Published in final edited form as:
Cultur Divers Ethnic Minor Psychol. 2011 July ; 17(3): 309–316. doi:10.1037/a0023883.

A Multigroup Confirmatory Factor Analysis of the Patient Health Questionnaire-9 among English- and Spanish-speaking Latinas

Erin L. Merz[1], **Vanessa L. Malcarne**[1,2,3], **Scott C. Roesch**[1,2], **Natasha Riley**[4], and **Georgia Robins Sadler**[1,3,5]

[1]SDSU/UCSD Joint Doctoral Program in Clinical Psychology, San Diego State University

[2]Department of Psychology, San Diego State University

[3]Rebecca and John Moores UCSD Cancer Center, UCSD School of Medicine

[4]Vista Community Clinic, UCSD School of Medicine

[5]Department of Surgery, UCSD School of Medicine

Abstract

Depression is a significant problem for ethnic minorities that remains understudied partly due to a lack of strong measures with established psychometric properties. One screening tool, the Patient Health Questionnaire-9 (PHQ-9), which was developed for use in primary care has also gained popularity in research settings. The reliability and validity of the PHQ-9 has been well established among predominantly Caucasian samples, in addition to many minority groups. However, there is little evidence regarding its utility among Hispanic Americans, a large and growing cultural group in the United States. In this study, we investigated the reliability and structural validity of the PHQ-9 in Hispanic American women. A community sample of 479 Latina women from southern California completed the PHQ-9 in their preferred language of English or Spanish. Cronbach's alphas suggested that there was good internal consistency for both the English- and Spanish-language versions. Structural validity was investigated using multigroup confirmatory factor analysis (CFA). Results support a similar one-factor structure with equivalent response patterns and variances among English- and Spanish-speaking Latinas. These results suggest that the PHQ-9 can be used with confidence in both English and Spanish versions to screen Latinas for depression.

Keywords

PHQ-9; depression; Latino; assessment; multigroup confirmatory factor analysis

The World Health Organization predicts that by the year 2030 major depression will be the second greatest cause of illness, disability, and death in the world and the leading cause of

disease burden in high-income countries (Lopez & Mathers, 2006). In the general community, 6.7% of individuals are estimated to be clinically depressed (Kessler, Chiu, Demler, & Walters, 2005). Depression is also very common in primary care medical settings (Simon & VonKorff, 1995; Wolf & Hopko, 2008).

Proactive screening and identification of depressed individuals is essential in reducing the impact of this global public health concern. Programs such as National Depression Screening Day have been useful in case identification in the general population (Greenfield et al., 2000; Jacobs, 1995). Moreover, the United States Preventive Services Task Force has recommended that all patients seeking medical care be screened for depression (U.S. Preventive Services Task Force, 2002). Although depression is common and treatable in primary care (Gill & Dansky, 2003; Spitzer et al., 1994), many cases are overlooked or misdiagnosed (Coyne, Schwenk & Fechner-Bates, 1995; Diez-Quevedo, Rangil, Sanchez-Planell, Kroenke, & Spitzer, 2001; Wolf & Hopko, 2008).

Diagnostic interviews are the gold standard in identification of depression; however, they are time-consuming and therefore unsuitable in settings where visits are short and/or screening must be brief. Consequently, self-report screening questionnaires, which can be completed and scored quickly, have gained popularity. These brief, easily administered tools can also serve important roles in research settings.

Although self-report surveys can be very useful for depression screening it cannot be assumed that a measure will accurately assess the construct it intends to, when diverse ethnic and language groups were not included in the measure's standardization (Bravo, 2003; Padilla & Medina, 1996). Many self-report measures have been validated on predominantly Caucasian samples, calling their cross-cultural applicability into question (Allen & Walsh, 2000).

In order to be used in a new group, measures that are developed using one standard group must be translated linguistically and conceptually/meaningfully, then empirically proven to measure the construct equivalently (Geisinger, 1994). That is, the measure must be made invariant for use between groups. Adaptations may be needed for cross-cultural applications across nations (e.g., Mexico vs. the United States), subcultures with a nation (e.g., Latinos within the United States), or cultural adaptations within a language (e.g., English- or Spanish-speaking Latinos within the United States; Allen & Walsh, 2000; Geisinger, 1994).

Once a measure is translated and/or adapted, its validity must be empirically proven, as the consequences of using nonequivalent measures can yield case-finding misclassifications and inaccurate epidemiological estimates. Advances in culturally sensitive assessments have resulted from recognition that depression occurs and negatively influences human quality of life across all races and ethnicities (Murray & Lopez, 1997) and that different cultural groups may define, experience, and communicate depression in different ways (e.g., Riolo, Nguyen, Greden, & King, 2005; Simon, VonKorff, Piccinelli, Fullerton, & Ormel, 1999).

With more than 35 million persons, Latinos are the largest and fastest growing minority group in the United States (U.S. Census Bureau, 2002). Moreover, there is evidence that Latinos experience high rates of depression (Chung et al., 2003; Mendelson, Rehkopf, &

Kubzansky, 2008; Riolo et al., 2005). Several reports indicate that immigrant stressors (e.g., acculturation, poverty) render Latinos particularly at risk for depression (Alegria et al., 2007; Sue & Chu, 2003; Torres, 2010; Vega, Sribney, Aguilar-Gaxioloa & Kolody, 2004).

Because of Latinos' increasing populace, and particular vulnerability to depression, it is important that common screening measures are valid to use among Latinos. However, little is known regarding the utility of these instruments among this group. One self-report measure of depression that has been widely used in screening, primary care, and research is the Patient Health Questionnaire-9 (PHQ-9; Spitzer, Kroenke, & Williams, 1999). The PHQ-9 takes approximately five minutes to fill out, and one minute to score (Kroenke, Spitzer, & Williams, 2001). It contains 9 items that parallel the diagnostic criteria for depression outlined by the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition – Text Revision (DSM-IV-TR; American Psychiatric Association, 2000). Respondents describe their experience of nine symptoms (anhedonia, depressed mood, sleep difficulties, fatigue, changes in appetite, feelings of worthlessness or guilt, difficulty concentrating, motor agitation, suicidality) over the previous two weeks. Each of the symptoms is rated on a 4-point scale as having occurred *not at all* (0), on *several days* (1), on *more than half the days* (2), or *nearly every day* (3) over the previous two weeks (Kroenke et al., 2001; Spitzer et al., 1999). Summed scores range from 0–27; larger scores reflect a greater endorsement of depressive symptoms. A recommended clinical cut-point (≥ 10) is used to identify respondents who should be evaluated further with a diagnostic assessment to determine their depression status (DeJesus, Vickers, Melin, & Williams, 2007; Kroenke et al., 2001; Kroenke, Spitzer, Williams, & Löwe, 2010; Spitzer et al., 1999).

The PHQ-9 has been embraced in screening, primary care, and research settings due to its brevity and excellent psychometric properties (Kroenke et al., 2001; Kroenke & Spitzer, 2002; Kroenke et al., 2010; Maizels, Smitherman, & Penzien, 2006; Spitzer et al., 1999). The original validation studies conducted in the United States by Spitzer and colleagues (1999, 2000) were of 3,000 patients from primary medical care clinics and 3,000 patients from obstetrics-gynecology care clinics. These reports suggest that the PHQ-9 has good internal consistency reliability (α = .89, α = .86, respectively), and good criterion validity with the clinical diagnostic interview as gold standard (r = .84, r = .79, respectively; Spitzer et al., 1999, Spitzer et al., 2000). Among 6,000 patients pooled, the PHQ-9 discriminated well between depressed and nondepressed individuals at the clinical cut-off of total score ≥ 10, with good sensitivity (88.0%) and specificity (88.0%) values (Kroenke et al., 2001).

Other studies have confirmed that the PHQ-9 has excellent test-retest reliability (e.g., Kroenke et al., 2001, Pinto-Meza et al., 2005, Patten & Schopflocher, 2009), and good construct criterion validity with structured diagnostic interviews (e.g., Spitzer et al., 1999; Löwe et al., 2004; Watnick, Wang, Demadura, & Ganzini., 2005) and with other depression questionnaires such as the Beck Depression Inventory-II (e.g., Dum, Pickren, Sobell, & Sobell, 2008, Hepner, Hunter, Edelen, Zhou, & Watkins, 2009, Martin, Rief, Klaiberg, & Braehler, 2006). For a recent review of the psychometric properties of the PHQ-9, see Kroenke and colleagues (2010).

The structural construct validity of the PHQ-9 has received some empirical attention. In general, any psychological measure should reveal the same simple structure that is defined a priori, and this structure should be reproducible across different samples and populations (Cicchetti, 1994; Clark & Watson, 1995; Floyd & Widaman, 1995; Reise, Waller, & Comrey, 2000). Because the PHQ-9 aims to measure the single construct of depression, it is generally hypothesized that a one-factor solution will fit the data. The majority of studies examining the structural validity of the PHQ-9 have yielded strong support for a single underlying factor (Cameron, Crawford, Lawton & Reid, 2008; Cannon et al., 2007; Dum et al., 2008; Hepner et al., 2009; Huang et al., 2006; Kendel et al., 2010; Lamoureux et al., 2009; Monahan et al., 2008).

The PHQ-9 has been widely embraced for cross-cultural use. It has been translated into many languages including Arabic (Becker, Al Zaid, & Al Faris, 2002), Brazilian Portuguese (Osório, Mendes, Crippa, & Loureiro, 2009), Chinese (Lubetkin, Jia, & Gold, 2003; Yeung et al., 2008), Dutch (Persoons, Luyckx, Desloovere, Vandenberghe, & Fischler, 2003), French (Carballeira et al., 2007; Dumont et al., 2005), German, (Löwe et al., 2004), Italian (Mazzotti et al., 2003; Picardi et al., 2005), Korean (Donnelly, 2007; Han et al., 2008), Malay (Azah et al., 2005), Spanish (Spain: Diez-Quevedo et al., 2001; United States: Huang, Chung, Kroenke, Delucchi, & Spitzer, 2006; Huang, Chung, Kroenke, & Spitzer, 2006), Swahili (Omoro, Fann, Weymuller, Macharia, & Yueh, 2006), Thai (Lotrakul, Sumrithe, & Saipanish, 2008), Twi, (Weobong et al., 2009), and Turkish (Corapcioglu & Ozer, 2004). The existence of many linguistic versions of the PHQ-9 makes it a particularly useful resource for application in non English-speaking groups in the United States, with its extremely diverse population.

Due to its established psychometric properties and availability of both English and Spanish language versions, the PHQ-9 is a particularly encouraging candidate for use with Latinos (Huang et al., 2006). Given the large Latino population in the United States, it is surprising that there is a paucity of studies evaluating the construct validity of the PHQ-9 in this group. To date, there has been only one published study that examined the construct validity of the PHQ-9 among Latinos, and this study focused on structural validity (Huang et al., 2006). Using a mostly female (97.8%), and mixed English- (26.4%) and Spanish-speaking (73.6%) sample derived from the original validation studies (Spitzer et al., 1999, Spitzer et al., 2000), Huang and colleagues examined the factor structure of the PHQ-9 via exploratory factor analysis. A one-factor solution, which explained 39.6% of the overall variance in the test scores, was derived. The internal consistency reliability of the PHQ-9 in this sample was good (α= .80; Huang et al., 2006).

Although this provides support for the unidimensionality of the PHQ-9 among this mostly female Latino sample, the psychometric properties of the English and Spanish-language versions of the PHQ-9 were not examined separately, and thus it remains unknown whether language preference will influence the structural construct validity of this measure. As first and second-generation Latinos in the United States become more acculturated, their language preference is likely to match that of their current sociocultural environment (Lara, Gamboa, Kahramanian, Morales, & Bautista, 2005). It is therefore important to ensure that the internal consistency reliability and factor structure of the PHQ-9 is maintained across

English and Spanish language versions administered to Latinos with different language preferences.

The present study assessed the structural validity of the English and Spanish versions of PHQ-9 among Latinos with different language preferences. To our knowledge, no published study has examined the equivalence of the PHQ-9 across English- and Spanish-speaking Latinos.

Method

Participants

Participants were a community sample of 479 women who self-identified as Hispanic American (English-speaking = 245, Spanish-speaking = 234) from San Diego County. Sample characteristics are described in Table 1. The ages of participants ranged from 18 to 80. In general, English-speaking respondents were born in the United States (63.7%) and had some college education (40.0%). Spanish-speaking respondents were mostly born in Mexico (87.6%) and had less than a high school education (72.5%).

Measure

Patient Health Questionnaire-9 (Spitzer et al., 1999)—The PHQ-9, described above, is a 9-item self-report measure that assesses depression on a 4-point scale (from 0 = *not at all* to 3 = *nearly every day*). Total scores range from 0–27 with higher scores denoting a greater endorsement of depressive symptoms, and scores ≥ 10 indicating that a respondent may be depressed.

Procedure

This study used data that were collected as part of a larger cross-sectional community-based study examining English- and Spanish-language instruments for potential use in a subsequent separate study evaluating the impact of an educational intervention to promote participation of Hispanic Americans in clinical trials. Research protocols were approved by Institutional Review Boards at the University of California, San Diego and San Diego State University. Recruitment strategies included face-to-face meetings at community sites and word-of-mouth. Individuals who agreed to participate were given the survey packet in their preferred language. Participants all gave written consent for their anonymous participation in the study. Survey completion took approximately one hour, and participants received $20 for their participation.

Data analysis—Confirmatory factor analysis (CFA) represents a theory-driven approach to test the a priori factor structure and goodness-of-fit between competing models. As described previously, it has been reliably demonstrated that a one-factor solution is the most appropriate fit for PHQ-9 data among several demographic groups. However, a more robust investigation of measurement properties involves making comparisons across demographic groups to ensure that these properties are retained, regardless of group characteristics. The statistical method used to test measurement and structural invariance across groups is multigroup CFA. Using this technique, separate models for each group were simultaneously

estimated, with equality constraints imposed upon relevant model parameters between groups, and change in model fit between nested models was tested. A series of multigroup CFA models replicating the factor structure of the PHQ-9 were fit to the data for English and Spanish-speaking participants in EQS 6.1 (Bentler, 2004). Following the methods of Vandenberg and Lance (2000), three multigroup CFA models (configural invariance, metric invariance, factor variance invariance) were fit to the PHQ-9 data.

The configural invariance model, which is the least restrictive, tests whether English and Spanish-speaking groups have the same factor structure across groups, with no equality constraints imposed. This provides the baseline value for the comparison of a more constrained model. If data support the same factor loading pattern across groups (i.e., the constructs are not substantially different across groups), this model may be tested against a more restrictive model (Vandenberg & Lance, 2000). The metric invariance model, which is more restrictive, tests whether each item on the PHQ-9 loads equivalently onto the same factor for both English and Spanish groups by constraining each item's factor loading to equivalence between language groups. If the metric invariance model fits well, then the associations between each item and the overall depression factor are the same regardless of language. Poorer model fit reflects either genuine differences, or bias in response patterns between groups. The metric invariance model is then compared to the baseline model. If this comparison demonstrates improved fit, then it is determined that the metric invariance model fits the data best. The factor variance invariance model, which is the most restrictive, tests whether English- and Spanish-language PHQ-9 factors have equivalent variability. Improved model fit from the previous step indicates that each language group has the same range of the continuum of scores. Conversely, worsened model fit indicates that the range of one group is truncated, due to real group differences or measurement bias.

Mardia's coefficients for the current sample revealed evidence of multivariate nonnormality for both English (114.84) and Spanish (94.55) groups (Mardia, 1970). Data that depart from multivariate normality can inflate the χ^2 maximum likelihood ratio, therefore the Satorra-Bentler scaled χ^2 (S-Bχ^2) was used to evaluate model fit (Satorra & Bentler, 2001). Because χ^2 tests may be unsatisfactory to determine model fit descriptive indices were also utilized to determine model fit (see Tanaka, 1993). In the current study we employed (a) the robust comparative fit index (CFI; Bentler, 1990), with a value greater > .95 indicating that a model fits well and values > .90 indicating a model is plausible; and (b) the standardized root mean square residual (SRMR; Hu & Bentler, 1999), with values < .05 indicating that a model fits well and values < .08 indicating that a model is plausible; and (c) the robust root mean square error of approximation (RMSEA; Steiger, 1990), with values < .05 indicating that a model fits well and values < .08 indicating that a model is plausible. A model was determined to fit well if two of the three criteria were met. For all individual model level parameters (e.g., factor loadings, factor variances), a significance level of .05 was employed. After each constraint was added, imposing more cross-group equality, a S-Bχ^2 difference test (Δ S-Bχ^2; Satorra & Bentler, 2001) was performed between the less restrictive and more restrictive model. This test uses scaling correction factors developed by Satorra and Bentler (2001), which allow for model comparison between nested models when data depart from normality. A significant test value ($p > .05$) indicates that the two models are not equivalent across English and Spanish groups.

Results

Preliminary analyses

The means and standard deviations for each item in English and Spanish are reported in Table 3. English-language total scores ranged from 0 to 23 ($M = 4.43$, $SD = 4.25$). Spanish-language total scores ranged from 0 to 27 ($M = 4.58$, $SD = 4.87$). Mean PHQ-9 scores were not significantly different between groups, $t(477) = -.356$, $p > .05$. In total, 12.9% ($n = 62$) of all participants in the sample met criteria for depression with scores ≥ 10, the clinical cut-off. Twenty-seven of these were English-speaking (11.0% of the English-speaking subsample) and 35 were Spanish-speaking (15.0% of the Spanish-speaking subsample). The proportion of respondents in each language group that met the clinical threshold was not significantly different (χ^2 [1] = 1.65, $p > .05$). Thirty-four (7.1%) respondents endorsed item 9 (suicidal ideation), with approximately half in each language group (English: 16, Spanish: 18). The proportion of respondents in each language group who endorsed this item was not significantly different (χ^2 [1] = .245, $p > .05$).

Reliability

Cronbach's alphas were calculated for the English and Spanish groups. In the current sample, the internal consistency was good for English ($\alpha = .84$) and Spanish ($\alpha = .85$) versions.

Multigroup CFA models

Configural invariance—First, we examined configural invariance by fitting the one-factor solution to the data for English- and Spanish-speaking groups. In this model, the factor loadings were freely estimated; no parameter estimates were constrained to equality across groups. The data are presented in Table 2. The one-factor solution provided a good fit to the data. All factor loadings for both groups were significant (Table 3), providing further evidence for configural invariance.

Metric invariance—Next, we examined metric invariance by constraining factor loadings to equivalence across English- and Spanish-speaking groups (Table 2). The metric invariance model fit well overall, which indicates that response patterns between groups were equivalent. The fit of this constrained model was then compared to that of the unconstrained configural invariance model. As shown in Table 2, a $\Delta S\text{-}B\chi^2$ test comparing the full metric invariance model with the configural model revealed that adding equality constraints to the factor loadings did not compromise model fit ($\Delta S\text{-}B\chi^2 = 14.59$, $df = 8$, $p = .068$). Because of this, invariance testing of individual factor loadings was not needed.

Factor Variance invariance—In the final step, equality constraints on the factor variance were added to the metric invariance model (Table 2). The factor variance invariance model fit well overall, which indicates that both English and Spanish groups yield the same range on the continuum of PHQ-9 scores. When the factor variance invariance model was compared to the metric invariance model, no significant difference was noted ($\Delta S\text{-}B\chi^2 = 1.61$, $df = 1$, $p = .204$). This suggests that the factor variance invariance model, which is the most parsimonious, is the best fit to the data. That is, for both English

and Spanish groups, there was equal variability of scores among the single-factor with (mostly) equivalent factor loadings.

Discussion

Overall, the PHQ-9 appears to have good internal consistency and structural validity when used in English and Spanish among Latinas. Internal consistency coefficients in the current study were similar between language groups and consistent with previous studies (e.g., Spitzer et al., 1999, Spitzer et al., 2000). Moreover, mean scores did not differ significantly between language groups. Both English and Spanish group means were similar to a report of population epidemiological data derived from the general community in Germany (M = 3.56; Rief, Nanke, Klaiberg, & Braehler, 2004) and another predominantly female, Latino sample (M = 4.7; Huang et al., 2006). The sample used by Rief and colleagues (2004) was approximately half male, which may explain the slightly lower group average. The proportion of patients who met the clinical threshold was also similar to that reported in these previous studies (Huang et al., 2006; Rief et al., 2004). Because all data were collected anonymously, respondents who scored above the clinical threshold were not able to be identified and referred for further assessment.

In this sample, the PHQ-9 demonstrated configural, metric, and factor variance equivalence. The same underlying one-factor structure was revealed for both English and Spanish groups, which is consistent with previous findings (e.g., Cameron et al., 2008, Huang et al., 2006). It should be noted that although all of the items loaded significantly onto the overall factor, item 9, which assesses suicidality, yielded the lowest factor loading. This is unsurprising, considering that item 9 does not discriminate well between depressed and non-depressed persons (Kroenke & Spitzer, 2002) and previous reports have also shown a lower factor loading for this item (Cameron et al., 2008, Cannon et al., 2007, Huang et al., 2006). This item is rarely endorsed, and is not a sensitive indicator of depression, however it should be retained due to its high-stakes clinical importance (e.g., Duffy et al., 2008, Huang et al., 2006; Kroenke et al., 2001; Kroenke & Spitzer, 2002). Any affirmative response on this item, regardless of duration should be immediately evaluated for suicide risk and appropriate action (e.g., hospitalization) be taken for those in imminent danger (DeJesus et al., 2007; Duffy et al., 2008). Although a small proportion of participants did endorse this item, the proportion was similar to previous reports (e.g., Huang et al., 2006; Rief et al., 2004), and, as noted earlier, because all data were collected anonymously, it was not possible to make participant referrals.

The final model, in which the factor variances of English and Spanish were constrained to equivalence, showed that each group's scores yielded the same variability across the depression continuum. That is, neither group yielded a truncated range of scores. In sum, these data suggest that, in Latinas, and for both English- and Spanish-language versions of the PHQ-9, there is a single underlying factor, items load equivalently onto that factor in general, and that there is an invariant range of total scores that make up that factor.

The implications of these findings are that the current version of the PHQ-9 can be recommended as an appropriate measure of depression, and the original recommended

clinical cut-point (≥ 10) may be retained, among English- and Spanish-speaking Latinas. After a rigorous test of model fit in each group, it appears that the overall construct of depression is being measured equivalently in Latinas, regardless of the language of administration. These findings, considered in light of the United States Preventive Services Task Force's recommendation for clinicians to offer routine screening for depression in their clinical practices (U.S. Preventive Services Task Force, 2002), suggest that the PHQ-9 can serve as a brief and low-burden screen for depression in Latinas, in both English and Spanish versions.

This study contributes to the growing literature on the psychometric properties of the PHQ-9 by exploring the structural construct validity in English and Spanish among Latinas. However, the study has several limitations. All respondents were community women who were largely of Mexican heritage, limiting the generalizability of findings to women of other Hispanic cultures. Diagnostic interviews were not administered, making it impossible to determine the case-finding characteristics of the PHQ-9 among Latinas. Another limitation is that because all study data were collected anonymously, the women who exceeded the clinical cut-point could not be followed up or referred.

In order to be certain of the psychometric properties and clinical implications of the PHQ-9, future research is needed. The factor structure should be tested with Mexican American men and other Latino sub-groups in the United States (e.g., Puerto Ricans, Cubans). Sensitivity and specificity rates within Latino men should also be established. Future studies should also assess the usefulness of the PHQ-9 in clinical populations.

Acknowledgments

The authors would like to acknowledge the following sources of support for this study: California Breast Cancer Research Program's 13AB-3500, 14BB-2601; NCI grants U54 CA132384, U54 CA132379, U56 CA92079, U56 CA92081; P60 MD000220-07; and P30 CA023100-23.

The research team wishes to thank all the Latinas who gave of their time to participate in this research study with the hope of helping to improve the well-being of their community. The research team also wishes to acknowledge other individuals who assisted with this study: Arianna Aldridge-Gerry, Viridiana Conde, Courtney Carter, Sheila LaHousse, Elva Leal, and Manpreet Mumman.

References

Alegria M, Shrout PE, Woo M, Guarnaccia P, Sribney W, Vila D, Polo A, Cao Z, Mulvaney-Day N, Torres M, Canino G. Understanding differences in past year psychiatric disorders for Latinos living in the U.S. Social Science and Medicine. 2007; 65:214–230. [PubMed: 17499899]

Allen, J.; Walsh, JA. A construct-based approach to equivalence: Methodologies for cross-cultural/multicultural personality assessment research. In: Dana, RH., editor. Handbook of cross-cultural and multicultural personality assessment. Personality and clinical psychology series. Mahwah, NJ: Lawrence Erlbaum Associates; 2000. p. 63-85.

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, DSM-IV-TR, 4th ed. Text Revision. Washington, DC: American Psychiatric Association; 2000.

Azah MN, Shah ME, Juwita S, Bahri IS, Rushidi WM, Jamil YM. Validation of the Malay Version Brief Patient Health Questionnaire (PHQ-9) among adult attending family medicine clinics [Abstract]. International Medical Journal. 2005; 12:259–263.

Becker S, Al Zaid K, Al Faris E. Screening for somatization and depression in Saudi Arabia: A validation study of the PHQ in Primary Care. International Journal of Psychiatry in Medicine. 2002; 32:271–283. [PubMed: 12489702]

Bentler PM. Comparative fit indexes in structural models. Psychological Bulletin. 1990; 107:238–246. [PubMed: 2320703]

Bentler, PM. EQS 6.1: Structural equations program manual. Encino, CA: Multivariate Software, Inc.; 2004.

Bentler PM, Chou CH. Practical issues in structural modeling. Sociological Methods & Research. 1987; 16:78–117.

Bravo, M. Instrument development: Cultural adaptations for ethnic minority research. In: Bernal, G.; Trimble, JE.; Burlew, AK.; Leong, FT., editors. Handbook of racial and ethnic minority psychology. Thousand Oaks, CA: Sage; 2003. p. 220-236.

Cameron IM, Crawford JR, Lawton K, Reid IC. Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. British Journal of General Practice. 2008; 58:32–36. [PubMed: 18186994]

Cannon DS, Tiffany ST, Coon H, Scholand MB, McMahon WM, Leppert MF. The PHQ-9 as a brief assessment of lifetime major depression. Psychological Assessment. 2007; 19:247–251. [PubMed: 17563207]

Carballeira Y, Dumont P, Borgacci S, Rentsch D, de Tonnac N, Archinard M, Andreoli A. Criterion validity of the French version of Patient Health Questionnaire (PHQ) in a hospital department of internal medicine. Psychology and Psychotherapy: Theory, Research and Practice. 2007; 80:69–77.

Chung H, Teresi J, Guarnaccia P, Meyers BS, Holmes D, Bobrowitz T, Eimicke JP, Ferran E. Depressive symptoms and psychiatric distress in low income Asian and Latino primary care patients: Prevalence and recognition. Community Mental Health Journal. 2003; 39:33–46. [PubMed: 12650554]

Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment. 1994; 6:284–290.

Clark LA, Watson D. Constructing validity: Basic issues in objective scale development. Psychological Assessment. 1995; 7:309–319.

Comrey, AL.; Lee, HB. A first course in factor analysis, Second Edition. Hillsdale, New Jersey: Lawrence Erlbaum Associates; 1992.

Corapcioglu A, Ozer GU. Adaptation of revised Brief PHQ (Brief-PHQ-r) for diagnosis of depression, panic disorder and somatoform disorder in primary healthcare settings [Abstract]. International Journal of Psychiatry in Clinical Practice. 2004; 8:11–18. [PubMed: 24937578]

Coyne JC, Schwenk TL, Fechner-Bates S. Nondetection of depression by primary care physicians reconsidered. General Hospital Psychiatry. 1995; 17:3–12. [PubMed: 7737492]

DeJesus RS, Vickers KS, Melin GJ, Williams MD. A system-based approach for depression management in primary care using the Patient Health Questionnaire-9. Mayo Clinic Proceedings. 2007; 82:1395–1402. [PubMed: 17976360]

Diez-Quevedo C, Rangil T, Sanchez-Planell L, Kroenke K, Spitzer RL. Validation and utility of the patient health questionnaire in diagnosing mental disorders in 1003 general hospital Spanish in-patients., 63. Psychosomatic Medicine. 2001; 63:679–686. [PubMed: 11485122]

Donnelly PL. The use of the Patient Health Questionnaire-9 Korean Version (PHQ-9K) to screen for depressive disorders among Korean Americans. Journal of Transcultural Nursing. 2007; 18:324–330. [PubMed: 18092395]

Duffy FF, Chung H, Trivedi M, Rae D, Regier DA, Katzelnick DJ. Systematic use of patient-rated severity monitoring: Is it helpful and feasible in clinical psychiatry? Psychiatric Services. 2008; 58:1148–1154. [PubMed: 18832500]

Dum M, Pickren J, Sobell LC, Sobell M. Comparing the BDI-II and the PHQ-9 with outpatient substance abusers. Addictive Behaviors. 2008; 33:381–387. [PubMed: 17964079]

Dumont P, Andreoli A, Borgacci S, Carballeira Y, Rentsch D, de Tonnac N, Archinard M. Quick detection of depression: A significant clinical issue. Swiss Medical Review. 2005; 5:344–346.

Floyd FJ, Widaman KF. Factor analysis in the development and refinement of clinical assessment instruments. Psychological Assessment. 1995; 8:286–299.

Geisinger KF. Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. Psychological Assessment. 1994; 6:304–312.

Gill JM, Dansky BS. Use of an electronic medical record to facilitate screening for depression in primary care. The Primary Care Companion to the Journal of Clinical Psychiatry. 2003; 5:125–128.

Greenfield SF, Reizes JM, Muenz LR, Kopans B, Kozloff RC, Jacobs DG. Treatment for depression following the 1996 National Depression Screening Day. The American Journal of Psychiatry. 2000; 157:1867–1869. [PubMed: 11058488]

Han C, Jo SA, Kwak JH, Pae CU, Steffens D, Jo I, Park MH. Validation of the Patient Health Questionnaire-9 Korean version in the elderly population: the Ansan Geriatric study. Comprehensive Psychiatry. 2008; 49:218–223. [PubMed: 18243897]

Hepner KA, Hunter SB, Edelen MO, Zhou AJ, Watkins K. A comparison of two depressive symptomatology measures in residential substance abuse treatment clients. Journal of Substance Abuse Treatment. 2009; 37:318–325. [PubMed: 19359127]

Hoyle, RH. Confirmatory factor analysis. In: Tinsley, HE.; Brown, SD., editors. Handbook of applied multivariate statistics and mathematical modeling. New York: Academic Press; 2000. p. 465-497.

Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling. 1999; 6:1–55.

Huang FY, Chung H, Kroenke K, Delucchi KL, Spitzer RL. Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. Journal of General and Internal Medicine. 2006; 21:547–552.

Huang FY, Chung H, Kroenke K, Spitzer RL. Racial and ethnic differences in the relationship between depression severity and functional status. Psychiatric Services,. 2006; 57:498–503. [PubMed: 16603745]

Jacobs DG. National Depression Screening Day: Educating the public, reaching those in need of treatment, and broadening professional understanding. Harvard Review of Psychiatry. 1995; 3:156–159. [PubMed: 9384943]

Kendel F, Wirtz M, Dunkel A, Lehmkuhl E, Hetzer R, Regitz-Zagrosek V. Screening for depression: Rasch analysis of the dimensional structure of the PHQ-9 and the HADS-D. Journal of Affective Disorders. 2010; 122:241–246. [PubMed: 19665236]

Kessler RC, Chiu WT, Demler O, Walters EE. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. Archives of General Psychiatry. 2005; 62:617–627. [PubMed: 15939839]

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 JB. The PHQ-9: Validity of a brief depression severity measure. Journal of General Internal Medicine. 2001; 16:606–613. [PubMed: 11556941]

Kroenke K, Spitzer RL, Williams JB, Löwe B. The Patient Health Questionnaire somatic, anxiety, and depressive symptom scales: A systematic review. General Hospital Psychiatry. 2010; 32:345–359. [PubMed: 20633738]

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]

Lara M, Gamboa C, Kahramanian MI, Morales LS, Bautista DH. Acculturation and Latino health in the United States: A review of the literature. Annual Review of Public Health. 2005; 26:367–397.

Lopez AD, Mathers CD. Measuring the global burden of disease and epidemiological transitions: 2002-2030. Annals of Tropical Medicine and Parasitology. 2006; 100:481–499. [PubMed: 16899150]

Lotrakul M, Sumrithe S, Saipanish R. Reliability and validity of the Thai version of the PHQ-9. 2008 Jun 20. from BMC Psychiatry, 8: http://www.biomedcentral.com/1471-244X/8/46.

Löwe B, Spitzer RL, Gräfe K, Kroenke K, Quenter A, Zipfel S, Buchholz C, Witte S, Herzog W. Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. Journal of Affective Disorders. 2004; 78:131–140. [PubMed: 14706723]

Lubetkin EI, Jia H, Gold MR. Depression, anxiety, and associated health status in low-income Chinese patients. American Journal of Preventive Medicine. 2003; 24:354–360. [PubMed: 12726874]

Maizels M, Smitherman TA, Penzien DB. A review of screening tools for psychiatric comorbidity in Headache Patients. Headache. 2006; 46:S98–S109. [PubMed: 17034404]

Mardia KV. Measures of multivariate skewness and kurtosis with applications. Biometrika. 1970; 36:519–530.

Martin A, Rief W, Klaiberg A, Braehler E. Validity of the Brief Patient Health Questionnaire Mood Scale (PHQ-9) in the general population. General Hospital Psychiatry. 2006; 28:71–77. [PubMed: 16377369]

Mazzotti E, Fassone G, Picardi A, Sagoni E, Ramieri I, Lega D, Camaioni D, Abeni D, Pasquini P. The Patient Health Questionnaire (PHQ) for the screening of psychiatric disorders: a validation study versus the Structured Clinical Interview for DSM-IV axis I (SCID-I) [Abstract]. Italian Journal of Psychopathology. 2003; 9:235–242.

Meade AW, Bauer DJ. Power and precision in confirmatory factor analytic tests of measurement invariance. Structural Equation Modeling. 2007; 14:611–635.

Mendelson T, Rehkopf DH, Kubzansky LD. Depression among Latinos in the United States: a meta-analytic review. Journal of Consulting and Clinical Psychology. 2008; 76:355–366. [PubMed: 18540730]

Mitchell, RJ. Path analysis: pollination. In: Scheiner, SM.; Gurevitch, J., editors. Design and Analysis of Ecological Experiments. New York, NH: Chapman and Hall, Inc.; 1993. p. 211-231.

Monahan PO, Shacham E, Reece M, Kroenke K, Ong'or WO, Omollo O, Yebei VN, Ojwang C. Validity/reliability of PHQ-9 and PHQ-2 depression scales among adults living with HIV/AIDS in Western Kenya. Journal of General and Internal Medicine. 2008; 24:189–197.

Omoro SA, Fann JR, Weymuller EA, Macharia IM, Yueh B. Swahili translation and validation of the Patient Health Questionnaire-9 in the Kenyan head and neck cancer patient population. International Journal of Psychiatry in Medicine. 2006; 36:367–381. [PubMed: 17236703]

Osório FD, Mendes AV, Crippa JA, Loureiro SR. 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. Perspectives in Psychiatric Care. 2009; 45:216–227. [PubMed: 19566694]

Padilla, AM.; Medina, A. Suzuki, LA.; Meller, PJ.; Ponterotto, JG. Handbook of Multicultural Assessment: Clinical, Psychological, and Educational Applications. San Francisco, CA: Jossey-Bass Publishers; 1996. Cross-cultural sensitivity in assessment. Using tests in culturally appropriate ways; p. 3-28.

Patten SB, Schopflocher D. Longitudinal epidemiology of major depression as assessed by the Brief Patient Health Questionnaire (PHQ-9). Comprehensive Psychiatry. 2009; 50:26–33. [PubMed: 19059510]

Persoons P, Luyckx K, Desloovere C, Vandenbenberghe J, Fischler B. Anxiety and mood disorders in otorhinolaryngology outpatients presenting with dizziness: Validation of the self-administered PRIME-MD Patient Health Questionnaire and epidemiology. General Hospital Psychiatry. 2003; 25:316–323. [PubMed: 12972222]

Picardi A, Adler DA, Abeni D, Chang H, Pasquini P, Rogers WH, Bungay KM. Screening for depressive disorders in patients with skin diseases: A comparison of three screeners. Acta Dermato-Venereologica. 2005; 85:414–419. [PubMed: 16159733]

Pinto-Meza A, Serrano-Blanco A, Peñarrubia MT, Blanco E, Haro JM. Assessing depression in primary care with the PHQ-9: Can it be carried out over the telephone? Journal of General and Internal Medicine. 2005; 20:738–742.

Reise SP, Waller NG, Comrey AL. Factor analysis and scale revision. Psychological Assessment. 2000; 12:287–297. [PubMed: 11021152]

Rief W, Nanke A, Klaiberg A, Braehler E. Base rates for panic and depression according to the Brief Patient Health Questionnaire: a population-based study. Journal of Affective Disorders. 2004; 82:271–276. [PubMed: 15488257]

Riolo SA, Nguyen TA, Greden JF, King CA. Prevalence of depression by race/ethnicity: Findings from the National Health and Nutrition Examination Survey III. American Journal of Public Health. 2005; 95:998–1000. [PubMed: 15914823]

Satorra A, Bentler PM. A scaled difference chi-square test statistic for moment structure analysis. Psychometrika. 2001; 66:507–514.

Schumacker, RE.; Lomax, RG. A beginner's guide to structural equation modeling, Second edition. Mahwah, NJ: Lawrence Erlbaum Associates; 2004.

Simon GE, Von Korff M, Piccinelli M, Fullerton C, Ormel J. An international study of the relation between somatic symptoms and depression. New England Journal of Medicine. 1999; 341:1329–1335. [PubMed: 10536124]

Simon G, VonKorff M. Recognition and management of depression in primary care. Archives of Family Medicine. 1995; 4:99–105. [PubMed: 7842160]

Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: The PHQ primary care study. Journal of the American Medical Association. 1999; 282:1737–1744. [PubMed: 10568646]

Spitzer RL, Williams JB, Kroenke KH, Hornyak R, McMurray L. 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]

Spitzer RL, Williams JB, Kroenke K, Linzer M, deGruy FV, Hahn SR, Brody D, Johnson JG. Utility of a new procedure for diagnosing mental disorders in primary care: the PRIME-MD 1000 study. Journal of the American Medical Association. 1994; 277:1749–1756. [PubMed: 7966923]

Steiger JH. Structural model evaluation and modification: An interval estimation approach. Multivariate Behavioral Research. 1990; 25:173–180.

Stevens, R. Applied multivariate statistics for the social sciences (3rd ed.). Mahwah, NJ: Erlbaum; 1996.

Sue S, Chu JY. The mental health of ethnic minority groups: Challenges posed by the Supplement to the Surgeon General's Report on Mental Health. Culture, Medicine, and Psychiatry. 2003; 27:447–465.

Tanaka, J. Bollen, JKA. Testing structural equation models. Newbury Park, CA: Sage; 1993. Multifaceted conceptions of fit in structural equation models; p. 10-39.

Torres L. Predicting levels of Latino depression: Acculturation, acculturative stress, and coping. Cultural Diversity and Ethnic Minority Psychology. 2010; 16:256–263. [PubMed: 20438164]

U.S. Census Bureau. Overview of Race and Hispanic Origin: Census 2000 Brief. Washington, DC:: U.S. Census Bureau; 2002.

U.S. Census Bureau. [Retrieved July 1, 2010] The American Community - Hispanics: 2004. 2007. from http://www.census.gov/prod/2007pubs/acs-03.pdf

U.S. Preventive Services Task Force. Screening for Depression: Recommendations and Rationale. Rockville, MD: Agency for Healthcare Research and Quality; 2002.

Vandenberg RJ, Lance CE. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational Research Methods. 2000; 3:4–69.

Vega WA, Sribney WM, Aguilar-Gaxiola SA, Kolody B. 12-month prevalence of DSM–III–R psychiatric disorders among Mexican Americans: Nativity, social assimilation, and age determinants. The Journal of Nervous and Mental Disease. 2004; 192:532–541. [PubMed: 15387155]

Watnick S, Wang PL, Demadura G, Ganzini L. Validation of 2 depression screening tools in dialysis patients. American Journal of Kidney Diseases. 2005; 46:919–924. [PubMed: 16253733]

Weobong B, Akpalu B, Doku V, Owusu-Agyei S, Hurt L, Kirkwood B, Prince M. The comparative validity of screening scales for postnatal common mental disorder in Kintampo, Ghana. Journal of Affective Disorders. 2009; 113:109–117. [PubMed: 18614241]

Wolf NJ, Hopko DR. Psychosocial and pharmacological interventions for depressed adults in primary care: A critical review. Clinical Psychology Review. 2008; 28:131–161. [PubMed: 17555857]

Yates BT, Taub J. Assessing the costs, benefits, cost-effectiveness, and cost-benefit of psychological assessment: We should, we can, and here's how. Psychological Assessment. 2003; 15:478–495. [PubMed: 14692844]

Yeung A, Fung F, Yu SC, Vorono S, Ly M, Wu S, Fava M. Validation of the Patient Health Questionnaire-9 for depression screening among Chinese Americans. Comprehensive Psychiatry. 2008; 49:211–217. [PubMed: 18243896]

Table 1

Sample characteristics

	English (n = 245)	Spanish (n = 234)
Age[1]	36.96 (13.74)	39.96 (12.34)
Education[2]		
Less than high school	32 (13.1%)	169 (72.5%)
High school/trade school	88 (35.9%)	40 (17.2%)
Some college/Associates degree	98 (40.0%)	13(5.6%)
Bachelor's degree	22 (9.0%)	8 (3.4%)
Postgraduate	5 (2.0%)	3 (1.3%)
Employment Status[2]		
Employed	187 (76.3%)	96 (42.3%)
Unemployed	6 (2.4%)	19 (8.4%)
Homemaker	14 (5.7%)	94 (41.4%)
Student/retired/disabled	38 (15.5%)	18 (7.9%)
Marital Status[2]		
Married	124 (50.6%)	129 (55.1%)
Single	121 (49.4%)	105 (44.9%)
Country of birth[2]		
United States	156 (63.7%)	10 (4.3%)
Mexico	74 (30.2%)	205 (87.6%)
Other	15 (6.1%)	19 (8.1%)

Note.

[1] *M* (*SD*),

[2] *n* (*p*)

Table 2

Goodness of Fit Statistics from configural invariance models for English and Spanish versions of the PHQ-9

	Model	S-Bχ^2	*df*	*p*	Reference Model #	Δ S-Bχ^2	Δ*df*	Δ *p*	CFI[1]	SRMR[2]	RMSEA[2]
1.	Configural	79.38	54	.422					.952	.053	.037
2.	Metric	92.87	62	.007	1	14.59	8	.068	.939	.079	.046
3.	Factor Variance	94.51	63	.006	2	1.61	1	.204	.938	.093	.046

Note. CFI = robust comparative fit index; SRMR = standardized root mean square residual; RMSEA = robust root mean square error of approximation;

[1] Plausible fit > .90, Good fit > .95;

[2] Plausible fit < .08, Good fit < .05

Table 3

Descriptive statistics and unstandardized factor loadings from baseline models for English and Spanish version of the PHQ-9

	PHQ-9 item	English (*n* = 245)		Spanish (*n* = 234)	
		Factor loadings	***M* (*SD*)**	**Factor loadings**	***M* (*SD*)**
1.	Little interest or pleasure in doing things[1]	1.00*	.45 (.72)	1.00*	.67 (.90)
2.	Feeling down, depressed, or hopeless	1.26*	.48 (.72)	1.03*	.59 (.79)
3.	Trouble falling or staying asleep, or sleeping too much	1.07*	.79 (.91)	.98*	.72 (.92)
4.	Feeling tired or having little energy	1.02*	.92 (.76)	1.18*	.85 (.86)
5.	Poor appetite or overeating	1.10*	.67 (.80)	1.04*	.69 (1.01)
6.	Feeling bad about yourself—or that you are a failure or have let yourself or your family down	1.06*	.42 (.71)	.81*	.42 (.80)
7.	Trouble concentrating on things, such as reading the newspaper or watching television	1.01*	.39 (.74)	.54*	.28 (.62)
8.	Moving or speaking so slowly that other people could have noticed. Or the opposite - being so fidgety or restless that you have been moving around a lot more than usual	.63*	.20 (.51)	.51*	.24 (.61)
9.	Thoughts that you would be better off dead, or of hurting yourself in some way	.30*	.10 (.43)	.40*	.12 (.48)

Note.

[1]The factor loading for the first item was fixed to 1 to set the metric for the latent variable;

*ps < .05