MEASUREMENT INVARIANCE OF THE PATIENT HEALTH QUESTIONNAIRE-9 (PHQ-9) DEPRESSION SCREENER IN U.S. ADULTS ACROSS SEX, RACE/ETHNICITY, AND EDUCATION LEVEL: NHANES 2005-2014

by

Jay Sunil Patel

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

Department of Psychological Sciences

Indianapolis, Indiana

December 2017

THE PURDUE UNIVERSITY GRADUATE SCHOOL

STATEMENT OF COMMITTEE APPROVAL

Dr. Jesse C. Stewart, Chair

Department of Psychology

Dr. Kevin L. Rand

Department of Psychology

Dr. Melissa A. Cyders

Department of Psychology

Approved by:

Dr. Nicholas J. Grahame

Head of the Graduate Program

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Author: Patel, Jay, Sunil MS
Institution: Purdue University
Degree Received: December 2017
Title: Measurement Invariance of the Patient Health Questionnaire-9 (PHQ-9) Depression Screener in U.S. Adults across Sex, Race/Ethnicity, and Education Level: NHANES 2005-2014.
Major Professor: Jesse Stewart

Importance: Despite its widespread use in clinical settings and in behavioral medicine research, little is known about the psychometric performance of the PHQ-9 across major U.S. sociodemographic groups. Thus, utilizing a large sample representative of the U.S. population and confirmatory factor analysis (CFA), we determine the factor structure and measurement invariance of the PHQ-9 across groups based on sex, race/ethnicity, and education level.

Objective: Our objective was to address key knowledge gaps by definitively determining the factor structure and measurement invariance of the PHQ-9 across major U.S. sociodemographic groups based on sex, race/ethnicity, and education level.

Design: The continuous National Health and Nutrition Examination Survey (NHANES) is a cross-sectional, epidemiologic study designed to assess the health and nutritional status of the U.S. population. We examined data from the 2005-2014 survey years.

Setting: NHANES is uses a stratified multistage probability sampling approach to enroll civilian, non-institutionalized adults and children in the U.S.

Participants: For our final sample, we selected the 26,202 adult respondents with no missing PHQ-9 data. The factors of interest were sex (49.3% men, 50.7% women), race/ethnicity (48.9% non-Hispanic White, 23.7% non-Hispanic Black, 17.8% Mexican American, 9.7% other Hispanic), and education level (9.9% less than 9^{th} grade, 16.6% 9^{th}-12^{th} grade but no diploma,

23.7% high school graduate/GED or equivalent, 28.9% some college or Associate's degree, 20/8% college graduate or above).

Main Outcome(s) and Measure(s): The Patient Health Quessionnaire-9 (PHQ-9)

Results: Results revealed that the best solution for the PHQ-9 consists of a cognitive/affective factor (items 1. anhedonia, 2. depressed mood, 6. feelings of worthlessness, 7. concentration difficulties, 8. psychomotor disturbances, and 9. thoughts of death) and a somatic factor (items 3. sleep disturbance, 4. fatigue, and 5. appetite changes; RMSEA = 0.034, RMSEA 90% CI = 0.032–0.036, TLI = 0.984, CFI = 0.988). To evaluate measurement invariance, we then conducted single-group and multiple-group CFAs to carry out the 5 steps of measurement invariance testing. Dimensional, configural, weak factorial, strong factorial, and strict factorial invariance was established for the PHQ-9 across the sex, race/ethnicity, and education level groups, as all models demonstrated close fit and the ΔCFI was < 0.010 for all steps.

Conclusions and Relevance: Using a U.S. representative sample, we determined that a two-factor solution for the PHQ-9 with a cognitive/affective factor and a somatic factor is invariant across sex, race/ethnicity, and education level groups. Therefore, clinically, the PHQ-9 is an acceptable measure to utilize in major U.S. sociodemographic groups, extending the use of this depression screener from the primary care clinic to the community. Additionally, we show that PHQ-9 cognitive/affective and somatic subscale scores have the same meaning and can be compared across major U.S. sociodemographic groups and provide a consistent, evidence-based approach to computing PHQ-9 subscale scores to be used in future studies.

INTRODUCTION

Major depressive disorder (MDD) is a top public health problem due to its high prevalence, chronicity, and serious ramifications. The lifetime prevalence of MDD in the U.S. is 16%.[15] In primary care settings within the U.S., the point prevalence of depressive disorders is 16-19%.[16,17] The course of MDD is often chronic, with a 15-year recurrence rate of 35% in the general population and 85% in mental health care settings.(Hardeveld, Spijker, De Graaf, Nolen, & Beekman, 2010) The serious ramifications of MDD include increased disability, chronic illness, mortality, and societal costs. Depressive disorders are the second leading cause of years lived with disability worldwide.(Ferrari et al., 2013) In addition, meta-analyses demonstrate that depression increases the risk of type 2 diabetes by 38%,(Rotella & Mannucci, 2013) coronary artery disease by 30%,(Rotella & Mannucci, 2013) and dementia by 85%.(Rotella & Mannucci, 2013) Given these findings, it is no surprise that depression is a predictor of increased mortality in various samples, including community samples and patients with coronary artery disease, cancer, and kidney disease.(Cuijpers et al., 2014) Finally, the total annual cost of depression has increased by $127 billion in just 10 years.(Greenberg et al., 2003; Greenberg, Fournier, Sisitsky, Pike, & Kessler, 2015)

These alarming observations have motivated recommendations to improve the detection and management of depression, in part, by routinely administering depression screeners. To illustrate, in 2016, the U.S. Preventive Services Task Force recommended depression screening for all people aged 18+ years in primary care, including pregnant and postpartum women.(Siu et al., 2016) Essentially, it is recommended that every adult receiving care in clinical settings be screened for depression at least once using a validated screener. One widely used depression

screener, highly recommended for use in primary care settings, is the Patient Health Questionnaire-9 (PHQ-9).(Nease & Malouin, 2003)

Despite its widespread use in clinical practice and research, surprisingly little is known about the PHQ-9's psychometric performance across major U.S. sociodemographic groups, such as those based on sex, race/ethnicity, and education level. An advanced statistical approach for evaluating an instrument's psychometric performance across groups is measurement invariance testing. If measurement invariance is established through this testing, it would indicate that PHQ-9 assesses the same construct in the selected groups and that observed group differences reflect true group differences. Thus, meaningful comparisons involving the PHQ-9 across the groups could be made. However, if measurement invariance is not established, it would raise serious concerns regarding whether the PHQ-9 assesses the same construct and whether it is meaningful and justifiable to compare PHQ-9 scores across the selected groups. Use of a depression screener shown *not* to possess measurement invariance could result in under-detection or over-detection of depression based on group membership. Under-detection would likely lead to under-treatment of depression in certain groups, whereas over-detection would likely lead to the wasting of limited depression treatment resources.

Although a few investigations have examined the measurement invariance of the PHQ-9 across sociodemographic groups, these studies have been limited in two key ways. One, they have not examined a consistent factor structure, with some using a one-factor solution(Baas et al., 2011; Cameron, Crawford, Lawton, & Reid, 2013; Crane et al., 2010; Huang, Chung, Kroenke, Delucchi, & Spitzer, 2006; Merz, Malcarne, Roesch, Riley, & Sadler, 2011) and one non-U.S. sample using a two-factor solution.(Petersen et al., 2015) Two, they have utilized select samples – namely, primary care patients,(Huang et al., 2006) people with HIV,(Crane et

al., 2010) Latina women,(Merz et al., 2011) and non-U.S. samples.(Baas et al., 2011; Cameron et al., 2013; Petersen et al., 2015) Consequently, it is not known which PHQ-9 factor structure provides the best fit across major U.S. sociodemographic groups and whether the PHQ-9 can be used in these groups without bias. To address these important knowledge gaps, the present study examined a large, diverse sample representative of the U.S. adult population and used a state-of-the-art analytic approach to definitively determine the factor structure and measurement invariance of the PHQ-9 across major U.S. sociodemographic groups based on sex, race/ethnicity, and education level.

METHODS

Study Design and Sample

The continuous National Health and Nutrition Examination Survey (NHANES) is a cross-sectional, epidemiologic study conducted by the National Center for Health Statistics of the Centers for Disease Control and Prevention to assess the health and nutritional status of the U.S. population. Using a stratified multistage probability sampling approach, NHANES enrolls a nationally representative sample of approximately 5,000 civilian, non-institutionalized adults and children each year. Non-Hispanic Blacks and Hispanics are among the groups oversampled. Those selected to participate are initially interviewed in their homes by trained personnel, who administer sociodemographic and health-related questionnaires using computer-assisted technology. One to two weeks after the household interview, respondents are asked to visit a Mobile Examination Center (MEC) to complete additional interviews, examinations, and laboratory assessments. The NHANES website (www.cdc.gov/nchs/nhanes.htm) provides further details regarding the study design and sample.

In this report, we examined NHANES data from all of the survey years in which the PHQ-9 was administered and data had been released (2005-2014). From the total sample for these survey years (N = 50,965), we first selected all respondents aged 18 years and older (n = 30,295), given that evidence suggests that depression is experienced and expressed differently in children and adolescents (e.g., with increased irritability and suicidal ideation) than in adults.(Association, 2013; Rohde, Lewinsohn, Klein, Seeley, & Gau, 2012) Next, we excluded the 4,093 respondents with missing PHQ-9 data, leaving a final sample of 26,202 adults (see Table 1 for respondent characteristics). For our analyses examining race/ethnicity, our sample was 24,014, as we further excluded the 634 respondents in the non-Hispanic Asian group

(because the restricted response patterns in this smaller subsample led to problems with model convergence) and the 1,554 respondents in the other/multi-racial group (because the highly heterogeneous nature of this subsample would cloud interpretation of any results). For our analyses examining education level, our sample was 26,182, as we further excluded the 20 respondents with missing education level data. The institutional review board at Indiana University-Purdue University Indianapolis (IUPUI) approved this study.

Measures

Depressive Symptoms

During the MEC interview, the PHQ-9 was administered to assess depressive symptom severity over the last two weeks (see p. 35 of Appendix A for items).(Kroenke, Spitzer, & Williams, 2001a) Respondents indicated, on a 0-3 scale, the frequency with which they experienced the following symptoms of MDD: (1) anhedonia, (2) depressed mood, (3) sleep disturbance, (4) fatigue, (5) appetite changes, (6) low self-esteem, (7) concentration problems, (8) psychomotor retardation/agitation, and (9) suicidal ideation. Total scores range from 0 to 27, with scores ≥10 representing clinically significant depressive symptoms.(Kroenke & Spitzer, 2002) The PHQ-9 demonstrates high internal consistency and good sensitivity and specificity for identifying cases of MDD in community samples.(Kroenke & Spitzer, 2002; Kroenke, Spitzer, & Williams, 2001b; Manea, Gilbody, & McMillan, 2012; Patten & Schopflocher, 2009; Wittkampf, Naeije, Schene, Huyser, & van Weert, 2007; Zuithoff et al., 2010)

Sociodemographic Factors

Data regarding sex, race/ethnicity, and education level were collected by NHANES personnel during the household interview (see pp. 36-41 of Appendix A for questions and response options). Sex was coded by NHANES personnel as either male or female.

Race/ethnicity was assessed by two questions: (1) “Do you consider yourself to be Hispanic, Latino, or Spanish origin?” (yes, no, don’t know, refused), and (2) “What race do you consider yourself to be?” (American Indian or Alaskan Native, Asian, Black or African American, Native Hawaiian or Pacific Islander, White, other, don’t know, refused). Using information from these questions, NHANES personnel classified respondents into five groups (non-Hispanic White, non-Hispanic Black, Mexican American, other Hispanic, other/multi-racial) for the 2005-2010 survey years and six groups (non-Hispanic Asian was added) for the 2011-2014 survey years. In this report, we examined the following groups: non-Hispanic White, non-Hispanic Black, Mexican American, and Other Hispanic.

Education level was assessed by the question: “What is the highest grade or level of school you have completed or the highest degree you have received?” Response options were: never attended/kindergarten only, 1st grade, 2nd grade, 3rd grade, 4th grade, 5th grade, 6th grade, 7th grade, 8th grade, 9th grade, 10th grade, 11th grade, 12th grade (no diploma), high school graduate, GED or equivalent, some college (no degree), associate degree (from occupational, technical, or vocational program), associate degree (academic program), bachelor’s degree, master’s degree, professional school degree, doctoral degree, refused, and don’t know. Using information from this question, NHANES personnel classified respondents into the following groups: those aged 20+ years – less than 9th grade, 9th-12th grade with no diploma, high school graduate/GED or equivalent, some college or associate degree, college graduate or above; those aged 18-19 years – never attended/kindergarten only, grade level ranging from 1st to 12th grade with no diploma, high school graduate, GED or equivalent, or more than high school. We used the categories for respondents aged 20+ years to reclassified the education level of respondents aged 18-19 years.

Data Analysis

We performed confirmatory factor analyses (CFAs) to determine the factor structure and measurement invariance of the PHQ-9 across major U.S. sociodemographic groups. All analyses were conducted using MPlus. Consistent with current recommendations,(Rhemtulla, Brosseau-Liard, & Savalei, 2012) we utilized the means and variance adjusted weighted least squares (WLSMV) estimation method, given the ordinal nature of the PHQ-9 indicators variables.

To determine the factor structure of the PHQ-9 in U.S. adults, we conducted five single-group CFAs on our full final sample (n = 26,202), each of which examined a plausible measurement model that has received prior support in the literature (Model 1,(Cameron, Crawford, Lawton, & Reid, 2008; Dum, Pickren, Sobell, & Sobell, 2008; Huang et al., 2006; Kocalevent, Hinz, & Brähler, 2013) Model 2,(Chilcot et al., 2013; Elhai et al., 2012; Krause, Bombardier, & Carter, 2008; Petersen et al., 2015) Model 3,(Krause et al., 2008; Petersen et al., 2015) Model 4,(Elhai et al., 2012; Petersen et al., 2015) Model 5(Kalpakjian et al., 2009; Krause, Reed, & McArdle, 2010; Richardson & Richards, 2008); see Table 2). In selecting the best fitting model, we considered model fit indices and current depression theory. Both absolute (root mean square error of approximation [RMSEA]) and relative (comparative fit index [CFI] and Tucker-Lewis index [TLI]) fit indices were incorporated into our decision making.

To determine whether the PHQ-9 allows for meaningful comparisons across major U.S. sociodemographic groups, we carried out the five steps of measurement invariance testing described by Gregorich(Gregorich, 2006) by conducting single-group and multiple-group CFAs. First, we evaluated *dimensional invariance* (i.e., equivalence in the number of latent factors across groups) by separately fitting the selected model determined by the single-group CFAs in the full final sample in each sex (men, women), race/ethnicity (non-Hispanic White, non-

Hispanic Black, Mexican American, Other Hispanic), and education level (less than 9th grade, 9th-12th grade with no diploma, high school graduate/GED or equivalent, some college or associate degree, college graduate or above) group using a single-group CFA approach. Second, we evaluated *configural invariance* (i.e., equivalence in the links between the latent factors and the item sets across groups) by simultaneous fitting the selected model in the groups within each sociodemographic factor (e.g., the model was fit to men and women at the same time) using a multiple-group CFA approach. Third, we evaluated *weak factorial invariance* (i.e., equivalence in the meaning of the latent factors across groups) by imposing equality constraints on the factor loadings of the multiple-group CFAs evaluating configural invariance. Fourth, we evaluated *strong factorial invariance* (i.e., equivalence in the systematic influences on item responses unrelated to the latent factors across groups) by further imposing equality constraints on item thresholds of the multiple-group CFAs evaluating weak factorial invariance. Fifth, we evaluated *strict invariance* (i.e., equivalence in the item error estimates unexplained by the latent factors across groups) by further imposing equality constraints on item residual variances of the multiple-group CFAs evaluating strong factorial invariance.

To determine whether measurement invariance held at each of the five steps, we used a cut point of 0.010 for CFI change, which is a relative fit index that examines the distance between the worst fitting model to the hypothesized model.(Little, Bovaird, & Card, 2012) Specifically, if CFI change was < 0.010 from one step to the next (e.g., from configural invariance to weak factorial invariance), we concluded that measurement invariance held for the latter step (e.g., weak factorial invariance). Conversely, if CFI change was ≥ 0.010, we concluded that measurement invariance was not established for that step. Consistent with current recommendations,(Cheung & Rensvold, 2002) we selected CFI change as the fit index on which

to base our decisions because the only other alternative for nested model testing with WLSMV estimation – the X^2 difference test – is known to be overly sensitive in larger samples,(Brown, 2014; Cheung & Rensvold, 2002) as is the case here.

RESULTS

Depressive Symptoms and Sociodemographic Factors

The mean PHQ-9 total score for our final sample was 3.19 (*SD* = 4.28), falling in the minimal depression range. Even so, approximately 9% of respondents had a PHQ-9 total score ≥ 10, which is indicative of clinically significant depressive symptoms.(Kroenke & Spitzer, 2002) As is presented in Table 1, the mean PHQ-9 total score for each group (2.63-3.92) also fell in the minimal depression range, and the percentage of respondents with clinically significant depressive symptoms ranged from 3.8-13.4% across the groups. The mean age of our final sample was 47.4 years (*SD* = 18.9). As is shown in Table 1, just over half of the sample were women and non-White, and there was good representation across the education levels.

PHQ-9 Factor Structure

We conducted five single-group CFAs – each examining a plausible measurement model that has received empirical support – to determine the factor structure of the PHQ-9 in U.S. adults. As is shown in Table 2, all five models demonstrated close fit overall, as the RMSEAs fall within the 0.01-0.05 range and the TLIs and CFIs fall within the 0.95-0.99 range. However, the fit indices were consistently better for the two-factor models (Models 2-5; RMSEA = 0.034-0.038, TLI = 0.979-0.984, CFI = 0.985-0.989) versus the one-factor model (Model 1; RMSEA = 0.046, TLI = 0.970, CFI = 0.977). This suggests that depressive symptoms, as measured by the PHQ-9, are best conceptualized as having two, rather than one, distinct symptom clusters – cognitive/affective and somatic.

Because the two-factor models demonstrated similar fit indices, we turned to previous research on depression to guide our selection of the best model. As can be seen in Table 2, Models 2-5 differ with respect to which factor the psychomotor disturbances item and

concentration difficulties item load. However, past work defining somatic symptoms indicates that these two symptoms are not common physical symptoms experienced during psychological distress.(Simon, Gater, Kisely, & Piccinelli, 1996) Additionally, past work specific to depression suggests that the six somatic symptoms of depression are disordered eating, body image problems, fatigue, breathing difficulties, sleep disturbances, and aches and pains.(Dozois, Dobson, & Ahnberg, 1998; Silverstein & Patel, 2011) Three of these symptoms (appetite changes, fatigue, and sleep disturbance) are measured by the PHQ-9. Therefore, we selected Model 2 (presented in Figure 1) as the best model. Of note, our selected model is in line with previous research on the Beck Depression Inventory-II, which has been found to have a two-factor solution with a somatic factor structure similar to what we report here.(Dozois et al., 1998)

PHQ-9 Measurement Invariance across Sex, Race/Ethnicity, and Education Level

We conducted a series of single-group and multiple-group CFAs to carry out the five steps of measurement invariance testing and determine whether the PHQ-9 allows for meaningful comparisons across sex, race/ethnicity, and education level groups in U.S. adults.

Dimensional and Configural Invariance

To evaluate dimensional invariance, we separately fit Model 2 to each sociodemographic group of interest. Separate single-group CFAs for men (RMSEA = 0.034, TLI = 0.984, CFI = 0.988) and women (RMSEA = 0.038, TLI = 0.983, CFI = 0.988); for the four race/ethnicity groups of non-Hispanic White (RMSEA = 0.039, TLI = 0.982, CFI = 0.987), non-Hispanic Black (RMSEA = 0.032, TLI = 0.989, CFI = 0.992), Mexican American (RMSEA = 0.032, TLI = 0.989, CFI = 0.992), and Other Hispanics (RMSEA = 0.035, TLI = 0.987, CFI = 0.991); and for the five education level groups of less than 9^{th} grade (RMSEA = 0.030, TLI = 0.990, CFI = 0.993), 9^{th} to 12^{th} grade – no diploma (RMSEA = 0.037, TLI = 0.982, CFI = 0.987), high school

graduate/GED equivalent (RMSEA = 0.037, TLI = 0.985, CFI = 0.989), some college or associate degree (RMSEA = 0.034, TLI = 0.986, CFI = 0.990), and college graduate or above (RMSEA = 0.029, TLI = 0.982, CFI =0.987) all demonstrated close fit. These results indicate that there is an equivalent number of latent factors (two) across these groups.

To evaluate configural invariance, we simultaneously fit Model 2 to the groups within each sociodemographic factor. Specifically, we ran three multiple-group CFAs – one for sex, one for race/ethnicity, and one for education level. As is shown in Table 3, the models for sex (RMSEA = 0.036, TLI = 0.984, CFI = 0.988), race/ethnicity (RMSEA = 0.036, TLI = 0.985, CFI = 0.989), and education level (RMSEA = 0.034, TLI = 0.985, CFI = 0.989) all demonstrated close fit. Furthermore, as can be seen in Table 4, the factor loadings across the groups were similar. These results indicate that there is equivalence in the links between the PHQ-9 latent factors and the PHQ-9 items sets across the groups. Our dimensional and configural invariance testing indicates that the selected two-factor model for the PHQ-9 with cognitive/affective and somatic symptom clusters (see Figure 1) exists across the examined sex, race/ethnicity, and education level groups.

Weak, Strong, and Strict Factorial Invariance

Because dimensional and configural invariance of the PHQ-9 was established, we proceeded with weak, strong, and strict factorial invariance testing. To evaluate weak factorial invariance, we equated the factor loadings across the sociodemographic groups of the three multiple-group CFAs that evaluated configural invariance (see Table 3). After imposing this constraint, the models for sex (RMSEA = 0.028, TLI = 0.990, CFI = 0.992), race/ethnicity (RMSEA = 0.030, TLI = 0.989, CFI = 0.991), and education level (RMSEA = 0.027, TLI = 0.991, CFI = 0.992) again all demonstrated close fit. In addition, our nested model testing

comparing the configural invariance models to the weak factorial invariance models showed that CFI change was < 0.010 for all three models (CFI change range = 0.002 to 0.004), demonstrating that weak factorial invariance was established. These results indicate that the two latent factors carry equivalent meaning across groups.

To evaluate strong factorial invariance, we further equated the item thresholds of the three multiple-group CFAs that evaluated weak factorial invariance. The models for sex (RMSEA = 0.025, TLI = 0.992, CFI = 0.991), race/ethnicity (RMSEA = 0.027, TLI = 0.992, CFI = 0.989), and education level (RMSEA = 0.027, TLI = 0.991, CFI = 0.987) all demonstrated close fit. Moreover, the CFI change from the weak to strong factorial invariance models was < 0.010 for all three models (CFI change range = -0.005 to -0.001), showing that strong factorial invariance was established. These results indicate that there is equivalence in the systematic influences on item responses related to the two latent factors across groups. Therefore, mean PHQ-9 scores can be meaningfully compared between these groups without bias.

To evaluate strict factorial invariance, we equated the error loadings of the three multiple group CFAs that evaluated strong factorial invariance. The models of sex (RMSEA = 0.027, TLI = 0.991, CFI = 0.991), race/ethnicity (RMSEA = 0.026, TLI = 0.992, CFI = 0.991), and education (RMSEA = 0.024, TLI = 0.993, CFI = 0.991) demonstrated close fit. Moreover, the CFI change from the weak to strong factorial invariance models was < 0.010 for all three models (CFI change range = 0.000 to 0.004), showing that strict factorial invariance was established. These results indicate that there is equivalence in the item error estimates unexplained by the two latent factors across groups.

Altogether, our weak, strong, and strict factorial invariance testing yields three conclusions. One, by establishing weak invariance, we demonstrate that the PHQ-9

cognitive/affective and somatic symptom clusters as specified in our two-factor model in Figure 1 have the same meaning across sex, race/ethnicity, and education level groups in U.S. adults. Two, as stated by Gregorich (2006), establishing strong and strict factorial invariance indicates that it is defensible to compare PHQ-9 observed means and variances/covariances across the examined sex, race/ethnicity, and education level groups. In other words, the PHQ-9 allows for meaningful comparisons in depressive symptoms across major sociodemographic groups in the U.S. Three, our use of a two-factor solution for the PHQ-9 demonstrates that subscale scores for somatic and cognitive/affective items are defensible to compare across the examined sex, race/ethnicity, and education level groups.

DISCUSSION

Our objective was to address key knowledge gaps by definitively determining the factor structure and measurement invariance of the PHQ-9 across major U.S. sociodemographic groups based on sex, race/ethnicity, and education level. Regarding factor structure, we determined that the best solution for the PHQ-9 consists two distinct symptom clusters. This finding indicates that depressive symptom severity, as measured by the PHQ-9, is best conceptualized as having a cognitive/affective symptom cluster consisting of anhedonia, depressed mood, feelings of worthlessness, concentration difficulties, psychomotor disturbances, and thoughts of death and a somatic symptom cluster consisting of sleep disturbance, fatigue, and appetite changes. Concerning measurement invariance, we determined that this two-factor solution for the PHQ-9 is invariant across sex, race/ethnicity, and education level groups in U.S. adults. This finding indicates that PHQ-9 cognitive/affective and somatic symptom clusters have the same meaning in these groups and that it is defensible to compare PHQ-9 total and subscale scores across these groups. Ultimately, the PHQ-9 is acceptable to use in major sociodemographic groups in the U.S., as it allows for meaningful comparisons in total depressive symptoms and depressive symptom clusters with minimal risk of measurement bias.

Previous research examining the PHQ-9 factor structure has yielded mixed results. Results of some studies have supported a one-factor solution, with all of the PHQ-9 items loading on a single latent factor.(Cameron et al., 2008; Dum et al., 2008; Huang et al., 2006; Kocalevent et al., 2013) However, these studies have used an inappropriate technique called principal component analysis, which often produces one-factor solutions for measures with a lower number of items, like the PHQ-9.(Brown, 2006) In addition, studies supporting a two-factor solution over a one-factor solution have not agreed upon the best two-factor solution. For

instance, two CFA studies have supported a two-factor model where the somatic factor consists of the sleep disturbance, fatigue, and appetite changes items,(Krause et al., 2008)-(Chilcot et al., 2013) whereas another CFA study has supported a two-factor model where the somatic factor consists of those three items plus the concentration difficulties and psychomotor changes items.(Elhai et al., 2012) Finally, one other CFA study has supported the latter two-factor, five-item somatic factor model with the addition of two residual correlations – one between the sleep disturbance and fatigue items and one between the feelings of worthlessness and thoughts of death items.(Petersen et al., 2015) The two-factor solution we report is identical to that of the first two studies.(Krause et al., 2008)-(Chilcot et al., 2013) Importantly, we extend these previous results by being the first study (a) to examine the PHQ-9 factor structure in large sample representative of the U.S. population and (b) to justify our two-factor solution using both our empirical findings and past theoretical work.

More importantly, we demonstrate that the PHQ-9 exhibits measurement invariance across major U.S. sociodemographic groups and, thus, is an acceptable depression screener to use in the general U.S. population regardless of sex, race/ethnicity, and education level. Our findings are in line with two U.S. studies that utilized multiple indicators, multiple causes (MIMC) modeling and concluded that the PHQ-9 is acceptable to use in two different clinical samples. The first study,(Huang et al., 2006) using a sample of 5,053 primary care patients, found minimal differential item functioning in Chinese American and Latino groups when compared to the non-Hispanic White group. The authors concluded that the differential item functioning was not severe enough to have a clinically meaningful impact in primary care settings. Similarly, the second study(Crane et al., 2010) found minimal differential item functioning between African Americans and Whites in sample of 1,467 HIV-infected patients,

also noting that it was not severe enough to have a clinically meaningful impact. Our findings are also consistent with the one U.S. study that utilized multiple-group CFA. That study involved 479 Latina women and supported PHQ-9 measurement invariance between English-speaking and Spanish-speaking Latina women.(Merz et al., 2011) Of note, all three of these prior studies using a U.S. sample have only used a one-factor solution, limiting the generalizability of their results, as the PHQ-9 is better conceptualized as a two-factor solution. Importantly, our findings extend these prior results in select samples to the U.S. adult population and major sociodemographic groups within this population.

In terms of clinical and public health implications, this is the first study to demonstrate that the PHQ-9 is an acceptable depression measure to use in major sociodemographic groups in the U.S. general population. Because the PHQ-9 does not show measurement bias based on sex, race/ethnicity, or education level and allows for meaningful comparisons across these groups, it is unlikely that using this brief depression screener will result in either (a) under-detection of depression and under-treatment or (b) over-detection of depression and the wasting of limited depression treatment resources based on sociodemographic group membership. Consequently, our findings provide strong support the use of the PHQ-9 for depression screening and for assessing depressive symptom levels in general clinical settings, such as primary care, as well as in communities at large, which may prove useful for population health efforts.

In terms of research implications, this is the first study to validate a two-factor solution for the PHQ-9 in a U.S. representative sample and to show that the PHQ-9 cognitive/affective and somatic subscale scores have the same meaning and can be compared across major U.S. sociodemographic groups. This validation of a two-factor solution in U.S. adults is an important advance for depression research because it provides a consistent, evidence-based way for

investigators to compute PHQ-9 subscale scores in future studies. Specifically, the cognitive/affective score should be computed as the sum of items 1, 2, 6, 7, 8, and 9, and the somatic score should be computed as the sum items 3, 4, and 5. Recent investigations using the PHQ-9 have begun to examine whether depressive symptom clusters differentially predict various health-related outcomes.(Case & Stewart, 2014; Holzapfel et al., 2008; Smolderen et al., 2009; Vrany, Berntson, Khambaty, & Stewart, 2016) Unfortunately, the approach to computing PHQ-9 subscale scores varies across these studies, which clouds interpretation. The use of consistent, evidence-based PHQ-9 subscale scores in these growing literatures will facilitate comparisons of findings across studies as well as future meta-analytic efforts.

Our study has important strengths, such as a large sample representative of the U.S. population and the use of the gold-standard approach for multiple-group invariance testing. However, it also has some limitations. First, due to restricted response patterns in the smaller groups, we were unable to examine some major race/ethnicities in the U.S., such as non-Hispanic Asians. Subsequent studies, perhaps utilizing future NHANES cohorts, should evaluate the measurement invariance of the PHQ-9 in these groups. Second, we could not use the X^2 difference test for nesting model testing because it is known to be too sensitive (i.e., detects significant differences between models when those differences are not meaningful) with larger sample sizes,(Cheung & Rensvold, 2002) like we have here. Therefore, consistent with current recommendations,(Cheung & Rensvold, 2002) we used CFI change, which is a less sensitive index for nesting model testing. Use of this less sensitive index could lead to the opposite problem of failing to detect meaningful differences between models. However, the similarity in the absolute and relative fit indices (all indicating close fit) and the factor loadings (Table 4) across groups suggests that we did not fail to detect meaningful differences in this study.

In conclusion, using a U.S. representative sample, we determined that a two-factor solution for the PHQ-9 with a cognitive/affective factor and a somatic factor is invariant across sex, race/ethnicity, and education level groups. In the clinical and public health domains, we demonstrate that the PHQ-9 is an acceptable measure to utilize in major sociodemographic groups in the U.S. general population, extending the use of this depression screener from the primary care clinic to the community. In the research domain, we show that PHQ-9 cognitive/affective and somatic subscale scores have the same meaning and can be compared across major U.S. sociodemographic groups and provide a consistent, evidence-based approach to computing PHQ-9 subscale scores to be used in future studies.

REFERENCES

Association, A. P. (2013). *DSM 5*. American Psychiatric Association.

Baas, K. D., Cramer, A. O. J., Koeter, M. W. J., Van De Lisdonk, E. H., Van Weert, H. C., & Schene, A. H. (2011). Measurement invariance with respect to ethnicity of the Patient Health Questionnaire-9 (PHQ-9). *Journal of Affective Disorders*, *129*(1–3), 229–235. https://doi.org/10.1016/j.jad.2010.08.026

Brown, T. A. (2006). Confirmatory factor analysis for applied research. *New York: Guilford*.

Brown, T. A. (2014). *Confirmatory factor analysis for applied research*. Guilford Publications.

Cameron, I. M., Crawford, J. R., Lawton, K., & Reid, I. C. (2008). Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. *British Journal of General Practice*, *58*(546), 32–36.

Cameron, I. M., Crawford, J. R., Lawton, K., & Reid, I. C. (2013). Differential item functioning of the HADS and PHQ-9: an investigation of age, gender and educational background in a clinical UK primary care sample. *Journal of Affective Disorders*, *147*(1–3), 262–8. https://doi.org/10.1016/j.jad.2012.11.015

Case, S. M., & Stewart, J. C. (2014). Race/ethnicity moderates the relationship between depressive symptom severity and C-reactive protein: 2005–2010 NHANES data. *Brain, Behavior, and Immunity*, *41*, 101–108.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255.

Chilcot, J., Rayner, L., Lee, W., Price, A., Goodwin, L., Monroe, B., … Hotopf, M. (2013). The factor structure of the PHQ-9 in palliative care. *Journal of Psychosomatic Research*, *75*(1), 60–64.

Crane, P. K., Gibbons, L. E., Willig, J. H., Mugavero, M. J., Lawrence, S. T., Schumacher, J. E., … Crane, H. M. (2010). Measuring depression levels in HIV-infected patients as part of routine clinical care using the nine-item Patient Health Questionnaire (PHQ-9), *22*(7). https://doi.org/10.1080/09540120903483034

Cuijpers, P., Vogelzangs, N., Twisk, J., Kleiboer, A., Li, J., & Penninx, B. W. (2014). Comprehensive meta-analysis of excess mortality in depression in the general community versus patients with specific illnesses. *American Journal of Psychiatry*.

Dozois, D. J. A., Dobson, K. S., & Ahnberg, J. L. (1998). A psychometric evaluation of the Beck Depression Inventory–II. *Psychological Assessment*, *10*(2), 83.

Dum, M., Pickren, J., Sobell, L. C., & Sobell, M. B. (2008). Comparing the BDI-II and the PHQ-9 with outpatient substance abusers. *Addictive Behaviors*, *33*(2), 381–387.

Elhai, J. D., Contractor, A. A., Tamburrino, M., Fine, T. H., Prescott, M. R., Shirley, E., … others. (2012). The factor structure of major depression symptoms: a test of four competing models using the Patient Health Questionnaire-9. *Psychiatry Research*, *199*(3), 169–173.

Ferrari, A. J., Charlson, F. J., Norman, R. E., Patten, S. B., Freedman, G., Murray, C. J. L., … Whiteford, H. A. (2013). Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010, *10*(11).

Greenberg, P. E., Fournier, A.-A., Sisitsky, T., Pike, C. T., & Kessler, R. C. (2015). The economic burden of adults with major depressive disorder in the United States (2005 and 2010). *The Journal of Clinical Psychiatry*, *76*(2), 155–162.

Greenberg, P. E., Kessler, R. C., Birnbaum, H. G., Leong, S. A., Lowe, S. W., Berglund, P. A., & Corey-Lisle, P. K. (2003). The economic burden of depression in the United States: how did it change between 1990 and 2000? *Journal of Clinical Psychiatry*, *64*(12), 1465–1475.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, *44*(11 Suppl 3), S78.

Hardeveld, F., Spijker, J., De Graaf, R., Nolen, W. A., & Beekman, A. T. F. (2010). Prevalence and predictors of recurrence of major depressive disorder in the adult population. *Acta Psychiatrica Scandinavica*, *122*(3), 184–191.

Holzapfel, N., Müller-Tasch, T., Wild, B., Jünger, J., Zugck, C., Remppis, A., … Löwe, B. (2008). Depression profile in patients with and without chronic heart failure. *Journal of Affective Disorders*, *105*(1), 53–62.

Huang, F. Y., Chung, H., Kroenke, K., Delucchi, K. L., & Spitzer, R. L. (2006). Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. *Journal of General Internal Medicine*, *21*(6), 547–552. https://doi.org/10.1111/j.1525-1497.2006.00409.x

Kalpakjian, C. Z., Toussaint, L. L., Albright, K. J., Bombardier, C. H., Krause, J. K., & Tate, D. G. (2009). Patient Health Questionnaire-9 in spinal cord injury: an examination of factor structure as related to gender. *The Journal of Spinal Cord Medicine*, *32*(2), 147.

Kocalevent, R.-D., Hinz, A., & Brähler, E. (2013). Standardization of the depression screener Patient Health Questionnaire (PHQ-9) in the general population. *General Hospital Psychiatry*, *35*(5), 551–555. https://doi.org/10.1016/j.genhosppsych.2013.04.006

Krause, J. S., Bombardier, C., & Carter, R. E. (2008). Assessment of depressive symptoms during inpatient rehabilitation for spinal cord injury: is there an underlying somatic factor when using the PHQ? *Rehabilitation Psychology*, *53*(4), 513.

Krause, J. S., Reed, K. S., & McArdle, J. J. (2010). Factor structure and predictive validity of somatic and nonsomatic symptoms from the Patient Health Questionnaire-9: a longitudinal study after spinal cord injury. *Archives of Physical Medicine and Rehabilitation*, *91*(8), 1218–1224.

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: a new depression diagnostic and severity measure. *Psychiatr Ann*, *32*(9), 1–7.

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001a). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001b). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Little, T. D., Bovaird, J. A., & Card, N. A. (2012). *Modeling contextual effects in longitudinal studies*. Routledge.

Manea, L., Gilbody, S., & McMillan, D. (2012). Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. *Canadian Medical Association Journal*, *184*(3), E191–E196.

Merz, E. L., Malcarne, V. L., Roesch, S. C., Riley, N., & Sadler, G. R. (2011). A multigroup confirmatory factor analysis of the Patient Health Questionnaire-9 among English-and Spanish-speaking Latinas. *Cultural Diversity and Ethnic Minority Psychology*, *17*(3), 309.

Nease, D. E., & Malouin, J. M. (2003). Depression screening: a practical strategy. *Journal of Family Practice*, *52*(2), 118–126.

Patten, S. B., & Schopflocher, D. (2009). Longitudinal epidemiology of major depression as assessed by the Brief Patient Health Questionnaire (PHQ-9). *Comprehensive Psychiatry*, *50*(1), 26–33.

Petersen, J. J., Paulitsch, M. A., Hartig, J., Mergenthal, K., Gerlach, F. M., & Gensichen, J. (2015). Factor structure and measurement invariance of the Patient Health Questionnaire-9 for female and male primary care patients with major depression in Germany. *Journal of Affective Disorders*, *170*, 138–142.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org/10.1037/a0029315

Richardson, E. J., & Richards, J. S. (2008). Factor structure of the PHQ-9 screen for depression across time since injury among persons with spinal cord injury. *Rehabilitation Psychology*, *53*(2), 243.

Rohde, P., Lewinsohn, P. M., Klein, D. N., Seeley, J. R., & Gau, J. M. (2012). Key characteristics of major depressive disorder occurring in childhood, adolescence, emerging adulthood, and adulthood. *Clinical Psychological Science*, 2167702612457599.

Rotella, F., & Mannucci, E. (2013). Depression as a risk factor for diabetes: a meta-analysis of longitudinal studies. *The Journal of Clinical Psychiatry*, *74*(1), 31.

Silverstein, B., & Patel, P. (2011). Poor response to antidepressant medication of patients with depression accompanied by somatic symptomatology in the STAR* D Study. *Psychiatry Research*, *187*(1), 121–124.

Simon, G., Gater, R., Kisely, S., & Piccinelli, M. (1996). Somatic symptoms of distress: an international primary care study. *Psychosomatic Medicine*, *58*(5), 481–488.

Siu, A. L., Bibbins-Domingo, K., Grossman, D. C., Baumann, L. C., Davidson, K. W., Ebell, M., … Kemper, A. R. (2016). Screening for depression in adults: US Preventive Services Task Force recommendation statement. *Jama*, *315*(4), 380–387.

Smolderen, K. G., Spertus, J. A., Reid, K. J., Buchanan, D. M., Krumholz, H. M., Denollet, J., … Chan, P. S. (2009). The association of cognitive and somatic depressive symptoms with depression recognition and outcomes after myocardial infarction. *Circulation: Cardiovascular Quality and Outcomes*, *2*(4), 328–337.

Vrany, E. A., Berntson, J. M., Khambaty, T., & Stewart, J. C. (2016). Depressive symptoms clusters and insulin resistance: race/ethnicity as a moderator in 2005–2010 NHANES data. *Annals of Behavioral Medicine*, *50*(1), 1–11.

Wittkampf, K. A., Naeije, L., Schene, A. H., Huyser, J., & van Weert, H. C. (2007). Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review. *General Hospital Psychiatry*, *29*(5), 388–395.

Zuithoff, N. P. A., Vergouwe, Y., King, M., Nazareth, I., van Wezep, M. J., Moons, K. G. M., & Geerlings, M. I. (2010). The Patient Health Questionnaire-9 for detection of major depressive disorder in primary care: consequences of current thresholds in a crosssectional study. *BMC Family Practice*, *11*(1), 98.

TABLES

Table 1
Characteristics of NHANES Respondents

	Frequency (%)	PHQ-9 mean (SD)	PHQ-9 ≥ 10, %	PHQ-9 somatic mean (SD)	PHQ-9 cognitive/affective mean (SD)
Sex (*n* = 26,202)					
Women	13,295 (50.7)	3.74 (4.60)	11.2	2.05 (2.24)	1.69 (2.82)
Men	12,907 (49.3)	2.63 (3.84)	6.4	1.41 (1.90)	1.22 (2.38)
Race/Ethnicity (*n* = 24,014)					
Non-Hispanic White	11,738 (48.9)	3.16 (4.17)	8.4	1.79 (2.09)	1.36 (2.54)
Non-Hispanic Black	5,687 (23.7)	3.22 (4.41)	9.3	1.73 (2.18)	1.49 (2.68)
Mexican American	4,270 (17.8)	3.14 (4.20)	8.9	1.62 (2.04)	1.52 (2.60)
Other Hispanic	2,319 (9.7)	3.82 (4.98)	12.5	1.91 (2.24)	1.91 (3.15)
Education Level (*n* = 26,182)					
Less than 9th grade	2,602 (9.9)	3.74 (4.99)	13.4	1.82 (2.30)	1.93 (3.14)
9th to 12th grade (no diploma)	4,352 (16.6)	3.92 (4.82)	12.5	2.00 (2.28)	1.92 (3.03)
High school graduate/GED equivalent	6,214 (23.7)	3.30 (4.31)	9.1	1.80 (2.13)	1.50 (2.63)
Some college or associate degree	7,571 (28.9)	3.20 (4.22)	8.6	1.79 (2.12)	1.41 (2.54)
College graduate or above	5,443 (20.8)	2.23 (3.19)	3.8	1.35 (1.70)	0.87 (1.88)

Note. NHANES = National Health and Nutrition Examination Survey. PHQ-9 = Patient Health Quesionnaire-9.

Table 2
Single-Group Confirmatory Factor Analysis Models and Fit Indices Evaluating Factor Structure of the Patient Health Questionnaire-9 (PHQ-9) in U.S. Adults

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
1. Anhedonia	Depression	Cognitive/Affective	Cognitive/Affective	Cognitive/Affective	Cognitive/Affective
2. Depressed Mood	Depression	Cognitive/Affective	Cognitive/Affective	Cognitive/Affective	Cognitive/Affective
3. Sleep Disturbance	Depression	Somatic	Somatic	Somatic	Somatic
4. Fatigue	Depression	Somatic	Somatic	Somatic	Somatic
5. Appetite Changes	Depression	Somatic	Somatic	Somatic	Somatic
6. Feelings of Worthlessness	Depression	Cognitive/Affective	Cognitive/Affective	Cognitive/Affective	Cognitive/Affective
7. Concentration Difficulties	Depression	Cognitive/Affective	Cognitive/Affective	Somatic	Cognitive/Affective
8. Psychomotor Disturbances	Depression	Cognitive/Affective	Somatic	Somatic	Both
9. Thoughts of Death	Depression	Cognitive/Affective	Cognitive/Affective	Cognitive/Affective	Cognitive/Affective
Chi-Square (χ2)	1523.52	806.756	1028.398	881.285	780.152
Degrees of Freedom	27	26	26	26	25
***p*-value**	0.00	0.00	0.00	0.00	0.00
RMSEA	0.046	0.034	0.038	0.035	0.034
RMSEA 90% Confidence Interval	0.044 - 0.048	0.032 - 0.036	0.036 - 0.040	0.033 - 0.037	0.032 - 0.036
TLI	0.970	0.984	0.979	0.982	0.984
CFI	0.977	0.988	0.985	0.987	0.989

Note. N = 26,202. For absolute fit indices (i.e., RMSEA), exact fit = 0.00, close fit = 0.01-0.05, acceptable fit = 0.05-0.08, mediocre fit = 0.08-0.10, and poor fit = greater than 0.10. For relative fit indices (i.e., TLI and CFI), exact fit = 1.00, close fit = 0.95-0.99, acceptable fit = 0.90-0.95, mediocre fit = 0.85-0.90, and poor fit = less than 0.85. RMSEA = root mean square error of approximation. TLI = Tucker-Lewis index. CFI = comparative fit index.

Table 3

Multiple-Group Confirmatory Factor Analysis Models and Fit Indices Evaluating Measurement Invariance of the Patient Health Questionnaire-9 (PHQ-9) across Sex, Race/Ethnicity, and Education Level in U.S. Adults

		χ2	df	Δχ2	*p*-value	RMSEA	RMSEA 90% CI	TLI	CFI	ΔCFI
Sex *n* = 26,202	**Configural**	924.80	52	---	---	0.036	0.034 - 0.038	0.984	0.988	---
	Weak	679.40	59	32.037	0.000	0.028	0.026 - 0.030	0.990	0.992	0.004
	Strong	746.82	84	102.027	0.000	0.025	0.023 - 0.026	0.992	0.991	-0.001
	Strict	766.89	75	746.821	0.000	0.027	0.025 - 0.028	0.991	0.991	0.000
Race/ Ethnicity *n* = 24,014	**Configural**	909.09	104	---	---	0.036	0.034 - 0.038	0.985	0.989	---
	Weak	809.06	125	138.020	0.000	0.030	0.028 - 0.032	0.989	0.991	0.002
	Strong	1045.45	200	345.989	0.000	0.027	0.025 - 0.028	0.992	0.989	-0.002
	Strict	856.80	173	264.653	0.000	0.026	0.024 - 0.027	0.992	0.991	0.002
Education Level *n* = 26,182	**Configural**	907.44	130	---	---	0.034	0.032 - 0.036	0.985	0.989	---
	Weak	764.09	158	88.853	0.000	0.027	0.025 - 0.029	0.991	0.992	0.003
	Strong	1210.92	258	592.955	0.000	0.027	0.025 - 0.028	0.991	0.987	-0.005
	Strict	876.80	222	393.994	0.000	0.024	0.022 - 0.025	0.993	0.991	0.004

Note. We used the two-factor Model 2 shown in Figure 1. For absolute fit indices (i.e., RMSEA), exact fit = 0.00, close fit = 0.01-0.05, acceptable fit = 0.05-0.08, mediocre fit = 0.08-0.10, and poor fit = greater than 0.10. For relative fit indices (i.e., TLI and CFI), exact fit = 1.00, close fit = 0.95-0.99, acceptable fit = 0.90-0.95, mediocre fit = 0.85-0.90, and poor fit = less than 0.85. For nested model testing, a change in CFI ≥ 0.010 signifies that measurement invariance was not established. RMSEA = root mean square error of approximation. CI = confidence interval. TLI = Tucker-Lewis index. CFI = comparative fit index.

Table 4

Factor Loadings from the Multiple-Group Confirmatory Factor Analyses Evaluating Configural Invariance

		Sex		Race/Ethnicity				Education Level				
Latent Variable	**PHQ-9 Item**	**Men**	**Women**	**Non-Hispanic White**	**Non-Hispanic Black**	**Mexican American**	**Other Hispanic**	**Less than 9th grade**	**9th to 12th grade (no diploma)**	**High school graduate/ GED equivalent**	**Some college or associate degree**	**College graduate or above**
Cognitive/Affective	1. Anhedonia	0.80	0.82	0.84	0.75	0.75	0.80	0.76	0.76	0.78	0.84	0.85
	2. Depressed Mood	0.89	0.88	0.89	0.87	0.86	0.87	0.88	0.86	0.88	0.89	0.88
	6. Feelings of Worthlessness	0.84	0.83	0.84	0.82	0.82	0.84	0.82	0.83	0.86	0.84	0.81
	7. Concentration Difficulties	0.75	0.79	0.77	0.79	0.76	0.84	0.78	0.74	0.80	0.78	0.77
	8. Psychomotor Disturbances	0.76	0.75	0.73	0.79	0.78	0.84	0.79	0.77	0.73	0.74	0.68
	9. Thoughts of Death	0.78	0.75	0.76	0.800	0.73	0.79	0.71	0.74	0.76	0.79	0.77
Somatic	3. Sleep Disturbance	0.73	0.70	0.69	0.79	0.73	0.77	0.74	0.74	0.72	0.73	0.66
	4. Fatigue	0.79	0.80	0.81	0.80	0.81	0.79	0.80	0.79	0.81	0.82	0.80
	5. Appetite Changes	0.72	0.77	0.76	0.77	0.73	0.76	0.74	0.75	0.75	0.77	0.73

FIGURE

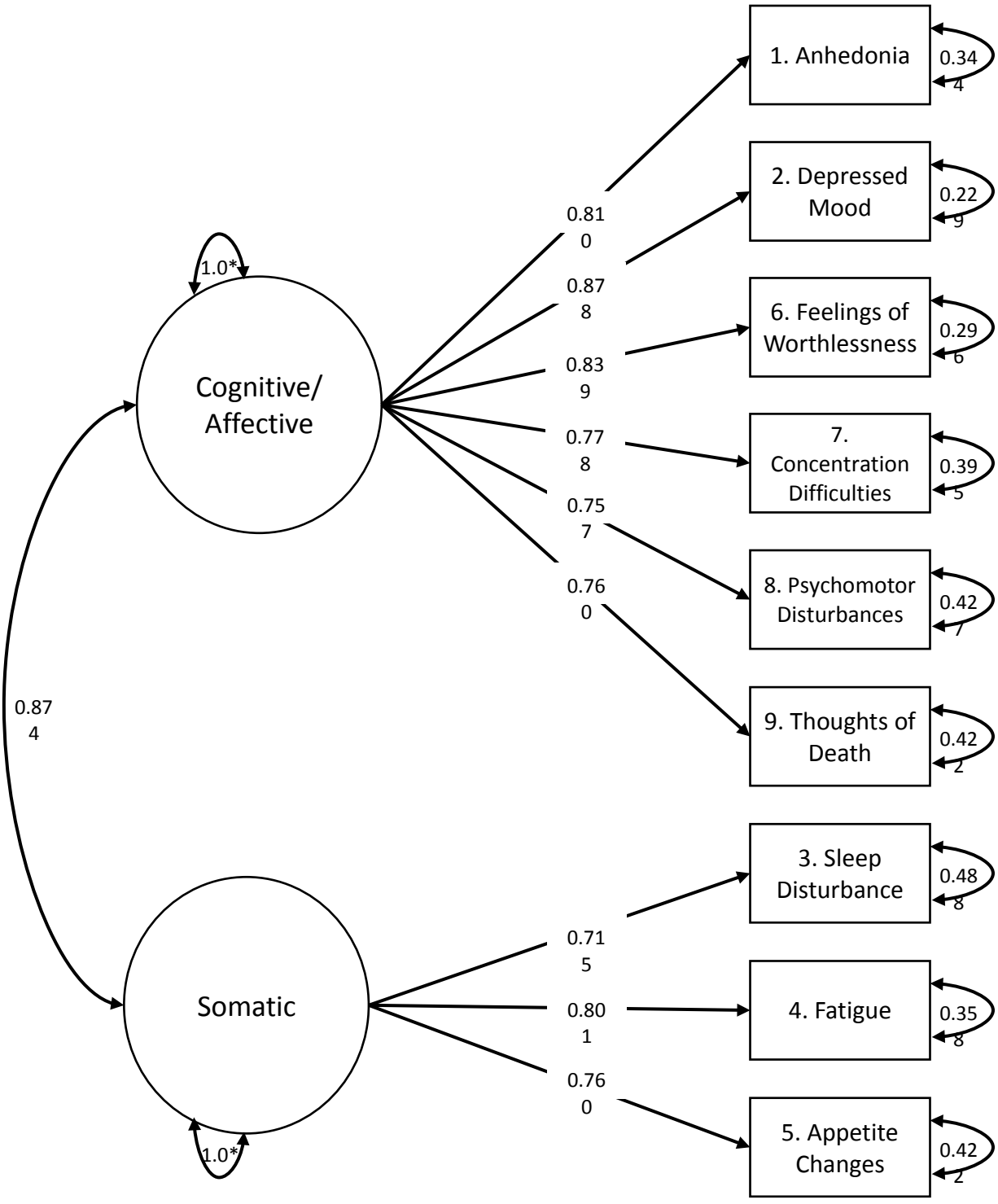


Figure 1: Two-factor Measurement Model of the Patient Health Quesionnaire-9 (PHQ-9)

On the right, the boxes represent the PHQ-9 items (indicator variables). Circular arrows that point back to the indicator variables represent item error variances. Moving to the left, unidirectional linear arrows pointing from circles to boxes represent standardized factor loadings. The circles represent latent factors. Circular arrows that point back to the latent factors represents latent variances (fixed to 1.0 for identification purposes). The bidirectional arrow between latent factors represents a standardized covariance coefficient.

APPENDIX

The Patient Health Questionnaire-9

Over the last 2 weeks, how often have you been bothered by any of the following problems. Circle the number that applies.	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself – or that you are a failure or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
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	0	1	2	3
9. Thoughts that you would be better off dead or of hurting yourself in some way	0	1	2	3
Column Totals	0	+ ☐	+ ☐	+ ☐
			Total Score	☐

If you checked off any problem, how difficult have these problems made it for you to do your work, take care of things at home, or get along with other people	Not difficult at all	Somewhat difficult	Very Difficult	Extremely difficult
	☐	☐	☐	☐

Developed by Drs. Robert L. Spitzer, Janet B.W. Williams, Kurt Kreonke and colleagues with an educational grant from Pfizer Inc.

In-Home Race/Ethnicity Interview Questions and Procedures 2005-2010

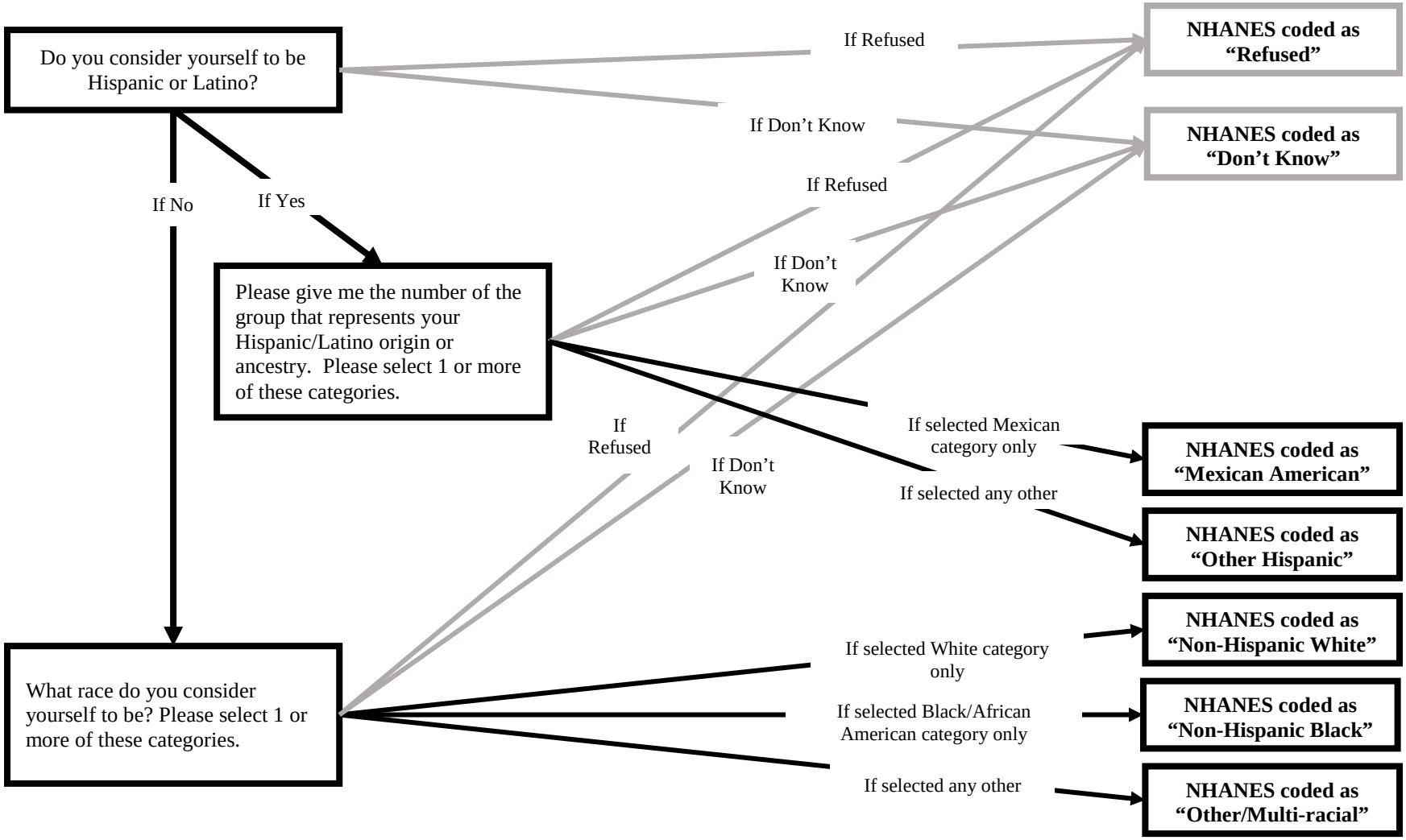
Do you consider yourself to be Hispanic or Latino?
Response options
Yes
No
Refused
Don't Know

Please give me the number of the group that represents your Hispanic/Latino origin or ancestry. Please select 1 or more of these categories.	
Response options	
Mexican	Paraguayan
Puerto Rican	Peruvian
Cuban	Uruguayan
Dominican Republic	Venezuelan
Costa Rican	Other South American
Guatemalan	Spaniard
Honduran	Spanish American
Nicaraguan	Hispano/Hispana
Panamanian	Hispanic/Latino
Salvadoran	Other Hispanic Latino
Other Central American	Spaniard
Argentinean	Spanish American
Bolivian	Hispano/Hispana
Chilean	Hispanic/Latino
Colombian	Other Hispanic Latino
Ecuadorian	Refused
	Don't Know

What race do you consider yourself to be? Please select 1 or more of these categories.
Response options
White
Black/African American
Indian (American)
Alaska Native
Native Hawaiian
Guamanian
Samoan
Other Pacific Islander
Asian Indian
Chinese
Filipino
Japanese
Korean
Vietnamese
Other Asian
Some Other Race
Refused
Don't Know

Source: https://www.cdc.gov/nchs/nhanes/nhanes2009-2010/questionnaires09_10.htm

NHANES Coding



In-Home Race/Ethnicity Interview Questions and Procedures 2011-2014

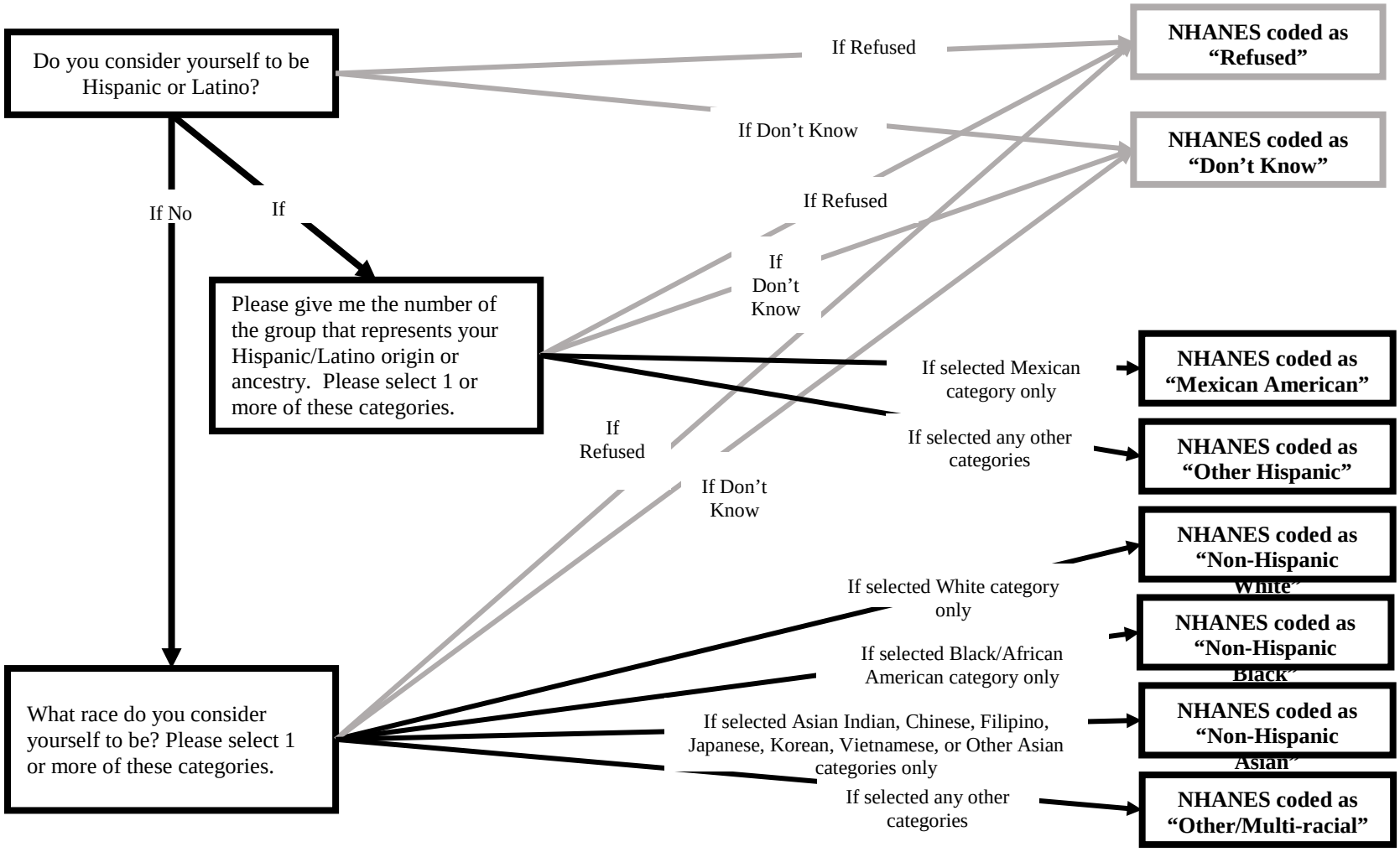
Do you consider yourself to be Hispanic or Latino?
Response options
Yes
No
Refused
Don't Know

Please give me the number of the group that represents your Hispanic/Latino origin or ancestry. Please select 1 or more of these categories.	
Response options	
Mexican	Paraguayan
Puerto Rican	Peruvian
Cuban	Uruguayan
Dominican Republic	Venezuelan
Costa Rican	Other South American
Guatemalan	Spaniard
Honduran	Spanish American
Nicaraguan	Hispano/Hispana
Panamanian	Hispanic/Latino
Salvadoran	Other Hispanic Latino
Other Central American	Spaniard
Argentinean	Spanish American
Bolivian	Hispano/Hispana
Chilean	Hispanic/Latino
Colombian	Other Hispanic Latino
Ecuadorian	Refused
	Don't Know

What race do you consider yourself to be? Please select 1 or more of these categories.
Response options
White
Black/African American
Indian (American)
Alaska Native
Native Hawaiian
Guamanian
Samoan
Other Pacific Islander
Asian Indian
Chinese
Filipino
Japanese
Korean
Vietnamese
Other Asian
Some Other Race
Refused
Don't Know

Source: https://www.cdc.gov/nchs/nhanes/nhanes2011-2012/questionnaires11_12.htm

NHANES Coding



In-Home Education Interview Questions and Procedures 2005-2014

What is the highest grade or level of school you have completed or the highest degree you have received?	
Response options	
Never Attended/Kindergarten Only 1st Grade 2nd Grade 3rd Grade 4th Grade 5th Grade 6th grade 7th Grade 8th Grade 9th Grade 10th Grade 11th Grade 12th Grade, No Diploma	High School Graduate GED or Equivalent Some College, No Degree Associate Degree: Occupational, Technical, or Vocational Program Associate Degree: Academic Program Bachelor's Degree Master's Degree Professional School Degree Doctoral Degree Refused Don't Know

NHANES Coding