RESEARCH ARTICLE

Measuring depression severity in global mental health: comparing the PHQ-9 and the BDI-II [version 1; peer review: 1 approved with reservations, 1 not approved]

Benedict Weobong [1-3], Helen A. Weiss [4], Isobel M. Cameron[5], Simon Kung[6], Vikram Patel [1,7], Steven D. Hollon [8]

[1]Adult Mental Health, Sangath, Socorro Village, Bardez-Porvorim, Goa, 403501, India
[2]College of Health Sciences, School of Public Health, Department of Social and Behavioural Sciences, University of Ghana, Accra, Accra, LG 25, Ghana
[3]Centre for Global Mental Health, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, London, London, WC1E 7HT, UK
[4]MRC Tropical Epidemiology Group, Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, London, WC1E 7HT, UK
[5]School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Foresterhill, Foresterhill, AB25 2ZD, UK
[6]Department of Psychiatry and Psychology, Mayo Clinic, Rochester, Minnesota, 55905, USA
[7]Department of Global Health and Social Medicine, Harvard Medical School, Boston, Boston, USA
[8]Department of Psychology, Vanderbilt University, Nashville, TN, USA

v1 **First published:** 28 Dec 2018, **3**:165 (https://doi.org/10.12688/wellcomeopenres.14978.1)
Latest published: 28 Dec 2018, **3**:165 (https://doi.org/10.12688/wellcomeopenres.14978.1)

Abstract

Background: We recently completed a randomised controlled trial in Goa India in which we observed a pattern of discordance with our two primary outcome measures; the Beck Depression Inventory (BDI-II) classified patients as moderately severe at the end of treatment, whilst the Patient Health Questionnaire (PHQ-9) classified these same patients as being only mildly depressed. The aim of this study is to explore whether the disparity between these two measures is seen in other settings.
Method: The relationship between BDI-II and PHQ-9 scores was compared between the Indian trial and two other studies (from United Kingdom and United States) that administered both measures to patients. Linear regression was used to quantify the non-concordance between the two measures across studies. Patients were classified by severity category on the BDI-II and PHQ-9, respectively, and relationship assessed using chi-square test. We further quantified the proportion assigned a higher category on the BDI-II than the PHQ-9 and assessed the difference in prevalence between studies using a test of proportions.
Results: Correlations between PHQ-9 and BDI-II were high and similar across studies (India: r=0.79; UK: r=0.87; US: r=0.77). Regression coefficients were similar across studies, but the predicted BDI-II mean score was significantly higher in the India study (24.3) compared to the US (20.5) or UK (20.8) studies. India participants had poorer outcomes on the BDI-II than the PHQ-9 and this difference was significant relative to both the UK (prevalence difference (PD): -15.9%; p<0.0001) and US studies (PD:

Open Peer Review

Reviewer Status ✗ ?

	Invited Reviewers	
	1	**2**
version 1 28 Dec 2018	✗ report	? report

1 **William Mellick**, Medical University of South Carolina (MUSC), Charleston, USA

2 **Kurt Kroenke**, Indiana University School of Medicine, Indianapolis, USA

Any reports and responses or comments on the article can be found at the end of the article.

-15.8%; p<0.0001).
Conclusions: The BDI-II and PHQ-9 measures are highly correlated, but the BDI-II tends to assign high severity scores in an Indian sample compared to UK/US samples. Where it is necessary to read items to patients, it seems likely that the PHQ-9 is a more accurate measure given that the BDI-II is longer and more complex.

Keywords
depression, severity, measurement, global mental health, BDI-II, PHQ-9

Corresponding author: Benedict Weobong (bkweobong@gmail.com)

Author roles: Weobong B: Conceptualization, Data Curation, Formal Analysis, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Weiss HA**: Data Curation, Formal Analysis, Methodology, Writing – Review & Editing; **Cameron IM**: Data Curation, Methodology, Writing – Review & Editing; **Kung S**: Data Curation, Methodology, Writing – Review & Editing; **Patel V**: Data Curation, Funding Acquisition, Methodology, Writing – Review & Editing; **Hollon SD**: Conceptualization, Methodology, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This research was funded by a Wellcome Trust Senior Research Fellowship grant to VP [091834]. BW is supported through an Intermediate Research Fellowship from the Wellcome Trust/India Alliance [502680].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Weobong B, Weiss HA, Cameron IM *et al.* **Measuring depression severity in global mental health: comparing the PHQ-9 and the BDI-II [version 1; peer review: 1 approved with reservations, 1 not approved]** Wellcome Open Research 2018, **3**:165 (https://doi.org/10.12688/wellcomeopenres.14978.1)

First published: 28 Dec 2018, **3**:165 (https://doi.org/10.12688/wellcomeopenres.14978.1)

Introduction

We recently completed a randomised controlled trial in Goa India comparing a culturally-adapted version of behavioural activation called the Healthy Activity Program (HAP) (Chowdhary *et al.*, 2016) plus Enhanced Usual Care (EUC) delivered by lay counsellors to EUC alone (Patel *et al.*, 2017). The authors found that HAP plus EUC was superior to EUC alone in treating moderate to severe depression both at the short-term (3-months post-randomization) (Patel *et al.*, 2017) and long-term (12-months post-randomization) (Weobong *et al.*, 2017) in general practice settings. Both primary outcome measures of depression, the revised Beck Depression Inventory (BDI-II) (Beck *et al.*, 1996) and the Patient Health Questionnaire (PHQ-9) (Spitzer *et al.*, 1999) showed superiority of the HAP plus EUC over EUC at both of these time-points. However, we observed a pattern of discordance in terms of depression severity between our two depression measures at both 3 months and again at 12 months; the modal patient was at the low end of the moderate range of severity on the BDI-II, whereas the same patient was indicated as having only mild residual symptoms on the PHQ-9. The aim of this study is to explore this discrepancy, since it has implications for how effective HAP is seen in absolute terms and as both measures are widely used.

We therefore searched the literature for other studies that administered both measures to the same participants and found two, one conducted in the United Kingdom (UK) (Cameron *et al.*, 2011) and the other in the United States (US) (Kung *et al.*, 2013). We contacted the lead authors of both studies and invited them to join us in investigating this discrepancy by virtue of sharing their patient level data and both compiled. If one of the measures is problematic, then perhaps it should not be used in other cultures. This is particularly important given there are growing concerns regarding the validity of measures for assessing severity of depression (Cameron *et al.*, 2008; Cameron *et al.*, 2011; Reddy *et al.*, 2010), and little by way of evidence on the objective psychometric comparison of these outcome measures.

Global mental health depends on the use of culturally appropriate measures if we are to accurately assess the burden of depression, and more importantly improve treatment plans/decision-making. In this paper, we address two questions: whether the discrepancy in terms of absolute scores observed in the India trial is similar in the other two UK and US studies, and whether the proportion of patients for whom the BDI-II score observed is classified in a higher severity category than the PHQ-9 score differs across the studies.

Methods

Only studies that used both the BDI-II and PHQ-9 as measures were eligible for the analysis in this paper. Both measures are endorsed by the National Institute for Health and Clinical Excellence to measure baseline depression severity and responsiveness to treatment in primary care (Smarr & Keefer, 2011).

Approvals were obtained for the collection and use of the primary data (including additional studies such as this study) for each of the studies. Consent was also provided by all participants in each of the studies involved in this analysis. For India, ethics approval was sought from the Indian Council of Medical Research, the Sangath Institutional Review Board (IRB), and the London School of Hygiene and Tropical Medicine. For UK, ethics approval was sought from the North of Scotland Research Ethics Committee. For US, ethics approval was sought from Mayo Clinic Department of Psychiatry and Psychology IRB.

Participants

The Indian study consisted of 438 participants (a subset seen at 3 and 12 months outcome time-points) of either gender aged 18–65 with probable diagnoses of moderately severe and severe depression based on PHQ-9 scores greater than 14 at baseline (Patel *et al.*, 2017). The BDI-II was not administered at baseline. Participants were all drawn from a parallel arm comparison of HAP plus EUC to EUC alone conducted in 10 primary health centres in the state of Goa on the west coast of India. All scores were drawn from the 3- and 12-month post-treatment assessments at the end of the trial.

The UK sample consisted of 267 participants of either gender aged 16 and above with diagnoses of depression as ascertained by their general practitioner (Cameron *et al.*, 2011). The study compared the performance of three different self-report measures of depression (the BDI-II and the PHQ-9) with a widely used clinician-rated instrument - the Hamilton Rating Scale for Depression (Hamilton, 1960). Participants who could not read the self-report measures because they were illiterate were ineligible for the study.

The US sample consisted of 625 depressed participants of either gender, aged 18–76 years (338 inpatients and 287 outpatients) (Kung *et al.*, 2013). The BDI-II and PHQ-9 were collected as part of routine clinical care and analysed retrospectively to compare their performance in that setting. As in the UK sample both scales were self-administered in English by participants who could read.

Measures

The BDI-II consists of 21 items covering a number of symptoms of depression. Each of the 21 items assess a different symptom with four different response options each a full sentence long. For example, the first item “Sad” is followed by response options ranging from: “0 – I do not feel sad.” “1 – I feel sad much of the time.” “2 – I am sad all the time.” “3 – I am so sad or unhappy that I can’t stand it” with total scores found by summing the highest response to each given item. The BDI-II has strong psychometric properties and historically is the most widely used self-report outcome measure of depression in trials. The BDI-II defines symptom severity at four levels recommended by Beck (Beck *et al.*, 1996), and in reference to the structured clinical interview for the Diagnostic and Statistical Manual of Mental Disorders, Third edition (Spitzer *et al.*, 1999): 0–13 Minimal Depression; 14–19 Mild Depression; 20–28 Moderate Depression; 29–63 Severe Depression. However, these were based on a sample drawn from a primary care site in University of Pennsylvania and may not generalize effectively to other primary

care settings, particularly in Low and Middle-Income Countries (LMIC) (Cameron *et al*., 2011).

The PHQ-9 is a structured questionnaire that enquires after the nine symptom-based criteria for a diagnosis of DSM-IV and DSM-5 depression. The instrument presents a common stem "Over the past two weeks how often have you been bothered by any of the following problems?" and then follows with nine specific questions such as "Little interest or pleasure in doing things". Each item is rated on a single four-point scale from "not at all" to "nearly every day" and total scores are summed across the items. Like the BDI-II, the PHQ-9 has been found to have good sensitivity and specificity (Kroenke *et al*., 2001) and is coming into increasing widespread use as a measure of depression severity. The PHQ-9 defines symptom severity at five levels recommended by Kroenke (Kroenke *et al*., 2001): 1–4 Minimal Depression; 5–9 Mild Depression; 10–14 Moderate Depression; 15–19 Moderately to Severe Depression; 20–27 Severe Depression.

Procedures

Both scales were administered as self-report instruments in the UK and US studies, the standard means of administration, and included all 21 items on the BDI-II. In the India study, because the vast majority of the participants were illiterate, study personnel read the items to the participants and recorded their responses in the three major local languages in the study area (Konkani/Marathi/Hindi). This followed a rigorous forward and back translation process consistent with the five major criteria for cross-cultural equivalence in psychiatric research: content equivalence, semantic equivalence, technical equivalence, criterion equivalence and conceptual equivalence (Flaherty *et al*., 1988). A forward translation was first completed by trained and experienced field researchers and these translations reviewed by a clinical psychologist fluent in the three local languages, together with senior and more experienced research team members, at the second stage. Where there were disagreements between the clinician and senior research team members on the quality of the forward translation, these were discussed with a psychiatrist with experience of working in both the UK and India-Goa, to advise on the concepts captured by the original English wording of each item to guide the choice of local language expressions. The draft consensus translation was then back-translated into English by a bilingual independent non-mental health professional, following which further modifications were made on the basis of the back-translation, if required. The item inquiring about interest in sex was omitted from the BDI-II in India so as not to offend participants.

Statistical analyses

We first estimated the reliability of each measure using Cronbach's alpha. Following this, we compared scores using Pearson product-moment correlation statistics in order to ascertain whether both measures were assessing the same construct of depression. In order to address our first objective regarding the observed discrepancy between the BDI-II and PHQ-9 scores in the India trial, we first examined the association between scores on the two measures within each study using linear regression. Following this, we assessed whether there was evidence of moderation by study by fitting an interaction term. We then used the predicted BDI-II score and modelled what this would be for participants with PHQ-9 score of 10 (moderate depression) for each study. Finally, we assessed if the intercepts differed between the three studies and generated scatter plots of the fitted values for the BDI-II and PHQ-9 for each sample. Effect sizes are reported as regression coefficients (with 95% CI) for the increase in BDI-II score for each unit increase in PHQ-9 score. In addition, to address our second objective patients were classified with respect to the prespecified depression severity categorical outcomes on each measure and rates of discordance compared across the studies, and the association was assessed with the chi-square statistic. We further explored the number and proportion with a higher category on the BDI-II than the PHQ-9 for each PHQ-9 category. We ruled out the possibility of temporal effects on the observed discrepancy in the India trial at the 3-month endpoint, by repeating the regression analysis using follow-up data of the same participants on the BDI-II and PHQ-9 at 12-months post-enrolment. Sensitivity analyses were conducted after dropping the sex item on the UK and US studies. Statistical analyses were conducted using STATA 15.

Results

A detailed description of the conduct of each study is provided in the respective publications (Cameron *et al*., 2011; Kung *et al*., 2013; Patel *et al*., 2017).

Reliability

The internal scale consistency (Cronbach's alpha) for the BDI-II and PHQ-9 were high in each study (India: BDI-II=0.91; PHQ-9=0.86; UK: BDI-II=0.94; PHQ-9=0.92; US: BDI-II=0.90; PHQ-9=0.83).

Construct validity

Correlations between BDI-II and PHQ-9 were high in each study (India: r=0.79, 95%CI 0.75-0.82; UK: r=0.87, 95%CI 0.84-0.90; US: r=0.77. 95%CI 0.73-0.80).

Score distribution

At the 3-month end-point for the India study, the regression coefficients were similar for the three studies (India: β=1.58, 95%CI 1.47-1.70; UK: β =1.58, 95%CI 1.47-1.70; US: β=1.48, 95%CI 1.39-1.58), and there was no evidence of moderation by study (p=0.32). As would be expected given differences in the scales, scores on the BDI-II were higher than on the PHQ-9 in each of the studies, but more so in the India study than in the other two (Figure 1). For example, at a PHQ-9 score of 10 (moderate depression) in the India study, the BDI-II mean score was 24.3 (95% CI 23.5, 25.1), and this was significantly different from the UK study 20.8 (95%CI 19.6, 21.9) and the US 20.5 (95%CI 19.5, 21.4).

Similar results were observed at the 12-month end-point for the India sample; the regression coefficient increased slightly to (β=1.67, 95%CI 1.56, 1.78) but the greater discrepancy between scores in the India study compared to the UK and US was maintained (Figure 2). At a PHQ-9 score of 10 (moderate

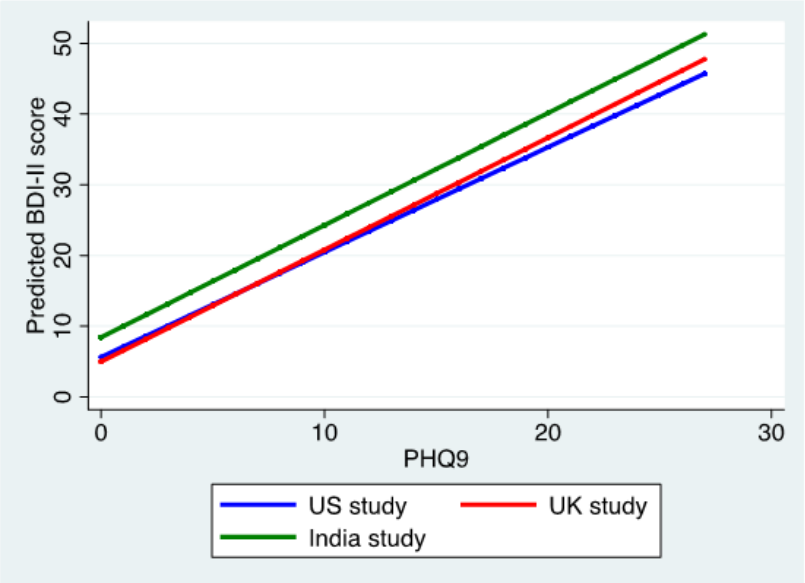

Figure 1. Scatter plot with fitted regression lines of BDI-II and PHQ-9 scores of the three studies (comparison with 3-month outcome data in India trial). Plot of regression model fitted with interaction term i.e. allowing slope of PHQ-9 with BDI-II to differ by study.

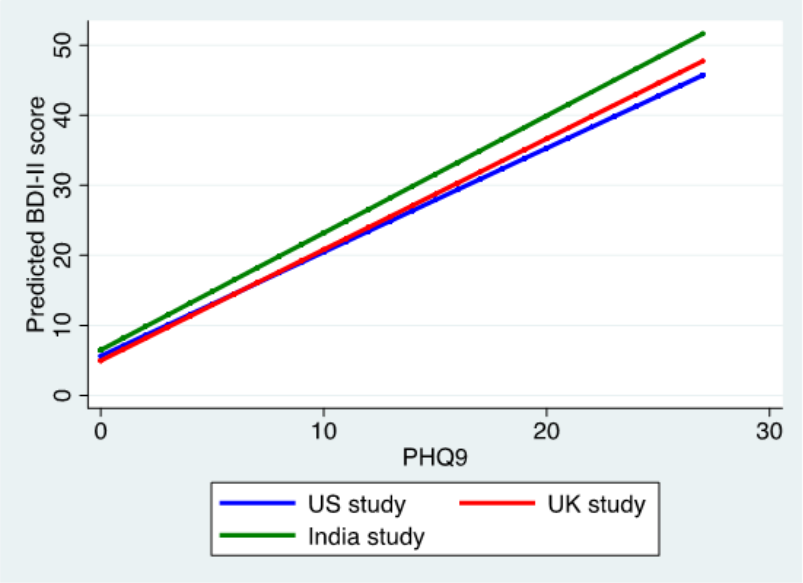

Figure 2. Scatter plot with fitted regression lines of BDI-II and PHQ-9 scores of the three studies (comparison with 12-month outcome data in India trial). Plot of regression model fitted with interaction term i.e. allowing slope of PHQ-9 with BDI-II to differ by study.

depression) in the India sample, the BDI-II mean score was 23.2 (95% CI 22.4, 23.9), still significantly different from the UK and US studies.

Severity banding

Table 1a–Table 1c show the cross-classification of individual participants on each of the categorical values used to describe absolute outcomes on the respective measures. As can be seen from Table 1a–Table 1c, the participants in the India sample were more likely to be classified as having poorer outcomes on the BDI-II than the PHQ-9 for each severity band of the PHQ-9, and this was significantly different between both the India and UK samples (prevalence difference (PD): -15.9%, 95% CI -23.2%, -8.7%; p<0.0001) and the India and US samples (PD: -15.8%, 95% CI -21.9%, -9.5%; p<0.0001).

Results were similar when the sex item was dropped from the UK and US studies.

Table 1a. Comparison of depression severity categories of the BDI-II versus the PHQ-9 in the India sample (N=438) showing row percentages.

PHQ-9 Categories	BDI-II Categories					
	Minimal Depression (score 0-13)	Mild Depression (score 14-19)	Moderate Depression (score 20-28)	Severe Depression (score 29-63)	Total PHQ-9 Category	*n (%) BDI-II higher category than PHQ-9
Minimal Depression (score 1-4)	103 (73%)	**20 (14%)**	**13 (09%)**	**5 (04%)**	141 (32%)	38/141 (26.9%)
Mild Depression (score 5-9)	**20 (24%)**	13 (15%)	**27 (32%)**	**24 (29%)**	84 (19%)	51/84 (60.7%)
Moderate Depression (score 10-14)	**6 (07%)**	12 (14%)	27 (31%)	**43 (48%)**	88 (20%)	43/88 (48.8%)
Moderately to Severe Depression (score 15-19)	**3 (04%)**	**0 (00%)**	**11 (14%)**	**64 (82%)**	78 (18%)	64/78 (82.1%)
Severe Depression (score 20-27)	**0 (00%)**	**0 (00%)**	**1 (02%)**	46 (98%)	47 (11%)	-
Total BDI	132 (30%)	45 (10.3%)	79 (18%)	182 (42%)	438 (100%)	**196/391 (50.1%)**

*Excluded the highest severity band of the PHQ-9 from this analysis because there was a disproportionate distribution of severe depression as assessed by the PHQ-9; the proportion was much higher in the US sample (40%) compared to the India (11%) and UK (17%) studies

Table 1b. Comparison of depression severity categories of the BDI-II versus the PHQ-9 in the UK sample (N=222) showing row percentages.

PHQ-9 Categories	BDI-II Categories					
	Minimal Depression (score 0-13)	Mild Depression (score 14-19)	Moderate Depression (score 20-28)	Severe Depression (score 29-63)	Total PHQ-9 Category	*n (%) BDI-II higher category than PHQ-9
Minimal Depression (score 1–4)	35 (88%)	**4 (10%)**	**1 (02%)**	**0 (00%)**	40 (18%)	5/40 (12.5%)
Mild Depression (score 5–9)	**12 (22%)**	27 (49%)	**14 (25%)**	**2 (04%)**	55 (25%)	16/55 (29.1%)
Moderate Depression (score 10–14)	**2 (04%)**	**11 (23%)**	23 (48%)	**12 (25%)**	48 (22%)	12/48 (25.0%)
Moderately to Severe Depression (score 15–19)	**2 (05%)**	**1 (02%)**	**12 (29%)**	**27 (64%)**	42 (19%)	27/42 (64.3%)
Severe Depression (score 20–27)	**0 (00%)**	**0 (00%)**	**0 (00%)**	37 (100%)	37 (17%)	-
Total BDI	51 (23%)	43 (19%)	50 (23%)	78 (35%)	222 (100%)	**60/185 (32.4%)**

*Excluded the highest severity band of the PHQ-9 from this analysis because there was a disproportionate distribution of severe depression as assessed by the PHQ-9; the proportion was much higher in the US sample (40%) compared to the India (11%) and UK (17%) studies.

Table 1c. Comparison of depression severity categories of the BDI-II versus the PHQ-9 in the US sample (N=625) showing row percentages.

PHQ-9 Categories	BDI-II Categories					
	Minimal Depression (score 0-13)	Mild Depression (score 14-19)	Moderate Depression (score 20-28)	Severe Depression (score 29-63)	Total PHQ-9 Category	*n (%) BDI-II higher category than PHQ-9
Minimal Depression (score 1-4)	30 (86%)	4 (11%)	1 (03%)	0 (00%)	35 (6%)	5/35 (14.3%)
Mild Depression (score 5-9)	28 (42%)	17 (25%)	16 (24%)	6 (09%)	67 (11%)	22/67 (32.8%)
Moderate Depression (score 10-14)	13 (14%)	16 (17%)	29 (31%)	35 (38%)	93 (15%)	35/93 (37.6%)
Moderately to Severe Depression (score 15-19)	5 (03%)	8 (05%)	40 (27%)	97 (65%)	150 (24%)	97/150 (64.7%)
Severe Depression (score 20-27)	1 (01%)	4 (01%)	25 (09%)	250 (89%)	280 (45%)	-
Total BDI	77 (12%)	49 (08%)	111 (18%)	388 (62%)	625 (100%)	159/345 (46.1%)

*Excluded the highest severity band of the PHQ-9 from this analysis because there was a disproportionate distribution of severe depression as assessed by the PHQ-9; the proportion was much higher in the US sample (40%) compared to the India (11%) and UK (17%) studies.

Discussion

Patients reported higher severity scores on the BDI-II relative to the PHQ-9 in our India sample than they did in either the UK or the US. We think this reflects differences in the method of administration across the studies; in India, we read the translated local language version items to our patients whereas in both the UK and the US studies literate patients read the items themselves. The BDI-II is a relatively complex instrument that requires participants to hold four different options in memory before giving a response to each item whereas the PHQ-9 requires only that the participants respond with the same simple frequency rating to each of its nine items. The BDI-II is sometimes criticized for being too transparent to respondents and thus easily faked by those wishing to present themselves in a favourable or unfavourable light, but that same critique is as likely to apply to the PHQ-9 as the BDI-II (Wang & Gorenstein, 2013).

The fact that correlations were high and comparable across the samples suggests that both measures were assessing the same underlying construct of depression, but the fact that scores on the BDI-II were higher relative to the PHQ-9 in our India study than in the other two studies suggests that absolute scores on the BDI-II are inflated relative to the PHQ-9. Given that participants in LMICs are often illiterate and would require interviewer administration, the PHQ-9 might be preferred over the BDI-II as a measure of depression severity. The PHQ-9 is also easily accessible as it is free, whereas the BDI-II is only available on purchase.

The strengths of this investigation include the cross-cultural approach and large sample sizes from well-designed studies. We acknowledge some limitations including our inability to account for the potential confounding effect of order of administration of the two measures (for the India and US studies) and other factors such as social desirability, educational attainment and sex of respondents (Cronbach, 1990). That being said order of administration was largely constant in the India study and though this may not have been the case for the US study, the large sample size and the randomness of which measure was completed first means it is unlikely order effects accounted for the differences. Moreover, comparison of the discordance among categorical responses on the two measures was complicated by the fact that the PHQ-9 defines five categories of depression while the BDI-II defines only four (the former adds a "moderately severe to severe" category). However, that difference in categorization was consistent across the studies and should not have contributed to differences in concordance. Additionally, the level of depression severity in the three studies may have influenced our findings. For example, the patients from the US study were either from the "Mood Clinic" or "Mood Disorder Unit" meaning they were referred or admitted for depression treatment. This might explain why the US sample had a disproportionately higher band of severely depressed PHQ-9. We however dealt with this by dropping this category from the analysis comparing the severity bands between the BDI-II and PHQ-9 in all three studies. Furthermore, we adjusted for study in our regression analysis. Finally, even though we strictly adhered to principles of cross-cultural psychiatric research in adapting the BDI-II and PHQ-9 in the India study, we are unable to completely rule out loss of meaning in translation. Admittedly, this limitation would apply to both measures though the BDI-II would pose more translation challenges given its relative

complexity. Dropping the BDI-II item on sex (sexual desire) in the India study could have offset its psychometric properties, and more importantly for this analysis meant that the samples may have been incomparable. However, dropping one item (sex) would likely result in lowering the mean score on the BDI-II in the India study but in sensitivity analysis we observed in the prediction model that in the India study the BDI-II scored people higher compared to the PHQ-9, and this was significantly higher compared to the UK/US studies. This suggests the robustness of our findings without the sex item, and we posit that the differences observed would have been much stronger if the sex item were maintained in the India study.

It is possible that it was the PHQ-9 that was problematic in our sample and not the BDI-II. Ours is the first study to examine head-to-head the severity categorisation of the PHQ-9 and BDI-II, comparing studies from high versus low and middle-income settings. The PHQ-9 is the simpler measure and places fewer demands on short-term memory than the BDI-II. Administering both measures orally in literate samples and seeing if that inflates absolute scores on the BDI-II relative to the PHQ-9 could resolve this issue. Such a study would be relatively easy to conduct and is encouraged given the reported concerns regarding the validity of measures for assessing severity of depression (Cameron *et al.*, 2008; Hansson *et al.*, 2009; Reddy *et al.*, 2010). Until such a study is done we have reservations about interpreting absolute values on the BDI-II and prefer to use the PHQ-9 instead. It may be the case that the PHQ-9 is more suitable as interviewer-administered in illiterate populations given findings from a study in Spain that the PHQ-9 performed similarly when read out over the phone compared to self-administration (Pinto-Meza *et al.*, 2005). What this could mean is that studies where illiteracy is a concern, particularly in LMICs, researchers might be well advised to use the less complicated PHQ-9 than the BDI-II if the scales must be read to illiterate participants. Both appear to be valid measures of the underlying construct when participants read and complete the scales themselves. Further work is required to assess their performance when read out to participants.

Ethical statement

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.

Data availability

The dataset that underpins the analysis in this paper is hosted in Sangath Data Repository (Sangath Data Bank) and can be made available upon request. In accordance with ethical constraints established when obtaining participant consent, data can only be made available to interested parties on condition that they sign an agreement stating that they will protect participant confidentiality. To request access, please submit a data request application using the online application form at http://www.sangath.in/application-form-for-requesting-datasets/.

Grant information

This research was funded by a Wellcome Trust Senior Research Fellowship grant to VP [091834]. BW is supported through an Intermediate Research Fellowship from the Wellcome Trust/India Alliance [502680].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We acknowledge the generous partnership and support of the Directorate of Health Services of the Government of Goa.

References

Beck AT, Steer RA, Ball R, *et al.*: **Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients.** *J Pers Assess.* 1996; **67**(3): 588–97.
PubMed Abstract | Publisher Full Text

Cameron IM, Cardy A, Crawford JR, *et al.*: **Measuring depression severity in general practice: discriminatory performance of the PHQ-9, HADS-D, and BDI-II.** *Br J Gen Pract.* 2011; **61**(588): e419–26.
PubMed Abstract | Publisher Full Text | Free Full Text

Cameron IM, Crawford JR, Lawton K, *et al.*: **Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care.** *Br J Gen Pract.* 2008; **58**(546): 32–6.
PubMed Abstract | Publisher Full Text | Free Full Text

Chowdhary N, Anand A, Dimidjian S, *et al.*: **The Healthy Activity Program lay counsellor delivered treatment for severe depression in India: systematic development and randomised evaluation.** *Br J Psychiatry.* 2016; **208**(4): 381–8.
PubMed Abstract | Publisher Full Text | Free Full Text

Cronbach LJ: **Essentials of psychological testing.** New York, Harper and Row. 1990.
Reference Source

Flaherty JA, Gaviria FM, Pathak D, *et al.*: **Developing instruments for cross-cultural psychiatric research.** *J Nerv Ment Dis.* 1988; **176**(5): 257–63.
PubMed Abstract | Publisher Full Text

Hamilton M: **A rating scale for depression.** *J Neurol Neurosurg Psychiatry.* 1960; **23**: 56–62.
PubMed Abstract | Publisher Full Text | Free Full Text

Hansson M, Chotai J, Nordstöm A, *et al.*: **Comparison of two self-rating scales to detect depression: HADS and PHQ-9.** *Br J Gen Pract.* 2009; **59**(566): e283–8.
PubMed Abstract | Publisher Full Text | Free Full Text

Kroenke K, Spitzer RL, Williams JB: **The PHQ-9: validity of a brief depression severity measure.** *J Gen Intern Med.* 2001; **16**(9): 606–13.
PubMed Abstract | Publisher Full Text | Free Full Text

Kung S, Alarcon RD, Williams MD, *et al.*: **Comparing the Beck Depression Inventory-II (BDI-II) and Patient Health Questionnaire (PHQ-9) depression measures in an integrated mood disorders practice.** *J Affect Disord.* 2013; **145**(3): 341–3.
PubMed Abstract | Publisher Full Text

Patel V, Weobong B, Weiss HA, *et al.*: **The Healthy Activity Program (HAP), a lay counsellor-delivered brief psychological treatment for severe depression, in primary care in India: a randomised controlled trial.** *Lancet.* 2017; **389**(10065): 176–185.
PubMed Abstract | Publisher Full Text | Free Full Text

Pinto-Meza A, Serrano-Blanco A, Penarrubia MT, *et al.*: **Assessing depression in primary care with the PHQ-9: can it be carried out over the telephone?** *J Gen*

Intern Med. 2005; **20**(8): 738–42.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Reddy P, Philpot B, Ford D, *et al.*: **Identification of depression in diabetes: the efficacy of PHQ-9 and HADS-D.** *Br J Gen Pract.* 2010; **60**(575): e239–45.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Smarr KL, Keefer AL: **Measures of depression and depressive symptoms: Beck Depression Inventory-II (BDI-II), Center for Epidemiologic Studies Depression Scale (CES-D), Geriatric Depression Scale (GDS), Hospital Anxiety and Depression Scale (HADS), and Patient Health Questionnaire-9 (PHQ-9).** *Arthritis Care Res (Hoboken).* 2011; **63**(Suppl 11): S454–66.
PubMed Abstract | **Publisher Full Text**

Spitzer RL, Kroenke K, Williams JB: **Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire.** *JAMA.* 1999; **282**(18): 1737–44.
PubMed Abstract | **Publisher Full Text**

Wang YP, Gorenstein C: **Psychometric properties of the Beck Depression Inventory-II: a comprehensive review.** *Braz J Psychiatr.* 2013; **35**(4): 416–31.
PubMed Abstract | **Publisher Full Text**

Weobong B, Weiss HA, McDaid D, *et al.*: **Sustained effectiveness and cost-effectiveness of the Healthy Activity Programme, a brief psychological treatment for depression delivered by lay counsellors in primary care: 12-month follow-up of a randomised controlled trial.** *PLoS Med.* 2017; **14**(9): e1002385.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ✗ ?

Version 1

Reviewer Report 15 April 2020

https://doi.org/10.21956/wellcomeopenres.16337.r34767

?

Kurt Kroenke
Indiana University School of Medicine, Indianapolis, IN, USA

This is a useful paper comparing two of the more commonly used depression measures in both practice as well as research. Strengths include comparing the PHQ-9 and BDI across studies from both the UK and US, examining differences in terms of severity classification, and testing the correspondence between the two measures. Comments are below. The comments with tables is available here.

1. In Tables 1a and 1b, a row examining the proportion of cases in which the PHQ-9 category was > than the BDI category should be added because it shows that the proportion of cases where the PHQ-9 category was > than the BDI category was comparable to the proportion of cases where BDI category was > than PHQ-9 category. This partly explains why continuous score were strongly correlated. This is frequently the case with ordinal categories on different scales; categories are operationally chosen rather the cut-points being determined by some separate criterion standard. The additional document shows Table 1a with the PHQ-9 row added.

2. An even more important point is that an important proportion of the disagreement between PHQ-9 and BDI categories is due to the fact that the PHQ-9 has 5 categories and the BDI has 4 categories. Below is shown a revised version of Table 1a with the two PHQ-9 categories of moderately severe and severe collapsed into a single severe category. This revised table shows even greater agreement between the PHQ-9 and BDI, when each has 4 categories. The rationale for lumping moderately severe with severe is that the treatment recommendations are relatively similar between these two groups (see comment #3 below. I would consider adding table below to the paper.

3. References showing treatment recommendations are relatively similar for PHQ-9 thresholds of 15 and 20, potentially justifying collapsing moderately severe and severe categories.
    1. Trangle M, Gursky J, Haight R, Hardwig J, Hinnenkamp T, Kessler D, Mack N, Myszkowski M. Institute for Clinical Systems Improvement. Adult Depression in Primary Care. Updated March 2016.
    2. MacArthur Foundation Initiative on Depression and Primary Care Toolkit. www.depression-primarycare.org

3. Chen TM, Huang FY, Chang C, Chung H. Using the PHQ-9 for depression screening and treatment monitoring for Chinese Americans in primary care. Psychiatric Services. 2006 57(7):976-981[1].

4. Methods, p. 4, it is stated: “Like the BDI-II, the PHQ-9 has been found to have good sensitivity and specificity (Kroenke *et al*., 2001) and is coming into increasing widespread use as a measure of depression severity.” There has been much research done since the original 2001 study that was cited, and 1 or 2 updated meta-analyses regarding the sensitivity and specificity should be cited (see below). Also, the last part of the sentence should be revised to: “… and is among the most widely-used measures of depression severity.”
    1. Levis B, Benedetti A, Thombs BD. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. BMJ 2019;365:l1476[2].
    2. Mitchell AJ, Yadegarfar M, Gill J, Stubbs B. Case finding and screening clinical utility of the Patient Health Questionnaire (PHQ-9 and PHQ-2) for depression in primary care: a diagnostic meta-analysis of 40 studies. BJPsych Open. 2016;2(2):127-138[3].

5. Minor point: Abstract, Conclusion. Replace “accurate” with either “efficient” or “practical” since a measure easier to read to patients does not mean that measure is more accurate.

References

1. Chen TM, Huang FY, Chang C, Chung H: Using the PHQ-9 for depression screening and treatment monitoring for Chinese Americans in primary care.*Psychiatr Serv*. 2006; **57** (7): 976-81 PubMed Abstract | Publisher Full Text
2. Levis B, Benedetti A, Thombs B: Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. *BMJ*. 2019. Publisher Full Text
3. Mitchell AJ, Yadegarfar M, Gill J, Stubbs B: Case finding and screening clinical utility of the Patient Health Questionnaire (PHQ-9 and PHQ-2) for depression in primary care: a diagnostic meta-analysis of 40 studies.*BJPsych Open*. 2016; **2** (2): 127-138 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
No source data required

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: I am a developer of the PHQ-9. However, the measure is public domain (free to use) so I have no financial conflicts of interest.

Reviewer Expertise: Depression trials and measures, psychometrics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 16 January 2019

https://doi.org/10.21956/wellcomeopenres.16337.r34539

William Mellick
Department of Psychiatry and Behavioral Sciences, Addiction Sciences Division (ASD), Medical University of South Carolina (MUSC), Charleston, SC, USA

In this manuscript, the authors utilized 3 dataset to examine the concordance between PHQ-9 and BDI-II scores. Associations were examined dimensionally as were severity categories. Across samples, the two measures were highly correlated which was similar in each sample. The authors found that BDI-II scores were higher in the India study versus US and UK studies. Additionally, the India study sample had higher scores on the BDI-II versus PHQ-9.

I think this manuscript presents an interesting research question, however there are key aspects that dampen my enthusiasm for it's publication in present form. These include a lack of detail about study samples, questions about analyses, and the framing of the study/conclusions drawn go beyond the statistical methods/results employed.

General comments:
1) The authors, at various points, state that the analyses performed indicate that the BDI-II and PHQ-9 are measuring the same construct. This is a false conclusion that the methods employed do not answer. Simply because scores are correlated does not mean that the same construct is being studied. If the authors wish to ask that question, I believe they will have to examine latent variables, i.e., via structural equation modeling.
2) The participant samples for each respective study are described in detail that prohibits definitive conclusions from the analyses. For instance, the authors, in the results section, refer readers to the original papers which is insufficient.

Introduction:
1) Were the US and UK studies treatment studies? The UK sample is described as participants with diagnoses of depression but what diagnoses in particular? and how were they evaluated? Though the measures described (i.e., including Hamilton rating scale) indicate symptoms, I don't think you can say that they define diagnoses.
2) In the second paragraph, "compiled" should be "complied." Typo.

3) Following sentence, "If one of the measures is problematic..." is a very broad statement. Please be precise because psychometrically this could mean many a things.

Procedures:
1) Methodologically, it seems troublesome that in the India sample self-report measures (designed to be self-administered) were answered via question and answer. The authors state in the limitations that socially desirable responses may have been elicited, which is good, but it remains a big concern.

2) In the statistical analyses, please define "CI."

Results:
1) when comparing within-group scores on the PHQ-9 and BDI-II, why not mean center them? That would presumably show that the difference between scores within the India study sample is true and not simply a function of range of scores for each measure.

2) For figures 1 and 2, please use a different legend as it is very difficult to distinguish groups in black and white printouts.

3) For the PHQ-9, what is the clinical/practical significance of falling into the minimal depression versus mild depression categories?

Discussion:
1) The conclusions drawn, from the reader's standpoint, feel like they are hard to justify without further detail of the samples from which the data are derived.

2) The points made about the BDI-II requiring more working memory because of 4-item choices requires justification/citation.

3) Beginning of second paragraph: correlation coefficients do not say anything about the underlying construct. This is problematic in present form.

4) Similarly, higher BDI-II versus PHQ-9 scores in the India study but not others may be sample-dependent.

5) The comment that if the "sex" item on the BDI-II were included in the India study that observed differences would be much stronger seems like a stretch. Do people, in generally, commonly endorse the "sex" item? Are there high rates of sexual dysfunction in the India study sample? I would think in practice people would be reluctant to answer honestly on that item, particularly in an interview format.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
No

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Depression and bipolar disorder, with a recent publication examining the measurement invariance of the BDI-II across race/ethnicity in inpatient adolescents from a U.S. sample.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.