Published in final edited form as:
Nurs Res. 2010 ; 59(6): 380–388. doi:10.1097/NNR.0b013e3181f84ee9.

Comparison of Depressive Symptom Severity Scores in Low-Income Women

Shawn M. Kneipp, PhD, ARNP[Associate Professor],
College of Nursing, Department of Healthcare Environments & Systems, University of Florida, Gainesville, Florida

John A. Kairalla, PhD[Assistant Professor],
University of Florida, College of Medicine, Department of Epidemiology and Health Policy Research, Division of Biostatistics, Gainesville, Florida

Jeanne Marie R. Stacciarini, PhD[Assistant Professor],
University of Florida, College of Nursing, Department of Healthcare Environments & Systems, Gainesville, Florida

Deidre Pereira, PhD[Assistant Professor], and
University of Florida, College of Public Health & Health Professions, Department of Clinical and Health Psychology, Gainesville, Florida

M. David Miller, PhD[Professor]
University of Florida, College of Education, Department of Research Methods and Evaluation, Gainesville, Florida

Abstract

Background—The Beck Depression Inventory-II (BDI-II) and the Patient Health Questionnaire-9 (PHQ-9) are considered reliable and valid for measuring depressive symptom severity and screening for a depressive disorder. Few studies have examined the convergent or divergent validity of these two measures, and none have been conducted among low-income women – even though rates of depression in this group are extremely high. Moreover, variation in within-subject scores suggests these measures may be less comparable in select subgroups.

Objective—We sought to compare these two measures in terms of construct validity, and examine whether within-subject differences in depressive symptom severity scores could be accounted for by select characteristics in low-income women.

Method—In a sample of 308 low-income women, construct validity was assessed using a multitrait-monomethod matrix approach, between-instrument differences in continuous symptom severity scores were regressed on select characteristics using backward stepwise selection, and differences in depressive symptom classification were assessed using the Mantel-Haenszel test.

Results—Convergent validity was high (r_s = 0.80, $p < .001$). Among predictors that included age, race, education, number of chronic health conditions, history of depression, perceived stress, anxiety, and/or the number of generalized symptoms, none explained within-subject differences in depressive symptom scores between the BDI-II and PHQ-9 ($p > .05$, $R^2 < 0.04$). Similarly, there was consistency in depressive symptom classification (X^2 = 172 and 172.6, p < .0001).

Correspondence: Shawn M. Kneipp, PhD, ARNP, Associate Professor, College of Nursing, Department of Healthcare Environments & Systems, PO Box 100187, University of Florida, Gainesville, FL 32610-0187, Phone: (352) 273-6417, Fax: (352) 273-6577, skneipp@ufl.edu.

Discussion—These findings demonstrate the BDI-II and PHQ-9 perform similarly among low-income women in terms of depressive symptom severity measurement and classifying levels of depressive symptoms, and do not vary across subgroups based on select demographics.

Keywords

depression; reliability and validity; disadvantaged population

---

Research related to major depressive disorder (MDD) and the role of depressive symptoms on health status and other disease outcomes has burgeoned in the past decade. Depression is the second leading cause of disability worldwide (World Health Organization, 2009), is highly prevalent in the United States population (Kessler et al., 2003), and has significant personal, family, and society costs (U.S. Department of Health and Human Services, 1999). Moreover, emerging findings show depression as a robust, consistent predictor for negative morbidity and mortality outcomes from a number of other chronic disease states (e.g., cardiovascular disease, diabetes mellitus) (Lustman & Clouse, 2005; Ramasubbu & Patten, 2003; van der Kooy et al., 2007).

Currently, investigators and clinicians can choose from a number of available instruments to assess for different aspects of depression, such as the degree of symptom severity or the presence of a diagnosis of depression. For research purposes, several of these instruments have adequate validity and reliability to measure depressive symptoms, while for clinical purposes, several, but not all of these instruments are considered excellent screening tools with high predictive validity for a diagnosis of Major Depressive Disorder (MDD). Available instruments include, but are not limited to, the Centers for Epidemiology Studies Depression Scale (CES-D; Radloff, 1977; Radloff & Terri, 1986), the Patient Health Questionnaire-9 (PHQ-9; Kroenke, Spitzer, & Williams, 2001), the Beck Depression Inventory-II (BDI-II; Beck, Steer, & Brown, 1996), the Hamilton Depression Rating Scale-17 (HAMD-17; Hamilton, 1960, 1967), and/or the Hamilton Depression Rating Scale-7 (HAMD-7; Maier, Heuser, Philipp, Frommberger, & Demuth, 1988; Maier et al., 1988; McIntyre et al., 2005). With this wide array of psychometrically or clinically valid instruments to select from, it is not surprising that the measures of depression vary widely across studies, making it difficult to compare study findings. Moreover, some instruments – such as the CES-D and the BDI – are used with consistent scoring practices for measuring depressive symptoms for research, but may need to have scoring thresholds adjusted for screening purposes in select primary care settings to achieve adequate predictive values for a depressive disorder (Zich, Attkisson, & Greenfield, 1990). When there is this type of ‘utility incongruence’ in the applicability and usefulness of depression measures used for research versus clinical purposes, it can make the translation of research findings into clinical practice even more challenging.

At least two of these depression measures are considered both valid and reliable for the measurement of depressive symptoms in the general population, have repeatedly demonstrated positive predictive values suitable for screening individuals in clinical settings (Watnick, Wang, Demadura, & Ganzini, 2005), and are used widely in studies to measure depression. These include the Beck Depression Inventory – Second Edition (BDI-II) (Beck et al., 1996), and the Patient Health Questionnaire – 9 (PHQ-9) (Kroenke et al., 2001). The BDI-II is a widely used measure of depressive symptoms. Composed of 21 items, the BDI-II yields a single score with a range of 0 to 63. Standardized score ranges for categorical levels of depressive symptoms are: 0-13 = minimal, 14-19 = mild, 20-28 = moderate, and 29-63 = severe. Items in the BDI-II were developed to reflect cognitive (C), affective (A), and somatic (S) components of depression (Beck et al., 1996), and map closely onto *Diagnostic and Statistical Manual of Mental Disorders- Fourth Edition* (DSM-IV) criteria for a Major

Depressive Episode (American Psychiatric Association, 2000). In methodological studies with diverse populations including African Americans (Gary & Yarandi, 2004; Grothe et al., 2005), predominantly Caucasian Canadian college students (Beck et al., 1996), chemically dependent inpatients (Buckley, Parker, & Heggie, 2001), and depressed geriatric outpatients (Steer, Rissmiller, & Beck, 2000), the BDI-II factor structure generally reflects these three intended domains within two factors; however, findings have varied, with more recent studies pointing to a more complex, as opposed to simple, two-factor structure (Kneipp, Kairalla, Stacciarini, & Pereira, 2009; Thombs, Ziegelstein, Beck, & Pilote, 2008; Ward, 2006).

The PHQ-9 is a 10-item self-report questionnaire used to screen for depression using DSM-IV criteria. Items 1 through 9 are used to calculate a symptom severity measure, with a range of 0 to 27 (Kroenke et al., 2001). Item 10 assesses the extent to which depressive symptoms interfere with day-to-day functioning, and can assist clinicians in determining whether symptoms meet the DSM-IV criteria for MDD, but is not included in the symptom severity score. Score ranges for categorizing levels of depressive symptoms are as follows: 0-4 = minimal, 5-9 = mild, 10-14 = moderate, 15-19 = moderately severe, and 20-27 = severe. The PHQ-9 has emerged consistently as a one-factor structure based on factor analysis findings across diverse populations, including substance abusers (Dum, Pickren, Sobell, & Sobell, 2008), African Americans, Chinese Americans, Latinos, and Caucasians (Huang, Chung, Kroenke, Delucchi, & Spitzer, 2006). Convergent validity of the PHQ-9 has been demonstrated in primary care settings (Martin, Rief, Klaiberg, & Braehler, 2006), dialysis patients (Watnick et al., 2005), substance abusers (Dum et al., 2008), patients with depression (Lee, Schulberg, Raue, & Kroenke, 2007), and Nigerian university students (Adewuya, Ola, & Afolabi, 2006).

Only four studies were found directly comparing measurement characteristics of the BDI-II and PHQ-9 (Dbouk, Arguedas, & Sheikh, 2008; Dum et al., 2008; Hepner, Hunter, Edelen, Zhou, & Watkins, 2009; Watnick et al., 2005). Among these, Watnick et al. (2005) found both the BDI-II and PHQ-9 demonstrated acceptable degrees of clinical validity based on receiver operator characteristic (ROC) curves using the *Structured Clinical Interview for Depression* (SCID) as the gold standard. The remaining three studies demonstrated a high degree of convergent validity between BDI-II and PHQ-9 symptom severity scores, with correlations ranging between 0.76 and 0.84. Two of these studies were conducted among substance abusing populations (Dum et al., 2008; Hepner et al., 2009), one was in a sample of dialysis patients (Watnick et al., 2005), and the last was in a sample of Hepatitis C patients (Dbouk et al., 2008). All of these studies were conducted in predominantly male or Caucasian samples, and only one reported on socioeconomic status (SES) whereby the mean years of education in the sample was 13.8 (Dum et al., 2008). Overall, these findings are of limited use to health disparities researchers interested in measuring depressive symptoms among predominantly low-income, predominantly female, and/or predominantly minority populations.

Study Purpose

Given the burgeoning interest in health disparities research that either focuses on depression as an outcome or as a mediating or moderating variable, and the inability to generalize existing measurement comparison findings to this population, we sought to compare key aspects of these measures in ways that are meaningful to both clinicians and scientists given the data available from an ongoing randomized clinical trial (RCT). Specifically, the aims of this study were to (a) compare convergent and divergent validity of the BDI-II and the PHQ-9 among low-income women, (b) examine whether demographic and other health-related factors predict differences in the standardized symptom severity scores of these two

measures, and (c) examine the extent to which BDI-II and PHQ-9 categorical symptom severity classifications were similar within subjects.

Theoretical Framework—Given the aims of this work focus on measurement, psychometric theory provides a guide for the analyses presented here. Concern over the development of psychological measures – and thus the origin of psychometric theory – has been traced back to the early 1900s (Strauss & Smith, 2009). Broadly conceived, psychometric theory addresses the measurement and quantification of psychological constructs. These constructs take the form of variables in studies conducted routinely in the social and health sciences, and, ultimately, the soundness of scientific knowledge claims rests upon the use of valid and reliable measures to represent them. As a major component of psychometric theory, the core principles of construct validation were first proposed by Cronbach and Meehl (1955) and extended to the multitrait-multimethod methodology by Campbell and Fiske (1959). The most recent version of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) also regard construct validity as the most important consideration in test use and interpretation and continue to emphasize the focus of theory and theory testing in construct validation. Campbell and Fiske's (1959) early efforts defined a means for using convergent and divergent validity analyses within a multitrait-multimethod matrix (MTMM) methodology (Strauss & Smith, 2009), which allows for partitioning covariance between measures into "method" covariance (e.g., dependent on data collection approach such as written questionnaire or interview) and "trait", or construct, covariance (e.g., dependent solely on construct attributes) (DeVellis, 1991). Since that time, construct validation methods using the MTMM methodology have advanced through the application of structural equation modeling (SEM). The basic premise of MTMM methodology remains a standard approach for instrument development, as does the acceptance of construct validation as an ongoing process, rather than a singular outcome from any one investigation (American Educational Research Association, 1999; Strauss & Smith, 2009).

The aims of this study apply select aspects of construct validation using a multitrait-*mono*method approach. This approach allows the evaluation of convergent and divergent validity of the PHQ-9 and BDI-II based on multiple *trait* (i.e., construct) comparisons with each other and theoretically distinct constructs (perceived stress, using the Perceived Stress Scale, and anxiety, using the Beck Anxiety Inventory). This approach does not, however, examine constructs across *methods*, which requires data be collected in different formats (e.g., written self-report, telephone interview). In the RCT from which data were collected, measures were completed in a self-report, written questionnaire format across study participants, precluding multimethod comparisons. In addition to adhering to traditional aspects of construct validity, the analyses focuses "inward" by examining not only the variance shared between measures of theoretically similar constructs (i.e., the convergence of the PHQ-9 and the BDI-II), but also on the variance that is *not* shared. Based on reported correlations between the PHQ-9 and BDI-II of $r = 0.76$ to 0.84, there remains a moderate amount of variance unaccounted for ($1 - r^2$, or 29% - 43%). We wanted to further examine whether the within-subject similarity of scores differed across women based on select characteristics. That is, within a sample of low-income women, we wanted to evaluate whether there are select characteristics that would explain why some women had remarkably similar levels of depressive symptom severity based on BDI-II and PHQ-9 scores, while others had very different scores on these two measures. Finally, from a clinical perspective, cut-off criteria for different categories of depressive symptomatology based on PHQ-9 and BDI-II scores are used often to guide further evaluation and treatment decisions, and it would be clinically useful to know whether women would be similarly classified on each

measure based on their severity score (e.g., based on ‘mild’, ‘moderate’, etc. cutoff criteria specified within each instrument).

Methods

Sample

Participants in this study included 308 low-income women enrolled in an ongoing randomized controlled trial (RCT) to address health disparities among women receiving public assistance benefits through the Temporary Assistance for Needy Families (TANF) program. The study was approved by the university Health Science Center Institutional Review Board. Women met eligibility criteria for enrollment in the ongoing study if they were: receiving TANF benefits, between 18 and 60 years of age, not currently employed, not receiving disability benefits, not pregnant, could read and speak English, and had at least one chronic health condition. Chronic health conditions were defined broadly, and included conditions ranging from relatively self-limiting problems such as seasonal allergies to more severe chronic health conditions, such as diabetes, hypertension, or other forms of cardiovascular disease. Conditions were identified by self-report during an extensive physical and mental health assessment using a battery of standardized, valid and reliable, self-administered health and health-related measures, including the BDI-II and PHQ-9. Data for these analyses were collected between February 27, 2006 and September 24, 2008, were collected at the study enrollment visit, and represent baseline values unaffected by the RCT intervention.

Measures

As noted, demographic and other variables were measured by self-report. In addition to select demographic and health variables, data were collected using self-report that are directly relevant to the analyses presented here – including perceived stress, general health, anxiety, and the number of chronic health conditions.

Perceived stress was measured using the 14-item Perceived Stress Scale (PSS) (Cohen, Tyrrell, & Smith, 1993). It has documented validity and reliability (test-retest correlation coefficients of $r = 0.85$, and Cronbach alphas of 0.79-0.86) in low-income populations (Kneipp et al., 2009; Tuck & Wallace, 2000). Scores range from 0 to 56, with higher scores indicating higher levels of stress. General health was measured using the Short Form-12 Version 2.0 (SF-12v2) (Ware, Kosinski, Turner-Bowker, & Gandek, 2002). The SF-12v2 is based on the SF-36, which has been used extensively in diverse populations. The SF-12v2 provides summary scores for physical and mental health that have demonstrated validity and reliability ($\alpha > 0.80$; test-retest coefficients = 0.78 and 0.60, respectively), and predict > 90% of the variance in SF-36 summary scores. The SF-12v2 captures eight health dimensions – one being general health, which was used to measure general health in this study. Anxiety was measured using the 21-item Beck Anxiety Inventory (BAI). The BAI has a score range of 0-63, has demonstrated validity and reliability ($\alpha > 0.90$, test-retest coefficient = 0.75), and has been used in diverse populations (Beck & Steer, 1990). Finally, participants provided a self-report of the number of chronic health conditions, and were instructed to base their responses on the number of health problems that had been diagnosed by a health care provider.

Data Analysis

Analyses were conducted using Stata/SE Version 10.0 and SAS Version 9.1.3 statistical software.

Missing data—Of the n = 308 women enrolled, not more than 11% had missing data on the BDI-II, PHQ-9, PSS, or BAI. Missing data were at the item level, meaning participants were missing data on select *items*, or *questions*, within a particular instrument, as opposed to not having answered any of the questions on a given measure. Among the relatively small proportion with missing data, the vast majority (>88%) had ≤ 2 items missing on any one instrument. There were no differences in participants with respect to age, race, marital status, number of children, or number of chronic conditions by complete vs. incomplete data, which suggests data were missing at random (Little & Rubin, 2002). Using a multinomial logistic regression prediction model (Horton & Laird, 2001), missing data were imputed at the item level. Predictor selection for imputing missing items within each measure was based on the distribution of the construct measured by select demographics and other health-related measures or non-missing items in the instrument with which it correlated most strongly and consistently over time (Kneipp et al., 2009). A consistent imputation model was applied to impute items within instruments; so, for example, when age, race, and the number of chronic health conditions were highly correlated with th e measure in which items were being imputed, these same predictors were consistently included in the imputation model *for every missing item within the same measure* to both reduce variability and to avoid introducing bias in the imputed items.

Convergent and Divergent Validity—Symptom severity scores included the sum of all 21 items in the BDI-II and items 1 through 9 on the PHQ-9 using standard scoring procedures for each (Beck et al., 1996; The John D. & Catherine T. MacArthur Foundation's Initiative on Depression & Primary Care, 2004). Based on our prior confirmatory factor analysis findings, which indicate the BDI-II in this population is comprised of a complex factor structure that includes General Depression (i.e., the full-scale, standard BDI-II that includes all 21 items), Cognitive (subscale), and Somatic (subscale) domains (Kneipp et al., 2009), these were included in the multitrait-monomethod analysis. Given the skewed distribution of the BDI-II domain scores and the PHQ-9 symptom severity scores (Shapiro Wilk test for normality, $p < .001$), Spearman's rho was used to evaluate convergent and divergent validity using a multitrait-monomethod approach.

Differences in Standardized BDI-II and PHQ-9 Scores—The BDI-II and PHQ-9 have very different possible symptom severity score ranges (0-63 and 0-27, respectively). To meaningfully examine difference scores between symptom severity findings within subjects, scores for each instrument were standardized to z-scores with means of zero and standard deviations (SD) of one. Standardizing variables in this way allowed for a direct comparison of scores and the calculation of difference scores in the BDI-II and PHQ-9 within subjects. Difference scores were calculated by subtracting the standardized BDI-II score from the standardized PHQ-9 score.

A stepwise, backward elimination approach was used to regress the standardized BDI-II / PHQ-9 difference score (Std_PBDiff) on select predictors given the exploratory nature of the inquiry and the lack of a comprehensive or cohesive theory to more specifically guide model building (Hair, Black, Babin, & Anderson, 2005). Demographic variables that tend to consistently relate to a wide variety of health outcomes (i.e., age, race, and education), factors that could reasonably be expected to account for differences in depression measure scores based on theoretical and measurement construct considerations (i.e., a history of depression, number of chronic conditions, and the total number of mental and physical symptoms experienced), and variables that have been associated strongly with depressive symptoms in prior studies with the same population (i.e., perceived stress) (Kneipp, Welch, Wood, Yucha, & Yarandi, 2007) were included in the full, initial regression model (Little & Rubin, 2002).

With the standardized BDI-II / PHQ-9 difference score (Std_PBDiff) as the dependent variable, data were examined using standard procedures to assure multiple regression assumptions could be met, and to determine whether influential cases were present. The assumptions of normality, linearity, and independence were met (Meyers, Gamst, & Guarino, 2006). Three outliers were identified to be highly influential based on regression diagnostic plots, which included studentized residuals of +3 / -3, leverage > (2k+2)/n (k = number of predictors), a DFITS values > $2\sqrt{(k/n)}$, and DFBETA values > $2\sqrt{(n)}$. The primary concern with outliers was their influence on regression beta coefficients. Highly influential outliers were identified by examining both DFITS and DFBETA values in Stata/SE 10.0, which are considered composite indicators of outlier status and leverage. Both provide a measure of the difference between the regression coefficient when the *i*th observation is included and excluded in the model, albeit on different scales (StataCorp, 2007). Based on a comprehensive examination of regression diagnostics findings, these outliers were likely to have undue influence on regression coefficients for the majority of predictors included in the model. Importantly, the decision was made to remove these outliers from the regression analyses *a priori* (before running the final regression model) and *without* examining whether the regression findings were similar or different with the outliers 'in' or 'out' of the model. In so doing, the intent was to make sound analytic decisions based on standard regression diagnostic approaches (Belsley, Kuh, & Welsch, 2004), and avoid the introduction of bias that can occur when different findings from different models (i.e., 'with' and 'without' outliers) are compared – and a final analytic approach is selected – *a posteriori*. Once these three outliers were removed, there were no additional extreme influential cases.

BDI-II and PHQ-9 Symptom Severity Classification—Both the BDI-II and the PHQ-9 use a set of cut-off scores to classify levels of depressive symptoms that range from "minimal" to "severe" (Arnau, Meagher, Norris, & Bramson, 2001; Kroenke et al., 2001). However, the BDI-II reflects four levels of symptom severity while the PHQ-9 includes five levels (the addition of a "moderate-severe" category), preventing direct classification comparisons between instruments.

Due to the clinical significance of the depressive classification categories, we wanted to examine whether a trend association of symptom severity classification between the BDI-II and PHQ-9 was present – that is, as the level of BDI-II classification increased, did the level of classification similarly increase on the PHQ-9 schema? To examine this question, the Mantel-Haenszel test for linear association was applied (Mantel, 1963), which takes advantage of the ordered categorical classifications to test for a linear trend between the depression classification schemes. Symptom classifications were based on the raw (i.e., the continuous, non-z-score transformed) symptom severity summary scores as outlined in the scoring manuals for each measure (Beck et al., 1996; MacArthur Foundation, 2004). A component of the Mantel-Haenszel test is the choice of scores to use for the within measure categories. Although different choices usually have little effect on the results, sometimes an imbalance in category frequency can lead to different conclusions. We followed the advice of Agresti (2002) and performed a sensitivity analysis by running the results with two reasonable approaches to score selection: equally spaced and standardized midranks, and compared the results.

Results

Sample Demographics

Study participants were predominantly Black/African American (55.6%) and unmarried (90.6%), averaged 29.5 years of age, had an average of 2.4 children, and lived on less than $600 per month (including cash assistance from TANF, food stamps, and any other cash

received). Over one-third had less than a 12th grade education (35%), with another third (33%) reporting some college or technical training. The mean number of chronic conditions reported was 2.6; depressive symptom severity scores were 19.6 (SD = 12.2) for the BDI-II and 10.6 (SD = 6.5) for the PHQ-9. Sample demographics are detailed in Table 1.

Convergent and Divergent Validity

Examination of the multitrait-monomethod matrix (Table 2) demonstrates convergent validity and some evidence for divergent validity given the pattern of correlations between constructs. As expected, the correlation coefficient for the BDI-II General Depression domain and the PHQ-9 based on Spearman's rho (r_s) was high, at 0.80 ($p < .001$). While the BDI-II General Depression and BDI-II Cognitive domains were most highly correlated (0.91, $p<0.001$), this was not surprising, given the BDI-II General Depression score is comprised in part by the Cognitive and Somatic domain items. Notably, the BDI-II General Depression domain and the PHQ-9 were more highly correlated than either of these constructs were with the PSS or BAI, and the BDI-II General Depression domain correlated more highly with the PHQ-9 than either the BDI-II Cognitive or Somatic subscale domains. These findings suggest that, among the multiple traits examined, the standard BDI-II (General Depression) and the PHQ-9 demonstrate a fairly high degree of convergent validity. There was some support for divergent validity, as well, as correlations between the two depression measures and the PSS and BAI – which were not intended to measure depressive symptoms – were lower (ranging from 0.60 to 0.69) than those observed between the BDI-II and PHQ-9 ($r_s = 0.80$).

Predictors of BDI-II and PHQ-9 Score Differences

Following z-score standardization and the removal of the three outliers, mean BDI-II and PHQ-9 scores were -0.014 (SD = 0.99) and -0.005 (SD = 1.00), respectively. Differences in within-person standardized BDI-II and PHQ-9 scores ranged from -1.78 to 1.76, with negative difference scores indicating the BDI-II standardized score was higher than the PHQ-9 standardized score, and positive difference scores indicating the PHQ-9 standardized score was higher than the BDI-II standardized score. Findings from the backward stepwise regression are detailed in Table 3. Each predictor included in a model was tested by itself (a single degree of freedom test) with the exception of race, which was tested in each model using a two degree of freedom test of overall race effect for Black/African American or Other. Predictors with the least significance were dropped one by one from the model in the following order: age ($p = 0.90$), anxiety ($p = 0.87$), number of chronic health conditions ($p = 0.88$), number of symptoms ($p = 0.34$), depression history ($p = 0.29$), race ($p = 0.23$), and education level ($p = 0.08$). This model reduction approach resulted in perceived stress as the final remaining predictor, although this was also not statistically significant ($p = 0.08$). In addition to the lack of statistically significant predictors, the amount of variance explained by the full model and thus each reduced model did not exceed 3.7%, indicating no subset of predictors explained the within-person differences between standardized BDI-II and PHQ-9 symptom severity scores.

BDI-II and PHQ-9 Symptom Severity Classification

Frequencies for symptom severity categories of the BDI-II and PHQ-9 are shown in Table 4. Overall, based on the category distributions, the PHQ-9 seems to be concentrated more in the middle categories, with the highest two being Mild (26%) and Moderate (27%). In contrast, the BDI-II scores seemed to be distributed more in the extremes with the highest two categories being None/Minimal (35%) and Severe (24%). In order to test the question of whether a significant trend exists based on the categorical designations, we utilized the Mantel-Haenszel test for linear association. The tests performed with equally spaced scoring and standardized ranks yielded little difference. With equal spacing, the chi-square equaled

172.6 ($p < .0001$) and with standardized midranks, the chi-square equaled 172.0 ($p < .0001$). This is evidence there is a significant correlation between the categorical designations of the two depressive scoring systems. That is, individuals falling into higher categories in one instrument would tend to rate higher in the other, and vice versa.

Discussion

The BDI-II and PHQ-9 are highly regarded measures of depressive symptoms with well-documented validity and reliability for use in both research and clinical settings. For screening purposes, both have acceptable predictive values when compared to the *Structured Clinical Interview for Depression* (SCID), which is considered a gold standard for diagnosing depression (Watnick et al., 2005). For research purposes, each is commonly used to measure depressive symptom severity. In terms of construct validity, the PHQ-9 has consistently demonstrated a one-factor structure consistently across studies with diverse patient and non-patient groups (Dum et al., 2008; Hepner et al., 2009; Huang et al., 2006). The BDI-II has demonstrated less consistency in factor structure across studies, and more recent analyses suggest its factor structure is a complex, higher-ordered structure in both low-income women (Kneipp et al., 2009) and other populations with varying health conditions (Thombs et al., 2008). Our finding of high convergent validity of BDI-II and PHQ-9 symptom severity scores in this population ($r_S = 0.80$) similarly parallels prior findings in the literature, which range from 0.76 to 0.84 (Dbouk et al., 2008; Dum et al., 2008; Hepner et al., 2009). In terms of convergent validity, correlation coefficients within the range typically found between the BDI-II and PHQ-9 clearly indicate there is theoretical and/or construct similarity between the constructs being measured; yet, there remains some unaccounted for variance – in this study, for example, 46% -- in the within-person scores between these measures.

The major findings of this study were in demonstrating parallel support for the construct validity of the BDI-II and the PHQ-9 with low-income women. The parallel validation evidence included those findings supporting convergent and divergent validity based on the multitrait-monomethod matrix analyses and the strong relationship of the classification systems. The inability to explain within-person differences in standardized scores between the BDI-II and PHQ-9 based on a number of demographic and other health characteristics among low-income women also suggests the parallel interpretations of the two instruments. This latter finding makes a unique contribution to the existing literature by discerning the extent to which the PHQ-9 can be used in place of the BDI-II in studies with low-income women when the goal is to measure depressive symptom severity. More specifically, this finding suggests that differences in the within-subject BDI-II and PHQ-9 symptom severity scores appear to be unrelated to salient, and potentially confounding, demographic, psychosocial, and health factors examined here.

Overall, our findings should instill some degree of confidence that these two instruments can be used interchangeably as measures of depressive symptom severity for many studies with samples similar to ours, suggesting that factors other than validity concerns may be used to select between the two instruments. For example, the PHQ-9 is available to the public at no cost, involves less respondent burden than the BDI-II, and scored is more easily in clinical settings; as such, the PHQ-9 may have distinct advantages over the BDI-II in circumstances in which financial constraints, respondent burden, or time pressures at data collection occasions are concerns. There is a more recently developed version of the BDI – the BDI-Fast Screen (BDI-FS) – which can be administered in a short period of time (Beck, Steer, & Brown, 2000). Although the BDI-FS has demonstrated good sensitivity and specificity in a pain clinic population (Poole, Bramwell, & Murphy, 2009), it has had less predictive value among older stroke patients (Healey, Kneebone, Carroll, & Anderson, 2008), and testing in

other populations is limited. Other factors should drive decision-making around the adoption of one measure over another. For example, the BDI-II may be preferred in studies with subjects for whom suicidal intent/plan is a concern, as the BDI-II more directly assesses a more 'active', or 'direct' form of suicidal ideation (e.g., references to "killing myself"), while the PHQ-9 assesses a more 'passive' form of suicidal ideation (e.g., a reference to "[thoughts of] being better off dead or of hurting yourself in some way"). Finally, between the two measures examined here, the PHQ-9 may have less potential confounding of general, depression-related somatic symptoms with those that co-exist with other disease states (e.g., fatigue following a myocardial infarction) (Thombs et al., 2008), given that proportionally fewer PHQ-9 items assess somatic symptoms of depression compared to the BDI-II (Kneipp et al., 2009). These and other additional considerations may be more or less important in the measurement selection process depending on the specific aims and unique qualities of any given study.

There are, of course, limitations of this study and its findings that should be noted. First, 11% of the sample had missing data. Although missing data were minimal, apparently random, and able to be corrected with accepted imputation procedures, this remains a study limitation. Second, these findings are not generalizable beyond women who are low-income, Caucasian or African American/Black, and who have at least one identified chronic health condition (as defined broadly within our sample). Third, although we included predictors of BDI-II and PHQ-9 within-person score differences based on prior findings and theoretical considerations, the set of predictors included was clearly not exhaustive, and the possibility remains that a relevant predictor of these differences was not accounted for in our regression models.

Overall, these findings provide investigators interested in depression among low-income women additional data upon which to base measurement selection decisions. For researchers seeking a measure of depressive symptoms, these findings demonstrate the BDI-II and PHQ-9 perform similarly in this population. These findings make an important contribution to the measurement literature to advance health disparities research.

Acknowledgments

Thank you to the women in the Welfare Transition Program for participation in this study, service providers and administrative personnel in the FloridaWorks Welfare Transition Program for their ongoing support, and the National Institutes of Health/National Institute of Nursing Research, contract grant number R01NR009406 for research support.

References

Adewuya AO, Ola BA, Afolabi OO. Validity of the patient health questionnaire (PHQ-9) as a screening tool for depression amongst Nigerian university students. Journal of Affective Disorders 2006;96(1-2):89–93. [PubMed: 16857265]

Agresti, A. Categorical data analysis. 2nd. Hoboken, NJ: John Wiley & Sons; 2002.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. Standards for educational and psychological testing. Washington DC: American Educational Research Association; 1999.

American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th - text revision. Washington, DC: Author; 2000.

Arnau RC, Meagher MW, Norris MP, Bramson R. Psychometric evaluation of the Beck Depression Inventory-II with primary care medical patients. Health Psychology 2001;20(2):112–119. [PubMed: 11315728]

Beck, AT.; Steer, RA. Beck anxiety inventory manual. San Antonio, TX: The Psychological Corporation, Harcourt Brace & Company; 1990. p. 23

Beck, AT.; Steer, RA.; Brown, GK. BDI-II manual. 2nd. San Antonio, TX: Harcourt Brace & Company; 1996.

Beck, AT.; Steer, RA.; Brown, GK. BDI-II fast screen for medical patients manual. London, United Kingdom: The Psychological Corporation; 2000.

Buckley TC, Parker JD, Heggie J. A psychometric evaluation of the BDI-II in treatment-seeking substance abusers. Journal of Substance Abuse Treatment 2001;20(3):197–204. [PubMed: 11516588]

Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin 1959;56(2):81–105. [PubMed: 13634291]

Cohen S, Tyrrell DA, Smith AP. Negative life events, perceived stress, negative affect, and susceptibility to the common cold. Journal of Personality and Social Psychology 1993;64(1):131–140. [PubMed: 8421249]

Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychological Bulletin 1955;52(4): 281–302. [PubMed: 13245896]

Dbouk N, Arguedas MR, Sheikh A. Assessment of the PHQ-9 as a screening tool for depression in patients with chronic hepatitis C. Digestive Diseases and Sciences 2008;53(4):1100–1106. [PubMed: 17934817]

DeVellis, RF. Scale development: Theory and applications. London, United Kingdom: SAGE Publications; 1991.

Dum M, Pickren J, Sobell LC, Sobell MB. Comparing the BDI-II and the PHQ-9 with outpatient substance abusers. Addictive Behaviors 2008;33(2):381–387. [PubMed: 17964079]

Gary FA, Yarandi HN. Depression among southern rural African American women: A factor analysis of the Beck Depression Inventory-II. Nursing Research 2004;53(4):251–259. [PubMed: 15266164]

Grothe KB, Dutton GR, Jones GN, Bodenlos J, Ancona M, Brantley PJ. Validation of the Beck Depression Inventory-II in a low-income African American sample of medical outpatients. Psychological Assessment 2005;17(1):110–114. [PubMed: 15769232]

Hair, JF.; Black, B.; Babin, B.; Anderson, RE. Multivariate data analysis. 6th. Upper Saddle River, NJ: Prentice Hall; 2005.

Hamilton M. A rating scale for depression. Journal of Neurology, Neurosurgery, and Psychiatry 1960;23:56–62.

Hamilton M. Development of a rating scale for primary depressive illness. The British Journal of Social and Clinical Psychology 1967;6(4):278–296. [PubMed: 6080235]

Healey AK, Kneebone II, Carroll M, Anderson SJ. A preliminary investigation of the reliability and validity of the Brief Assessment Schedule Depression Cards and the Beck Depression Inventory-Fast Screen to screen for depression in older stroke survivors. International Journal of Geriatric Psychiatry 2008;23(5):531–536. [PubMed: 18008393]

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(3):318–325. [PubMed: 19359127]

Horton NJ, Laird NM. Maximum likelihood analysis of logistic regression models with incomplete covariate data and auxiliary information. Biometrics 2001;57(1):34–42. [PubMed: 11252616]

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 Internal Medicine 2006;21(6):547–552. [PubMed: 16808734]

The John D. & Catherine T. MacArthur Foundation's Initiative on Depression & Primary Care. Depression management tool kit. Hanover, NH: Dartmouth College & Duke University; 2004.

Kessler RC, Berglund P, Demler O, Jin R, Koretz D, Merikangas KR, et al. The epidemiology of major depressive disorder: Results from the National Comorbidity Survey Replication (NCS-R). JAMA 2003;289(23):3095–3105. [PubMed: 12813115]

Kneipp SM, Welch DP, Wood CE, Yucha CB, Yarandi H. Psychosocial and physiological stress among women leaving welfare. Western Journal of Nursing Research 2007;29(7):864–883. [PubMed: 17630386]

Kneipp SM, Kairalla JA, Stacciarini J, Pereira D. The Beck Depression Inventory II factor structure among low-income women. Nursing Research 2009;58(6):400–409. [PubMed: 19918151]

Kroenke K, Spitzer RL, Williams JB. The PHQ-9: Validity of a brief depression severity measure. Journal of General Internal Medicine 2001;16(9):606–613. [PubMed: 11556941]

Lee PW, Schulberg HC, Raue PJ, Kroenke K. Concordance between the PHQ-9 and the HSCL-20 in depressed primary care patients. Journal of Affective Disorders 2007;99(1-3):139–145. [PubMed: 17049999]

Little, RJ.; Rubin, DB. Statistical analysis with missing data. 2nd. Hoboken, NJ: John Wiley & Sons; 2002.

Lustman PJ, Clouse RE. Depression in diabetic patients: The relationship between mood and glycemic control. Journal of Diabetes and Its Complications 2005;19(2):113–122. [PubMed: 15745842]

Maier W, Heuser I, Philipp M, Frommberger U, Demuth W. Improving depression severity assessment--II. Content, concurrent and external validity of three observer depression scales. Journal of Psychiatric Research 1988;22(1):13–19. [PubMed: 3397905]

Maier W, Philipp M, Heuser I, Schlegel S, Buller R, Wetzel H. Improving depression severity assessment--I. Reliability, internal validity and sensitivity to change of three observer depression scales. Journal of Psychiatric Research 1988;22(1):3–12. [PubMed: 3397908]

Mantel N. Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. Journal of the American Statistical Association 1963;58:690–700.

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(1):71–77. [PubMed: 16377369]

McIntyre RS, Konarski JZ, Mancini DA, Fulton KA, Parikh SV, Grigoriadis S, et al. Measuring the severity of depression and remission in primary care: Validation of the HAMD-7 scale. CMAJ 2005;173(11):1327–1334. [PubMed: 16301700]

Meyers, LS.; Gamst, G.; Guarino, AJ. Applied multivariate research: Design and interpretation. Thousand Oaks, CA: SAGE Publications; 2006.

Poole H, Bramwell R, Murphy P. The utility of the Beck Depression Inventory Fast Screen (BDI-FS) in a pain clinic population. European Journal of Pain 2009;13(8):865–869. [PubMed: 19010075]

Radloff LS. The CES-D Scale: A self-report depression scale for research in the general population. Applied Psychological Measurement 1977;1(3):385–401.

Radloff LS, Terri L. Use of the Center for Epidemiological Studies-Depression Scale with older adults. Clinical Gerontologist 1986;5(1-2):119–136.

Ramasubbu R, Patten SB. Effect of depression on stroke morbidity and mortality. Canadian Journal of Psychiatry 2003;48(4):250–257.

StataCorp. Stata base reference manual. Vol. 3. College Station, TX: Stata Press; 2007.

Steer RA, Rissmiller DJ, Beck AT. Use of the Beck Depression Inventory-II with depressed geriatric inpatients. Behaviour Research and Therapy 2000;38(3):311–318. [PubMed: 10665163]

Strauss ME, Smith GT. Construct validity: Advances in theory and methodology. Annual Review of Clinical Psychology 2009;5:1–25.

Thombs BD, Ziegelstein RC, Beck CA, Pilote L. A general factor model for the Beck Depression Inventory-II: Validation in a sample of patients hospitalized with acute myocardial infarction. Journal of Psychosomatic Research 2008;65(2):115–121. [PubMed: 18655855]

Tuck I, Wallace D. Chronic fatigue syndrome: A woman's dilemma. Healthcare for Women International 2000;21(5):457–466.

United States Department of Health and Human Services. Mental health: A report of the Surgeon General. Rockville, MD: Author; 1999.

Van der Kooy K, van Hout H, Marwijk H, Marten H, Stehouwer C, Beekman A. Depression and the risk for cardiovascular diseases: Systematic review and meta analysis. International Journal of Geriatric Psychiatry 2007;22(7):613–626. [PubMed: 17236251]

Ward LC. Comparison of factor structure models for the Beck Depression Inventory--II. Psychological Assessment 2006;18(1):81–88. [PubMed: 16594815]

Ware, JE.; Kosinski, M.; Turner-Bowker, DM.; Gandek, B. How to score version 2 of the SF-12 Health Survey. Lincoln, RI: QualityMetric Incorporated; 2002.

Watnick S, Wang PL, Demadura T, Ganzini L. Validation of 2 depression screening tools in dialysis patients. American Journal of Kidney Diseases 2005;46(5):919–924. [PubMed: 16253733]

World Health Organization. Mental health: Depression. Author; 2009.

Zich JM, Attkisson CC, Greenfield TK. Screening for depression in primary care clinics: The CES-D and the BDI. International Journal of Psychiatry in Medicine 1990;20(3):259–277. [PubMed: 2265888]

Table 1

Sample Demographics (*n* = 308)

Characteristic	%	Mean	SD
Age		29.5	7.8
Race			
Black/African American	55.6		
White	41.3		
Other	3.1		
Ethnicity			
Hispanic	4.7		
Education Level			
<12th Grade	35.2		
High School Diploma/GED	25.6		
Some College/Technical Training	33.8		
College Degree	1.8		
Number of Children		2.4	1.3
Mean Age of Children		7.0	5.1
Marital Status			
Unmarried	90.6		
Married	9.4		
Total Household Income		$556.40	$426.84
Number of Chronic Conditions		2.6	1.6
Total Reported Symptoms		8.0	6.8
PSS [0-56]*		29.9	7.2
PHQ-9 [0-27]*		10.6	6.5
BDI-II [0-63]*		19.6	12.2
SF-12 GH [0-100]*		40.9	22.5

Notes.

* [ ] Range of possible scores;

PSS = Perceived Stress Scale; PHQ-9 = Patient Health Questionnaire-9 Item, Symptom Severity Score; BDI-II = Beck Depression Inventory – II; SF-12 GH = Short Form – 12 General Health

Table 2

Multitrait-Monomethod Correlation Matrix*

Traits (Constructs)	PHQ-9	BDI-II General Depression	BDI-II Cognitive	BDI-II Somatic	PSS	BAI
PHQ-9	1.0					
BDI-II General Depression	0.80	1.0				
BDI-II Cognitive	0.70	0.91	1.0			
BDI-II Somatic	0.75	0.88	0.69	1.0		
PSS	0.61	0.69	0.64	0.62	1.0	
BAI	0.61	0.60	0.56	0.61	0.46	1.0

*All $p < .05$; Spearman's rho.

Table 3

Predictors of Standardized PHQ-9 and BDI-II Difference Scores

Predictor Variables	Full Model	Reduced Model 1	Reduced Model 2	Reduced Model 3	Reduced Model 4	Reduced Model 5	Reduced Model 6	Final Model
Age	0.0006†	*Dropped*	---	---	---	---	---	---
	(0.005)€							
	0.90							
Race							*Dropped*	---
Black / African American	0.069	0.069	0.069	0.071	0.068	0.056		
	(0.075)	*(0.075)*	*(0.074)*	*(0.073)*	*(0.073)*	*(0.073)*		
	0.36	0.35	0.35	0.33	0.36	0.44		
Other	0.220	0.220	0.221	0.222	0.231	0.235		
	(.140)	*(0.140)*	*(0.139)*	*(0.139)*	*(0.138)*	*(0.138)*		
	0.12	0.12	0.11	0.11	0.10	0.09		
Overall Race Effect p value	0.26	0.26	0.25	0.24	0.23	0.23		
Education	0.038	0.038	0.039	0.038	0.042	0.040	0.040	*Dropped*
	(0.024)	*(0.023)*	*(0.023)*	*(0.023)*	*(0.023)*	*(0.23)*	*(0.023)*	
	0.11	0.10	0.10	0.10	0.07	0.08	0.08	
Number of Chronic Health Conditions	-0.004	-0.004	-0.003	*Dropped*	---	---	---	---
	(0.019)	*(0.019)*	*(0.018)*					
	0.85	0.85	0.88					
Depression History	0.079	0.079	0.080	0.076	0.081	*Dropped*	---	---
	(0.081)	*(0.081)*	*(0.081)*	*(0.076)*	*(0.075)*			
	0.34	0.33	0.32	0.31	0.29			
Number of Symptoms	0.005	0.005	0.006	0.005	*Dropped*	---	---	---
	(0.007)	*(0.007)*	*(0.006)*	*(0.005)*				

Predictor Variables	Full Model	Reduced Model 1	Reduced Model 2	Reduced Model 3	Reduced Model 4	Reduced Model 5	Reduced Model 6	Final Model
	0.42	0.41	0.37	0.34				
BAI Total Score	0.0006	0.0007	*Dropped*	---	---	---	---	---
	(0.004)	*(0.004)*						
	0.89	0.87						
PSS Score	-0.012	-0.012	-0.012	-0.012	-0.010	-0.008	-0.009	-0.008
	(0.006)	*(0.006)*	*(0.005)*	*(0.005)*	*(0.005)*	*(0.005)*	*(0.005)*	*(0.005)*
	0.04*	0.03*	0.03*	0.03*	0.05	0.08	0.06	0.08
Constant	0.090	0.107	0.100	0.092	0.076	0.064	0.136	0.257
	(0.238)	*(0.189)*	*(0.184)*	*(0.175)*	*(0.174)*	*(0.174)*	*(0.159)*	*(0.144)*
Observations	305	305	305	305	305	305	305	305
R-squared	0.037	0.037	0.037	0.036	0.033	0.030	0.020	0.010

†Unstandardized β coefficients;

€Standard errors *italicized* in parentheses;

*p = < .05;

**p < .001.

Least significant variable dropped from model at each step until all predictors statistically significant.

Number of Symptoms based on standard primary care Review of Systems checklist; BAI = Beck Anxiety

Inventory; PSS = Perceived Stress Scale.

Table 4

Distribution of PHQ-9 and BDI-II Symptom Severity Categories[†*]

	PHQ-9 Symptom Categories				
BDI-II Symptom Categories	***Minimal***	***Mild***	***Moderate***	***Moderate Severe***	***Severe***
Minimal	52 (49.1%) (81.2%)	41 (38.7%) (51.9%)	12 (11.3%) (14.5%)	1 (0.9%) (2.4%)	0
Mild	11 (16.4%) (17.2%)	23 (34.3%) (29.1%)	27 (40.3%) (32.5%)	5 (7.5%) (11.9%)	1 (1.5%) (2.7%)
Moderate	1 1.7%) (1.6%)	12 (20.3%) (15.2%)	27 (45.8%) (32.5%)	15 (25.4%) (35.7%)	4 (6.8%) (10.8%)
Severe	0	3 (4.1%) (3.8%)	17 (23.3%) (20.5%)	21 (28.8%) (50.0%)	32 (43.8%) (86.5%)

[†]Mantel-Haenszel test for linear association with equally spaced and standardized midranks scoring: X^2=172.6 and 172.0 respectively, df=1, $p < .0001$

[*]n, (row %), (column %)