lanka; Tarlai Kalan, Pakistan; Cluj-Napoca, Romania; Worcester, South Africa; Koforidua, Ghana; and Kingston, Jamaica, and include at least one country from each major world region as defined by the World Health Organisation (WHO). Since the time of site selection, Romania's classification has changed to 'high income country'. Within each site, pregnant women were recruited via primary health clinics during their standard antenatal care visits. The goal was not to achieve an identical sampling strategy in each site, but rather - given significant differences in health systems and client accessibility- to utilise a locally adapted sampling strategy to achieve maximum comparability of samples across sites, Full details of the sampling procedures within each site are provided in Supplementary Materials.

Inclusion criteria were that women were in the third trimester of pregnancy (29-40 weeks of gestation), aged over 18, and having their main residence within one of the defined geographical regions of the EBLS-FR study. Further participant characteristic information is provided in Table 1.

Data collection

Data, including the PHQ-9, were collected by trained fieldworkers using computer-assisted personal interviews (CAPI). The data analysed in the current study were part of a broader questionnaire that collected information on participant health, wellbeing, adversity exposure, feelings about their pregnancy, reproductive history, attitudes towards the pregnancy and future parenting, and social support.

Ethics

Prior to data collection, the study underwent ethical review within the University of Cambridge as well as from the ethics committees of the research institutes in each site. Informed consent was collected from all participants.

Sample size

A full discussion of the sample size and minimum detectable effects is provided in Supplementary Materials. The available sample size was based on the maximum that could be attained within resource constraints. Monte Carlo power analyses were used to explore the minimum detectable effect sizes for the primary analyses of the present study, i.e., the measurement invariance analyses and suggested. We used a n=150 per group and 8 groups for these analyses. The basic population model was based on an 8-group multi-group CFA of a scale similar the PHQ-9. It assumed 8 groups and 9 items with standardised loadings of .70, factor variances of 1, and thresholds of -1,0,1 and WLSMV estimation was used. Different levels of non-invariance ($\Delta\lambda$ = 0.10 to 0.40) were then introduced into a single item in a single group (the $\Delta\lambda$ are on a standardized scale). The power to detect non-invariance for a single loading was based on the percentage of replications for which the difference between the non-invariant loading and the corresponding loadings in the other groups ($\Delta\lambda$) was statistically significant. Results are shown in the Table S12 in Supplementary Materials. These suggest that from $\Delta\lambda$~.35 and up our study is well-powered to detect individual loading invariance violations Key output files for these analyses are available at: https://osf.io/eg635/. Given that our goal was to detect non-invariance that could have substantively important biasing effects on cross-country comparisons, this was judged acceptable for the purposes of the present study.

Measures

Patient Health Questionnaire (PHQ-9)

The PHQ-9 measures depressive symptoms experienced in the last two weeks based on DSM-IV-TR criteria for depression. The nine items refer to the following symptoms: anhedonia, dysphoria, sleep disturbances, fatigue, changes in eating, low self-esteem, concentration difficulties, hypo- or hyper-active behaviours, and thoughts of suicide or self-harm. Responses were recorded on a four-point Likert-type scale with response options: 0= not at all; 1= several days; 2= more than half the days; 3 =nearly every day. It was, however, necessary to recode the responses as 0= not at all 1= several days (encompassing the 'several days', 'more than half the days' and the 'nearly every day' options from the original response scale) for the 'thoughts' item as no participant in the Romanian group selected the two highest response options and no participant in the Philippines group selected the highest response option for this item. These differences in response scale across items can be accommodated within the latent factor models used and facilitated the inclusion of the 'thoughts' items in the factorial validity, reliability, and measurement invariance analyses.

While the psychometric properties of PHQ-9 have been extensively researched, the measure has undergone limited validation in LMICs, including in the eight countries of EBLS (Carroll et al., 2020). Two previous studies have examined the psychometric properties of the PHQ-9 in a sample of pregnant women in Ghana, finding that PHQ-9 scores were associated with disability scores but there was some evidence of potential multi-dimensionality and internal consistency values fell below conventional cut-offs of .70 (Barthel et al., 2015; Bindt et al., 2012). A previous study in Pakistan found good sensitivity, specificity and reliability of the PHQ-9 in an antenatal population (Gallis et al., 2018). We could not locate any published studies assessing the PHQ-9 in antenatal samples in any of the other countries included in our sample. In Sri Lanka, however, Hanwella et al. (2014) found good sensitivity and specificity and strong internal consistency for a Sinhala version of the PHQ-9 in a sample of 75 participants diagnosed with major depressive disorder and gender-matched controls . In South Africa, PHQ-9 validation studies (Aggarwal et al., 2017; Bhana et al., 2015; Cholera et al., 2014), have had mixed results. While internal consistency values have generally been adequate, sensitivity was low in one study of primary care patients (Bhana et al., 2015) but better in another (Cholera et al., 2014). In Vietnam, the measure showed evidence of good factorial, convergent and

external validity, and excellent reliability in a sample of sexual minority women (Nguyen et al., 2016). No studies based in the Philippines could be identified; however, a study of Filipino migrant domestic workers in China found acceptable internal consistency, evidence for convergent and divergent validity, and reasonable sensitivity and specificity at an optimal cut-point of 6 for the PHQ-9 (Garabiles et al., 2019).

Nomological net measures

To construct a nomological net (Cronbach & Meehl, 1955), we selected several measures that, from previous research, would be expected to be differentially related to PHQ-9 scores, including: maternal stress, well-being, suicidality, aggression, and self-control. We expected particularly close associations between depressive symptoms and stress, well-being, and suicidality, allowing us to test convergent validity with depressive symptoms. Conversely, we expected aggression and self-control to be more closely related to one another than to maternal depressive symptoms, allowing us to test divergent validity. Details of these measures are provided below, including the omega total internal consistency values and 95% CIs in the current sample. Where items are on a 4-point scale or less these were calculated using the parameters of a categorical one-factor CFA(Green & Yang, 2009) . Where items are on a 5-point scale or above they were calculated using a CFA estimated using robust maximum likelihood estimation. The confidence intervals are based on bias corrected and accelerated bootstrapping. All were estimated using the ci.reliability function in MBESS for R statistical software (see Kelley & Pornprasertmanit, 2016).

Maternal stress was measured using the *Perceived Stress Scale* (Cohen, 1988)*;* a 10-item measure capturing the extent to which participants felt under stress in the previous month (example item: 'Feeling that difficulties were piling up so high that I could not overcome them'). Item responses were recorded on a four-point Likert-type scale from *not at all* to *nearly every day* and averaged to provide an overall stress score. Higher scores indicated higher levels of stress. Omega reliabilities and 95% confidence intervals (95% CI) for the Perceived Stress Scale scores in each country were: Ghana=.51 (95% CI = .10 to .70); Jamaica= .84 (95% CI = .75 to .89); Pakistan= .82 (95% CI =.75 to .87); Philippines=.81; Romania=.88 (95% CI = .82 to .91); South Africa= .89 (95% CI = .55 to 1.00); Sri Lanka =.95 (95% CI = .75 to 1.00); Vietnam= .85 (95% CI = .72 to .92).

Well-being was measured using the *WHO-5 Wellbeing Index* (Topp et al., 2015), comprising 5 items capturing previous fortnight well-being (example item: 'I have felt cheerful and in good spirits'). Responses were recorded on a 6-point Likert-type scale from *at no time* to *all of the time* and summed, then linearly rescaled to provide overall scores on a 0 to 100 scale (with higher scores representing better well-being). This rescaling was to facilitate comparison with other samples. Omega reliabilities for the WHO-5 Wellbeing Index scores in each country were: Ghana=.86 (95%CI = .82 to .90); Jamaica=.75 (95% CI = .67 to .81) ; Pakistan=.88 (95% CI = .84 to .91); Philippines=.87 (95% CI = .83 to .90); Romania=.77 (95% CI = .64 to .83); South Africa=.78 (95% CI = .70 to .83); Sri Lanka=.85 (95% CI = .80 to .88); Vietnam=.82 (95% CI = .76 to .86).

Aggression was measured using an adapted version of the *Brief Aggression Questionnaire* (Webster et al., 2014). The version administered in EBLS-FR includes 12 items (example item: 'sometimes I fly off the handle for no good reason') capturing different forms of aggression (physical aggression, verbal aggression, anger, and hostility). Responses were recorded on a 5-point Likert-type scale from *never* to *always*. Item scores were averaged to provide overall aggression scores (with higher scores representing more aggression). Omega reliabilities for the aggression scores in each country were: Ghana=.78 (95% CI = .72 to .82); Jamaica=.74 (95% CI = .65 to .80); Pakistan=.55 (95% CI = .36 to 1); Philippines=.72 (95% CI = .61 to .78); Romania=.67 (95% CI = .51 to .77); South Africa= .70 (95% CI =); Sri Lanka=.68 (95% CI = .56 to .75); Vietnam=.66 (95% CI = .54 to .73).

Self-control was measured using an adapted version of the *Brief Self-control Scale* (Tangney et al., 2004). The version administered in EBLS-FR includes 8 items (example item 'I am good at resisting temptation'), with responses recorded in a 5-point Likert-type scale from *not at all* to *very much*. Individual item responses were averaged to form overall self-control scores. Omega reliabilities for the self-control scores in each country were: Ghana=.73 (95% CI = .66 to .79); Jamaica=.73 (95% CI =.66 to .79); Pakistan=.62 (95% CI = .44 to .72); Philippines=.52 (95% CI = .03 to .60); Romania=.70 (95% CI = .61 to .76); South Africa= .58 (95% CI = .43 to .67); Sri Lanka=.59 (95% CI =); Vietnam=.57 (95% CI = .44 to .67).

Translation

All measures were initially available in English, from which they were translated into 9 different languages, guided by the best practice translation recommendations provided by WHO: https://www.who.int/substance_abuse/research_tools/translation/en/. The languages were: Urdu (Pakistan), Afrikaans and IsiXhosa (South Africa), Romanian (Romania), Filipino (Tagalog) (the Philippines), Sinhala and Tamil (Sri Lanka), Vietnamese (Vietnam) and Twi (Ghana), selected based on the size of the respective linguistic group among the population of pregnant women in eight study sites. The English version was used in Jamaica. In some cases, the measures had previously been translated into the relevant local language; however, even in these cases we conducted our own translation in an effort to maximise the comparability across the sites. The process of translation, adaptation, and cross-cultural validation of the instruments aimed to harmonize among eight studies sites and the original version. The translation process involved two independent translators producing two forward-translated versions of the questionnaire. An expert panel meeting in each site resolved any discrepancies between two versions and generated a harmonised translated version to ensure natural and acceptable common language for the broadest audience. Fieldworkers in eight sites were provided with training to address any ambiguities and misunderstandings.

Measure pre-testing

All versions of the questionnaires were pre-tested using a convenience sample of participants (n=5-10 women in each site) who were similar in background to the target population. The pre-testing led to the detection and correction of minor problems.

Statistical Procedure

Internal consistency reliability

Omega total (McDonald, 1999) was used to compute the internal consistency reliability of the PHQ-9 within each country sample, using the *psych* package for R statistical software (Revelle, 2017). Unlike Cronbach's alpha, omega total does not assume that the relations between each item and the underlying construct being measured (here a depression latent variable) are equal (McNeish, 2018). It is, therefore, more appropriate for computing internal consistency for measures such as the PHQ-9 where this assumption is unlikely to hold.

Nomological network

Nomological network analysis can be used to evaluate whether the observed pattern of associations between scores from a focal measure (here the PHQ-9) and others matches that expected by theory (Cronbach & Meehl, 1955). Composite scores were created for the PHQ-9 and all other nomological network measures. Within each country sample, raw Pearson correlations (i.e., unadjusted for other measures) were used to estimate the associations between all measures. These associations were visualised using diagrams produced using qgraph (Epskamp et al., 2012), which represents each score as a node and each correlation as an edge. The thickness of edges is proportional to the strength of association between nodes. This facilitated a descriptive comparison of the way in which PHQ-9 scores relate to other scores across the country samples. A statistical comparison of nomological networks was conducted using Jennrich's test of the equality of correlation matrices (Jennrich, 1970). This method compares the differences between two matrices to the averages of two matrices using a chi-square test. We did not correct for multiple comparisons in applying these tests to the pairwise combinations of matrices because, given the goal of identifying possible cross-country non-equivalence. we were more concerned with having adequate power to detect possible differences than about type 1 errors.

Relative importance analysis

To provide complementary information on the relations between depressive symptoms and related factors, we also conducted a linear regression and relative importance analysis in each of the country samples (Grömping, 2006). In this we predicted depressive symptom scores from the same measures in the nomological network using a linear regression model and quantified the 'relative importance' of each predictor using the LMG statistic. This statistic provides a measure of the average sequential R^2 contribution of each predictor over all different orderings of entry into the model (to account for the fact that different orderings produce different sequential R^2 values for a given predictor). Bootstrapping was used to estimate the variability in the estimates.

Factorial validity and measurement invariance

We began by evaluating whether the proposed one-factor structure of the PHQ-9 provided a good fit to the data in each country individually. That is, we used CFA to establish whether a single

factor could be considered sufficient to describe structure of the PHQ-9. This was also a pre-requisite for being included in the subsequent cross-country measurement invariance analysis (see below) which required a common configural factor structure across all sites.

A CFA model was fit for each country using weighted least squares means and variances (WLSMV) to take account of the ordered categorical nature of the items. Scaling and identification were achieved by fixing the latent variable variance to 1. Models were judged to fit well if comparative fit index (CFI) and Tucker-Lewis Index (TLI) were >.95 and Root Mean Square Error of Approximation (RMSEA) and Standardised Root Mean Square Residual (SRMR) were <.08 (Hu & Bentler, 1999; Schermelleh-Engel et al., 2003). Where models did not fit well, modification indices (MI) and expected parameter changes (EPCs) were inspected to identify possible sources of poor fit. In general, given the exploratory nature of the current study, we allowed for the inclusion of residual covariances indicated by MIs and EPCs as they may provide insights into country-specific item relations.

Assuming reasonably fitting one factor CFAs could be achieved in each country individually, we proceeded to test measurement invariance across countries. While universal model fit cut-offs are difficult to define due to the fact that they are sensitive to a range of modelling situation features beyond model mis-specification, we adopted commonly used heuristics of CFI and TLI ≥.95, RMSEA ≤ .05 and SRMR ≤.08 to indicate a well-fitting model (e.g., Hu & Bentler, 1999; Schermelleh-Engel et al., 2003)

Any countries for which a one-factor model was not supported were not included in this analysis. Our configural model comprised a multi-group one-factor model. One country served as the reference group and the mean and variance of its PHQ-9 latent factor were fixed to 0 and 1 respectively and scale factors were fixed to 1. As there was no strong a priori reason to use any country over the others as the reference group, we used the country that came first alphabetically as the reference group. In addition, the loading of the first indicator was constrained to be equal across all countries, one threshold per item was fixed equal across groups and a second threshold for the first item was fixed equal across groups. The latent factor means and variances of all countries other than the reference group were freely estimated. Together these constraints provided latent variable scaling

and identification for the configural model. Configural invariance was judged to hold if this initial multi-group model showed reasonable fit according to CFI, TLI, RMSEA and SRMR.

Next, we tested metric invariance by constraining the factor loadings to be equal across all groups. Metric invariance was judged to hold based on the criteria recommended by Chen (2007), that is, if CFI did not decrease by more than .010, if RMSEA did not increase by more than .015, and SRMR did not increase by more than .030. If metric invariance did not hold, a partially invariant model was sought through iterative release of cross-country loading constraints. This was guided by inspection of MIs and EPCs.

Finally, scalar invariance was tested by adding cross-country invariance constraints on item thresholds. No constraints were placed on items that had previously shown a lack of metric invariance. Scalar invariance was judged to hold based on the criteria recommended by Chen (2007), that is, if CFI did not decrease by more than .010, if RMSEA did not increase by more than .015, and SRMR did not increase by more than .010. If these criteria were not met, iterative release of cross-country threshold constraints were used to attempt to identify a partially invariant scalar model.

All (multi-group) CFA models were fit in *Mplus 8.4* (Muthén & Muthén, 2015). The *MplusAutomation* package (Hallquist & Wiley, 2018) for R statistical software was also used to assist with managing the fitting, comparison, and reporting of models.

Data availability

Analysis code is available at: https://osf.io/eg635/. Other study materials and data are available on reasonable request to the first author.

Results

Descriptive statistics

Table 2 provides the item descriptive statistics for the PHQ-9 by country. These show some variation in overall PHQ-9 scores across the sites with the highest reported scores in Jamaica (M=9.0; SD=5.1) and the lowest in Romania (M=5.1, SD=3.5).

Internal consistency

Omega total reliability values and 95% confidence intervals (CIs) calculated using the bias an accelerated bootstrap method described in (Kelley & Pornprasertmanit, 2016) are provided in Table 1.

These suggested that the internal consistency reliabilities of the PHQ-9 scores were good in all countries.

Nomological network

Tables S1-S8 of Supplementary Materials provide the correlations among PHQ-9, stress, well-being, suicidality, self-control, and aggression scores. PHQ-9 scores were significantly associated with well-being, suicidal ideation, and stress scores in all eight sites; and with aggression and self-control scores in all sites except Ghana. Figure 1 visualises the patterns of associations between PHQ-9 scores and the other scores in the nomological net correlation matrix across countries. The *p*-values for the statistical comparisons of the nomological net correlation matrices between countries are provided in Table S9 of Supplementary Materials. The correlation matrix from Ghana differed significantly from that of all other countries. In addition, there were significant differences between Jamaica and Romania and Sri Lanka; between Pakistan and Sri Lanka; between the Philippines and Romania; between Romania and both South Africa and Sri Lanka; between South Africa and Sri Lanka; and between Sri Lanka and Vietnam.

Relative importance analysis

Tables S10 of Supplementary Materials provides the results of the relative importance analysis in each of the eight countries. Across all countries stress and wellbeing tended to be consistently higher in relative importance for predicting depressive symptoms than aggression or self-control; however, there were some exceptions (e.g., self-control had a higher LMG value than well-being in South Africa). Overall, the LMG values did not appear to be substantially different across countries and their 95% CIs overlapped for each predictor.

Factorial validity

The CFA models fit well to the data from the Philippines (CFI=.95, TLI=.94, RMSEA=.06, SRMR=.06), South Africa (CFI=.94, TLI=.92, RMSEA=.07, SRMR=.08), Sri Lanka (CFI=.95, TLI=.94, RMSEA= .09, SRMR=.07) and Vietnam (CFI=.98, TLI=.97, RMSEA=.06, SRMR=.09), therefore, no modifications were made to the models in these countries. The CFA model fit to the data from Pakistan did not fit well according to several criteria (CFI=.92, TLI=.89, RMSEA=.09,

SRMR=.10); however, there were no large MIs or EPCs. We, therefore, also made no further modifications to this model.

When fit to the data from Ghana, a one-factor CFA for the PHQ-9 did not initially fit well (CFI=.86, TLI=.81, RMSEA=.11, SRMR=.08). MIs and EPCs suggested that fit could be improved with the addition of a residual covariance between the 'energy' and 'sleep' items and/or the 'failure' and 'sleep' items. Including these residual covariances led to improved fit and a model that had suboptimal but acceptable fit (CFI=.93, TLI=.90, RMSEA=.08; SRMR=.07).

The CFA model fit to the data from Jamaica also fit poorly at first (CFI=.85, TLI=.81, RMSEA=.11, SRMR=.09). Similar to the Ghana model, MIs and EPCs suggested that fit could be improved with the addition of a residual covariance between the 'sleep' and 'energy' item and/or the 'failure' and 'energy' items. This improved the fit of the model; however, it remained suboptimal (CFI=.91, TLI=.87, RMSEA=.10, SRMR=.08). Nevertheless, we did not continue to add further residual covariances in order to control the risk of capitalising on chance.

The initial CFA model fit to the data from Romania also showed poor fit (CFI=.86, TLI=.85, RMSEA=.13, SRMR=.13), with MIs and EPCs indicating that fit could be improved with the addition of residual covariances between the 'energy' and 'sleep', 'energy' and 'failure' and/or 'failure' and 'hopelessness' items. All three residual covariances were included, leading to a well-fitting model according to CFI and TLI but still a poor fitting model according to RMSEA and SRMR (CFI=.94, TLI=.92, RMSEA=.09, SRMR=.12). Given that the CFI and TLI values were reasonable we did not make any further modifications to this model.

Cross-country invariance

Measurement invariance was tested only across the four countries where the original PHQ-9 factor structure was supported (i.e., Philippines, South Africa, Sri Lanka, and Vietnam), with the Philippines serving as the reference group. The focus on these countries only was based on the results from the individual CFAs which showed that the one-factor CFA model was not an adequate description of the data in the remaining countries (Ghana, Jamaica, Romania, and Pakistan). While the fit for the countries included in the measurement invariance analysis did not all show good fit according to conventional criteria, they were judged to show fit that was sufficiently good for the

purposes of further exploring cross-country measurement differences in a measurement invariance analysis. The configural model fit well (CFI=.955, TLI=.942, RMSEA=.072, SRMR=.074).

Fit declined with the addition of metric invariance constraints (CFI=.926, TLI=.921, RMSEA=.084, SRMR=.080), with the magnitude of the declines suggesting a lack of invariance by Chen's (2007) criteria. Iterative removal of the cross-country loadings constraints on the following items were necessary to achieve partial metric invariance (CFI=.947, TLI=.941, RMSEA=.063, SRMR=.079): 'sleep' item and energy item in South Africa; 'concentration' item in Vietnam; 'thoughts' item in the Philippines; and 'sleep' item in Sri Lanka.

Scalar invariance constraints were added to the partially metric invariant model. In addition, to resolve an estimation issue a cross-group equality constraint was placed on the third threshold of the 'concentration' item in the Vietnam group. The addition of scalar invariance constraints led to a decline in model fit (CFI=.900, TLI=.915, RMSEA=.087, SRMR=.086). It required the iterative removal of 14 cross-group threshold equality constraints to achieve model fit that was within Chen's (2007) criteria (CFI=.938, TLI=.942, RMSEA=.072, SRMR=.083). Given that more than half the thresholds had to be freed to achieve a reduction in fit less than that required to stay within Chen's (2007) criteria (adding to the constraints not placed on the 5 items that did not show metric invariance) it was concluded that even partial scalar invariance could not be achieved (Pokropek et al., 2019). As such, the partially invariant metric model was judged as the optimal model for this data.

Table 3 provides loading parameter estimates and variances from the final model (partial metric). We do not report threshold/scale factor and mean differences because of the failure to achieve even partial scalar invariance. The estimates suggest that the 'sleep' and 'energy' items were more strongly related to the depressive symptoms latent variable in South Africa as compared to the other countries. The 'sleep' item was also somewhat more related to the depressive symptoms latent variable in Sri Lanka as compared to the Philippines and Vietnam. In addition, the 'concentration' item was more strongly related to the depression latent variable in Vietnam compared with the other countries. As this parameter estimate was large compared to the other loadings, we checked for possible estimation errors by re-estimating the model using different start values and by checking that the corresponding standardised estimate was not a Heywood case. We also checked that other

parameter estimates were similar with versus without the cross-group parameter constraints on this loading (see: https://osf.io/my5nv/). The large magnitude of this loading on the unstandardised scale likely reflects the skewed distribution of responses on this item in this group, with the vast majority of respondents endorsing the 1st and 2nd response option. Finally, the suicidal ideation item was less strongly related to the depression latent variable in the Philippines than in the other countries.

Supplementary Analyses

To provide further illumination on the poor CFA fit in the Ghana, Jamaica, Pakistan, and Romania samples and to guide the development of more suitable factor models, post-hoc exploratory factor analyses were conducted. Parallel analysis with principal components analysis (PA-PCA), the minimum average partial (MAP) test, and visual inspection of a scree plot were used to guide factor retention. Full details are provided in Supplementary Materials. In brief, despite poor fit in the initial one-factor CFAs there was no strong evidence for multi-dimensionality in the Ghana, Pakistan, and Romania sample based on follow-up EFAs. A two-factor model was, however, supported in the Jamaica sample. The two factors captured cognitive-motivational and physical symptoms of depression respectively.

Discussion

Reliable, valid, and preferably brief measures of perinatal depressive symptoms are important for a robust global health evidence base to underpin the monitoring, treatment, mitigation and prevention of depression. Such evidence is needed especially in low resource settings. This study of psychometric properties of the PHQ-9 across eight diverse cultures found the internal consistency to be adequate in all settings. However, the proposed single-factor structure fit well for only half of the countries sampled (Philippines, South Africa, Sri Lanka, and Vietnam).

Even for these countries, differences emerged in multi-group measurement invariance analyses. Here, only partial metric invariance could be achieved, with the 'sleep', 'concentration', 'energy' and 'thoughts' items showing stronger relations to overall depressive symptom levels in some countries compared to others. Specifically, concentration problems were more strongly related to the depression latent variable in Vietnam; suicidal ideation was less strongly related to the depression latent variable in Philippines; sleep problems were more strongly related to the depression latent variable in Sri

Lanka and South Africa (especially the latter); and energy issues were more strongly related to the depressive symptoms latent variable in South Africa. Measurement invariance analyses thus further suggested differences in the content of the PHQ-9 depressive symptoms latent construct across cultures.

Broadly, the nomological network associations for the PHQ-9 scores were similar across all countries, with the direction of associations with other relevant measures consistent across all sites. All associations were in the direction predicted by theory, providing indirect evidence of construct validity (Cronbach & Meehl, 1955). The magnitude of associations, however, sometimes varied considerably across sites and there were a number of statistically significant differences in the correlation matrices between sites. For example, the correlation between PHQ-9 and aggression scores varied from r=.09 in Ghana up to r=.47 in South Africa. These analyses suggested that there were some differences in how depressive symptoms clusters with other related constructs, consistent with the idea that antenatal depressive symptoms may vary in their meaning and implications across cultures. However, the relative importance analysis suggested that the confidence intervals for the relative importance for each covariate across countries overlapped.

Taken together, these findings show important strengths of the PHQ-9 scores in relation to construct validity but also highlight the need for caution when applying the PHQ-9 to compare levels of antenatal depressive symptoms across diverse cultural contexts. Cross-cultural differences in the dimensionality of the PHQ-9 have previously been reported, suggesting that in some contexts it is better characterised using a one-factor model, and in others, a two-factor model (Carroll et al., 2020; Shin et al., 2020). Shin et al. (2020), for example, found evidence that a two-factor model with 'affective-somatic' and 'cognitive' dimensions better described the PHQ-9 in a normative Korean sample than a one-factor model, similar to the findings in this study for Jamaica, while the one factor model fit best in the other countries A lack of direct comparisons of the PHQ-9 using a common data collection protocol across prior studies makes it difficult to establish the extent to which these differences simply reflect methodological differences across studies. Our results thus provide important evidence of possible cross-country differences in cultural understandings and manifestations of antenatal depressive symptoms

The presence of differences in the measurement of depressive symptoms across these diverse contexts is consistent with previous research suggesting that there are substantive differences in the manner in which distress-related concepts are understood, manifested, and accepted across cultures (Kohrt et al., 2014, 2016). The lack of factorial validity for some countries (Pakistan, Jamaica, Ghana and Romania) is also consistent with previous findings suggesting that when depressive symptom measures are developed in Western contexts and then applied in low- and middle-income countries, their psychometric properties in these latter countries are poorer than in the contexts in which they were developed (Shrestha et al., 2016). The reasons for the lack of factorial validity in these new contexts is not entirely clear and it is likely that they differ by context. Pakistan, for example, has been shown to have relatively high rates of antenatal depression compared to the broader region and global averages (Mahendran et al., 2019), an observation confirmed in our study. However, considerable stigma and false beliefs about mental illnesses have been reported in this context (Atif et al., 2021) which may undermine the reliable reporting of symptoms in measures such as the PHQ-9 that rely on brief, self-reports. For Jamaica; however, there was evidence for possible multi-dimensionality in the PHQ-9 based in follow-up EFAs. These suggested that in this context antenatal depressive symptoms are experienced and/or reported in a more differentiated manner in which cognitive/motivational versus physical symptoms are more distinct from one another. There is some precedent for this as similar dimensions were supported for the PHQ-9 in a Korean sample (Shin et al., 2020). The reason why this distinction was observed particularly in the Jamaica sample of our study is not clear. The mental health context for pregnant women in Jamaica can, however, diverge substantially from that of the women other countries sampled. For example, it is common for women to have children with multiple partners over their reproductive years, for relationships to be less stable, and for less partner (practical and emotional) support to potentially be available (see e.g., Bernard et al., 2018). Further research will be required to establish whether cultural norms such as these play a role in measurement differences in antenatal depression symptoms across the world.

Among the countries that we could compare in our measurement invariance analyses we found several additional differences. We found, for example, that the PHQ-9 suicidal thoughts item was less strongly related to the depression latent variable in the Philippines than in other countries.

Previous research has also found cultural variations in the strength of relations between suicidal ideation and a latent depression variable. For example, Dere et al. (2015) found that the suicidal ideation item of the Hopkins symptom checklist 15-item depression scale (HSCL-15) had a low item response theory discrimination parameter in a sample from Thailand compared with others from diverse regions of the world. To help explant this finding, they noted that while depression is more associated with suicidal ideation in HICs, impulse control disorders may be a more important driver of suicidal ideation in some LMICs.

The low loading of the suicidal ideations item in the Philippines may reflect the strong stigma against suicidality in this context, as well as high levels of family-orientation and religiosity deterring individuals from suicidality (or reporting it) despite distress. Indeed, the Philippines has the lowest suicide mortality rate in the Western Pacific region (*World Health Statistics, 2020*) and studies have suggested that religiosity and spirituality are important factors that prevent contemplating and attempting suicide in the Philippines (Bautista et al., 2017; Colucci et al., 2010; Estrada et al., 2019; Redaniel et al., 2011). Future research using techniques such as cognitive interviewing with local respondents could help illuminate the nature of this and the other measurement differences identified in the current study and suggest how the PHQ-9 and measures like it can be adapted to be more comparable across different contexts. These would also help disentangle measurement differences related to influential factors, including cultural understandings of the concept of depression; stigma and associated reporting biases; local applicability of specific items; and issues related to translation and the training of interviewers.

The above considerations regarding cultural influences on specific symptoms underline the importance of ensuring the local relevance of measures of antenatal depressive symptoms and suggests that substantial adaptation may sometimes be necessary to ensure that mental health measures produce scores with validity in in different contexts. However, the ability to use a measure for cross-country comparisons is an important competing consideration as the need for a substantial adaptation can undermine the possibility of valid comparisons. Fortunately, when using latent variable methods, provided that there is a core set of cross-culturally comparable items, some items can vary across contexts without compromising comparability (Pokropek et al., 2019). Further development

work would be valuable to identify an optimal set of cross-culturally invariant items; however, our results suggest that the 'pleasure', 'eating' and 'movement' items represent good candidates as they showed no strong evidence of differing in their functioning across the sites included in the current study. The addition of further cross-culturally invariant items would be beneficial to increase the pool of comparable items. Locally relevant items could then be drawn from cultural concepts of distress, defined as the way that cultural groups 'experience, understand, and communicate suffering, behavioural problems, or troubling thoughts and emotions' to optimise functioning in local contexts (Kohrt et al., 2014). Kohrt et al., (2016) found that augmenting the PHQ-9 with cultural concepts of distress improved the detection of depression in a Nepali sample. This suggests that the approach is promising; however, they did not address comparisons across countries, leaving this as an important area for future work.

Another potential contributing factor to the cross-cultural differences may relate to the suboptimal construction of items. For example, though best practices in item generation (e.g., Lietz, 2010) suggest that items should not be double-barrelled, some PHQ-9 items contain such double-barrelled statements (e.g., sleeping too much or too little). Suboptimal construction in the source language can make it more difficult to translate items in a manner that retains conceptual equivalence (Acquadro et al., 2018). The construction of depression measures to be used across the world should ensure optimal item construction and consider translatability at an early stage of development (de Jong et al., 2018).

Limitations

It is important to consider the limitations of the current study. First, though our measurement protocol was identical and all participants were recruited from antenatal care pathways, there were some differences in the sampling methods across studies due to differences in the structure of healthcare systems and other local factors that precluded the application of an identical sampling method. Thus, a lack of comparability could reflect different compositions in samples across the sites. Second, measurement invariance analyses can identify the presence of non-invariance but provide limited insights into its cause. They cannot, for example, differentiate between non-invariance due to methodological issues such as translation problems or differences in interviewer skills, and more

substantive drivers of non-invariance, such as different cultural understandings of mental health constructs. To differentiate these factors, techniques such as the culture, comprehension, and translation bias techniques suggested by (Bader et al., 2021) are necessary. However, it was not possible to apply this method in the current dataset because it requires sufficiently powered statistical comparisons between groups differing in language but not culture; in culture but not language; and in both language and culture. Indeed, though multiple languages are spoken in some sites, our small sample size meant we were not able to examine invariance across languages within sites

Our sample size, though large overall, was relatively small within each country, though simulation studies (see Supplementary Materials) suggest that non-invariance can be detected well with group sizes of 100, except for small effect sizes (Kim & Yoon, 2011). As such, any non-invariance missed due to our sample sizes was not likely to have been of a substantively important magnitude.

We were unable to assess the sensitivity and specificity of PHQ-9 across contexts as we did not include (ICD-10) gold standard measures of depression as part of our assessment battery. Given the potential measurement differences we identified in the current study, further work to establish (possibly different) optimal cut-points across different country contexts will be important to ensure the PHQ-9 functions optimally for detecting depression.

Our nomological network analysis was also based on measures that have not had their cross-country measurement invariance established but to facilitate analysis of the reasonably large set of variables involved given the modest country-specific sample sizes, we computed composite scores (as opposed to used latent variable models) that implicitly assume measurement invariance. This means that differences across countries in the associations between PHQ-9 and other measures will reflect measurement differences in both the PHQ-9 and the other measures. Some of the measures (e.g., self-control) also had low reliabilities, which will attenuate their associations with PHQ-9 scores. Future studies could examine and model the cross-country invariance in these other variables too using a multi-group latent variable model for all constructs, to help disentangle the different sources of differences in associations across countries. Similarly, we had only cross-sectional data meaning that we could focus only on the correlations between rather than directional effects of these variables on

one another. Future studies with repeated measures of these constructs could construct longitudinal nomological networks using techniques such as graphical vector autoregression (e.g., Epskamp, 2020).

It is also important to note that our conclusions are confined to the PHQ-9 measure of depression and while they might suggest cross-cultural difference in depressive symptoms, this conclusion may vary depending on the instruments used. There are a very large number of measures of depressive symptoms available, with variable similarity in content (Fried, 2017). It would, therefore, be informative to evaluate cross-country differences in depressive symptom measures in other commonly used measures.

Finally, we relied on a particular measurement framework for assessing depressive symptoms across countries, namely a latent variable model that assumes that depressive symptoms are observable manifestations of an underlying latent depression variable. However, there is recognition - in network analyses approaches in particular - that to explain symptom inter-relations there is no need to posit an underlying latent variables for mental health constructs and instead, symptoms may be directly and indirectly connected to and influence one another (Borsboom & Cramer, 2013). Indeed, if the latent variable assumption is incorrect then this may have contributed to the suboptimal fit in some countries. Further, network analysis approaches to mental health can yield valuable information about which particular symptoms may be the highest priority intervention targets owing to their high levels of connectedness to other symptoms (Speyer et al., 2021). Future investigations of cross-cultural differences in the PHQ-9 using network approaches are, therefore, promising for yielding further valuable information about how depressive symptoms may manifest and be captured by psychometric instruments across the world.

Future Directions

The current study focused on the reliability/validity of the PHQ-9 within, and cross-country measurement invariance of the PHQ-9, across antenatal populations in eight low- and middle-income countries. Important extensions to this work would include an evaluation of measures of depressive symptoms across a wider range of country contexts and in other types of respondents. Depressive symptoms during pregnancy are not necessarily equivalent to those experienced at other stage of life.

Indeed, pregnancy represents an important life event associated with social, relationship, and hormonal changes, as well as other pregnancy-specific risk factors such as unwanted or coerced pregnancy and prenatal intimate partner violence that can impact mental health (Atif et al., 2021; Brown et al., 2021; Murray et al., 2018; Yin et al., 2021). Further, due to the fact that pregnancy symptoms (e.g., appetite and sleep changes) can mask or be mistaken for some of the symptoms of antenatal depression, there are challenges in accurately detecting and quantifying symptoms during this period. A particularly valuable future direction would therefore be to examine the measurement invariance of the PHQ-9 across the pre-conception, antenatal, and post-partum periods and beyond in diverse country contexts in order to evaluate the extent to which it is a suitable tool for tracking changes in depressive symptoms associated with the not only pregnancy but the periods that precede it and follow it. Such a tool would be invaluable for tracking the long-term effects of interventions to improve maternal mental health in different country contexts.

Conclusions

The PHQ-9 appears to yield reliable scores with high convergent validity for depressive symptoms across a set of culturally diverse contexts assessed in the current study. Not all countries, however, demonstrated factorial validity with respect to the hypothesised single-factor structure of the PHQ-9. Further, differences in factor structure and lack of measurement invariance may affect the use of scores for valid cross-country comparisons. To improve the cross-cultural measurement of depressive symptoms, future studies could aim to identify a pool of cross-culturally invariant items in established measures to facilitate robust cross-country comparisons, and develop and test new candidate 'universal' items alongside items capturing culturally-specific manifestations of depression. Greater use of latent variable models with non-invariance modelled could facilitate valid comparisons where some items are shown to be non-comparable across contexts.

References

Acquadro, C., Patrick, D. L., Eremenco, S., Martin, M. L., Kuliś, D., Correia, H., & Conway, K. (2018). Emerging good practices for translatability assessment (TA) of patient-reported outcome (PRO) measures. *Journal of Patient-Reported Outcomes*, *2*(1), 1–11.

Adler, N. E., Epel, E. S., Castellazzo, G., & Ickovics, J. R. (2000). Relationship of subjective and objective social status with psychological and physiological functioning: Preliminary data in healthy, White women. *Health Psychology*, *19*(6), 586.

Aggarwal, S., Taljard, L., Wilson, Z., & Berk, M. (2017). Evaluation of modified patient health questionnaire-9 teen in South African adolescents. *Indian Journal of Psychological Medicine*, *39*(2), 143.

Ali, G.-C., Ryan, G., & De Silva, M. J. (2016). Validated screening tools for common mental disorders in low and middle income countries: A systematic review. *PloS One*, *11*(6), e0156939.

Association, A. P. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Atif, M., Halaki, M., Raynes-Greenow, C., & Chow, C.-M. (2021). Perinatal depression in Pakistan: A systematic review and meta-analysis. *Birth*, *48*(2), 149–163.

Bader, M., Jobst, L. J., Zettler, I., Hilbig, B. E., & Moshagen, M. (2021). Disentangling the effects of culture and language on measurement noninvariance in cross-cultural research: The culture, comprehension, and translation bias (CCT) procedure. *Psychological Assessment*.

Barthel, D., Barkmann, C., Ehrhardt, S., Schoppen, S., Bindt, C., & Group, I. C. S. (2015). Screening for depression in pregnant women from Côte d' Ivoire and Ghana: Psychometric properties of the Patient Health Questionnaire-9. *Journal of Affective Disorders*, *187*, 232–240.

Bautista, A. D., Pacayra, E. E., Sunico-Quesada, C. R., Reyes, M. E. S., & Davis, R. D. (2017). The fizzling effect: A phenomenological study on suicidality among filipino lesbian women and gay men. *Psychological Studies*, *62*(3), 334–343.

Bernard, O., Gibson, R. C., McCaw-Binns, A., Reece, J., Coore-Desai, C., Shakespeare-Pellington, S., & Samms-Vaughan, M. (2018). Antenatal depressive symptoms in Jamaica associated with limited perceived partner and other social support: A cross-sectional study. *PloS One*, *13*(3), e0194338.

Berry, J. W. (1969). On cross-cultural comparability. *International Journal of Psychology*, *4*(2), 119–128.

Bhana, A., Rathod, S. D., Selohilwe, O., Kathree, T., & Petersen, I. (2015). The validity of the Patient Health Questionnaire for screening depression in chronic care patients in primary health care in South Africa. *BMC Psychiatry*, *15*(1), 118.

Biaggi, A., Conroy, S., Pawlby, S., & Pariante, C. M. (2016). Identifying the women at risk of antenatal anxiety and depression: A systematic review. *Journal of Affective Disorders*, *191*, 62–77.

Bindt, C., Appiah-Poku, J., Te Bonle, M., Schoppen, S., Feldt, T., Barkmann, C., Koffi, M., Baum, J., Nguah, S. B., & Tagbor, H. (2012). Antepartum depression and anxiety associated with disability in African women: Cross-sectional results from the CDS study in Ghana and Côte d'Ivoire. *PloS One*, *7*(10), e48396.

Borsboom, D., & Cramer, A. O. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*, 91–121.

Brown, R. H., Eisner, M., Walker, S., Tomlinson, M., Fearon, P., Dunne, M. P., Valdebenito, S., Hughes, C., Ward, C. L., & Sikander, S. (2021). The impact of maternal adverse childhood experiences and prenatal depressive symptoms on foetal attachment: Preliminary evidence from expectant mothers across eight middle-income countries. *Journal of Affective Disorders*, *295*, 612–619.

Carroll, H. A., Hook, K., Perez, O. F. R., Denckla, C., Vince, C. C., Ghebrehiwet, S., Ando, K., Touma, M., Borba, C. P., & Fricchione, G. L. (2020). Establishing Reliability and Validity for Mental Health Screening Instruments in Resource-Constrained Settings: Systematic Review of the PHQ-9 and Key Recommendations. *Psychiatry Research*, 113236.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504.

Cholera, R., Gaynes, B. N., Pence, B. W., Bassett, J., Qangule, N., Macphail, C., Bernhardt, S., Pettifor, A., & Miller, W. C. (2014). Validity of the patient health questionnaire-9 to screen for depression in a high-HIV burden primary healthcare clinic in Johannesburg, South Africa. *Journal of Affective Disorders*, *167*, 160–166.

Cohen, S. (1988). *Perceived stress in a probability sample of the United States.*

Colucci, E., Kelly, C. M., Minas, H., Jorm, A. F., & Nadera, D. (2010). Mental Health First Aid guidelines for helping a suicidal person: A Delphi consensus study in the Philippines. *International Journal of Mental Health Systems*, *4*(1), 1–9.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

de Jong, J. A., Dorer, B., Lee, S., Yan, T., & Villar, A. (2018). Overview of questionnaire design and testing. *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*, *115*.

Dere, J., Watters, C. A., Yu, S. C.-M., Bagby, R. M., Ryder, A. G., & Harkness, K. L. (2015). Cross-cultural examination of measurement invariance of the Beck Depression Inventory–II. *Psychological Assessment*, *27*(1), 68.

Ediriweera, H. W., Fernando, S. M., & Pai, N. B. (2012). Mental health literacy survey among Sri Lankan carers of patients with schizophrenia and depression. *Asian Journal of Psychiatry*, *5*(3), 246–250.

Epskamp, S. (2020). Psychometric network models from time-series and panel data. *Psychometrika*, 1–26.

Epskamp, S., Cramer, A. O., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, *48*(4), 1–18.

Estrada, C. A. M., Nonaka, D., Gregorio, E. R., Leynes, C. R., Del Castillo, R. T., Hernandez, P. M. R., Hayakawa, T., & Kobayashi, J. (2019). Suicidal ideation, suicidal behaviors, and attitudes towards suicide of adolescents enrolled in the Alternative Learning System in Manila, Philippines—A mixed methods study. *Tropical Medicine and Health*, *47*(1), 22.

Fekadu Dadi, A., Miller, E. R., & Mwanri, L. (2020). Antenatal depression and its association with adverse birth outcomes in low and middle-income countries: A systematic review and meta-analysis. *PloS One*, *15*(1), e0227323.

Fisher, J., Mello, M. C. de, Patel, V., Rahman, A., Tran, T., Holton, S., & Holmes, W. (2012). Prevalence and determinants of common perinatal mental disorders in women in low-and lower-middle-income countries: A systematic review. *Bulletin of the World Health Organization*, *90*, 139–149.

Flynn, H. A., Sexton, M., Ratliff, S., Porter, K., & Zivin, K. (2011). Comparative performance of the Edinburgh Postnatal Depression Scale and the Patient Health Questionnaire-9 in pregnant and postpartum women seeking psychiatric services. *Psychiatry Research*, *187*(1–2), 130–134.

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191–197.

Gallis, J. A., Maselko, J., O'Donnell, K., Song, K., Saqib, K., Turner, E. L., & Sikander, S. (2018). Criterion-related validity and reliability of the Urdu version of the patient health questionnaire in a sample of community-based pregnant women in Pakistan. *PeerJ*, *6*, e5185.

Garabiles, M. R., Lao, C. K., Yip, P., Chan, E. W., Mordeno, I., & Hall, B. J. (2019). Psychometric Validation of PHQ–9 and GAD–7 in Filipino Migrant Domestic Workers in Macao (SAR), China. *Journal of Personality Assessment*, 1–12.

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, *74*(1), 155–167.

Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software*, *17*(1), 1–27.

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in M plus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638.

Hanwella, R., Ekanayake, S., & de Silva, V. A. (2014). The validity and reliability of the Sinhala translation of the Patient Health Questionnaire (PHQ-9) and PHQ-2 screener. *Depression Research and Treatment*, *2014*.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29–29.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Hue, M. T., Nguyet Van, N. H., Nha, P. P., Vu, N. T., Duc, P. M., Van Trang, N. T., Thinh, P. T. N., Anh, L. N., Huyen, L. T., & Tu, N. H. (2020). Factors associated with antenatal depression among pregnant women in Vietnam: A multisite cross-sectional survey. *Health Psychology Open*, *7*(1), 2055102920914076.

Jennrich, R. I. (1970). An asymptotic χ2 test for the equality of two correlation matrices. *Journal of the American Statistical Association*, *65*(330), 904–912.

Kelley, K., & Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: Evaluation of methods, recommendations, and software for composite measures. *Psychological Methods*, *21*(1), 69.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling*, *18*(2), 212–228.

Kohrt, B. A., Jordans, M. J., Tol, W. A., Luitel, N. P., Maharjan, S. M., & Upadhaya, N. (2011). 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*, *11*(1), 127.

Kohrt, B. A., Luitel, N. P., Acharya, P., & Jordans, M. J. (2016). Detection of depression in low resource settings: Validation of the Patient Health Questionnaire (PHQ-9) and cultural concepts of distress in Nepal. *BMC Psychiatry*, *16*(1), 58.

Kohrt, B. A., Rasmussen, A., Kaiser, B. N., Haroz, E. E., Maharjan, S. M., Mutamba, B. B., De Jong, J. T., & Hinton, D. E. (2014). Cultural concepts of distress and psychiatric disorders: Literature review and research recommendations for global mental health epidemiology. *International Journal of Epidemiology*, *43*(2), 365–406.

Lietz, P. (2010). Research into questionnaire design: A summary of the literature. *International Journal of Market Research*, *52*(2), 249–272.

Mahendran, R., Puthussery, S., & Amalan, M. (2019). Prevalence of antenatal depression in South Asia: A systematic review and meta-analysis. *J Epidemiol Community Health*, *73*(8), 768–777.

McDonald, R. P. (1999). *Test theory: A unified treatment.*

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412.

Murray, A. L., Kaiser, D., Valdebenito, S., Hughes, C., Baban, A., Fernando, A. D., Madrid, B., Ward, C. L., Osafo, J., & Dunne, M. (2018). The intergenerational effects of intimate partner violence in pregnancy: Mediating pathways and implications for prevention. *Trauma, Violence, & Abuse*, 1524838018813563.

Muthén, L. K., & Muthén, B. (2015). Mplus. *The Comprehensive Modelling Program for Applied Researchers: User's Guide*, *5*.

Nguyen, T. Q., Bandeen-Roche, K., Bass, J. K., German, D., Nguyen, N. T. T., & Knowlton, A. R. (2016). A tool for sexual minority mental health research: The Patient Health Questionnaire (PHQ-9) as a depressive symptom severity measure for sexual minority women in Viet Nam. *Journal of Gay & Lesbian Mental Health*, *20*(2), 173–191.

Pokropek, A., Davidov, E., & Schmidt, P. (2019). A monte carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(5), 724–744.

Redaniel, M. T., Lebanan-Dalida, M. A., & Gunnell, D. (2011). Suicide in the Philippines: Time trend analysis (1974-2005) and literature review. *BMC Public Health*, *11*(1), 536.

Revelle, W. R. (2017). *psych: Procedures for personality and psychological research*.

Ryder, A. G., Yang, J., Zhu, X., Yao, S., Yi, J., Heine, S. J., & Bagby, R. M. (2008). The cultural shaping of depression: Somatic symptoms in China, psychological symptoms in North America? *Journal of Abnormal Psychology*, *117*(2), 300.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23–74.

Shin, C., Ko, Y.-H., An, H., Yoon, H.-K., & Han, C. (2020). Normative data and psychometric properties of the Patient Health Questionnaire-9 in a nationally representative Korean population. *BMC Psychiatry*, *20*, 1–10.

Shrestha, S. D., Pradhan, R., Tran, T. D., Gualano, R. C., & Fisher, J. R. (2016). Reliability and validity of the Edinburgh Postnatal Depression Scale (EPDS) for detecting perinatal common mental disorders (PCMDs) among women in low-and lower-middle-income countries: A systematic review. *BMC Pregnancy and Childbirth*, *16*(1), 1–19.

Sidebottom, A. C., Harrison, P. A., Godecker, A., & Kim, H. (2012). Validation of the Patient Health Questionnaire (PHQ)-9 for prenatal depression screening. *Archives of Women's Mental Health*, *15*(5), 367–374.

Speyer, L. G., Eisner, M., Ribeaud, D., Luciano, M., Auyeung, B., & Murray, A. L. (2021). Developmental relations between internalising problems and ADHD in childhood: A symptom level perspective. *Research on Child and Adolescent Psychopathology*, *In press*.

Spitzer, R. L., Williams, J. B., Kroenke, K., Hornyak, R., McMurray, J., & Group, P. H. Q. O.-G. S. (2000). Validity and utility of the PRIME-MD patient health questionnaire in assessment of 3000 obstetric-gynecologic patients: The PRIME-MD Patient Health Questionnaire Obstetrics-Gynecology Study. *American Journal of Obstetrics and Gynecology*, *183*(3), 759–769.

Stevanovic, D., Urbán, R., Atilola, O., Vostanis, P., Balhara, Y. S., Avicenna, M., Kandemir, H., Knez, R., Franic, T., & Petrov, P. (2015). Does the Strengths and Difficulties Questionnaire–self report yield invariant measurements across different nations? Data from the International Child Mental Health Study Group. *Epidemiology and Psychiatric Sciences*, *24*(4), 323–334.

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: An illustration using M plus and the lavaan/semTools packages. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 111–130.

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, *72*(2), 271–324.

Topp, C. W., Østergaard, S. D., Søndergaard, S., & Bech, P. (2015). The WHO-5 Well-Being Index: A systematic review of the literature. *Psychotherapy and Psychosomatics*, *84*(3), 167–176.

Valdebenito, S., Murray, A., Hughes, C., Băban, A., Fernando, A. D., Madrid, B. J., Ward, C., Osafo, J., Dunne, M., & Sikander, S. (2020). Evidence for Better Lives Study: A comparative birth-cohort study on child exposure to violence and other adversities in eight low-and middle-income countries-foundational research (study protocol). *BMJ Open*, *10*(10), e034986.

Webster, G. D., DeWall, C. N., Pond Jr, R. S., Deckman, T., Jonason, P. K., Le, B. M., Nichols, A. L., Schember, T. O., Crysel, L. C., & Crosier, B. S. (2014). The brief aggression questionnaire: Psychometric and behavioral evidence for an efficient measure of trait aggression. *Aggressive Behavior*, *40*(2), 120–139.

World health statistics 2020: Monitoring health for the SDGs, sustainable development goals. (n.d.). Retrieved November 22, 2020, from https://www.who.int/publications-detail-redirect/9789240005105

Yin, X., Sun, N., Jiang, N., Xu, X., Gan, Y., Zhang, J., Qiu, L., Yang, C., Shi, X., & Chang, J. (2021). Prevalence and associated factors of antenatal depression: Systematic reviews and meta-analyses. *Clinical Psychology Review*, *83*, 101932.

Zhong, Q., Gelaye, B., Rondon, M., Sánchez, S. E., García, P. J., Sánchez, E., Barrios, Y. V., Simon, G. E., Henderson, D. C., & Cripe, S. M. (2014). Comparative performance of patient health questionnaire-9 and Edinburgh Postnatal Depression Scale for screening antepartum depression. *Journal of Affective Disorders*, *162*, 1–7.

Table 1: Participant demographic information

	Ghana	Jamaica	Pakistan	Philippines	Romania	South Africa	Sri Lanka	Vietnam
N	145	152	147	154	150	150	152	150
Age	M=29.1	M=25.5	M=27.3	M=27.7	M=30.0	M=29.7	M=29.7	M=29.9
	SD=6.3	SD=5.7	SD=5.2	SD=6.0	SD= 4.6	SD=6.0	SD=5.5	SD=5.2
Subjective SES[a]	M=4.8	M=5.1	M=4.6	M=4.6	M=7.0	M=4.7	M=6.0	M=5.1
	SD=2.4	SD=1.8	SD=2.0	SD=1.9	SD=1.2	SD=2.0	SD=2.1	SD=1.4
%1st pregnancy	19	27	12	25	51	35	42	29
% clinically depressed[b]	5	13	11	8	2	8	4	3

Note. [a]SES= socioeconomic status based on the 10-point MacArthur Scale of Subjective Social Status, where higher scores mean higher perceived status (Adler et al., 2000). [b]Estimate based on percent scoring above PHQ-9 cut-off of 14.

Table 2: PHQ-9 item contents and descriptive statistics by country

Item	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
	Ghana		**Jamaica**		**Pakistan**		**Philippines**		**Romania**		**South Africa**		**Sri Lanka**		**Vietnam**	
Pleasure	1.89	0.94	2.32	1.07	2.10	1.19	2.07	0.97	1.65	0.59	1.65	0.59	1.70	0.75	1.65	0.61
Hopeless	1.77	0.96	2.03	1.04	1.91	1.18	1.80	0.95	1.27	0.52	1.27	0.52	1.63	0.80	1.33	0.58
Sleep	2.38	0.99	2.43	1.15	2.35	1.36	2.39	0.99	2.18	0.88	2.18	0.88	2.04	0.88	2.20	0.85
Energy	2.39	0.89	2.67	1.05	3.05	1.16	2.22	0.93	2.18	0.76	2.18	0.76	2.04	0.87	1.99	0.77
Eating	2.01	0.99	2.26	1.14	2.39	1.34	1.92	1.07	1.85	0.82	1.85	0.82	1.72	0.83	1.81	0.87
Failure	1.65	0.91	1.55	0.97	1.37	0.84	1.66	0.92	1.19	0.49	1.19	0.49	1.30	0.58	1.27	0.55
Concentrate	1.52	0.85	1.71	1.10	1.37	0.86	1.66	0.93	1.41	0.75	1.41	0.75	1.30	0.64	1.51	0.70
Moving	1.48	0.75	1.73	0.99	1.59	0.93	1.56	0.86	1.35	0.58	1.35	0.58	1.47	0.73	1.40	0.61
Thoughts	1.32	0.80	1.36	0.83	1.13	0.46	1.17	0.51	1.03	0.16	1.03	0.16	1.17	0.56	1.06	0.33

Total score	7.38	4.69	9.04	5.10	8.24	5.44	7.40	4.58	5.11	3.49	6.82	5.07	5.37	4.15	5.19	3.68
Omega reliability (95% CI)	.73 (.60-.80)		.71 (.53-.79)		.74 (.63-.80)		.77 (.68-.83)		.81 (.67-.88)		.74 (.62-.80)		.84 (.77-.88)		.82 (.70-.87)	

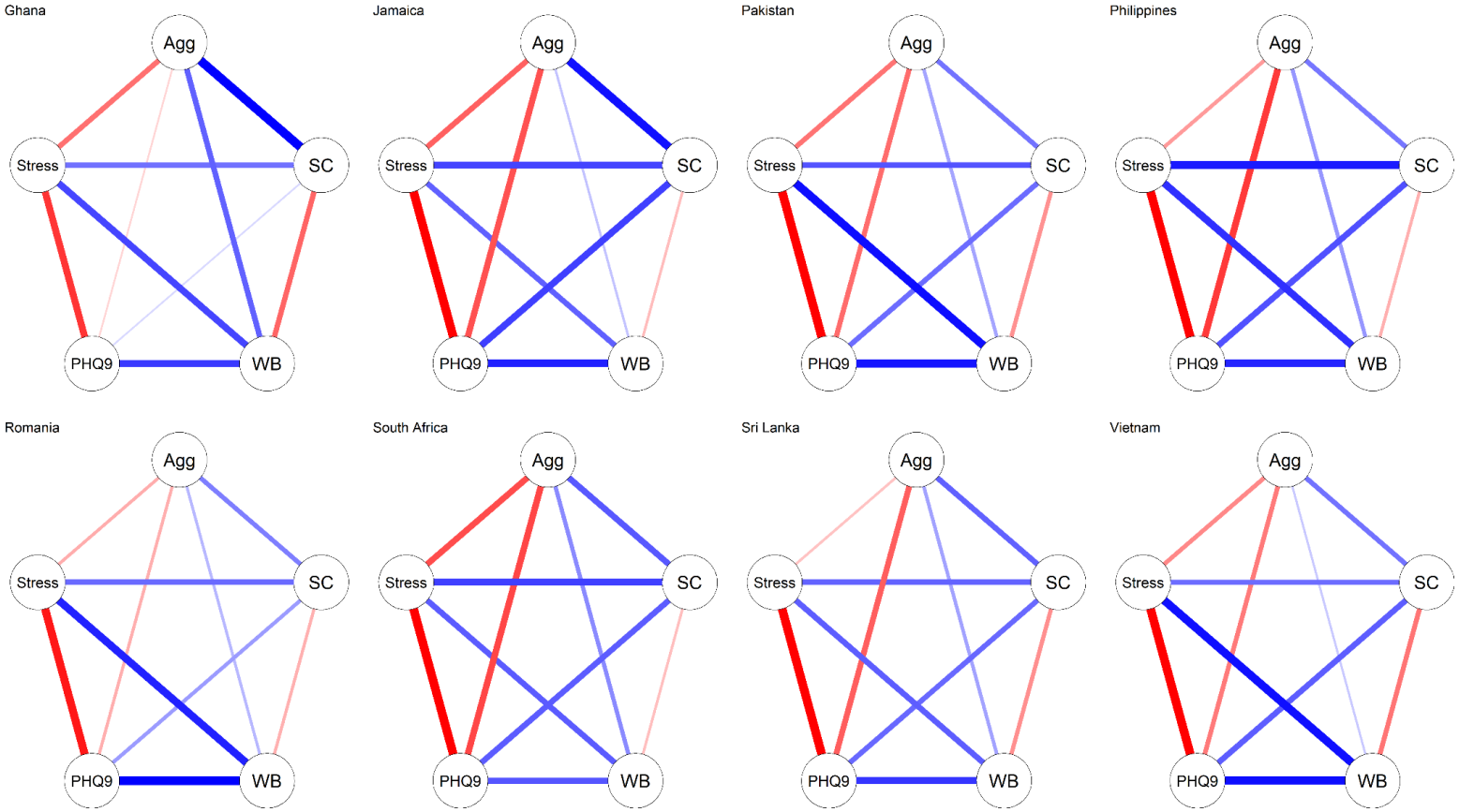
Note. Items are measured on a 4-point Likert-type scale with 1 = *not at all* to 4 = nearly every day. Categorial omega reliability 95% confidence intervals (CIs) were calculated using the bias corrected and accelerated bootstrap method described in (Kelley & Pornprasertmanit, 2016). The total score for the PHQ-9 is based on the sum of all item scores. This was done prior to collapsing some PHQ-9 item categories to facilitate measurement invariance analysis to provide comparability of scores with established standards and other studies using the PHQ-9.

Table 3: Unstandardized loading estimates for the four countries compared in a measurement invariance analysis

	Philippines			**South Africa**			**Sri Lanka**			**Vietnam**		
	Estimate	**SE**	***p***	**Estimate**	**SE**	***p***	**Estimate**	**SE**	***p***	**Estimate**	**SE**	***p***
Pleasure	0.416	0.061	<.001	0.416	0.061	<.001	0.416	0.061	<.001	0.416	0.061	<.001
Hopeless	0.705	0.056	<.001	0.705	0.056	<.001	0.705	0.056	<.001	0.705	0.056	<.001
Sleep	0.464	0.070	<.001	**1.334**	**0.316**	**<.001**	**0.789**	**0.170**	**<.001**	0.464	0.07	<.001
Energy	0.582	0.062	<.001	**1.370**	**0.438**	**.002**	0.582	0.062	<.001	0.582	0.062	<.001
Eating	0.427	0.080	<.001	0.427	0.080	<.001	0.427	0.08	<.001	0.427	0.08	<.001
Failure	0.635	0.066	<.001	0.635	0.066	<.001	0.635	0.066	<.001	0.635	0.066	<.001
Concentration	0.600	0.071	<.001	0.600	0.071	<.001	0.600	0.071	<.001	**10.522**	**40.667**	**.796**
Moving	0.681	0.074	<.001	0.681	0.074	<.001	0.681	0.074	<.001	0.681	0.074	<.001
Thoughts	**0.525**	**0.124**	**<.001**	1.104	0.166	<.001	1.104	0.166	<.001	1.104	0.166	<.001

Note. Table shows both invariant and non-invariant loadings across countries. Non-invariant parameters are in bold. Standardized estimates are provided in Supplementary Materials.

Figure 1: Nomological networks for PHQ-9 scores in each country



Note. Agg= aggression; SC= self-control; WB= well-being; PHQ9= PHQ-9 scores. Red indicates a positive association and blue indicates a negative association, with edge thicknesses proportional to the magnitude of the Pearson correlations between variables (represented as nodes).