HHS Public Access

Author manuscript

Depress Anxiety. Author manuscript; available in PMC 2020 September 01.

Published in final edited form as:

Depress Anxiety. 2019 September ; 36(9): 813–823. doi:10.1002/da.22940.

Measurement Invariance of the Patient Health Questionnaire-9 (PHQ-9) Depression Screener in U.S. Adults Across Sex, Race/Ethnicity, and Education Level: NHANES 2005-2016

Jay S. Patel, MS[1], **Youngha Oh, MEd**[2], **Kevin L. Rand, PhD**[1], **Wei Wu, PhD**[1], **Melissa A. Cyders, PhD**[1], **Kurt Kroenke, MD**[3,4,5], **Jesse C. Stewart, PhD**[1]

[1]Department of Psychology, Indiana University-Purdue University Indianapolis (IUPUI), Indianapolis, IN

[2]Educational Psychology, Research, Evaluation, Measurement, and Statistics (REMS), Texas Tech University, Lubbock, TX

[3]VA HSR&D Center for Health Information and Communication, Roudebush VA Medical Center, Indianapolis, IN

[4]Department of Medicine, Indiana University School of Medicine, Indianapolis, IN

[5]Regenstrief Institute, Indianapolis, IN

Abstract

Background: Despite its popularity, little is known about the measurement invariance of the Patient Health Questionnaire-9 (PHQ-9) across U.S. sociodemographic groups. Use of a screener shown not to possess measurement invariance could result in under-detection/treatment of depression, potentially exacerbating sociodemographic disparities in depression. Therefore, we assessed the factor structure and measurement invariance of the PHQ-9 across major U.S. sociodemographic groups.

Methods: U.S. population representative data came from the 2005–2016 National Health and Nutrition Examination Survey (NHANES) cohorts We conducted a measurement invariance analysis of 31,366 respondents across sociodemographic factors of sex, race/ethnicity, and education level.

Results: Considering results of single-group confirmatory factor analyses (CFAs), depression theory, and research utility, we justify a two-factor structure for the PHQ-9 consisting of a cognitive/affective factor and a somatic factor (RMSEA=0.034, TLI=0.985, CFI=0.989). Based on multiple-group CFAs testing configural, scalar, and strict factorial invariance, we determined that invariance held for sex, race/ethnicity, and education level groups, as all models demonstrated close model fit (RMSEA=0.025–0.025, TLI=0.985–0.992, CFI=0.986–0.991). Finally, for all steps ΔCFI was < -0.004, and ΔRMSEA was < 0.01.

Corresponding Author: Jesse C. Stewart, Ph.D., Department of Psychology, Indiana University-Purdue University Indianapolis (IUPUI), 402 N. Blackford St., LD 100E, Indianapolis, IN 46202. Telephone: (317) 274-6761. Fax: (317) 274-6756. jstew@iupui.edu.

Data Availability Statement: Data is publicly available at https://www.cdc.gov/nchs/nhanes/index.htm

Conclusions: We demonstrate that the PHQ-9 is acceptable to use in major U.S. sociodemographic groups and allows for meaningful comparisons in total, cognitive/affective, and somatic depressive symptoms across these groups, extending its use to the community. This knowledge is timely as medicine moves towards alternative payment models emphasizing high-quality and cost-efficient care, which will likely incentivize behavioral and population health efforts. We also provide a consistent, evidence-based approach for calculating PHQ-9 subscale scores.

Introduction

Depression is a top public health concern due to its high prevalence, chronicity, and grave ramifications. The lifetime prevalence of major depressive disorder (MDD) in the U.S. is 16% (Kessler et al., 2003). Depression also disproportionately affects people with lower socioeconomic status (Lorant et al., 2003). The course of MDD is often chronic, with a 15-year recurrence rate of 35% in the general population (Hardeveld, Spijker, De Graaf, Nolen, & Beekman, 2010). Its grave ramifications include increased disability, mortality, and societal costs. Depression is the second leading cause of disability (Ferrari et al., 2013), and associated with increased mortality risk (Cuijpers et al., 2014). Moreover, the total annual cost of depression has increased from $83 billion in 2000 to $210 billion in 2010 (Greenberg, Fournier, Sisitsky, Pike, & Kessler, 2015; Greenberg et al., 2003).

These alarming observations have motivated efforts to improve the detection and management of depression by routinely administering screeners. In 2016, the U.S. Preventive Services Task Force (USPSTF) recommended depression screening in the general adult population (Siu et al., 2016). Accompanying editorials underscored the importance of "population screening" (Thase, 2016) while also critiquing the USPSTF statement for not specifying "the ideal screening interval and the settings with highest potential yield" (Reynolds & Frank, 2016, pg. 189). One widely-used, self-report depression screener that was designed for primary care is the Patient Health Questionnaire-9 (PHQ-9; Hirschtritt & Kroenke, 2017). The PHQ-9 was identified as the first choice screener for adults in the accompanying editorial (Thase, 2016). In addition to a depression screener, the PHQ-9 has been validated as a continuous measure of depressive symptom severity (Kroenke, Spitzer, & Williams, 2001). The push for increased depression screening is occurring in the context of medicine's movement toward alternative payment models that emphasize high-quality and cost-efficient care (Ganguli & Ferris, 2018). It has been posited that behavioral health integration and population health management will be incentivized and thus more widely adopted under these emerging payment models (Burwell, 2015; Joynt Maddox, 2018; "Smarter Spending. Healthier People: Paying Providers for Value, Not Volume," 2015). These potentially profound changes to the healthcare system will amplify the need for brief depression screeners known to operate in an unbiased manner in major sociodemographic groups in the U.S., especially considering depression is associated with high healthcare utilization and costs (Bock et al., 2014; Kroenke & Unutzer, 2017; Luber et al., 2000; Prina et al., 2015). Moreover, measurement-based care is a key pillar in evidence-based behavioral health integration programs that have not only improved depression outcomes but also have proven cost-effective (Kroenke & Unutzer, 2017).

Despite its widespread use in research and clinical practice, surprisingly little is known about the psychometric performance of the PHQ-9 across major U.S. sociodemographic groups. An advanced statistical approach called multiple-group confirmatory factor analysis (CFA) can evaluate an instrument's psychometric performance across groups through measurement invariance testing. Measurement invariance is the statistical property of measurement that signifies the same underlying construct is being measured across groups. If measurement invariance is established, it would demonstrate that PHQ-9 assesses the same construct across U.S. sociodemographic groups and that observed differences in PHQ-9 scores among these groups reflect true group differences in depressive symptoms. Thus, we could conclude that the PHQ-9 is acceptable to use in major U.S. sociodemographic groups and that the PHQ-9 allows for meaningful comparisons in depressive symptoms across these groups. However, if measurement invariance is not established, it would raise serious concerns regarding the validity and utility of the PHQ-9 for population-level screening in the U.S. Widespread use of a depression screener shown *not* to possess measurement invariance would likely result in under-detection or over-detection of depression in certain groups. Under-detection would lead to under-treatment of depression, which could increase sociodemographic disparities in depression care and outcomes (Kim, 2014; Simpson, Krishnan, Kunik, & Ruiz, 2007), whereas over-detection would lead to the wasting of limited treatment resources.

Although a few investigations have examined the measurement invariance of the PHQ-9, these studies have been limited in two ways. One, they have not examined a consistent factor structure, with some claiming the PHQ-9 has one factor (Baas et al., 2011; Cameron, Crawford, Lawton, & Reid, 2013; Crane et al., 2010; Galenkamp, Stronks, Snijder, & Derks, 2017; González-Blanch et al., 2018; Huang, Chung, Kroenke, Delucchi, & Spitzer, 2006; Keum, Miller, & Inkelas, 2018; Merz, Malcarne, Roesch, Riley, & Sadler, 2011) and others claiming two factors (Petersen et al., 2015), making the generalizability of findings confusing. Two, they have utilized non-representative samples – namely, primary care patients (Huang et al., 2006), college students (Keum et al., 2018), 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; Galenkamp et al., 2017; González-Blanch et al., 2018; Petersen et al., 2015). Consequently, it is not known which PHQ-9 factor structure provides the best fit and justification across major U.S. sociodemographic groups and whether the PHQ-9 can be used in these groups without bias. To fill these gaps, we examined a large, diverse sample representative of the U.S. adult population and used a state-of-the-art analytic approach to determine the factor structure and measurement invariance of the PHQ-9 across sex, race/ethnicity, and education level.

Methods

Study Design and Sample

The continuous National Health and Nutrition Examination Survey (NHANES) is a cross-sectional, population-based study designed to assess the health and nutritional status of the U.S. population. Using stratified multistage probability sampling, NHANES enrolls a nationally representative sample of ~5,000 non-institutionalized civilians each year. Those

selected to participate are initially interviewed in their homes by trained personnel, who administer 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. See the NHANES website (www.cdc.gov/nchs/nhanes.htm) for further details.

We examined NHANES data from all survey years in which the PHQ-9 was administered (2005–2016). From the total sample (N=60,936), we selected all respondents aged 18+ years who had complete PHQ-9 data (n = 31,366; Table 1). This sample was used to determine the PHQ-9 factor structure and measurement invariance across sex. For our analyses across race/ethnicity, our sample was 30,179, as we excluded the 1,187 respondents in the other/multi-racial group (because this highly heterogeneous group would cloud result interpretation). For our analyses across education level, our sample was 31,344, as we excluded the 22 respondents with missing data. The study was approved by the local institutional review board.

Measures

Depressive Symptoms—The PHQ-9 was administered during the face-to-face MEC interview to assess depressive symptoms over the last two weeks (Kroenke et al., 2001). Respondents indicated, on a 0–3 scale, the frequency with which they experienced the following symptoms: (1) anhedonia, (2) depressed mood, (3) sleep disturbance, (4) fatigue, (5) appetite changes, (6) low self-esteem, (7) concentration problems, (8) psychomotor disturbances, and (9) suicidal ideation. Total scores range from 0 to 27, with scores ≥10 representing clinically significant depressive symptoms (Kroenke & Spitzer, 2002). Furthermore, the PHQ-9 is validated as a depressive symptom severity measure (total score 1–4: minimal depression, 5–9: mild depression, 10–14: moderate depression, 15–19: moderately severe depression, and 20–27: severe depression; Kroenke et al., 2001). The PHQ-9 demonstrates high internal consistency and good sensitivity and specificity for identifying cases of MDD (Hirschtritt & Kroenke, 2017; Kroenke et al., 2001; Mitchell, Yadegarfar, Gill, & Stubbs, 2016; Moriarty, Gilbody, McMillan, & Manea, 2015; Zuithoff et al., 2010).

Sociodemographic Factors—Sex, race/ethnicity, and education level data were collected by NHANES personnel during the household interview. 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?”, and (2) “What race do you consider yourself to be?” Using responses to 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 years and six groups (non-Hispanic Asian was added) for the 2011–2016 years.

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?” Using responses to this question, NHANES personnel classified respondents into the following groups: those aged 20+ years – less than 9^{th} grade, 9^{th}–12^{th} grade with no diploma, high school graduate/GED

or equivalent, some college or associate degree, or college graduate or above; those aged 18–19 years – never attended/kindergarten only, grade level ranging from 1st-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 reclassify respondents aged 18–19 years.

Data Analysis

We performed CFAs using M*Plus* software (Muthén & Muthén, 2015). Consistent with current recommendations (Rhemtulla, Brosseau-Liard, & Savalei, 2012), we used the means and variance adjusted weighted least squares (WLSMV) estimation method due to the ordinal scale . To determine the factor structure, we conducted five single-group CFAs on our full sample (n = 31,366), each of which examined a plausible model that has received support (Model 1, Cameron, Crawford, Lawton, & Reid, 2008; Dum, Pickren, Sobell, & Sobell, 2008; Galenkamp et al., 2017; González-Blanch et al., 2018; Huang et al., 2006; Keum et al., 2018; Kocalevent, Hinz, & Brähler, 2013; Model 2, Chilcot et al., 2013; Elhai et al., 2012; Krause, Bombardier, & Carter, 2008; Petersen et al., 2015; Model 3, Baas et al., 2011; 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 justifying our baseline model, we considered fit indices (root mean square error of approximation [RMSEA], Tucker-Lewis index [TLI], and comparative fit index [CFI]) and current depression theory. To assess model fit, we used the following guidelines: For absolute fit indices (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 (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 (Hu & Bentler, 1999).

To determine the PHQ-9's measurement invariance, we carried out the four steps described by Gregorich (2006) and Sass (2014) using single and multiple-group CFAs. In these steps, four models were tested sequentially, each representing a specific level of measurement invariance, going from the least to most restrictive level of invariance. First, we evaluated *dimensional invariance* (equivalence in the number of latent factors across groups) by separately fitting the selected baseline model to each sex, race/ethnicity, and education level group using a single-group CFA approach. Second, we evaluated *configural invariance* (the latent factors are indicated by the same items across groups) by simultaneously fitting the selected model to the groups within each sociodemographic factor (e.g., the model was simultaneously fit to men and women) using a multiple-group CFA approach. Third, we evaluated *scalar invariance* (equivalence in the meaning of the latent factors and in potential item response biases unrelated to the latent factors across groups) by imposing equality constraints on the factor loadings (correlation between the item and factor) and item thresholds (the ordinal equivalent of an item intercept) of the configural invariance model. Fourth, we evaluated *strict invariance* (equivalence in the item error estimates unexplained by the latent factors across groups) by further imposing equality constraints on item residual variances of the scalar invariance model.

Consistent with current recommendations, we used a < -0.004 change in CFI (ΔCFI), to determine whether measurement invariance held at each step (Rutkowski & Svetina, 2017).

If ΔCFI was < −0.004 from one step to the next (e.g., from configural to scalar invariance), indicating that adding more equality constraints did not substantially decrease model fit, we concluded that the latter model (e.g., scalar invariance) is not significantly worse. Conversely, if ΔCFI was ≥ −0.004, we selected the former model (e.g., configural invariance). As is also recommended, we used ΔRMSEA < 0.010, in conjunction with ΔCFI, for invariance testing (Rutkowski & Svetina, 2017).

Results

Depressive Symptoms and Sociodemographic Factors

The mean PHQ-9 total score was 3.20 (*SD* = 4.27), falling in the minimal depression range. Even so, 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 (2.23–3.92) also fell in the minimal depression range, and the percentage of respondents with clinically significant depressive symptoms ranged from 3.8–13.2%. The mean age was 47.5 years (*SD* =18.8). About half of the sample were women and non-White, and there was good representation across the education levels (see Table 1). Finally, all groups demonstrated a similar correlation between the somatic and cognitive/affective constructs.

PHQ-9 Factor Structure

We conducted single-group CFAs, assessing the factor structure for the PHQ-9. All five models demonstrated close model-data fit, as the RMSEAs fell within the 0.01–0.05 range and the TLIs and CFIs fell within the 0.95–0.99 range (Table 2). Based on the fit indices alone, all models were plausible; however, we advocate for a two-factor model. Because depression is a multifaceted disorder, the high correlation of the factors and the items most likely suggest a higher order factor of depression (Byrne, 2005), rather than a more simplistic one-factor solution. In addition, a two-factor model has research utility. Specifically, studies examining depressive symptoms clusters have found that they are differentially associated with various health-related outcomes (Case & Stewart, 2014; Holzapfel et al., 2008; Roest et al., 2013; Roest et al., 2011; Smolderen et al., 2009; Vrany, Berntson, Khambaty, & Stewart, 2015).

Because all two-factor models had similar model-data fit, we turned to current depression theory to guide our selection of the best model. Models 2–5 differ with respect to on which factor the psychomotor disturbances and concentration difficulties items load. Concentration difficulties refer to depression-related issues with attention (a domain of cognitive functioning) and are thus conceptualized as a cognitive symptom of depression (De Jonge, Mangano, & Whooley, 2007; Duivis, Vogelzangs, Kupper, de Jonge, & Penninx, 2013; Hoen et al., 2010; McIntyre et al., 2015). For this theoretical reason, we conclude it is more reasonable that the concentration difficulties item fall under the cognitive/affective factor, which rules out Model 4. Model 5, in which the psychomotor disturbances item loads on both factors, revealed that this item was a stronger indicator of the cognitive/affective factor (factor loading = 0.6) than the somatic factor (factor loading = 0.2). This finding is consistent with current views on the neurobiological underpinnings of depression-related

psychomotor disturbances (Drevets, 2001). Specifically, psychomotor disturbances are thought to arise from dysfunction in the same brain reward pathways that are thought to underlie anhedonia (Drevets, 2001), an affective symptom of depression (Treadway & Zald, 2011). Thus, the psychomotor disturbances item theoretically justified under the PHQ-9 cognitive/affective factor, ruling out Models 3 and 5. Therefore, we selected Model 2 (see Figure 1), which had factor loadings above 0.70, for subsequent testing.

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

To evaluate *dimensional invariance*, we separately fit Model 2 to each group using a single-group CFA approach. All models demonstrated close fit (Table 3); therefore, this model served as a baseline for subsequent invariance testing.

To evaluate *configural invariance*, we simultaneously fit Model 2 to the groups within each sociodemographic factor by running three multiple-group CFAs. The models for sex, race/ethnicity, and education level all demonstrated close fit (Table 4), and the factor loadings across the groups were similar (Table 5). These results demonstrate that the number of factors are indicated by the same pattern of item loadings across groups, meaning that Model 2 exists across sex, race/ethnicity, and education level groups.

To evaluate *scalar invariance*, we further equated the factor loadings and item thresholds across sociodemographic groups from the configural model. The models for sex, race/ethnicity, and education level all demonstrated close fit (Table 4). The ΔCFI (< -0.004) and ΔRMSEA (< 0.010) criteria were met for all three models, signifying that scalar invariance was established. These results demonstrate that there is equivalence in both the meaning of the latent factors and the systematic influences on item responses unrelated to the latent factors across groups.

To evaluate *strict invariance*, we equated the error variances of the three multiple group CFAs from the scalar model. Yet again, the models of sex, race/ethnicity, and education level all demonstrated close fit (Table 4). The ΔCFI (< -0.004) and ΔRMSEA (< 0.010) criteria were met for all three models, indicating that strict invariance was established. These results demonstrate that there is equivalence in the item error variances across groups.

Altogether, measurement invariance tests yield three conclusions. One, the PHQ-9 cognitive/affective and somatic factors as specified in Figure 1 carry the same meaning across sex, race/ethnicity, and education level groups in U.S. adults. Two, it is defensible to compare PHQ-9 observed means and variances/covariances across sex, race/ethnicity, and education level groups (Gregorich, 2006). In other words, the PHQ-9 allows for meaningful comparisons in depressive symptoms across these major U.S. sociodemographic groups with minimal risk of bias. Three, our use of a two-factor solution demonstrates that it is defensible to compare PHQ-9 cognitive/affective and somatic subscale scores across these groups.

Discussion

Our study examined the factor structure and measurement invariance of the PHQ-9 across major U.S. sociodemographic groups. First, we justified a two-factor solution for the PHQ-9 consisting of cognitive/affective and somatic factors. A subscale score for the *cognitive/affective factor* can be computed by summing the responses to items 1 (anhedonia), 2 (depressed mood), 6 (low self-esteem), 7 (concentration difficulties), 8 (psychomotor disturbances), and 9 (suicidal ideation). A subscale score for the *somatic factor* can be computed by summing items 3 (sleep disturbance), 4 (fatigue), and 5 (appetite changes). Future studies should consider standardizing subscale scores (i.e., *z* scoring) when comparing them.

Our two-factor solution differs from previous studies advocating for a one-factor solution (Dum et al., 2008; Galenkamp et al., 2017; González-Blanch et al., 2018; Huang et al., 2006; Keum et al., 2018; Kocalevent et al., 2013). Three of these studies (Dum et al., 2008; Huang et al., 2006; Kocalevent et al., 2013) used principal component analysis, which is mathematically inappropriate (Brown, 2014). The studies that used CFA did not consider the research utility of a two-factor solution. Specifically, investigations using the PHQ-9 have found that depressive symptom clusters are differentially associated with various health-related outcomes (Case & Stewart, 2014; Holzapfel et al., 2008; Roest et al., 2013; Roest et al., 2011; Smolderen et al., 2009; Vrany et al., 2015). Consistent with CFA recommendations (Brown, 2014), our study takes into consideration empirical evidence, current theory, and practical utility. Nonetheless, the PHQ-9 total score currently has high clinical utility, and future research is needed to compare the utility of the total and subscale scores in different clinical scenarios (e.g., screening versus monitoring).

Second, we established strict measurement invariance for the PHQ-9 across major U.S. sociodemographic groups. Therefore, the observed differences in the PHQ-9 total, cognitive/affective, and somatic scores among these groups reflect true group differences in these depressive symptoms. The findings we report support the conclusions that (a) the PHQ-9 is acceptable to use in major sociodemographic groups in the U.S. and that (b) the PHQ-9 allows for meaningful comparisons in total, cognitive/affective, and somatic depressive symptoms across these groups with minimal risk of bias. Although we tested measurement invariance using a two-factor solution, it is still appropriate to compute and use the single PHQ-9 total score for research, clinical, or other purposes across the examined groups, as the high correlation among the factors and items implies a higher order factor of depression (Byrne, 2005). However, it may be useful in future research to explicitly test this assumption.

As with all studies, ours is not without limitations. First, although we use a large U.S. representative sample, there were missing values which could limit the generalizability of findings. However, missingness in the sample was below the 10% guideline (Little, Jorgensen, Lang, & Moore, 2013). Second, due to the cross-sectional nature of the NHANES data, we were not able to assess measurement invariance overtime, as done by González-Blanch et al. (2018). Future studies are needed to determine whether measurement invariance of the selected two-factor solution for the PHQ-9 holds overtime. Third, age was

beyond the scope of this study, future investigations are needed to determine the factor structure and measurement invariance of the PHQ-9 across various categorizations for age groups.

Conclusion

We extend previous research validating the PHQ-9 as a depression screener and severity measure by determining its factor structure in U.S. adults and its measurement invariance across sex, race/ethnicity, and education level groups. Our findings demonstrate that the PHQ-9 is acceptable to use in major sociodemographic groups in the U.S. and allows for meaningful comparisons in total, cognitive/affective, and somatic depressive symptoms across these groups, extending the PHQ-9's use from the clinic to the community. This knowledge is especially useful and timely as medicine moves toward wider adoption of alternative payment models emphasizing high-quality and cost-efficient care, which will likely incentivize behavioral health integration and population health management. Finally, we also provide a consistent, evidence-based approach for computing PHQ-9 cognitive/affective and somatic subscale scores, which should facilitate comparisons of results across studies and future meta-analytic efforts.

Acknowledgments

Conflicts of Interest and Source of Funding: A portion of Dr. Stewart's time was supported by the National Heart, Lung, and Blood Institute under Award Number R01HL122245. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Baas KD, Cramer AO, Koeter MW, van de Lisdonk EH, van Weert HC, & Schene AH (2011). Measurement invariance with respect to ethnicity of the Patient Health Questionnaire-9 (PHQ-9). Journal of affective disorders, 129(1), 229–235. [PubMed: 20888647]

Bock J-O, Luppa M, Brettschneider C, Riedel-Heller S, Bickel H, Fuchs A, … Schäfer I (2014). Impact of depression on health care utilization and costs among multimorbid patients–results from the multicare cohort study. PLoS One, 9(3), e91973.

Brown TA (2014). Confirmatory factor analysis for applied research: Guilford Publications.

Burwell SM (2015). Setting value-based payment goals—HHS efforts to improve US health care. N Engl J Med, 372(10), 897–899. [PubMed: 25622024]

Byrne BM (2005). Factor analytic models: Viewing the structure of an assessment instrument from three perspectives. Journal of personality assessment, 85(1), 17–32. [PubMed: 16083381]

Cameron IM, Crawford JR, Lawton K, & Reid IC (2008). Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. Br J Gen Pract, 58(546), 32–36. [PubMed: 18186994]

Cameron IM, Crawford JR, Lawton K, & Reid IC (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), 262–268. [PubMed: 23218250]

Case SM, & Stewart JC (2014). Race/ethnicity moderates the relationship between depressive symptom severity and C-reactive protein: 2005–2010 NHANES data. Brain Behav Immun, 41, 101–108. [PubMed: 24859042]

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. [PubMed: 23751240]

Crane P, Gibbons L, Willig J, Mugavero M, Lawrence S, Schumacher J, … Crane H (2010). Measuring depression levels in HIV-infected patients as part of routine clinical care using the nine-item Patient Health Questionnaire (PHQ-9). AIDS care, 22(7), 874–885. [PubMed: 20635252]

Cuijpers P, Vogelzangs N, Twisk J, Kleiboer A, Li J, & Penninx BW (2014). Comprehensive meta-analysis of excess mortality in depression in the general community versus patients with specific illnesses. American Journal of Psychiatry, 171(4), 453–462. [PubMed: 24434956]

De Jonge P, Mangano D, & Whooley MA (2007). Differential association of cognitive and somatic depressive symptoms with heart rate variability in patients with stable coronary heart disease: findings from the Heart and Soul Study. Psychosomatic medicine, 69(8), 735. [PubMed: 17942844]

Diniz BS, Butters MA, Albert SM, Dew MA, & Reynolds CF (2013). Late-life depression and risk of vascular dementia and Alzheimer's disease: systematic review and meta-analysis of community-based cohort studies. The British Journal of Psychiatry, 202(5), 329–335. [PubMed: 23637108]

Drevets WC (2001). Neuroimaging and neuropathological studies of depression: implications for the cognitive-emotional features of mood disorders. Current opinion in neurobiology, 11(2), 240–249. [PubMed: 11301246]

Duivis HE, Vogelzangs N, Kupper N, de Jonge P, & Penninx BW (2013). Differential association of somatic and cognitive symptoms of depression and anxiety with inflammation: findings from the Netherlands Study of Depression and Anxiety (NESDA). Psychoneuroendocrinology, 38(9), 1573–1585. [PubMed: 23399050]

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

Elhai JD, Contractor AA, Tamburrino M, Fine TH, Prescott MR, Shirley E, … Galea S (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. [PubMed: 22698261]

Ferrari AJ, Charlson FJ, Norman RE, Patten SB, Freedman G, Murray CJ, … Whiteford HA (2013). Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. PLoS medicine, 10(11), e1001547.

Galenkamp H, Stronks K, Snijder MB, & Derks EM (2017). Measurement invariance testing of the PHQ-9 in a multi-ethnic population in Europe: the HELIUS study. BMC Psychiatry, 17(1), 349. [PubMed: 29065874]

Gan Y, Gong Y, Tong X, Sun H, Cong Y, Dong X, … Deng J (2014). Depression and the risk of coronary heart disease: a meta-analysis of prospective cohort studies. BMC Psychiatry, 14(1), 371. [PubMed: 25540022]

Ganguli I, & Ferris TG (2018). Accountable Care at the Frontlines of a Health System: Bridging Aspiration and Reality. Jama, 319(7), 655–656. [PubMed: 29228143]

González-Blanch C, Medrano LA, Muñoz-Navarro R, Ruíz-Rodríguez P, Moriana JA, Limonero JT, … Group PR (2018). Factor structure and measurement invariance across various demographic groups and over time for the PHQ-9 in primary care patients in Spain. PLoS One, 13(2), e0193356.

Greenberg PE, Fournier A-A, Sisitsky T, Pike CT, & Kessler RC (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. [PubMed: 25742202]

Greenberg PE, Kessler RC, Birnbaum HG, Leong SA, Lowe SW, Berglund PA, & Corey-Lisle PK (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. [PubMed: 14728109]

Gregorich SE (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. [PubMed: 17060839]

Hardeveld F, Spijker J, De Graaf R, Nolen W, & Beekman A (2010). Prevalence and predictors of recurrence of major depressive disorder in the adult population. Acta Psychiatrica Scandinavica, 122(3), 184–191. [PubMed: 20003092]

Hirschtritt ME, & Kroenke K (2017). Screening for depression. Jama, 318(8), 745–746. [PubMed: 28829850]

Hoen PW, Whooley MA, Martens EJ, Na B, van Melle JP, & de Jonge P (2010). Differential associations between specific depressive symptoms and cardiovascular prognosis in patients with stable coronary heart disease. Journal of the American College of Cardiology, 56(11), 838–844. [PubMed: 20813281]

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. [PubMed: 17512058]

Hu L. t., & Bentler PM (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal, 6(1), 1–55.

Huang FY, Chung H, Kroenke K, Delucchi KL, & Spitzer RL (2006). Using the patient health questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. J Gen Intern Med, 21(6), 547–552. [PubMed: 16808734]

Joynt Maddox KE (2018). Financial Incentives and Vulnerable Populations—Will Alternative Payment Models Help or Hurt? New England Journal of Medicine, 378(11), 977–979. [PubMed: 29539282]

Kalpakjian CZ, Toussaint LL, Albright KJ, Bombardier CH, Krause JK, & Tate DG (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), 147–156. [PubMed: 19569462]

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

Keum BT, Miller MJ, & Inkelas KK (2018). Testing the factor structure and measurement invariance of the PHQ-9 across racially diverse US college students. Psychological assessment.

Kim M (2014). Racial/ethnic disparities in depression and its theoretical perspectives. Psychiatric Quarterly, 85(1), 1–8. [PubMed: 23801269]

Kocalevent R-D, Hinz A, & Brähler E (2013). Standardization of the depression screener patient health questionnaire (PHQ-9) in the general population. Gen Hosp Psychiatry, 35(5), 551–555. [PubMed: 23664569]

Krause JS, Bombardier C, & Carter RE (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 JS, Reed KS, & McArdle JJ (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. [PubMed: 20684902]

Kroenke K, & Spitzer RL (2002). The PHQ-9: a new depression diagnostic and severity measure. Psychiatric annals, 32(9), 509–515.

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

Kroenke K, & Unutzer J (2017). Closing the false divide: sustainable approaches to integrating mental health services into primary care. J Gen Intern Med, 32(4), 404–410. [PubMed: 28243873]

Little TD, Jorgensen TD, Lang KM, & Moore EWG (2013). On the joys of missing data. Journal of Pediatric Psychology, 39(2), 151–162. [PubMed: 23836191]

Lorant V, Deliège D, Eaton W, Robert A, Philippot P, & Ansseau M (2003). Socioeconomic inequalities in depression: a meta-analysis. American journal of epidemiology, 157(2), 98–112. [PubMed: 12522017]

Luber MP, Hollenberg JP, Williams-Russo P, DiDomenico TN, Meyers BS, Alexopoulos GS, & Charlson ME (2000). Diagnosis, treatment, comorbidity, and resource utilization of depressed patients in a general medical practice. The International Journal of Psychiatry in Medicine, 30(1), 1–13. [PubMed: 10900557]

Luppino FS, de Wit LM, Bouvy PF, Stijnen T, Cuijpers P, Penninx BW, & Zitman FG (2010). Overweight, obesity, and depression: a systematic review and meta-analysis of longitudinal studies. Archives of General Psychiatry, 67(3), 220–229. [PubMed: 20194822]

McIntyre RS, Xiao HX, Syeda K, Vinberg M, Carvalho AF, Mansur RB, … Cha DS (2015). The prevalence, measurement, and treatment of the cognitive dimension/domain in major depressive disorder. CNS Drugs, 29(7), 577–589. [PubMed: 26290264]

Merz EL, Malcarne VL, Roesch SC, Riley N, & Sadler GR (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. [PubMed: 21787063]

Mitchell AJ, Yadegarfar M, Gill J, & Stubbs B (2016). Case finding and screening clinical utility of the Patient Health Questionnaire (PHQ-9 and PHQ-2) for depression in primary care: a diagnostic meta-analysis of 40 studies. British Journal of Psychiatry Open, 2(2), 127–138. [PubMed: 27703765]

Moriarty AS, Gilbody S, McMillan D, & Manea L (2015). Screening and case finding for major depressive disorder using the Patient Health Questionnaire (PHQ-9): a meta-analysis. Gen Hosp Psychiatry, 37(6), 567–576. [PubMed: 26195347]

Muthén L, & Muthén B (2015). Mplus. The comprehensive modelling program for applied researchers: user's guide, 5.

Petersen JJ, Paulitsch MA, Hartig J, Mergenthal K, Gerlach FM, & 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. [PubMed: 25240840]

Prina AM, Cosco TD, Dening T, Beekman A, Brayne C, & Huisman M (2015). The association between depressive symptoms in the community, non-psychiatric hospital admission and hospital outcomes: A systematic review. Journal of psychosomatic research, 78(1), 25–33. [PubMed: 25466985]

Reynolds CF, & Frank E (2016). US preventive services task force recommendation statement on screening for depression in adults: not good enough. JAMA Psychiatry, 73(3), 189–190. [PubMed: 26815331]

Rhemtulla M, Brosseau-Liard PÉ, & 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. [PubMed: 22799625]

Richardson EJ, & Richards JS (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.

Roest AM, Carney RM, Freedland KE, Martens EJ, Denollet J, & de Jonge P (2013). Changes in cognitive versus somatic symptoms of depression and event-free survival following acute myocardial infarction in the Enhancing Recovery In Coronary Heart Disease (ENRICHD) study. Journal of affective disorders, 149(1–3), 335–341. [PubMed: 23489396]

Roest AM, Thombs BD, Grace SL, Stewart DE, Abbey SE, & de Jonge P (2011). Somatic/affective symptoms, but not cognitive/affective symptoms, of depression after acute coronary syndrome are associated with 12-month all-cause mortality. Journal of affective disorders, 131(1–3), 158–163. [PubMed: 21159385]

Rotella F, & Mannucci E (2013). Depression as a risk factor for diabetes: a meta-analysis of longitudinal studies. The Journal of clinical psychiatry.

Rutkowski L, & Svetina D (2017). Measurement invariance in international surveys: Categorical indicators and fit measure performance. Applied Measurement in Education, 30(1), 39–51.

Sass DA, Schmitt TA, & Marsh HW (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. Structural Equation Modeling: A Multidisciplinary Journal, 21(2), 167–180.

Simpson SM, Krishnan LL, Kunik ME, & Ruiz P (2007). Racial disparities in diagnosis and treatment of depression: a literature review. Psychiatric Quarterly, 78(1), 3–14. [PubMed: 17102936]

Siu AL, Bibbins-Domingo K, Grossman DC, Baumann LC, Davidson KW, Ebell M, … Kemper AR (2016). Screening for depression in adults: US Preventive Services Task Force recommendation statement. Jama, 315(4), 380–387. [PubMed: 26813211]

Smarter Spending. Healthier People: Paying Providers for Value, Not Volume. (2015, 1/26/2015).

Smolderen KG, Spertus JA, Reid KJ, Buchanan DM, Krumholz HM, Denollet J, … Chan PS (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. [PubMed: 20031858]

Spitzer RL, Kroenke K, Williams JB, & Group P. H. Q. P. C. S. (1999). Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Jama, 282(18), 1737–1744. [PubMed: 10568646]

Spitzer RL, Williams JB, Kroenke K, Linzer M, deGruy FV 3rd, Hahn SR, … Johnson JG (1994). Utility of a new procedure for diagnosing mental disorders in primary care. The PRIME-MD 1000 study. Jama, 272(22), 1749–1756. [PubMed: 7966923]

Thase ME (2016). Recommendations for screening for depression in adults. Jama, 315(4), 349–350. [PubMed: 26813206]

Treadway MT, & Zald DH (2011). Reconsidering anhedonia in depression: lessons from translational neuroscience. Neuroscience & Biobehavioral Reviews, 35(3), 537–555. [PubMed: 20603146]

Vrany EA, Berntson JM, Khambaty T, & Stewart JC (2015). Depressive symptoms clusters and insulin resistance: race/ethnicity as a moderator in 2005–2010 NHANES data. Annals of Behavioral Medicine, 50(1), 1–11.

Zuithoff NP, Vergouwe Y, King M, Nazareth I, van Wezep MJ, Moons KG, & Geerlings MI (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. [PubMed: 21144018]

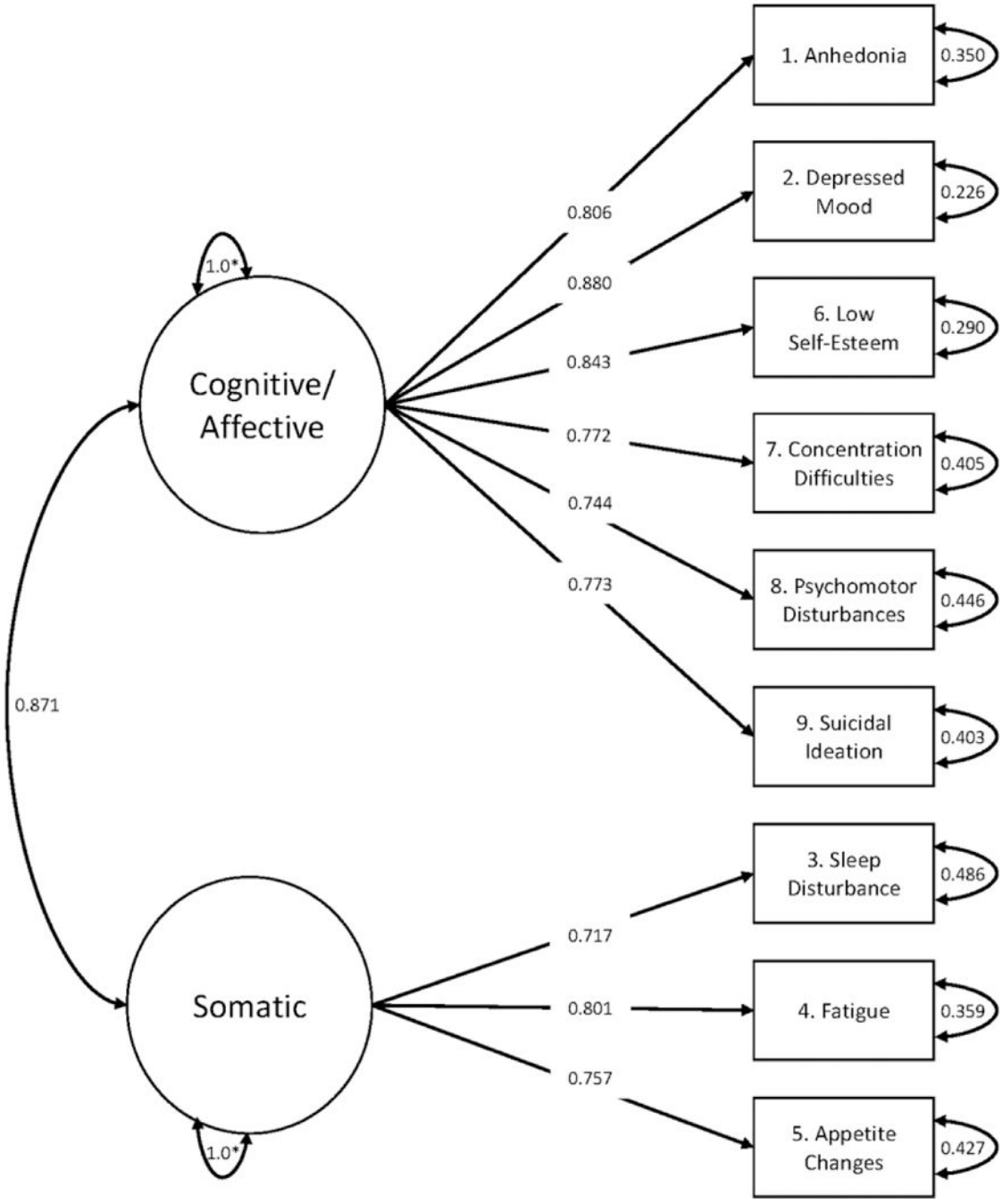

Figure 1:

Two-Factor Measurement Model of the Patient Health Questionnaire-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, as demonstrated by the asterisk). The bidirectional arrow between latent factors represents a correlation

Table 1

Characteristics of NHANES Respondents

	Frequency (%)	PHQ-9 Total Score mean (SD)	PHQ-9 Total Score ≥ 10 (%)	PHQ-9 Cognitive/Affective Score (range: 0–18) mean (SD)	PHQ-9 Somatic Score (range 0–9) mean (SD)	Correlation between Cognitive/Affective and Somatic Scores
Sex (*n* = 31,366)						
Women	15,935 (50.8)	3.72 (4.57)	11.0	1.67 (2.80)	2.05 (2.23)	0.65
Men	15,431 (49.2)	2.67 (3.87)	6.5	1.24 (2.40)	1.43 (1.90)	0.61
Race/Ethnicity (*n* = 30,179)						
Non-Hispanic White	13,455 (42.9)	3.19 (4.19)	8.4	1.38 (2.55)	1.81 (2.09)	0.63
Non-Hispanic Black	6,793 (21.7)	3.21 (4.35)	9.1	1.48 (2.65)	1.73 (2.16)	0.63
Non-Hispanic Asian	1,745 (5.6)	2.23 (3.07)	3.3	1.00 (1.90)	1.22 (1.58)	0.55
Mexican American	5,205 (8.6)	3.15 (4.20)	8.6	1.51 (2.59)	1.64 (2.05)	0.64
Other Hispanic	2,981 (12.2)	3.80 (4.93)	12.2	1.88 (3.11)	1.92 (2.24)	0.69
Education Level (*n* = 31,344)						
Less than 9th grade	3,156 (10.1)	3.77 (5.03)	13.2	1.95 (3.17)	1.82 (2.31)	0.68
9th to 12th grade (no diploma)	5,017 (16.0)	3.92 (4.79)	12.4	1.92 (3.01)	2.00 (2.27)	0.64
High school graduate/GED equivalent	7,410 (23.6)	3.31 (4.32)	9.0	1.51 (2.65)	1.81 (2.14)	0.63
Some college or associate degree	9,101 (29.0)	3.21 (4.22)	8.6	1.41 (2.54)	1.81 (2.12)	0.64
College graduate or above	6,660 (21.2)	2.26 (3.19)	3.8	0.89 (1.88)	1.37 (1.70)	0.59

Note. The cognitive/affective score was computed as the sum of items 1, 2, 6, 7, 8, and 9, and the somatic score was computed as the sum items 3, 4, and 5.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. Low Self-Esteem	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. Suicidal Ideation	Depression	Cognitive/Affective	Cognitive/Affective	Cognitive/Affective	Cognitive/Affective
Chi-Square ($\chi 2$)	1783.11	941.82	1182.37	994.99	910.98
Degrees of Freedom	27	26	26	26	25
***p*-value**	<0.001	<0.001	<0.001	<0.001	<0.001
RMSEA	0.046	0.034	0.038	0.034	0.034
RMSEA 90% Confidence Interval	0.044 – 0.047	0.032–0.035	0.036–0.040	0.033–0.036	0.032–0.036
TLI	0.972	0.985	0.981	0.984	0.985
CFI	0.979	0.989	0.986	0.988	0.989

Note. N = 31,366. For absolute fit indices (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 (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

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

		χ^2	Df	*p*-value	RMSEA	RMSEA 90% CI	TLI	CFI
Sex *n* = 31,366	**Women**	638.62	26	0.000	0.038	0.036–0.041	0.984	0.988
	Men	427.30	26	0.000	0.032	0.029-0.034	0.986	0.990
Race/Ethnicity *n* = 30,179	**Non-Hispanic White**	556.47	26	0.000	0.039	0.036-0.042	0.983	0.988
	Non-Hispanic Black	247.90	26	0.000	0.035	0.031-0.040	0.986	0.990
	Non-Hispanic Asian	79.24	26	0.000	0.034	0.026-0.043	0.982	0.987
	Mexican American	177.50	26	0.000	0.033	0.029-0.038	0.987	0.990
	Other Hispanic	113.44	26	0.000	0.034	0.027-0.040	0.988	0.991
Education Level *n* = 31,344	**Less than 9th grade**	98.54	26	0.000	0.030	0.024-0.036	0.990	0.993
	9th to 12th grade (no diploma)	213.14	26	0.000	0.038	0.033-0.043	0.981	0.986
	High school graduate/GED equivalent	266.67	26	0.000	0.035	0.032-0.039	0.986	0.990
	Some college or associate degree	350.29	26	0.000	0.037	0.034-0.041	0.984	0.989
	College graduate or above	159.61	26	0.000	0.028	0.024-0.032	0.986	0.990

Note. We used the two-factor Model 2 shown in Figure 1. For absolute fit indices (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 (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. CI = confidence interval. TLI = Tucker-Lewis index. CFI = comparative fit index

Table 4

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

		χ^2	df	$\Delta\chi^2$	*p*-value	RMSEA	RMSEA 90% CI	TLI	CFI	ΔCFI	ΔRMSEA
Sex *n* = 31,366	**Configural**	1048.15	52	---	<0.001	0.035	0.033–0.037	0.985	0.989	---	---
	Scalar	877.68	84	117.93	<0.001	0.025	0.023–0.026	0.992	0.991	0.002	−0.010
	Strict	907.43	75	54.46	<0.001	0.027	0.025–0.028	0.991	0.991	0.000	−0.008
Race/Ethnicity *n* = 30,179	**Configural**	1204.45	166	---	<0.001	0.032	0.030–0.034	0.988	0.989	---	---
	Scalar	1629.07	262	411.65	<0.001	0.029	0.028–0.031	0.990	0.986	−0.003	−0.003
	Strict	1419.80	235	287.26	<0.001	0.029	0.027–0.030	0.990	0.987	0.001	−0.003
Education Level *n* = 31,344	**Configural**	1092.39	130	---	<0.001	0.034	0.032–0.036	0.985	0.989	---	---
	Scalar	1507.24	258	700.51	<0.001	0.028	0.026–0.029	0.990	0.986	−0.003	−0.006
	Strict	1088.25	222	486.91	<0.001	0.025	0.023–0.026	0.992	0.990	0.004	−0.009

Note. We used the two-factor Model 2 shown in Figure 1. For absolute fit indices (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 (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 ≥ −0.004 or change in RMSEA ≥ 0.010 signifies that measurement invariance was not established for that step. RMSEA = root mean square error of approximation. CI = confidence interval. TLI = Tucker-Lewis index. CFI = comparative fit index.

Table 5

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

		Sex		Race/Ethnicity					Education Level				
Construct	**PHQ-9 Item**	**Men**	**Women**	**Non-Hispanic White**	**Non-Hispanic Black**	**Non-Hispanic Asian**	**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	0.79	0.82	0.84	0.74	0.70	0.74	0.80	0.75	0.75	0.77	0.83	0.86
	2. Depressed Mood	0.89	0.88	0.89	0.87	0.85	0.86	0.87	0.89	0.85	0.88	0.89	0.89
	6. Low Self-Esteem	0.84	0.84	0.85	0.83	0.77	0.83	0.83	0.84	0.83	0.85	0.85	0.81
	7. Concentration Difficulties	0.75	0.79	0.77	0.78	0.71	0.75	0.83	0.80	0.73	0.79	0.77	0.75
	8. Psychomotor Disturbances	0.74	0.74	0.72	0.78	0.65	0.76	0.82	0.81	0.77	0.73	0.73	0.66
	9. Suicidal Ideation	0.79	0.76	0.77	0.81	0.68	0.74	0.79	0.70	0.73	0.78	0.80	0.80
Somatic	3. Sleep Disturbance	0.74	0.69	0.70	0.78	0.67	0.75	0.77	0.76	0.74	0.72	0.73	0.66
	4. Fatigue	0.79	0.80	0.81	0.79	0.74	0.81	0.79	0.80	0.79	0.81	0.81	0.80
	5. Appetite Changes	0.73	0.77	0.76	0.76	0.72	0.72	0.76	0.74	0.75	0.75	0.77	0.73