NIH Public Access
Author Manuscript
J Affect Disord. Author manuscript; available in PMC 2015 April 01.

Published in final edited form as:
J Affect Disord. 2014 April ; 158: 148–153. doi:10.1016/j.jad.2014.02.012.

Cross-cultural Validity of the Spanish Version of PHQ-9 among Pregnant Peruvian Women: A Rasch Item Response Theory Analysis

Qiuyue Zhong[1], **Bizu Gelaye**[1], **Jesse R. Fann**[2], **Sixto E Sanchez**[3,4], and **Michelle A Williams**[1]

[1]Department of Epidemiology, Harvard School of Public Health, Boston, MA

[2]Departments of Psychiatry and Behavioral Sciences, Rehabilitation Medicine and Epidemiology, University of Washington, Seattle, WA

[3]Universidad San Martin de Porres, Lima, Peru

[4]Asociación Civil PROESA, Lima, Peru

Abstract

Objective—We sought to evaluate the validity of the Spanish language version of the Patient Health Questionnaire-9 (PHQ-9) depression scale in a large sample of pregnant Peruvian women using Rasch item response theory (IRT) approaches. We further sought to examine the appropriateness of the response formats, reliability and potential differential item functioning (DIF) by maternal age, educational attainment and employment status.

Methods—This cross-sectional study was conducted among 1,520 pregnant women in Lima, Peru. A structured interview was used to collect information on demographic characteristics and PHQ-9 items. Data from the PHQ-9 were fitted to the Rasch IRT model and tested for appropriate category ordering, the assumptions of unidimensionality and local independence, item fit, reliability and presence of DIF.

Results—The Spanish language version of PHQ-9 demonstrated unidimensionality, local independence, and acceptable fit for the Rasch IRT model. However, we detected disordered response categories for the original four response groups. After collapsing “more than half the days” and “nearly every day”, the response categories ordered properly and the PHQ-9 fit the Rasch IRT model. The PHQ-9 had moderate internal consistency (person separation index, PSI =

© 2014 Elsevier B.V. All rights reserved.

Correspondence and requests for reprint: Bizu Gelaye, PhD, MPH, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Ave, K505, Boston, MA 02115 USA, Telephone: 617-432-1071, Facsimile: 617-566-7805, bgelaye@hsph.harvard.edu.

AUTHOR CONTRIBUTIONS: BG and MAW conceived and designed the study. QZ and BG analyzed data for the study. All authors interpreted the data, critically revised the draft for important intellectual content, and gave final approval of the manuscript to be published.

CONFLICT OF INTEREST: The authors have no competing interests to declare.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

0.72). Additionally, the items of PHQ-9 were free of DIF with regard to age, educational attainment, and employment status.

Conclusions—The Spanish version of the PHQ-9 was shown to have item properties of an effective screening instrument. Collapsing rating scale categories and reconstructing three-point Likert scale for all items improved the fit of the instrument. Future studies are warranted to establish new cut-off scores and criterion validity of the three-point Likert scale response options for the Spanish language version of the PHQ-9.

Introduction

The Patient Health Questionnaire-9 (PHQ-9) is a very brief, easy to administer and interpret depression screening instrument (Kroenke et al., 2001). Because of its brevity and demonstrated good reliability and validity (Kroenke et al., 2001), the PHQ-9 is one of the most widely used depression screening instrument in primary care settings among racially and ethnically diverse populations. The PHQ-9 has generally been validated under the framework of classic test theory (CTT) or traditional psychometric approaches. However, these traditional techniques often provide inaccurate diagnosis and hide the heterogeneity that exists in each specific item (Hobart et al., 2007, Packham and MacDermid, 2013). These methods are now being complemented and, in some cases replaced, by item response theory (IRT) approaches and particularly by the application of the Rasch models.

Rasch IRT models are considered the most appropriate and robust methods to examine the measurement properties of rating scales such as the PHQ-9 (Hobart et al., 2007). Specifically, application of Rasch IRT models for analysis of the PHQ-9 provides an opportunity to identify and subsequently reduce the potential bias that may exist when using the depression screening instruments in new cultural settings. Consequently, results from analyses of Rasch models can be used to increase the validity and utility of depression screening results when the PHQ-9 is used in culturally diverse settings.

To date, only five studies have evaluated the properties of the PHQ-9 using Rasch IRT models (Gelaye et al., 2013; Williams et al., 2009; Smith et al., 2008; Smith et al., 2009; Lamoureux et al., 2009) and none included the Spanish language version of PHQ-9. Although the PHQ has been used in studies of pregnant women (Melville et al., 2010; Spitzer et al., 2000) its validity has not been confirmed using Rasch models. Given the high prevalence of depression in this population.(World Health Organization, 2008), it is important to evaluate the validity and utility of PHQ-9. In addition, symptomology characteristics of pregnant women are different from those of non-pregnant women (Yawn et al., 2009; Yonkers et al., 2009; Ross et al., 2003). Significant potential benefit will be gained to the mother and fetus/infant if depression is detected and managed (Adewuya et al., 2006).. Therefore the aim of this study is to validate the Spanish language version of the PHQ-9 among pregnant Peruvian women using Rasch IRT models, and to examine the appropriateness of the response format, psychometric validity and potential bias of items by age, education level and employment status. Clinically, the PHQ-9 can be used as a time-effective screening instrument to potentially identify pregnant women at risk for depression in resource limited clinical settings.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Methods

Study Population

This cross-sectional investigation was a part of Pregnancy Outcomes, Maternal and Infant Study (PrOMIS) Cohort. The PrOMIS Cohort is an ongoing prospective cohort study of pregnant women enrolled in prenatal care clinics at the Instituto Nacional Materno Perinatal (INMP) in Lima, Peru. The INMP is the main reference establishment for maternal and perinatal care operated by the Ministry of Health of the Peruvian government. Women who attended the INMP for their first prenatal care visit from February 2012 to March 2013 were recruited for this investigation. Pregnant women who were 18-49 years of age, with a gestational age ≤ 16 weeks, and who spoke and understood Spanish were eligible for inclusion in this cohort. Informed consent was provided for all participants. The institutional review boards of the INMP and the Harvard School of Public Health Office of Human Research Administration approved all procedures used in this study.

Data Collection and Variable Specification

Participants were interviewed in a private setting by trained research personnel using a structured questionnaire. Information regarding maternal socio-demographic, lifestyle characteristics, medical and reproductive histories and experience with symptoms of depression were collected from the structured questionnaire. Of the 1,810 women approached, 1,556 completed the interview. Due to missing information on the PHQ-9, 26 participants were excluded, hence a total of 1,520 participants were included in the present analysis.

The Patient Health Questionnaire (PHQ-9)

The PHQ-9 is a nine-item depression-screening instrument (Kroenke and Spitzer, 2002; Kroenke et al., 2001; Spitzer et al., 1999). It was developed based on the criteria from the *Diagnostic and Statistical Manual of Mental Disorders, Fourth edition* (DSM-IV). Each item was rated on the frequency of a depressive symptom in the two weeks prior to evaluation. The PHQ-9 assessed nine depressive symptoms experienced by participants including: 1) anhedonia, 2) depressed mood, 3) insomnia or hypersomnia, 4) fatigue or loss of energy, 5) appetite disturbances, 6) guilt or worthlessness, 7) diminished ability to think or concentrate, 8) psychomotor agitation or retardation, and 9) suicidal thoughts. The PHQ-9 score was calculated by assigning a score of 0, 1, 2, and 3, to the response categories of “not at all,” “several days,” “more than half the days,” and “nearly every day,” respectively.

Statistical Analysis

The Rasch IRT analysis is an item-based approach where ordinal observed item scores are transformed to linear measures representing the underlying latent construct (Bond and Fox, 2013). The Rasch IRT is based on a mathematical model where the probability for endorsing an item is a logistic function of the difference between the person's level of depression and the level of depression expressed by the item (item difficulty) (Bond and Fox, 2013; Lamoureux et al., 2009; Rasch, 1993; Forkmann et al., 2013). We applied the Andrich Rating Scale Model (RSM) to evaluate the PHQ-9 (Andrich, 1978b; Andrich, 1978a). A

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

single set of response thresholds was estimated for the nine items (Andrich, 1978b; Andrich, 1978a).

Evaluation of Rasch IRT Model Assumptions

Prior to estimating Rasch IRT models, we evaluated the following assumptions (1) ascending ordering of response categories, (2) local independence, (3) unidimensionality, (4) item fit and, (5) reliability as previously described by Forkmann and colleagues (Forkmann et al., 2013). First, we examined whether response category thresholds, namely, the transition points between adjacent response categories ("not at all," "several days," "more than half the days" and "nearly every day"), were properly ordered. Disordered categories indicate respondents' difficulty in discriminating between response categories. In such instances collapsing adjacent categories of disordered thresholds might improve overall fit to the model (Linacre, 2012; Pallant and Tennant, 2007). Next, we used principal component analysis (PCA) of the residuals to assess whether the assumption of unidimensionality was met. In order to satisfy unidimensionality assumption for IRT Rasch models the eigenvalues of the residuals should be less than 2.0. After the Rasch factor is taken into account, no further associations should occur (Lamoureux et al., 2009; Linacre, 2012). We then assessed the residual correlation matrix to identify local dependency. Local dependency occurs when the answer to one item is dependent on the answer to another item (Williams et al., 2009). An item correlation of 0.3 or higher is considered as evidence of local independency (Lambert et al., 2013).

We evaluated the item fit using infit mean square (MnSq) and outfit MnSq statistics. The infit MnSq was more sensitive to unexpected response of an individual's ability level (Bond and Fox, 2013; Linacre, 2012) and the outfit MnSq was more sensitive to the influence of outlying scores (Bond and Fox, 2013). The ideal expected value of the infit and outfit MnSq is 1. Values above this ideal statistics (MnSq>1) can be interpreted as demonstrating more variation between expected and observed data, indicating an underfit. When the fit statistics was less than the ideal value (MnSq<1), it suggested less variation between the expected and observed data, indicating an over fit (Bond and Fox, 2013). Fit statistics higher than 1.5 and below 0.5 indicate too much and too little variation in response patterns and should be considered for removal from the instrument to improve fit (Tennant and Conaghan, 2007). Therefore, in our study, we used mean square infit and outfit criteria of between 0.7 and 1.40 (more stringent criterion) to test model misfit (Tennant and Conaghan, 2007). Finally, we computed the person separation index (PSI) to evaluate internal consistency for PHQ-9 under the Rasch IRT model. Values above 0.7 are generally considered as acceptable (Fisher, 1992). Rasch IRT model analysis was performed using Winsteps software (Version 3.80.0, Chicago, Illinois).

Differential Item Functioning (DIF)

We completed DIF analysis, to determine the extent to which, if at all, maternal responses to individual PHQ-9 items differed according to maternal age (<35 versus ≥35 years), educational attainment (<12 versus ≥12 years of education) and employment status (no versus yes) (Forkmann et al., 2013; Dye et al., 2013; Forjaz et al., 2009).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

We used an R LORDIF package to evaluate DIF (Choi et al., 2011). LORDIF detects DIF using ordinal logistic regression procedures. Presence of DIF was evaluated across age, education level and employment status. We tested both uniform and non-uniform DIF based on likelihood ratio tests (Choi et al., 2011). Given the possibility of multiple comparison problems we used the Bonferroni adjusted type I error rate (α=0.05/9=0.0056). Additionally given that the statistical power of likelihood ratio test is sample size dependent (Choi et al., 2011), we examined the magnitude of DIF based on the effect size measure McFadden's pseudo R^2. We considered DIF to be as negligible (pseudo R^2 <0.13), moderate (pseudo R^2 between 0.13 and 0.26) and large (pseudo R^2>0.26) (Zumbo, 1999).

Results

Participant Characteristics

Participants' socio-demographic and reproductive characteristics are summarized in Table 1. Briefly, the mean age of participants was 28.0 years (standard deviation = 6.2 years). The majority of participants were Mestizo (75.5%), with at least 7 years of education (95.5%), and multigravida (68.2%). On average, participants were interviewed at a gestational age of 9.8 weeks (standard deviation = 3.4 weeks). More than half of study participants were not employed (56.5%), and reported that the current pregnancy was unplanned (57.6%).

Evaluation of Rasch Model Assumptions

In the initial analysis, the thresholds of four categories did not appear to increase monotonically. The threshold for categories “more than half the days” and “nearly every day” was lower than that of “several days” and “more than half the days” (Figure 1). Additionally, by examining the item category probability curve of the PHQ-9, “more than half the days” was never the most probable response category for any participant, indicating that “more than half the days” was redundant (Figure 1). To improve model fit, we collapsed “more than half the days” and “nearly every day.” After combining, the new item category probability curve had a smooth distribution with non-descending category thresholds. Based on the principal component analysis of the residuals, the eigenvalue of the first contrast after considering the Rasch factor was 1.6, hence the assumption of unidimensionality of PHQ-9 held in this context. The largest positive correlation was 0.12 between item 6 (low self-esteem) and item 9 (suicidal ideation). No pairs of items had correlation >0.3, hence again, the assumption of local independency held in this context. Before collapsing, the infit MnSq ranged from 0.62 to 1.47. For item 8 (psychomotor agitation or retardation), we observed evidence of model misfit as the outfit MnSq values were between 0.8 and 1.14, which are incompatible with the range of values considered to reflect a good fit (0.7-1.4) (Tennant and Conaghan, 2007). After collapsing response categories, the model fit was improved and both infit and outfit MnSq fell in the acceptable range of values. Item 9 (suicidal ideation) had the highest infit MnSq value, 1.36. The PSI for the current model was 0.72, reflecting a moderate internal consistency of PHQ-9. The item difficulties in logits ranged from -1.56 (the highest level of symptomatology) for item 4 (low energy) to 2.04 (the lowest level of symptomatology) for item 9 (suicidal ideation) after collapsing “more than half the days” and “nearly every day.” (Supplement Figure 1)

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Differntial Item Functioning

Responses for item 7 (concentration difficulty) and item 9 (suicidal ideation) differed according to educational status. However, both of the DIFs were negligible, with pseudo R^2 =0.0040 for item 7 and pseudo R^2 =0.0085 for item 9, respectively. Similarly, we detected a DIF by age. In participants with the same level of depression, younger women (< 35 years) respond in different ways compared to women >= 35 years old to item 1 (anhedonia) with *P*-value < 0.0001 (Supplement Figure 2). However, the magnitude for this DIF was also negligible with pseudo R^2 = 0.0079. No DIF was detected for employment status.

Discussion

The Spanish language version of PHQ-9 demonstrated unidimensionality, local independence, adequate fit for the Rasch IRT model and moderate reliability. However, we detected disordered response categories for the original four response categories (“not at all,” “several days,” “more than half the days,” and “nearly every day”). After collapsing the latter two categories, the response categories ordered properly and the PHQ-9 fit the Rasch IRT model. As it intended, the Spanish language version of the PHQ-9, when used among pregnant Peruvian women, was found to measure one underlying construct (i.e., depression). Additionally, we noted that each item of the PHQ-9 captured one unique feature of depression. Lastly, we found that maternal responses to the 9-items were not materially influenced by maternal characteristics such as age, educational attainment and employment status.

Our study results showing disordered response categories have been reported by others (Lamoureux et al., 2009; Williams et al., 2009). The presence of too many response categories or similar labeling of response options may result in disordered categories (Lamoureux et al., 2009). Collapsing response options has been one suggested solution to disordered categories (Lamoureux et al., 2009). Indeed, in our study as “more than half the days” was found to be a redundant response category. Hence, when we collapsed “more than half of the days” with “nearly every day,” the fit of the Rasch model improved substantially. Of note, Lamoureux (Lamoureux et al., 2009) suggested that “several days” and “more than half the days” were used interchangeably in people with visual impairment. This suggestion is inline with the categorical algorithm of the original PHQ-9 as suggested by Spitzer et al (Spitzer et al., 1999). The categorical algorithm dichotomizes the four response categories, for all PHQ-9 items except “suicidal ideation,” into two response categories: “not at all and several days” and “more than half the days and nearly every day” to define presence of depression (Spitzer et al., 1999). When too many categories existed, respondents needed to make subtle distinctions between response categories. Collapsing categories would not only improve model fit for Rasch IRT model but also reduce respondents' burden (Williams et al., 2009). However, collapsing would change the original cutoff scores for possible depression and depression severity (Lamoureux et al., 2009). Further validation studies are needed to confirm our observation, establish new clinically important cutoff scores, and examine criterion validity for this new three-point Likert scale instrument.

Another important point that merits discussion is the issue of the double-barreled items with opposite symptoms (Williams et al., 2009). These items include item 3 (too little or too

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

much sleep), item 5 (loss of appetite or overeating) and item 8 (psychomotor agitation or retardation). In our study, before collapsing the disordered categories, item 8 did not fit with the assumption of the Rasch IRT model. Williams et al (2009) in their study among spinal cord injury patients also detected misfit for item 8 (Williams et al., 2009). After collapsing response categories, however, we found no misfitting items, indicating that collapsing response categories would increase the item fit even in the presence of the double-barreled items. Some investigators suggest splitting of the double-barreled items into singular-dimensional items to improve instrument properties (Williams et al., 2009). Studies are needed to confirm and examine the extent to which splitting items of the Spanish language version PHQ-9 instrument would improve the instrument's psychometric properties (Gelaye et al., 2013).

The DIF analysis showed that responses to the 9 items of the PHQ-9 questionnaire were invariant for maternal characteristics such as maternal age, educational attainment and employment status. The lack of DIF along with the results of the Rasch IRT model support that the construct validity of the PHQ-9 across cross-cultural settings with different characteristics in assessing depressive symptoms.

Some limitations should be considered when interpreting results from this study. The Rasch model is a one-parameter IRT model that assumes the item discrimination is the same for all items. Under this assumption, easy to endorse items (for example, item 4, low energy) discriminate as well as difficult to endorse items (for example, item 9, suicidal ideation) for women with lower level of depressive symptoms (Harris, 1989), which is not always true. Future studies using 2- or 3- parameter IRT models to validate the PHQ-9 are warranted.

To our knowledge, this is the first time that the Spanish language version of the PHQ-9 has been subjected to a Rasch IRT analysis. Given the sample-independent property (Scheiblechner, 2009) of the Rasch IRT model analysis, this study provides strong evidence for the validity of the Spanish version of the PHQ-9. The sample size in our analysis also allowed for examining the nuances of item response bias. Moreover, the Rasch analysis enabled a detailed examination of the structure and operation of the PHQ-9 among sub-groups of women. The PHQ-9 demonstrated its appropriateness to be used in this population.

In summary, the application of the Rasch IRT model has provided further evidence to the reliability and validity of the Spanish language version of the PHQ-9 scale. Of note, the analysis showed that the items are unidimentional, locally independent, and reliable with acceptable item fit. The original four response categories, however, appear disordered. After collapsing response categories “more than half the days” and “nearly every day” to reconstruct a three-point Likert scale, the response options are properly ordered and the items fit the Rasch IRT model. Future studies are warranted to establish new cut-off scores and criterion validity of the three-point Likert scale response options. Finally, our DIF analyses showed that the PHQ-9 scale assesses depressive symptoms independent of age, educational attainment and employment status. Beyond providing evidence of validity, this analysis illustrates the added value of using the Rasch IRT model to inspect individual items and response options in cross-cultural settings.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors wish to thank the dedicated staff members of Asociacion Civil Proyectos en Salud (PROESA), Peru and Instituto Especializado Materno Perinatal, Peru for their expert technical assistance with this research.

ROLE OF FUNDING SOURCE: This research was supported by an award from the National Institutes of Health (NIH), the Eunice Kennedy Shriver Institute of Child Health and Human Development (R01-HD-059835). The NIH had no further role in study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the paper for publication

References

Adewuya AO, Ola BA, Dada AO, Fasoto OO. Validation of the Edinburgh Postnatal Depression Scale as a screening tool for depression in late pregnancy among Nigerian women. Journal of Psychosomatic Obstetrics and Gynecology. 2006; 27:267–272. [PubMed: 17225628]

Andrich D. Application of a psychometric rating model to ordered categories which are scored with successive integers. Applied Psychological Measurement. 1978a; 2:581–594.

Andrich D. A rating formulation for ordered response categories. Psychometrika. 1978b; 43:561–573.

Bond, TG.; Fox, CM. Applying the Rasch model: Fundamental measurement in the human sciences. Psychology Press; New York, NY: 2013.

Choi SW, Gibbons LE, Crane PK. lordif: An R package for detecting Differential Item Functioning using Iterative Hybrid Ordinal Logistic Regression/Item Rresponse Theory and Monte Carlo simulations. Journal of Statistical Software. 2011; 39:1–30. [PubMed: 21572908]

Dye DC, Eakman AM, Bolton KM. Assessing the validity of the dynamic gait index in a balance disorders clinic: an application of Rasch analysis. Physical Therapy. 2013; 93:809–818. [PubMed: 23449914]

Fisher WP. Reliability statistics. Rasch Measurement Transactions. 1992; 6:238.

Forjaz MJ, Rodriguez-Blázquez C, Martinez-Martin P. Rasch analysis of the hospital anxiety and depression scale in Parkinson's disease. Movement Disorders. 2009; 24:526–532. [PubMed: 19097178]

Forkmann T, Gauggel S, Spangenberg L, Brähler E, Glaesmer H. Dimensional assessment of depressive severity in the elderly general population: psychometric evaluation of the PHQ-9 using Rasch Analysis. Journal of Affective Disorders. 2013; 148:323–330. [PubMed: 23411025]

Gelaye B, Williams MA, Lemma S, Deyessa N, Bahretibeb Y, Shibre T, Wondimagegn D, Lemenhe A, Fann JR, Vander Stoep A, Andrew Zhou XH. Validity of the Patient Health Questionnaire-9 for depression screening and diagnosis in East Africa. Psychiatry Research. 2013; 210:653–661. [PubMed: 23972787]

Harris D. Comparison of 1-, 2-, and 3-Parameter IRT Models. Educational Measurement: Issues and Practice. 1989; 8:35–41.

Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. The Lancet Neurology. 2007; 6:1094–1105.

Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatric Annals. 2002; 32:1–7.

Kroenke K, Spitzer RL, Williams JBW. The PHQ-9. Journal of General Internal Medicine. 2001; 16:606–613. [PubMed: 11556941]

Lambert SD, Pallant JF, Boyes AW, King MT, Britton B, Girgis A. A Rasch analysis of the Hospital Anxiety and Depression Scale (HADS) among cancer survivors. Psychological Assessment. 2013; 25:379–390. [PubMed: 23356678]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Lamoureux EL, Tee HW, Pesudovs K, Pallant JF, Keeffe JE, Rees G. Can clinicians use the PHQ-9 to assess depression in people with vision loss? Optometry and Vision Science. 2009; 86:139–145. [PubMed: 19156007]

Linacre, JM. A User's Guide to WINSTEPS ® MINISTEP Rasch-Model Computer Programs Program Manual 3.80.0. Chicago, Illinois: 2012. Winsteps.com

Melville JL, Gavin A, Guo Y, Fan MY, Katon WJ. Depressive disorders during pregnancy: prevalence and risk factors in a large urban sample. Obstetrics and Gynecology. 2010; 116:1064–1070. [PubMed: 20966690]

World Health Organization. Maternal mental health and child health and development in low and middle income countries: report of the meeting held in Geneva, Switzerland, January 30 - February 1. Vol. 2008. World Health Organization Press; Geneva: 2008.

Packham T, Macdermid JC. Measurement properties of the Patient-Rated Wrist and Hand Evaluation: Rasch analysis of responses from a traumatic hand injury population. Journal of Hand Therapy. 2013; 26:216–224. [PubMed: 23561017]

Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). British Journal of Clinical Psychology. 2007; 46:1–18. [PubMed: 17472198]

Rasch, G. Probabilistic models for some intelligence and attainment tests. MESA Press; Chicago, Illinois: 1993.

Ross LE, Gilbert Evans SE, Sellers EM, Romach MK. Measurement issues in postpartum depression part 1: anxiety as a feature of postpartum depression. Archives of Women's Mental Health. 2003; 6:51–57.

Scheiblechner HH. Rasch and pseudo-Rasch models: suitableness for practical test applications. Psychological Test and Assessment Modeling. 2009; 51:181.

Smith AB, Rush R, Fallowfield LJ, Velikova G, Sharpe M. Rasch fit statistics and sample size considerations for polytomous data. BMC Medical Research Methodology. 2008; 8:33. [PubMed: 18510722]

Smith AB, Rush R, Wright P, Stark D, Velikova G, Sharpe M. Validation of an item bank for detecting and assessing psychological distress in cancer patients. Psychooncology. 2009; 18:195–199. [PubMed: 18677714]

Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. The Journal of the American Medical Association. 1999; 282:1737–1744.

Spitzer RL, Williams JB, Kroenke K, Hornyak R, Mcmurray J. Validity and utility of the PRIME-MD patient health questionnaire in assessment of 3000 obstetric-gynecologic patients: the PRIME-MD Patient Health Questionnaire Obstetrics-Gynecology Study. American Journal of Obstetrics and Gynecology. 2000; 183:759–769. [PubMed: 10992206]

Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Care and Research. 2007; 57:1358–1362. [PubMed: 18050173]

Williams RT, Heinemann AW, Bode RK, Wilson CS, Fann JR, Tate DG. Improving measurement properties of the Patient Health Questionnaire-9 with rating scale analysis. Rehabilitation Psychology. 2009; 54:198–203. [PubMed: 19469610]

Yawn BP, Pace W, Wollan PC, Bertram S, Kurland M, Graham D, Dietrich A. Concordance of Edinburgh Postnatal Depression Scale (EPDS) and Patient Health Questionnaire (PHQ-9) to assess increased risk of depression among postpartum women. Journal of the American Board of Family Medicine. 2009; 22:483–491. [PubMed: 19734393]

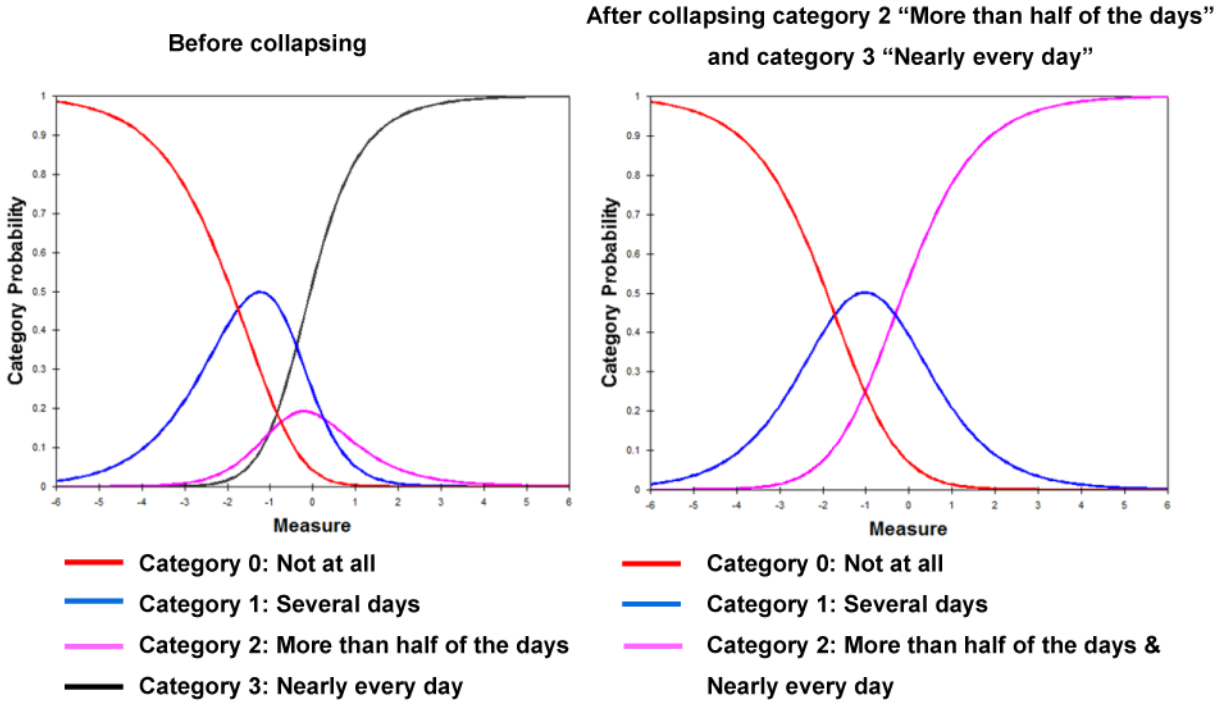
Yonkers KA, Smith MV, Gotman N, Belanger K. Typical somatic symptoms of pregnancy and their impact on a diagnosis of major depressive disorder. General Hospital Psychiatry. 2009; 31:327–333. [PubMed: 19555792]

Zumbo, BD. Directorate of Human Resources Research and Evaluation. Department of National Defense; Ottawa, ON: 1999. A handbook on the theory and methods of differential item functioning (DIF).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript



Summary of the Category Structure of the PHQ-9 before Collapsing Category 2 "More than half of the days" and Category 3 "Nearly every day"

Category	Observed count	Observed count %	Observed average measure	Outfit MnSq	Andrich Threshold
0	6719	49	-2.03	1.27	NONE
1	4043	30	-0.82	0.79	-0.91
2	959	7	0.10	1.15	1.09
3	1959	14	0.95	1.06	-0.18

Figure 1. Category Probability Curve of PHQ-9 (Before and After Collapsing Categories 2

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Sociodemographics and Reproductive Characteristics of the Study Population, Lima, Peru [a]

Characteristic	Participants (N=1520)	
	n	%
Age (mean ± standard devation)		28.0 ± 6.2
Age (years)		
18-20	89	5.9
20-29	859	56.5
30-34	306	20.1
>35	266	17.5
Education (years)		
<6	68	4.5
7-12	856	56.3
>12	596	39.2
Married/living with a partner	1238	81.4
Not employed	858	56.5
How hard to pay for medical care		
Very hard/Hard	321	21.1
Somewhat hard	504	33.2
Not very hard	692	45.5
How hard to pay for the very basics		
Very hard/Hard	286	18.8
Somewhat hard	491	32.3
Not very hard	742	48.8
Nulliparous	762	50.1
Primagravida	482	31.7
Mestizo	1148	75.5
Unplaned pregnancy	875	57.6
Gestational age at interview (mean ± standard devation)		9.8 ± 3.4

[a] Due to missing data, percentages may not add up to 100%

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 2

Item Hierarchy and Fit Statistics for the PHQ-9 Before and After Collapsing Response Categories 2 and 3 [a]

Item No.	Abbreviated Items	Item Difficulty in Logits	Model SE	Infit[b]		Outfit[b]	
				MnSq	Zstd	MnSq	Zstd
Before Collapsing							
9	Suicidal ideation	1.81	0.06	1.37	5.0	1.13	1.4
6	Low self-esteem	0.91	0.05	1.28	5.0	1.08	1.3
7	Concentration difficulty	0.79	0.05	1.35	6.2	1.14	2.2
8	Psychomotor agitation or retardation	0.76	0.05	1.47	8.1	1.08	1.3
2	Depressed mood	-0.34	0.03	0.85	-4.1	0.80	-4.6
3	Sleep problems	-0.59	0.03	1.01	0.2	1.02	0.4
1	Anhedonia	-0.84	0.03	0.85	-4.3	1.00	0.0
4	Low energy	-1.18	0.03	0.62	-9.9	0.69	-7.2
5	Appetite changes	-1.32	0.03	1.18	4.9	1.24	4.5
Mean		0.00	0.04	1.11	1.2	1.02	-0.1
SD		1.03	0.01	0.27	5.8	0.16	3.4
After Collapsing							
9	Suicidal ideation	2.04	0.07	1.36	5.5	1.12	1.1
6	Low self-esteem	1.06	0.05	1.19	4.1	1.03	0.5
8	Psychomotor agitation or retardation	1.05	0.05	1.27	5.7	1.04	0.6
7	Concentration difficulty	0.98	0.05	1.24	5.2	1.13	1.9
2	Depressed mood	-0.41	0.04	0.87	-4.0	0.84	-4.1
3	Sleep problems	-0.68	0.04	0.99	-0.3	1.05	1.2
1	Anhedonia	-1.02	0.04	0.86	-4.7	1.01	0.3
5	Appetite changes	-1.47	0.04	1.08	2.4	1.14	3.0
4	Low energy	-1.56	0.04	0.64	-9.9	0.69	-7.7
Mean		0.00	0.05	1.06	0.5	1.01	-0.4
SD		1.23	0.01	0.22	5.2	0.14	3.2

SD: Standard error, MnSq: mean square, Zstd: z-standardized

[a] Response category 0: Not at all; response category 1: Several days; response category 2: More than half of the days; response category 3: Nearly every day.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

[b] Infit MnSq value and outfit MnSq value should range between 0.6 to 1.4, and between 0.5 to 1.7, respectively.