HHS Public Access

Author manuscript

Psychiatry Res. Author manuscript; available in PMC 2021 September 01.

Published in final edited form as:

Psychiatry Res. 2020 September ; 291: 113236. doi:10.1016/j.psychres.2020.113236.

Establishing Reliability and Validity for Mental Health Screening Instruments in Resource-Constrained Settings: Systematic Review of the PHQ-9 and Key Recommendations

Haley A. Carroll[+,a,b,c], **Kimberly Hook**[+,*,a,b,c], **Oscar F. Rojas Perez**[a,b,c], **Christy Denckla**[b,c,d], **Christine Cooper Vince**[e], **Senait Ghebrehiwet**[a], **Kanako Ando**[f], **Mia Touma**[g], **Christina P.C. Borba**[a,b], **Gregory L. Fricchione**[c,h], **David C. Henderson**[a,b,c,d]

[a]Boston Medical Center, Department of Psychiatry, Boston, MA, USA [b]Boston University School of Medicine, Department of Psychiatry, Boston, MA, USA [c]Massachusetts General Hospital, Department of Psychiatry, Boston, MA, USA [d]Harvard T.H. Chan School of Public Health, Boston, MA, USA [e]University of Geneva, Department of Psychiatry, Geneva, Switzerland [f]Northeastern University, Boston, MA, USA [g]Boston University, Boston, MA, USA [h]Harvard Medical School

Abstract

Mental illness is one of the largest contributors to the global disease burden. The importance of valid and reliable mental health measures is crucial in order to accurately measure said burden, to quantify symptom improvement, and to ensure that symptoms are appropriately identified and quantified. This is of particular importance in low and middle-income countries (LMICs), where burden of mental illness is relatively high, and there is heterogeneity in linguistic, racial, and ethnic groups. Using the PHQ-9 as an illustrative example, this systematic review aims to provide an overview of existing work and highlight common validation and reporting practices. A systematic review of published literature validating the use of the PHQ-9 in LMICs as indexed in the PubMed and PsychInfo databases was conducted. The review included $n = 49$ articles (reduced from $n = 2,349$). This manuscript summarizes these results in terms of the frequency of reporting

[*] kimberly.hook@bmc.org; (617) 414-1955.

[+]Both authors contributed equally

CREDIT AUTHOR STATEMENT

Haley A. Carroll[+]: Conceptualization, Methodology, Formal analysis, Investigation, Writing – Original Draft. **Kimberly Hook**[+]: Conceptualization, Methodology, Formal analysis, Investigation, Writing – Original Draft. **Oscar F. Rojas Perez:** Writing – Original Draft. **Christy Denckla:** Conceptualization, Methodology, Formal analysis, Investigation. **Christine Cooper Vince:** Conceptualization, Methodology, Formal analysis, Investigation. **Senait Ghebrehiwet:** Conceptualization, Methodology, Writing – Original Draft. **Kanako Ando:** Data Curation. **Mia Touma:** Data Curation. **Christina Borba:** Supervision. **Gregory Fricchione:** Supervision. Funding acquisition. **David Henderson:** Supervision. Funding acquisition. ([+]both authors contributed equally)

Availability of data and materials: The dataset used and/or analyzed during the current study is available from the corresponding author on reasonable request.

Competing interests: The authors declare that they have no competing interests

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

on important procedures and the types of reliability and validity measured. Then, building off of the existing literature, we provide key recommendations for measure validation in LMICs, which can be generalized for any type of measure used in a setting in which it was not initially developed.

Keywords

global mental health; psychometrics; low- and middle-income countries

1. Introduction

Globally, major depressive disorder (MDD) is a significant contributor to the burden of disease and disability (Paykel et al., 2005; Wittchen & Jacobi, 2005; World Health Organization [WHO], 2004). During the past two decades, MDD was listed as the second leading cause of years lived with disability worldwide (De Silva et al., 2014; WHO, 2004). Estimates suggest MDD will become the leading cause of disability globally, especially in low-and middle-income countries (LMICs) where the availability and access to mental health services and resources are scarce (De Silva et al., 2014). While efforts to increase access to MDD treatment is actively ongoing worldwide, a key concern centers on the linguistic and cultural validity of screening and diagnostic instruments. Past studies have established differences in epidemiology of depressive disorders across countries, racial groups, and ethnic groups (Ayuso-Mateos et al., 2001; de Wit et al., 2008; Gonzalez et al., 2010; Hasin et al., 2005), and evidence of differing symptom presentations of depressive disorders in varied cultural contexts exists (Deisenhammer et al., 2012; Kirmayer & Young, 1998; Zayas & Gulbas, 2012). Therefore, it is critical to continue investigating the cross-linguistic and cross-cultural validation of self-report questionnaires. Such validation analyses are not only needed but are critical to ensure the accuracy of prevalence rates and symptom profiles of depressive disorders across linguistic, racial, and ethnic groups in LMICs.

The most frequently used screener for depression globally is the Patient Health Questionnaire depression scale (the nine item version is often referred to as the PHQ-9; Lowe et al., 2004a; Lowe et al., 2004b), which has been translated into over 70 different languages and dialects (Pfizer, 2013). The PHQ-9 is an extensively used measure designed to screen for the presence and severity of depression according to the criteria from the Diagnostic and Statistical Manual for Mental Disorders (DSM-IV-TR; American Psychiatric and American Psychiatric Association Task Force, 2000). Professional application of the PHQ-9 ranges widely from community mental health settings to primary care and personnel sites (Gilbody et al., 2007; Wennerstrom et al., 2011). As a part of its broadening clinical use, researchers and mental health providers have sought to adapt and validate the PHQ-9 for use across languages and diverse cultural groups in high-income countries (HICs) and LMICs (e.g., Huang et al., 2006; Lotrakul et al., 2008). In the United States (HIC), the PHQ-9 has been reported to improve the recognition of depression and has shown excellent psychometric properties (Kroenke et al., 2001; Kroenke & Spitzer, 2002, Kroenke et al., 2010). Due to its established psychometric properties and brevity, the PHQ-9 has been

widely embraced for cross-cultural use and translated into many languages including Arabic (Becker et al., 2002), Korean (Donnelly, 2007), Malay (Azah et al., 2005), Turkish (Corapcioglu & Ozer, 2004), Spanish (Huang et al., 2006), and many other languages. Although widely used, less is known about the psychometric properties of the increasingly popular PHQ-9 in LMICs.

Measure validation is key for several reasons, one of which is its ability to quantify symptom improvement following treatment; without validation, it is challenging to ensure that symptoms are appropriately captured and measured in varying contexts (Liu et al., 2011). Despite efforts to improve, integrate, and increase access to effective evidence-based treatments for depression, close attention is also being given to the challenges in detecting depression cross-culturally (Abas et al., 2013; WHO, 2008). The under-detection of depression in LMICs can limit and impact the development and/or availability of services. As a result, there is a significant need to improve the detection of depression in resource-limited countries that both lack sufficient number of mental health providers and cross-culturally validated screening and diagnostic instruments.

To date, a number of research teams have published studies documenting the comparability of different language versions of the PHQ-9 (Arthurs et al., 2012; Huang et al., 2014) and the use of the PHQ-9 across ethnic and racial groups (Arthurs et al., 2012; Crane et al., 2010; Hepner et al., 2008). For example, differences in item functioning were found between the English and Chinese versions of the PHQ-9 when assessing appetite, sleep, and psychomotor changes in a sample of primary care patients (Huang et al., 2006). Another study found differences between the English and French versions that assessed sleep, self-esteem, and anhedonia items in patients diagnosed with system sclerosis (Arthurs et al., 2012). Regarding the comparability across racial and ethnic communities, studies have found differences in items about low energy, sleep, and psychomotor changes between HIV-infected African Americans and non-Latinx Whites in the English version (Crane et al., 2010), as well as differences in the psychomotor changes item between Surinam Dutch and Native Dutch male primary care patients with the Dutch version of the PHQ-9 (Baas et al., 2011). These studies provide insight into how cultural and language differences can impede the accuracy of depression detection and the need for thorough cultural and linguistic validation of the PHQ-9 in LMICs.

In the present study, we conducted a systematic review of current validation efforts of the PHQ-9 in LMICs. The PHQ-9 was chosen as an illustrative example as it is a commonly used instrument, though the authors believe that the considerations describe below are equally important for any type of measure used in a setting in which it was not initially developed (e.g., from HICs to LMICs). This paper seeks to provide an overview of existing work and highlight common validation and reporting practices. Finally, the authors are aware that difficulties related to undertaking formal measure validation are common, due to factors such as competing interests between funders, researchers, and time restraints. In order to guide future work, authors will also provide key recommendations about measure validation in resource-constrained settings that attempt to balance these challenges.

2. Methods

2.1 Search strategy

In February 2019, we conducted a systematic review of published literature validating the use of the PHQ-9 in LMICs as indexed in the PubMed and PsychInfo databases. PubMed and PsychInfo were selected as they are databases known to index published peer-reviewed articles in the field of mental health (i.e., psychiatry, psychology). We did not include databases that indexed grey literature (e.g., Google Scholar), as the primary purpose of the present manuscript was focused on identifying practices of measure validation in peer reviewed articles. Table 1 displays the search terms used for this review. All abstracts returned by the database search were reviewed by two reviewers for possible inclusion. For articles identified as potentially relevant, full texts were retrieved and assessed for inclusion using the criteria outlined below. Though the search terms used were in English, the search was not restricted by language in an effort to include as many potentially relevant papers as possible. We followed PRISMA guidelines (Moher et al., 2009) to guide the search, analysis, and reporting of data.

2.2 Inclusion Criteria

2.2.1 Publication date.—Papers were eligible for inclusion in this review if they were published after 2001, as the PHQ-9 was first published in 2001 (Kroenke et al., 2001). Our search included relevant papers published between 2002 and the search date of the review (February 26, 2019).

2.2.2 Setting.—Studies were included if the research was conducted in LMICs as designated by World Bank classification (World Bank Country and Lending Groups, n.d.). In other words, the country must have been designated as either a low-income, lower-middle-income, or upper-middle-income economy by the World Bank at the time data in each respective study was collected. We included the identified LMICs in our search terms (n = 152), and then cross-checked the text of each article to confirm that the study was conducted in a country designated as a LMICs at the time of data collection.

2.2.3 Purpose.—To be eligible for inclusion, the primary purpose of the study must have been to validate or modify the PHQ-9 in a LMICs.

2.3 Exclusion Criteria

Given our interest in examining existing practices for validating measures in LMICs, we excluded papers where validity (e.g., face-validity) or reliability (e.g., Cronbach's alpha) may have been reported but validation was not the primary purpose of the paper.

2.4 Quality Appraisal

The quality of eligible studies was appraised using modified criteria described by Ali et al. (2016), who designed and used a standardized coding tool for data extraction regarding study design for cross cultural adaptation, reliability and validity. After adapting said coding tool for this study, a random sample of approximately 30% (n = 14) of the articles identified was coded independently by two reviewers, and percentage agreement was calculated for all

criteria. Agreement between reviewers was acceptable (above 95%), and Kappa was above 0.9. Where coders differed, a consensus was reached after reviewing extracted information among the coders.

3. Results

From the *n* = 2,349 articles first identified in the record search, we assessed *n* = 91 articles and found that *n* = 49 met our inclusion criteria (see Figure 1). There were *n* = 15 studies completed in Africa, *n* = 9 in the Americas, and *n* = 25 in Asia (see Table 2). No studies were conducted in Europe or Oceana.

The evaluation of the study design for cross cultural adaptation of the *n* = 49 articles (see Table 3) revealed infrequent reporting in manuscripts regarding procedures concerning the recruitment of research assistants from the country of study (6%, *n* = 3), if participants were fluent in the language (29%, *n* = 14), a formal assessment in language proficiency (2%, *n* = 1), pilot testing of the measure (20%, *n* = 10), and any involvement or engagement with the community (12%, *n* = 6). There were more frequent levels of reporting of medical comorbidities (31%, *n* = 15), method of measurement (59%, *n* = 29), methods regarding the training of the research assistants collecting data (43%, *n* = 21), accounting for participant educational attainment (67%, *n* = 33), and the translational procedures conducted (43%, *n* = 21).

Regarding reliability and validity (see Table 4), our analysis revealed infrequent levels of reported assessment of test-retest (25%, *n* = 12) and inter-rater (2%, *n* = 1) reliability, as well as discriminative (2%, *n* = 1) and predictive (4%, *n* = 2) validity. There were more frequent levels of criterion (59%, *n* = 29) and convergent validity reported (50%, *n* = 23), in addition to assessments of factor structure (43%, *n* = 21). We also found it was more common to report levels of internal consistency, or Cronbach's alpha (86%, *n* = 42), and sensitivity (76%, *n* = 37) and specificity (76%, *n* = 37).

4. Discussion

The present study evaluated the existing evidence for reliability and validity for the PHQ-9 in LMICs. The primary purpose of the manuscript was to provide an overview of practices commonly implemented in global mental health and to guide future validation studies on measures of mental health. We believe the practices represented here may translate to other validation studies, provide guidance for future research on gaps in the literature, and may offer guidelines for future validation studies. Table 5 presents an overview of recommendations for future measure validation studies.

This systematic review identified 49 studies which evaluated the validity and reliability of the PHQ-9 in LMICs. Of those studies, the majority took place in Asia (*n* = 25), with most studies conducted in China (*n* = 11). In many countries, only one validation study was conducted (i.e., Cameroon, Somalia, Zimbabwe, Nigeria, Ghana, Cote d'Ivoire, Haiti, Chile, Colombia, Pakistan, Sri Lanka, Nepal, Malaysia, Thailand, Vietnam, Lebanon, and Iran). The only country besides China with more than three validation studies was India (*n* = 6). No studies were found validating the PHQ-9 in LMICs in Europe and Oceania. This is

concerning as rates of depressive disorders are purported to be significant in these areas (Australian Bureau of Statistics, 2007; ESEMeD/MHEDEA 2000 Investigators, 2004; Petrea, 2012; Whiteford et al., 2013). In addition, lack of validation studies neglect potential cultural variations of depression existing in these geographical areas.

The review utilized modified guidelines of Ali et al. (2016) to assess the study design, as well as reliability and validity metrics, reported in each manuscript. Regarding the reported study methods, within the 49 studies assessed, we found infrequent levels (6%) of reporting on the country of origin of the research assistant(s) and the involvement of the community within the country of study (12%). Without local input, it is possible that measures from different contexts either under- or over -screen for symptoms; measures may also neglect to capture varying types of disease presentation (Gonzalez & Trickett, 2014; van de Vijver & Leung, 2000). Ideally, a tool to measure depression should incorporate local knowledge, in order to account for local understanding, expression, and language (see Recommendation 1, Table 5). In one example from our review, Kochhar and colleagues (2007) included local populations in a review panel to provide feedback on measure translation for seven languages spoken in India. Another viable option may be recruiting field workers from local communities (who are fluent in the languages of the community) to conduct data collection, as seen in a study from South Africa by Bhana and colleagues (2015). If these steps are not possible, authors may also identify key informants to provide basic insights into psychological concepts captured by measures or complete small-scale pilot tests to assess measure functioning. Other options include clearly reporting community involvement and identifying any lack of community involvement to provide additional clarity around limitations to the measure validation. Lastly, utilizing both quantitative and qualitative methods to triangulate data and inform adaption efforts is recommended (van de Vijver & Leung, 2000; see Recommendation 2, Table 5). For example, as seen in the work conducted by Pence and colleagues (2012) in Cameroon, one could conduct qualitative focus groups on a psychological measure in local populations, make suggested changes to the measure, and then confirm these changes with subsequent quantitative evaluation. Researchers could consider following the qualitative procedures within the design, implementation, monitoring, and evaluation model (Module 1, DIME) proposed by the Johns Hopkins University Bloomberg School of Public Health Applied Mental Health Research Group (2013). If this is not possible, an emphasis should be placed on recruiting patient and non-patient populations for qualitative feedback on the measure adaptation, as well as subsequent quantitative evaluation.

There were more frequent levels of reporting around both the training that research assistants and clinicians received (43%) and method of measurement (59%). Related to researcher training, it is well-established that measures may be impacted by administrator error (Hoyt, 2000). Well-described, consistent training procedures may limit this type of error that may otherwise impact the validity of measures. Additionally, many research groups administer the PHQ-9 orally, with few studies formally assessing how this might alter the psychometrics of a measure originally designed as a paper and pencil measure. Though this may be more appropriate for certain health care settings or patient populations (e.g., literacy), it remains important to clarify in what manner the measure is validated for future use of the measure. Existing literature notes changes in psychometric properties that

may occur when route of administration is altered, due to factors such as differences between auditory and visual processing or social desirability (Miller et al., 2015; Ruggeri et al., 2016). Clear reporting practices about administration, as well as discussion of any impacts of administration method, would improve transparency in measure validation studies (see Recommendation 3, Table 5). Within our literature review, some studies clearly reported if the measure was self-administered, such as in a study in Malaysia by Sherina and colleagues (2012), or if the measure was administered via in-person interviews, as in a validation study in rural Uganda conducted Nakku and colleagues (2016). Importantly, the study conducted by Nakku and colleagues (2016) also provided information on the training of the interviewers (2 day training) and their language fluency (English and Luganda). Researchers can address this recommendation by explicitly reporting methods used for measure administration in their manuscript. If researchers change the measure administration from the prior validated method (e.g., the classic PHQ9 paper and pencil to in-person interviewer), they can provide justification, describe adaptation procedures, and conduct validation studies on the changed measure.

Further, while there were more frequent levels of reporting within studies to their translation procedures (43%) and the education level of participants (67%), it was often unclear what language the measure had been administered in (English or otherwise), if the participants were fluent in that language (29%) or any measure of language proficiency (2%). Recommendations for thorough translation exist (WHO, n.d.), and ensuring that protocols for competent translation are utilized would strengthen future work. Noteworthy examples of thorough translation procedures come from Gelaye et al. (2013) and Bhana et al. (2015), who provided a discussion of the specific steps taken to ensure accurate transcriptions, including multiple rounds of forward and backward translation. If robust translation procedures are not followed, it is possible that measures may be inadvertently altered, resulting in changes from their original forms and alterations in the types of data that captured. Calling attention to possible discrepancies that may arise from translation is one way of drawing attention to these concerns, as Gothwal and colleagues (2014) reported in their manuscript, offers a way to acknowledge concerns about future use of an assessment tool. Issues related to language proficiency of participants are also key challenges that impact psychometric data (Abedi, 2002; Bauer & Alegría, 2010). For example, if measures are translated well, and yet respondents are unable to accurately respond to questions due to language challenges (e.g., inability to understand language used in measure or difficulty with nonverbal responding), measures will inherently misrepresent reported data (see Recommendation 4, Table 5). Multiple articles included in our study (e.g., Arrieta et al., 2017; Cholera et al., 2015) specifically stated that participants were fluent in the language of the measure or provided the same measure in various translations to ensure that participants would have appropriate options. Further, Arrieta et al. (2017) assessed and reported participant ability to read and write, which can both inform optimal administration route. Efforts to counter these measurement threats would aid future users of measures in better understanding the true performance of measures in a given context; alternatively, identifying these confounders in manuscripts can help other clinicians and researchers recognize possible threats to report validity and reliability data.

Another concern regarding validity of the measure relates to medical comorbidities, as depressive symptoms may arise or be influenced due to a host of other existing health conditions (e.g., thyroid disease, anemia, dementia; Hage & Azar, 2012; Eizadi-Mood et al., 2018; Muliyala & Varghese, 2010). Nevertheless, we found only moderate levels of reporting or controlling for medical comorbidities (31%). While brief measures may not be able to parse apart symptoms arising from purely medical versus purely psychiatric conditions, efforts that acknowledge and control for common disorders in a given population may be of value in validation studies. This greater attention to medical factors may better reflect true psychiatric illness, as opposed to symptoms that better reflect medical comorbidities. Indeed, some have advocated for emphasis on cognitive affective symptoms to better characterize major depressive disorder. Factor analysis of the 21-item Beck Depression Inventory (BDI)-II validates this instrument as a way to distinguish measures of somatic and cognitive-affective symptoms, and supports the existence of these two dimensions (Storch et al., 2004; Whisman et al., 2000).

Finally, our analyses revealed infrequent reported rates of pilot testing (20%). Pilot testing is important in the iterative process of measure development as it serves to test the feasibility of larger studies, allow for small scale testing of an intervention or measure, and make corrections prior to initiation of a full-scale study (Zailinawati et al., 2006). Regarding validity studies, pilot tests are also key in identifying biases that cannot be controlled for statistically (e.g. language proficiency). As a caveat, we did not contact the authors of the studies included in our analysis, and it is possible that pre-testing occurred and was not reported in the literature. Regardless, we posit that either more rigorous methods (i.e., pilot testing) or more thorough reporting of methods will improve the literature surrounding reliability and validity of measures such as the PHQ-9 in LMICs.

In reference to reliability, authors most commonly reported internal consistency (86%). While calculating Cronbach's alpha is relatively straightforward, it is also biased by certain factors (e.g., test length, test dimensionality) and may be better understood when presented alongside other psychometric data (Sijtsma, 2008; Tavakol & Dennick, 2011). We found somewhat low levels of test-retest reliability (25%) reported. Test-retest reliability is critical for measures captured at various time points (Duff, 2012), such as measuring changes in depression due to intervention or monitoring disease progression over time. The low levels of reported measurement of test-retest validity is concerning as many clinical trials utilize measures like the PHQ-9 (e.g., Muñoz-Navarro et al., 2017), and results are difficult to interpret (i.e., challenging to assess potential bias) without reporting test-retest reliability from data obtained within these trails. Finally, we found very low levels of inter-rater reliability (2%). Reporting inter-rater reliability reflects agreement among different individuals collecting data (McHugh, 2012); for verbal administration of measures, providing inter-rater reliability estimates is particularly important to ensure that systematic error stemming from differences between individuals is minimized (see Recommendation 5, Table 5). This may further ensure that minor variations in phrasing of assessment items or in body language that may influence scores can be corrected early and may indicates potential areas where research staff may need further training to ensure the reproducibility of assessment results (Velligan et al., 2011). While this statistic was not commonly reported in the articles included in this literature review, Adewuya et al. (2006) were clear in reporting

their inter-rater reliability among trained interviewers, offering an example for subsequent work. In sum, while reliability data is often present in some form, more robust reporting and thoughtful consideration of types of reliability to report (in light of modifications made to test administration) would strengthen the body of literature supporting the use of the PHQ-9 in global contexts.

Finally, considering validity, we found that the most commonly reported forms of validity were sensitivity (76%) and specificity (76%). Sensitivity and specificity are relatively easy to calculate and have provided data for meta-analytic reviews that have assessed use of the PHQ-9 across the globe (e.g., Levis et al., 2019). Nevertheless, considerations about the source of this data are needed. While some findings are based upon clinical diagnosis by a qualified mental health professional, other studies rely on using standardized clinical interviews, which may or may not have been validated in a given cultural context. If diagnosis is made using a tool that has not been validated, even a “gold standard” test may have limited utility and interpretability. Predictive validity was seldom reported (4%); of note, some studies indicated efforts to report predictive validity, though data reported was instead reflective of criterion validity. Conversely, criterion validity was reported in the majority of studies (59%). We also found moderate levels of convergent validity (50%) but very low levels of reported discriminant validity (2%). Finally, there were moderate levels of reported factor structure (43%), with past work frequently suggesting that the PHQ-9 is either a one or two factor measure (e.g., Chilcot et al., 2013; Gelaye et al., 2013; González-Blanch et al., 2018). These differences are indicative of the variations across cultures and suggest the importance of continued validation of this measure in identified populations.

Limitations

The present study is limited by several factors. First, we only included indexed journals and therefore may be missing literature from sources such as Google Scholar or similar grey literature. Additionally, while we did not reduce our sample based on language, we found only one study in Spanish and the remaining 48 articles were published in English. We posit that validation studies in other languages may have been missed in our search. Finally, our search did not result in any sources from Europe or Oceania. Yet, we believe that our search provided a comprehensive overview of the existing literature in LMICs, particularly as similar patterns of data reporting were clear within the included literature, and thus our included studies provide a sufficient basis for assessing current validation practices.

Conclusions

The administration of brief screening measures, such as the PHQ-9, is a useful method to quickly and simply assess change and suggest prevalence rates of various disorders. However, the usefulness of these measures rests on their ability to accurately screen for illness, and there is potential harm to both underreport and over-pathologize if measures are not adequately validated in the context it is used. While competing demands may make full measure validation studies challenging, future work that offers clarity and thoroughness in reporting data, as well as thoughtful identification of reliability and validity data that is most critical to a specific study, will offer subsequent users the most complete information regarding the appropriateness of adapted screening tools in varied contexts.

Appendix

Appendix: Table 1:

List of articles included in the systematic review

Region	Country	Author	Year	Title
Africa				
	Cameroon	Pence BW, Gaynes BN, Atashili J, O'Donnell JK, Tayong G, Kats D, Whetten R, Whetten K, Njamnshi AK, Ndumbe PM.	2012	Validity of an interviewer-administered patient health questionnaire-9 to screen for depression in HIV-infected patients in Cameroon.
	Kenya	Monahan PO, Shacham E, Reece M, Kroenke K, Ong'or WO, Omollo O, Yebei VN, Ojwang C.	2009	Validity/reliability of PHQ-9 and PHQ-2 depression scales among adults living with HIV/AIDS in western Kenya.
	Kenya	Omoro SA, Fann JR, Weymuller EA, Macharia IM, Yueh B.	2006	Swahili translation and validation of the Patient Health Questionnaire-9 depression scale in the Kenyan head and neck cancer patient population.
	Somalia	Nallusamy V, Afgarshe M, Shlosser H.	2016	Reliability and validity of Somali version of the PHQ-9 in primary care practice.
	Uganda	Nakku JEM, Rathod SD Kizza D, Breuer E3 Mutyaba K, Baron EC, Ssebunnya J, Kigozi F.	2016	Validity and diagnostic accuracy of the Luganda version of the 9-item and 2-item Patient Health Questionnaire for detecting major depressive disorder in rural Uganda.
	Uganda	Akena D, Joska J, Obuku EA, Stein DJ.	2013	Sensitivity and specificity of clinician administered screening instruments in detecting depression among HIV-positive individuals in Uganda.
	Ethiopia	Hanlon C, Medhin G, Selamu M, Breuer E, Worku B, Hailemariam M, Lund C, Prince M, Fekadu A.	2015	Validity of brief screening questionnaires to detect depression in primary care in Ethiopia.
	Ethiopia	Gelaye B, Williams MA, Lemma S, Deyessa N, Bahretibeb Y, Shibre T, Wondimagegn D, Lemenhe A, Fann JR, Vander Stoep A, Andrew Zhou XH.	2013	Validity of the Patient Health Questionnaire-9 for depression screening and diagnosis in East Africa.
	South Africa	Aggarwal S, Taljard L, Wilson Z, Berk M	2017	Evaluation of Modified Patient Health Questionnaire-9 Teen in South African Adolescents.
	South Africa	Cholera R, Gaynes BN, Pence BW, Bassett J, Qangule N, Macphail C, Bernhardt S, Pettifor A, Miller WC.	2014	Validity of the Patient Health Questionnaire-9 to screen for depression in a high-HIV burden primary healthcare clinic in Johannesburg, South Africa.
	South Africa	Bhana A, Rathod SD, 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.
	Zimbabwe	Chibanda D Verhey R, Gibson LJ, Munetsi E, Machando D, Rusakaniko S Munjoma R, Araya R, Weiss HA, Abas M	2016	Validation of screening tools for depression and anxiety disorders in a primary care population with high HIV prevalence in Zimbabwe.
	Nigeria	Adewuya AO1, Ola BA, Afolabi OO.	2006	Validity of the patient health questionnaire (PHQ-9) as a screening tool for depression amongst Nigerian university students.
	Ghana	Weobong B, Akpalu B, Doku V, Owusu-Agyei S, Hurt L, Kirkwood B, Prince M.	2009	The comparative validity of screening scales for postnatal common mental disorder in Kintampo, Ghana.
	Cote d'Ivoire and Ghana	Barthel D, Barkmann C, Ehrhardt S, Schoppen S, Bindt C; International CDS Study Group.	2015	Screening for depression in pregnant women from Cote d'Ivoire and Ghana: Psychometric

Region	Country	Author	Year	Title
				properties of the Patient Health Questionnaire-9.
America				
	Haiti	Marc LG, Henderson WR, Desrosiers A, Testa MA, Jean SE, Akom EE.	2014	Reliability and validity of the Haitian Creole PHQ-9.
	Mexico	Arrieta J, Aguerrebere M, Raviola G, Flores H, Elliott P, Espinosa A, Reyes A, Ortiz-Panozo E, Rodriguez-Gutierrez EG, Mukherjee J, Palazuelos D, Franke ' MF.	2017	Validity and Utility of the Patient Health Questionnaire (PHQ)-2 and PHQ-9 for screening and diagnosis of depression in the Rural Chiapas, Mexico: A cross-sectional Study
	Mexico	Familiar I, Ortiz-Panozo E, Hall B, Vieitez I, Romieu I, Lopez-Ridaura R, Lajous M.	2015	Factor structure of the Spanish version of the Patient Health Questionnaire-9 in Mexican women.
	Peru	Zhong Q, Gelaye B, Fann JR, Sanchez SE, Williams MA	2014	Cross-cultural validity of the Spanish version of PHQ-9 among pregnant Peruvian women: a Rasch item response theory analysis.
	Peru	Zhong Q, Gelaye B, Rondon M, Sanchez SE, Garcia PJ, Sanchez E, Barrios YV, Simon GE, Henderson DC, Cripe SM, Williams MA.	2014	Comparative performance of Patient Health Questionnaire-9 and Edinburgh Postnatal Depression Scale for screening antepartum depression.
	Chile	Baader T, Molina J, Venezian S, Rojas C, Farias R, Fierro-Freixenet C,	2012	Validacion y utilidad de la encuesta PHQ-9 (Patient Health Questionnaire) en el diagnostico de depresion en pacientes usuarios de atencion primaria en depresion en pacientes usuarios de atencion primaria en Chile
	Brazil	Santos IS, Tavares BF, Munhoz TN, Manzolli P, de Avila GB, Jannke E, Matijasevich A.	2017	Patient health Questionnaire-9 versus Edinburgh Postnatal Depression Scale in screening for major depressive episodes: a cross-sectional population-based study
	Brazil	de Lima Osorio F, Vilela Mendes A, Crippa JA, Loureiro SR.	2009	Study of the discriminative validity of the PHQ-9 and PHQ-2 in a sample of Brazilian women in the context of primary health care.
	Colombia	Cassiani-Miranda CA, Vargas-Hernandez MC, Perez-Anibal E, Herazo-Bustos MI, Hernandez-Carrillo M.	2017	[Reliability and dimensionality of PHQ-9 in screening depression symptoms among health science students in Cartagena, 2014].
Asia				
	China	Du N, Yu K, Ye Y, Chen S.	2017	Validity study of Patient Health Questionnaire-9 items for Internet screening in depression among Chinese university students.
	China	Liu ZW, Yu Y, Hu M, Liu HM, Zhou L, Xiao SY.	2016	PHQ-9 and PHQ-2 for Screening Depression in Chinese Rural Elderly.
	China	Feng Y, Huang W, Tian TF, Wang G, Hu C, Chiu HF, Ungvari GS, Kilbourne AM, Xiang YT.	2016	The psychometric properties of the Quick Inventory of Depressive Symptomatology-Self-Report (QIDS-SR) and the Patient Health Questionnaire-9 (PHQ-9) in depressed inpatients in China.
	China	Chin WY, Chan KT, Lam CL, Wong SY, Fong DY, Lo YY, Lam TP, Chiu BC.	2014	Detection and management of depression in adult primary care patients in Hong Kong: a cross-sectional survey conducted by a primary care practice-based research network.
	China	Xiong N Fritzsche K, Wei J, Hong X Leonhart R, Zhao X, Zhang L, Zhu L, Tian G8 Nolte S, Fischer F	2015	Validation of patient health questionnaire (PHQ) for major depression in Chinese

Region	Country	Author	Year	Title
				outpatients with multiple somatic symptoms: a multicenter cross-sectional study.
	China	Yu X, Tam WW, Wong PT, Lam TH, Stewart SM.	2012	The Patient Health Questionnaire-9 for measuring depressive symptoms among the general population in Hong Kong.
	China	Zhang Y, Ting R, Lam M, Lam J, Nan H, Yeung R, Yang W, Ji L, Weng J, Wing YK, Sartorius N, Chan JC.	2013	Measuring depressive symptoms using the Patient Health Questionnaire-9 in Hong Kong Chinese subjects with type 2 diabetes.
	China	Zhang YL, Liang W, Chen ZM, Zhang HM, Zhang JH, Weng XQ, Yang SC, Zhang L, Shen LJ, Zhang YL.	2013	Validity and reliability of Patient Health Questionnaire-9 and Patient Health Questionnaire-2 to screen for depression among college students in China.
	China	Wang W, Bian Q, Zhao Y, Li X, Wang W, Du J, Zhang G, Zhou Q, Zhao M.	2014	Reliability and validity of the Chinese version of the Patient Health Questionnaire (PHQ-9) in the general population.
	China	Chen S, Fang Y, Chiu H, Fan H, Jin T, Conwell Y.	2013	Validation of the nine-item Patient Health Questionnaire to screen for major depression in a Chinese primary care population.
	China	Chen S1, Chiu H, Xu B, Ma Y, Jin T, Wu M, Conwell Y.	2010	Reliability and validity of the PHQ-9 for screening late- life depression in Chinese primary care.
	Pakistan	Gholizadeh L, Ali Khan S, Vahedi F, Davidson PM.	2017	Sensitivity and specificity of Urdu version of the PHQ-9 to screen depression in patients with coronary heart disease
	India	Chowdhury AN, Ghosh S, Sanyal D.	2004	Bengali adaptation of brief patient health questionnaire for screening depression at primary care.
	India	Poongothai S, Pradeepa R, Ganesan A, Mohan V.	2009	Reliability and validity of a modified PHQ-9 item inventory (PHQ-12) as a screening instrument for assessing depression in Asian Indians (CURES-65).
	India	Patel V, Araya R, Chowdhary N, King M, Kirkwood B, Nayak S, Simon G, Weiss HA.	2008	Detecting common mental disorders in primary care in India: a comparison of five screening questionnaires.
	India	Gothwal VK, Bagga DK1 Bharani S, Sumalini R, Reddy SP.	2014	The patient health questionnaire 9: validation among patients with glaucoma.
	India	Ganguly, Samrat; Samanta, Moumita; Roy, Prithwish; Chatterjee, Sukanta; Kaplan, David W; Basu, Bharati	2013	Patient Health Questionnaire-9 as an effective tool for screening of depression among Indian adolescents.
	India	Kochhar PH, Rajadhyaksha SS, Suvarna VR.	2007	Translation and validation of brief patient health questionnaire against DSM IV as a tool to diagnose major depressive disorder in Indian patients.
	Sri Lanka	Hanwella R, Ekanayake S, de Silva VA.	2014	The Validity and Reliability of the Sinhala Translation of the Patient Health Questionnaire (PHQ-9) and PHQ-2 Screener.
	Nepal	Kohrt, Brandon A.; Luitel, Nagendra P.; Acharya, Prakash; Jordans, Mark J. D.	2016	Detection of depression in low resource settings: Validation of the Patient Health Questionnaire (PHQ-9) and cultural concepts of distress in Nepal.
	Malaysia	Sherina MS, Arroll B, Goodyear-Smith F.	2012	Criterion validity of the PHQ-9 (Malay version) in a primary care clinic in Malaysia.
	Thailand	Lotrakul M, Sumrithe S, Saipanish R.	2008	Reliability and validity of the Thai version of the PHQ-9.

Region	Country	Author	Year	Title
	Vietnam	Nguyen TQ, Bandeen-Roche K, Bass JK, German D, Nguyen NT, Knowlton AR.	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.
	Lebanon	Sawaya H, Atoui M, Hamadeh A, Zeinoun P, Nahas Z	2016	Adaptation and initial validation of the Patient Health Questionnaire - 9 (PHQ-9) and the Generalized Anxiety Disorder - 7 Questionnaire (GAD-7) in an Arabic speaking Lebanese psychiatric outpatient sample.
	Iran	Khamseh ME, Baradaran HR, Javanbakht A, Mirghorbani M, Yadollahi Z, Malek M.	2011	Comparison of the CES-D and PHQ-9 depression scales in people with type 2 diabetes in Tehran, Iran.

List of Abbreviations

MDD	Major Depressive Disorder
PHQ-9	Patient Health Questionnaire 9-item Version
LMICs	Low- and Middle-Income Countries
HICs	High-Income Countries

References

Abas M, Baingana F, Broadhead J, Iacoponi E, & Vanderpyl J, 2003 Common mental disorders and primary health care: current practice in low-income countries. Harv Rev Psychiatry, 11(3), 166–173. doi: 10.1080/10673220390217881 [PubMed: 12893507]

Abedi J 2002 Standardized achievement tests and English language learners: Psychometrics issues. Educational Assessment, 8(3), 231–257. doi: 10.1207/s15326977ea0803_02

Adewuya A, Ola B, & Afolabi O, 2006 Validity of the patient health questionnaire (PHQ-9) as a screening tool for depression amongst Nigerian university students. J Affect Disord, 96(1–2), 89–93. doi: 10.1016/j.jad.2006.05.021 [PubMed: 16857265]

Ali GC, Ryan G, & De Silva MJ, 2016 Validated screening tools for common mental disorders in low and middle-income countries: A systematic review. PloS One, 16, 11(6), e0156939.

American Psychiatric Association, 2000 Diagnostic and statistical manual of mental disorders: DSM-IV-TR. American Psychiatric Association; Washington, DC: 2000.

Arrieta J, Aguerrebere M, Raviola G, Flores H, Elliott P, Espinosa A…. Franke MF, 2017 Validity and utility of the Patient Health Questionnaire (PHQ)-2 and PHQ-9 for screening and diagnosis of depression in rural Chiapas, Mexico: A cross-sectional study. J Clin Psycho, 73(9), 1076–1090. 10.1002/jclp.22390

Arthurs E, Steele RJ, Hudson M, Baron M, & Thombs BD, 2012 Are scores on English and French versions of the PHQ-9 comparable? An assessment of differential item functioning. PLoS One, 7, 1–7. e52028.

Australian Bureau of Statistics, 2007 National survey of mental health and wellbeing: Summary of results. Retrieved from https://www.ausstats.abs.gov.au/ausstats/subscriber.nsf/0/6AE6DA447F985FC2CA2574EA00122BD6/$File/National%20Survey%20of%20Mental%20Health%20and%20Wellbeing%20Summary%20of%20Results.pdf

Ayuso-Mateos JL, Vázquez-Barquero JL, Dowrick C, Lehtinen V, Dalgard OS, Casey P, … & Wilkinson G, 2001 Depressive disorders in Europe: prevalence figures from the ODIN study. Br J Psychiatry, 179, 308–316. doi: 10.1192/bjp.179.4.308 [PubMed: 11581110]

Azah MN, M Shah ME, Juwita S, S Bahri I, Rushidi WM, & Jamil Y, 2005 Validation of the Malay version brief Patient Health Questionnaire (PHQ-9) among adult attending family medicine clinics. International Medical Journal-Tokyo, 12, 259–263.

Baas KD, Cramer AO, Koeter MW, Van De Lisdonk EH, Van Weert HC, & Schene AH, 2011 Measurement invariance with respect to ethnicity of the Patient Health Questionnaire-9 (PHQ-9). J Affect Disord, 129, 229–235. doi: 10.1016/j.ad.2010.08.026 [PubMed: 20888647]

Bauer A, & Alegría M, 2010 Impact of patient language proficiency and interpreter service use on the quality of psychiatric care: A systematic review. Psychiatr Serv, 61(8). doi: 10.1176/appi.ps.61.8.765

Becker S, Al Zaid K, & Al Faris E, 2002 Screening for somatization and depression in Saudi Arabia: a validation study of the PHQ in primary care. Int J Psychiatry Med, 32, 271–283. doi: 10.2190/XTDD-8L18-P9E0-JYRV [PubMed: 12489702]

Bhana A, Rathod SD, 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, 118 10.1186/s12888-015-0503-0 [PubMed: 26001915]

Chilcot J, Rayner L, Lee W, Price A, Goodwin L, Monroe B… Hotopf M, 2013 The factor structure of the PHQ-9 in palliative care. J Psychosom Res, 75(1), 60–64. doi: 10.1016/j.jpsychores.2012.12.012 [PubMed: 23751240]

Cholera R, Gaynes BN, Pence BW, Bassett J, Qangule N, Macphail C… Miller WC, 2014 Validity of the Patient Health Questionnaire-9 to screen for depression in a high-HIV burden primary healthcare clinic in Johannesburg, South Africa. J Affect Disord, 167, 160–166. 10.1016/j.jad.2014.06.003 [PubMed: 24972364]

Corapcioglu A, & Ozer GU, 2004 Adaptation of revised Brief PHQ (Brief-PHQ-r) for diagnosis of depression, panic disorder and somatoform disorder in primary healthcare settings. Int J Psychiat Clin, 8, 11–18. doi:101080/13651500310004452

Crane PK, Gibbons LE, Willig JH, Mugavero MJ, Lawrence ST, Schumacher JE, … & Crane HM, 2010 Measuring depression levels in HIV-infected patients as part of routine clinical care using the nine-item Patient Health Questionnaire (PHQ-9). AIDS Care, 22, 874–885. doi: 1.1080/0954012090483034 [PubMed: 20635252]

De Silva MJ, Lee L, Fuhr DC, Rathod S, Chisholm D, Schellenberg J, & Patel V, 2014 Estimating the coverage of mental health programmes: A systematic review. Int J Epidemiol, 43, 341–353. doi: 10.1093/ije/dyt191 [PubMed: 24760874]

Deisenhammer EA, Çoban-Başaran M, Mantar A, Prunnlechner R, Kemmler G, Alkın T, & Hinterhuber H, 2012 Ethnic and migrational impact on the clinical manifestation of depression. Soc Psychiatry Psychiatr Epidemiol, 47, 1121–1129. doi: 10.1007/s00127-011-0417-1 [PubMed: 21805303]

De Wit MA, Tuinebreijer WC, Dekker J, Beekman AJT, Gorissen WH, Schrier AC, … & Verhoeff AP, 2008 Depressive and anxiety disorders in different ethnic groups. Soc Psychiatry Psychiatr Epidemiol, 43, 905–912. doi: 10.1007/s00127-008-0382-5 [PubMed: 18587679]

Donnelly PL, 2007 The use of the patient health questionnaire–9 Korean version (PHQ-9K) to screen for depressive disorders among Korean Americans J. Transcult. Nurs, 18, 324–330. doi: 10.1177/1043659607305191 [PubMed: 18092395]

Duff K, 2012 Evidence-based indicators of neuropsychological change in the individual patient: Relevant concepts and methods. Arch Clin Neuropsychol, 27(3), 248–261. doi: 10.1093/arclin/acr120 [PubMed: 22382384]

Eizadi-Mood N, Ahmadi R, Babazadeh S, Yaraghi A, Sadeghi M, & Peymani P, 2018 Anemia, depression, and suicidal attempts in women: Is there a relationship? J Res Pharm Pract, 7(3), 136. doi: 10.4103/jrpp.jrpp_18_25 [PubMed: 30211238]

ESEMeD/MHEDEA 2000 Investigators, 2004 Prevalence of mental disorders in Europe: Results from the European Study of the Epidemiology of Mental Disorders (ESEMeD) project. Acta Psychiatr Scand, 109(Suppl. 420), 21–27.

Gelaye B, Williams M, Lemma S, Deyessa N, Bahretibeb Y, Shibre T… Zhou X-H, 2013 Validity of the Patient Health Questionnaire-9 for depression screening and diagnosis in East Africa. Psychiatry Res, 210(2), 653–661. doi: 10.1016/j.psychres.2013.07.015 [PubMed: 23972787]

Gilbody S, Richards D, Brealey S, & Hewitt C, 2007 Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. J Gen Intern Med, 22, 1596–1602. doi: 10.1007/s11606-007-0333-y [PubMed: 17874169]

González-Blanch C, Medrano L, Muñoz-Navarro R, Ruíz-Rodríguez P, Moriana J, Limonero J… Cano-Vindel A, 2018 Factor structure and measurement invariance across various demographic groups and over time for the PHQ-9 in primary care patients in Spain. PLoS One, 13(2), e0193356. doi: 10.1371/journal.pone.0193356 [PubMed: 29474410]

González HM, Tarraf W, Whitfield KE, & Vega WA, 2010 The epidemiology of major depression and ethnicity in the United States. J Psychiatr Res, 44, 1043–1051. doi: 10.1016/j.jpsychires.2010.03.017 [PubMed: 20537350]

Gonzalez J, & Trickett E, 2014 Collaborative measurement development as a tool in CBPR: Measurement development and adaptation within the cultures of communities. Am. J. Community Psychol, 54(1–2), 112–124. doi: 10.1007/s10464-014-9655-1 [PubMed: 24748283]

Gothwal V, Bagga D, Bharani S, Sumalini R, & Reddy S, 2014). The Patient Health Questionnaire-9: Validation among patients withgGlaucoma. PloS ONE, 9(7), e101295. doi: 10.1371/journal.pone.0101295 [PubMed: 24999659]

Hage M, & Azar S, 2012 The link between thyroid function and depression. Journal of Thyroid Research, 2012, 1–8. doi: 10.1155/2012/590648

Hasin DS, Goodwin RD, Stinson FS, & Grant BF, 2005 Epidemiology of major depressive disorder: results from the National Epidemiologic Survey on Alcoholism and Related Conditions. Arch. Gen. Psychiatry, 62, 1097–1106. doi: 10.1001/archpsyc.62.10.1097 [PubMed: 16203955]

Hepner KA, Morales LS, Hays RD, Edelen MO, & Miranda J, 2008 Evaluating differential item functioning of the PRIME-MD mood module among impoverished black and white women in primary care. Women's Health Issues, 18, 53–61. doi: 10.1016/j.whi.2007.10.001 [PubMed: 18069001]

Hoyt W, 2000 Rater bias in psychological research: When is it a problem and what can we do about it? Psychol Methods, 5(1), 64–86. doi: 10.1037/1082-989x.5.1.64 [PubMed: 10937323]

Huang FY, Chung H, Kroenke K, Delucchi KL, & Spitzer RL, 2006 Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. J Gen Intern Med, 21, 547–552. doi: 10.1111/j.1525-1497.2006.00409.x [PubMed: 16808734]

Johns Hopkins University Bloomberg School of Public Health; Applied Mental Health Research Group, 2013 Design, implementation, monitoring, and evaluation of mental health and psychosocial assistance programs for trauma survivors in low-resource countries: A user's manual for researchers and program implementers (adult version). http://hopkinshumanitarianhealth.org/assets/documents/VOT_DIME_MODULE1_FINAL.PDF (accessed 7 June 2020).

Kirmayer LJ, & Young A, 1998 Culture and somatization: clinical, epidemiological, and ethnographic perspectives. Psychosom Med, 60, 420–430. [PubMed: 9710287]

Kroenke K, & Spitzer RL, 2002 The PHQ-9: A new depression diagnostic and severity measure. Psychiatr Ann, 32, 509–515. doi: 10.3928/0048-5713-20020901-06

Kroenke K, Spitzer RL, & Williams JB, 2001 The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med, 16, 606–613. doi: 10.1046/j.15251497.2001.016009606.x [PubMed: 11556941]

Kroenke K, Spitzer RL, Williams JB, & Löwe B, 2010 The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen Hosp Psychiatry, 32, 345–359. doi: 10.1016/jgenhosppsych.2010.03.006 [PubMed: 20633738]

Levis B, Benedetti A, & Thombs BD, 2019 Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. BMJ, 365, l1476. doi: 10.1136/bmj.l1476 [PubMed: 30967483]

Liu SI, Yeh ZT, Huang HC, Sun FJ, Tjung JJ, Hwang LC, … & Yeh AWC, 2011 Validation of Patient Health Questionnaire for depression screening among primary care patients in Taiwan. Compr Psychiatry, 52(1), 96–101. doi: 10.1016/j.comppsych.2010.04.013 [PubMed: 21111406]

Lotrakul M, Sumrithe S, & Saipanish R, 2008 Reliability and validity of the Thai version of the PHQ-9. BMC Psychiatry, 8, 1–7. doi: 10.1186/1471-244x-8/46 [PubMed: 18173833]

Löwe B, Kroenke K, Herzog W, & Gräfe K, 2004a Measuring depression outcome with a brief self-report instrument: sensitivity to change of the Patient Health Questionnaire (PHQ-9). J Affect Disord, 81, 61–66. [PubMed: 15183601]

Löwe B, Spitzer RL, Gräfe K, Kroenke K, Quenter A, Zipfel S, … & Herzog W, 2004b Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. J Affect Disord, 78, 131–140. doi: 10.1016/s0165-0327(02)00237-9 [PubMed: 14706723]

McHugh M, 2012 Interrater reliability: The kappa statistic. Biochemia Medica, 276–282. doi: 10.11613/bm.2012.031 [PubMed: 23092060]

Miller P, Baxter S, Royer J, Hitchcock D, Smith A, & Collins K… Finney C, 2015 Children's social desirability: Effects of test assessment mode. Pers Individ Differ, 83, 85–90. doi: 10.1016/j.paid.2015.03.039

Moher D, Liberati A, Tetzlaff J, & Altman D, 2009 Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. BMJ, 339, b2535–b2535. doi: 10.1136/bmj.b2535 [PubMed: 19622551]

Muliyala KP, & Varghese M, 2010 The complex relationship between depression and dementia. Ann Indian Acad Neurol, 13(Suppl2), S69–S73. [PubMed: 21369421]

Muñoz-Navarro R, Cano-Vindel A, Medrano L, Schmitz F, Ruiz-Rodríguez P, Abellán-Maeso C… Hermosilla-Pasamar A, 2017 Utility of the PHQ-9 to identify major depressive disorder in adult patients in Spanish primary care centres. BMC Psychiatry, 17(1). doi: 10.1186/s12888-017-1450-8

Paykel ES, Brugha T, & Fryers T, 2005 Size and burden of depressive disorders in Europe. Eur Neuropsychopharmacol, 15, 411–423. doi: 10.1186/s40359-018-0238-z [PubMed: 15950441]

Petrea I, 2012 Mental health in former Soviet countries: From past legacies to modern practices. Public Health Reviews, 34(2). doi: 10.1007/bf03391673

Pfizer Inc. Patient health questionnaires (PHQ) screeners, official website. (2013). http://www.phqscreeners.com/overview.aspx?Screener=02_PHQ-9.

Ruggeri K, Maguire Á, Andrews J, Martin E, & Menon S, 2016 Are we there yet? Exploring the impact of translating cognitive tests for dementia using mobile technology in an aging population. Frontiers Aging Neurosci, 8. doi: 10.3389/fnagi.2016.00021

Sijtsma K, 2008 On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika, 74(1), 107–120. doi: 10.1007/s11336-008-9101-0 [PubMed: 20037639]

Storch EA, Roberti JW, Roth DA, 2004 Factor structure, concurrent validity, and internal consistency of the Beck Depression Inventory-Second Edition in a sample of college students. Depress. Anxiety, 19:187–189. [PubMed: 15129421]

Tavakol M, & Dennick R, 2011 Making sense of Cronbach's alpha. International Journal of Medical Education, 2, 53–33. [PubMed: 28029643]

van de Vijver FJR, & Leung K, 2000 Methodological issues in psychological research on culture. Cross-Cult. Psychol, 31(1), 33–51.

Velligan D, Lopez L, Castillo D, Manaugh B, Milam A, & Miller A 2011 Interrater reliability of using brief standardized outcome measures in a community mental health setting. Psychiatr Serv, 62(5), 558–560. doi: 10.1176/ps.62.5.pss6205_0558 [PubMed: 21532087]

Wennerstrom A, Vannoy SD, Allen CE, Meyers D, O'Toole E, Wells KB, & Springgate BF, 2011 Community-based participatory development of a community health worker mental health outreach role to extend collaborative care in post-Katrina New Orleans. Ethn Dis, 21, S1–45–51.

Whiteford H, Degenhardt L, Rehm J, Baxter A, Ferrari A, Erskine H… Vos T, 2013 Global burden of disease attributable to mental and substance use disorders: Findings from the Global Burden of Disease Study 2010. Lancet, 382(9904), 1575–1586. doi: 10.1016/s0140-6736(13)61611-6 [PubMed: 23993280]

Whisman MA, Perez JE, Ramel W, 2000 Factor structure of the Beck Depression Inventory-Second Edition (BDI-II) in a student sample. J. Clin. Psychol, 56:545–551. [PubMed: 10775046]

Wittchen HU, & Jacobi F, 2005 Size and burden of mental disorders in Europe—a critical review and appraisal of 27 studies. Eur Neuropsychopharmacol, 15, 357–376. doi: 10.1016/j.euroneuro.2005.04.012 [PubMed: 15961293]

World Bank Country and Lending Groups, n.d. Retrieved from https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-countryand-lending-groups

World Health Organization, 2008 Integrating mental health into primary care: a global perspective.

World Health Organization, 2004 The global burden of disease: 2004 update.

World Health Organization, n.d. Process of translation and adaptation of instruments. Retrieved from https://www.who.int/substance_abuse/research_tools/translation/en/

Zailinawati AH, Schattner P, & Mazza D, 2006 Doing a pilot study: Why is it essential? Malaysian Family Physician. 1(2&3), 70–73. [PubMed: 27570591]

Zayas LH, & Gulbas LE, 2012 Are suicide attempts by young Latinas a cultural idiom of distress?. Transcult Psychiatry, 49, 718–734. doi: 10.1177/1363461512463262 [PubMed: 23075802]

HIGHLIGHTS

- Measure validation in diverse settings is key to account for contextual nuance
- Consider impact of medical comorbidities, administration, language on psychometrics
- Thorough reporting of assessment procedures aids in assessing measure performance

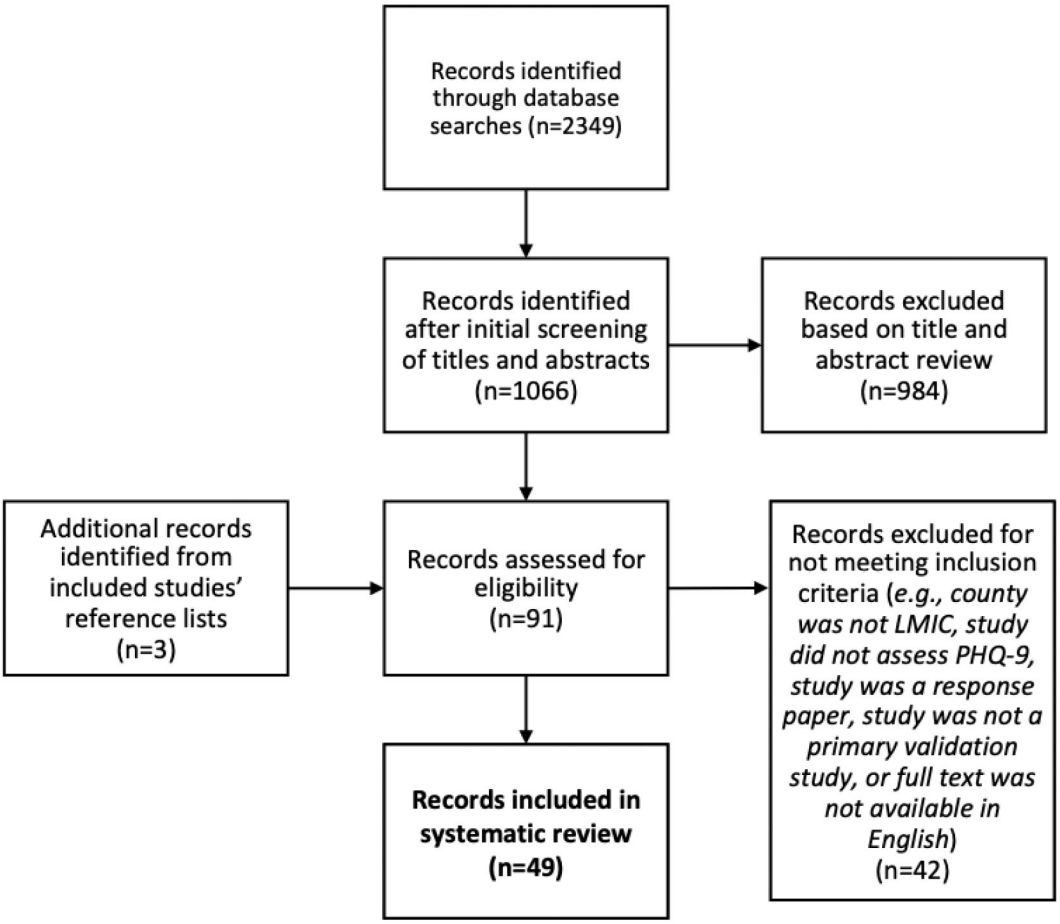

Figure 1.
Summary of inclusion process.

Table 1.

Search terms used in PubMed and PsychInfo.

Measure Classifiers	**“patient health questionnaire” or “phq” or “phq-9”**
Low and Middle Income Country Classifiers	afghan* or albania* or algeria* or american samoa* or angola* or antigua* or argentin* or armenia* or azerbaijan* or bangladesh* or belarus* or beliz* or benin* or bhutan* or bolivia* or bosnia* or brazil* or bulgaria* or burkina* or burundi* or cabo verde* or cambodia* or cameroon* or central africa republic* or chad* or chile* or china or chinese or colombia* or comoros or comorian or congo* or costa rica* or cote d’ivoire or ivory coast or djibouti* or cuba or cuban or dominica* or equador* or egypt* or el salvador or eritrea* or guinea* or ethiopia* or fiji* or gabon* or gambia* or georgia* or ghana* or grenad* or guatemala* or guyana* or haiti* or hondura* or hungar* or india* or indonesia* or iran* or iraq* or jamaica* or jordan* or kazakh* or kenya* or kiribati* or korea* or kosovo* or kyrgyz* or laos* or laotian* or latvia* or leban* or lesotho* or liberia* or libya* or lithuania* or macedonia* or malawi* or madagasca* or malay* or maldiv* or mali or marshall island* or malta* or mauri* or mauritius* or mexic* or micronesia* or moldova* or mongolia* or montenegr* or morocc* or mozambi* or myanm* or burm* or namibia* or nauru* or nepal* or nicaragua* or niger* or northern mariana* or oman* or pakistan* or palau* or panam* or paraguay* or peru* or philippin* or filipin* or poland* or polish* or puerto ric* or romania* or russia* or rwanda* or samoa* or sao tome* or senegal* or serbia* or seychell* or sierra leonne* or slovak* or solomon* or somalia* or south africa* or sri lanka* or lucia* or kitts* or vincent or grenadines or sudan* or surinam* or swazi* or syria* or tajikistan* or tanzania* or thai* or timor* or togo* or tonga* or trinidad* or tunisia* or turkey* or turkish* or turkmenistan* or tuvalu* or uganda* or ukrain* or uruguay* or uzbekistan* or vanuatu* or venezuela* or vietnam* or gaza* or yemen* or zambia* or zimbabwe* or serbia* or mayotte* or Sub-Sahara* or Sahara* or Africa* or SSA or Asia* or Pacific our South America* or Latin America* or Central America* or East Europe* or Eastern Europe* or low income countr* or middle income countr* or LIC or LICs or MIC or MICs

Table 2.

Number of validation studies by sub-region and country.

Region (# of Studies)		Countries Included (# of Studies)
Africa (15 studies, 10 countries)	Central (1)	Cameroon (1)
	East (7)	Kenya (2), Somalia (1), Uganda (2), Ethiopia (2)
	South (4)	South Africa (3), Zimbabwe (1)
	West (3)	Nigeria (1), Ghana (1), Cote d'Ivoire (1)
America (9 studies, 6 countries)	Caribbean (1)	Haiti (1)
	North (2)	Mexico (2)
	South (6)	Peru (2), Chile (1), Brazil (2), Colombia (1)
Asia (25 studies, 10 countries)	East (11)	China (11)
	South (9)	Pakistan (1), India (6), Sri Lanka (1), Nepal (1)
	Southeast (3)	Malaysia (1), Thailand (1), Vietnam (1)
	West (2)	Lebanon (1), Iran (1),
Total	49 studies	26 countries

* One study was in both Ghana and Cote d'Ivoire, though was counted only in Cote d'Ivoire.

Table 3.

Study design for cross cultural adaptation.

Coding Criteria: Cross Cultural Adaptation	N	%
Were medical comorbidities measured?	15	31
Was the method of measure administration reported?	29	59
Was research assistant/clinician training reported?	21	43
Was research assistant /data collector from country of study? (this information must be reported explicitly)	3	6
Were patients fluent in the language of measure? (this must be explicitly stated)	14	29
Was language proficiency assessed? (needs to be a formal assessment of language ability)	1	2
Was educational attainment assessed? (more detailed information about the level of education as a measured variable beyond simply acting as an inclusion criteria)	33	67
Were translation procedures reported?	21	43
Was pilot testing reported?	10	20
Was the local community involved in measure adaptation?	6	12

Table 4.

Evidence for reliability and validity.

Reported Reliability and Validity Data	N	%
Reliability		
Internal consistency	42	86
Test-retest	12	24
Inter-rater	1	2
Validity		
Criterion (measure related to gold standard)	29	59
Sensitivity	37	76
Specificity	37	76
Convergent (related to other things you would “expect”)	23	47
Discriminant (not related to things you would expect)	1	2
Predictive Validity (does measure predict related construct in future, longitudinal)	2	4
Factor Structure	21	43

Table 5.

Key recommendations for validation studies.

	Key Recommendations
1	Attend to context when validating measures, perhaps evidenced by using data collectors from country of origin or including the local community in measure adaption.
	Illustrative Example: Recruitment of field workers from local communities who are fluent in relevant languages and trained by psychologists (e.g., Bhana et al., 2015), or including local populations in a review panel to provide feedback on measure translation (e.g., Kochhar et al., 2007). Application to LMICs: Identify key informants to provide basic insights into psychological concepts, complete small-scale pilot tests to assess measure functioning, clearly report community involvement and identify any lack of community involvement. If not feasible, provide additional clarity around limitations to the measure validation.
2	Pursue multimethod validation that encompasses both qualitative and quantitative approaches. This is specifically important in terms of assessing local knowledge and understanding of mental health constructs.
	Illustrative Example: Conducted aualitative focus groups on measure adaptation in local populations and make suggested changes to measure and confirm changes with subsequent quantitative evaluation (e.g., Pence et al., 2012). Application to LMICs: Recruit patient and non-patient populations for qualitative and quantitative evaluation. Consider following the qualitative procedures within the design, implementation, monitoring, and evaluation model (Module 1, DIME, Johns Hopkins University Bloomberg School of Public Health Applied Mental Health Research Group, 2013).
3	Consider the psychometric impact of administering instruments verbally, particularly if measurers were designed as “pencil and paper” instruments. If verbal administration is indicated, discuss methods used to train staff.
	Illustrative Example: Reported if the measure was self-administered (e.g., Sherina et al, 2012) or interviewer administered and the procedures for training of interviewers (Nakku et al., 2016). Application to LMICs: Report methods used for measure administration clearly in manuscript, and if deciding to change measure administration from method in prior validation studies (e.g., paper and pencil to in-person interviewer) provide justification, describe adaptation procedures, and conduct validation on the changed measure.
4	Provide clear descriptions of limitations or caveats of using adapted measures (e.g., issues related to language fluency that may have impacted validation).
	Illustrative Example: Report specific steps taken to ensure thorough translation (e.g., Gelaye et al., 2013; Bhana et al., 2015). Note discrepancies or concerns about translation that may impact a measure’s use (Gothwal et al., 2014). Confirm that participants are fluent in the language in the measure and consider assessing participant’s ability to read and write (e.g., Arrieta et al., 2017; Cholera et al., 2015). Application to LMICs: Providing thorough descriptions of translation procedures, including adherence to international guidance, is suggested. Authors should recognize that while a measure may be translated into a dominant language in a given country, other cultural groups may lack knowledge of the dominant language or may use other preferred dialects. Ensure that participants’ ability to accurately respond to a measure is not limited by inability to read or write.
5	Report inter-rater reliability, particularly if instruments are administered verbally.
	Illustrative Example: Indicate inter-rater reliability among data collection staff to provide insight into any systematic error that arose during measure administration that may impact results (Adewuya et al., 2006). Application to LMICs: Issues related to literacy or cultural preferences may impact traditional delivery of measures (e.g., resulting in verbal administration of a traditional paper measure). Minor variations in phrasing of assessment items or in body language may influence scores. Reporting inter-rater reliability provides evidence that measures can be used with fidelity and also indicates potential areas where research staff may need further training to ensure the reproducibility of assessment results.