



Published in final edited form as:

*J Affect Disord.* 2019 July 01; 254: 59–68. doi:10.1016/j.jad.2019.05.017.

## The Measurement Invariance of the Patient Health Questionnaire-9 for American Indian Adults

Melissa L. Harry, PhD, MSW<sup>1,1</sup>, Stephen C. Waring, DVM, PhD<sup>1</sup>

<sup>1</sup>Essentia Health, Essentia Institute of Rural Health, 502 East Second Street, Duluth, MN 55805

### Abstract

**Background:** American Indian people have high suicide rates. However, little epidemiological data is available on depression prevalence, a suicide risk factor, in this population. Some research suggests that depression scales may perform differently for American Indian people. However, the Patient Health Questionnaire-9 (PHQ-9), a depression scale widely-used in clinical practice, had not been assessed for cross-cultural measurement invariance with American Indian people.

**Methods:** In this retrospective study of existing electronic health record (EHR) data in an upper Midwestern healthcare system, we assessed the measurement invariance of the standard one-factor PHQ-9 and five previously identified two-factor models for 4,443 American Indian and 4,443 Caucasian American adults (age  $\geq 18$ ) with a PHQ-9 in the EHR from 12/1/2005 to 12/31/2017. We also conducted subgroup analyses with adults ages  $\geq 65$ .

**Results:** Models showed good fits (e.g., CFI  $> 0.99$ , RMSEA  $< 0.05$ ) and internal consistency reliability (ordinal alpha  $> 0.80$ ). All models displayed measurement invariance between racial groups. Factor correlation was high for two-factor models, providing support for the one-factor model. American Indian adults had significantly higher odds of PHQ-9 total scores  $\geq 10$  and  $\geq 15$  than Caucasian American adults.

**Limitations:** Data came from a single healthcare system.

**Conclusions:** The PHQ-9 exhibited cross-cultural measurement invariance between American Indian and Caucasian American adults, supporting the PHQ-9 as a depression screening tool in this clinical care population. American Indian adults also had higher levels of depression than Caucasian Americans. Future research could confirm the generalizability of our findings to other American Indian populations.

<sup>1</sup>Melissa L. Harry, PhD, MSW, Essentia Institute of Rural Health, 502 East Second Street | 6AV-2, Duluth, MN 55805, Phone: 218-576-0714. Fax: 218-727-8159., Melissa.Harry@EssentiaHealth.org.

Author Statement

Contributors: MLH conceptualized the study, drafted the manuscript, and conducted the data analyses. SCW contributed to the conceptualization of the topic and development of the manuscript. All authors approved the final article.

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

**Declarations of interest:** none.

## Keywords

Adults; American Indian; cross-cultural; measurement equivalence; measurement invariance; Native American; Patient Health Questionnaire-9

---

## Introduction

Depression is the foremost cause of disability globally and plays a major role in death by suicide (World Health Organization, 2017). American Indian (the Indigenous population of the lower 48 U.S. states, also identifying as Native American) and Alaskan Native people have higher suicide rates compared to other populations (Suicide Prevention Resource Center, 2013). Recent research has found co-factors significantly associated with suicide for this group when compared to Caucasian (White) people, such as greater odds of positive alcohol toxicology in decedents, knowing others who have committed suicide, and living outside of a metropolitan area (Leavitt et al., 2018). Although depression is also a risk factor for suicide, little epidemiological evidence is available from the published literature or within national datasets regarding depression prevalence for American Indian people (Garrett et al., 2015). A recent systematic review and meta-analysis by Kisely and colleagues (2017), which compared depression prevalence between non-Indigenous and Indigenous peoples in the Americas (U.S., Canada, and Latin America), found lifetime depression prevalence rates were lower among Indigenous peoples and that depression diagnoses over a one-year period did not differ significantly. These included studies conducted with U.S. populations, namely Northern Plains and Southwestern tribes (Beals et al., 2005a; 2005b), pregnant American Indian women (Melville et al., 2010), adolescents (Costello et al., 1997), and convicted female DWI offenders (C'de Baca et al., 2004). However, other cited studies showed that American Indian people had significantly higher depression prevalence, specifically in samples representative of the U.S. population (Huang B. et al., 2006; Smith et al., 2006) and adults with diabetes (Li et al., 2008). All cited studies had limitations, foremost being employing depression measures not validated for use with Indigenous people (Kisely et al., 2017). Scales for assessing mental health symptomology have been predominately developed and tested in U.S. populations of European descent (Crockett et al., 2005). Yet tests of scale reliability and validity should also assess the cross-cultural invariance (or equivalence) of a scale between diverse groups (Cleary, 2013), such as through multiple group measurement invariance testing (Tran et al., 2017). Establishing cross-cultural measurement invariance, particularly what is referred to as scalar or strong invariance, is necessary for accurately comparing scale mean scores between different groups (Brown, 2015).

American Indian people are a heterogenous cultural group with 573 federally-recognized tribes (National Conference of State Legislatures, 2018). Previous research suggests that some groups of American Indian people may have culturally distinct expressions of depression, such as related to loneliness (Armenia et al., 2014; O'Neil, 2004). Intergenerational experiences of historical trauma from colonization, genocide, forced assimilation and relocation, and the concomitant loss of culture, family, land, language, and spirituality for American Indian, Alaskan Native, and Canadian First Nations people have

also been described as influencing depression symptomology in this population (e.g., Brave Heart and DeBruyn, 1998; Brown-Rice, 2013; Tucker et al., 2016; Whitbeck et al., 2002; Whitbeck et al., 2009). Specifically, Brave Heart and DeBruyn (1998) connected unresolved grief and trauma with high depression rates for American Indian people. More recent research suggests that the rumination and repetitive cognitive action of thinking about historical trauma (historical loss thinking) may negatively affect psychological wellbeing for American Indian people (Tucker et al., 2016).

The Patient Health Questionnaire-9 (PHQ-9) is a depression scale widely used in both primary care and other healthcare settings as a Healthcare Effectiveness Data and Information Set (HEDIS) measure (Kroenke et al., 2001; National Committee for Quality Assurance, 2018). However, we could not identify any studies that reported on the cross-cultural measurement invariance of the PHQ-9 with American Indian people. Research assessing the cross-cultural measurement invariance of the PHQ-9 presents mixed results for Kroenke et al.'s (2001) original one-factor PHQ-9 model. Galenkamp et al. (2017) found the one-factor PHQ-9 invariant across African Surinamese, Dutch, Ghanaian, Moroccan, South-Asian Surinamese, and Turkish adults in the Netherlands. The one-factor model was also cross-culturally invariant between Surinam Dutch and Dutch women and partially invariant for Surinam Dutch men (Baas et al., 2011). Here in the U.S., research supported a one-factor model while also reporting differential item functioning for some PHQ-9 items between non-Hispanic White, African American, Chinese American, and Latino primary care patients (Huang, F.Y. et al., 2006). Patel (2017) found a two-factor model with somatic and cognitive/affective factors invariant between non-Hispanic White, non-Hispanic Black, Mexican American, and other Hispanic groups and between genders. Keum et al. (2018) tested four separate PHQ-9 models between Asian American, African American, Latino/a American, and Caucasian American college students. These models included the standard one-factor model and three two-factor somatic and cognitive-affective models previously identified by Krause and colleagues (2008; 2010) and Richardson and Richards (2008). Keum et al. (2018) found that the one-factor model was cross-culturally measurement invariant and also the best fit for all four racial groups due to the better fitting two-factor models having high factor correlations ( $> 0.85$ ) within groups. Of note, Keum et al. only tested the one-factor model and the best fitting two-factor model for measurement invariance between groups. Furthermore, while the one-factor model was invariant between English or Spanish-speaking Latina women (Merz et al., 2011), a two-factor, seven-item PHQ-9 fit Latina college students best, and displayed cross-cultural measurement invariance with Caucasian American female students (Granillo, 2012). Exploratory factor analysis also produced a two-factor model for a community sample of African American, Hispanic/Latino/a, and African respondents, although the author did not assess cross-cultural measurement invariance (Morehead, 2012).

Other research on the factor structure of the PHQ-9 based on diagnosis, sex, or other grouping supports a variety of two-factor models with underlying somatic and non-somatic (cognitive/affective) latent factors. This includes research with patients with a range of psychiatric diagnoses (Beard et al., 2016), German men and women diagnosed with major depression (Petersen et al., 2014), British patients with persistent major depressive disorder (Guo et al., 2017), cancer patients (Hinz et al., 2016), palliative care patients (Chilcot et al.,

2013), patients with stable coronary heart disease (de Jonge et al., 2007), Army National Guard soldiers at risk for depression (Elhai et al., 2012), and people with spinal cord injuries (Krause et al., 2008; Krause et al., 2010; Richardson and Richards, 2008). However, one study with people with spinal cord injuries supported both one- and two-factor models, although the two-factor model was not congruent between male and female genders (Kalpakjian et al., 2009). Another study found both one- and two-factor models to be measurement invariant both over time and across several demographic characteristics (e.g., sex, age, marital status, employment, education) for primary care patients in Spain (González-Blanch et al., 2018). However, the authors reported high latent factor correlation (0.86) for the two-factor model, lending support to the one-factor model. Other studies exist that support the one-factor model by Kroenke et al. (2001). These include research conducted with patients with multiple sclerosis or spinal cord injury (Chung et al., 2015), outpatient substance abusers (Dum et al., 2008), and Chinese adolescents and adults living in Hong Kong (Yu et al., 2012). Research in primary care and obstetrics-gynecology on the PHQ-8, which drops item nine on suicidality, also supported a one-factor model (Kroenke et al., 2010). Together with the mixed results of cross-cultural measurement invariance testing, these disparate findings suggest that the latent factor structure of the PHQ-9 differs for some groups.

In the present study, we aimed to fill a gap in the literature by evaluating the cross-cultural measurement invariance of the PHQ-9 depression scale between American Indian and Caucasian American (non-Hispanic White) adults. Figure 1 presents the six PHQ-9 factor models assessed in this study. These include the standard one-factor model (Kroenke et al., 2001) and five two-factor models previously identified with other groups (de Jonge et al., 2007; Granillo, 2012; Krause et al., 2008; Krause et al., 2010; Richardson and Richards, 2008). Table S1 in the supplementary materials presents these and other two-factor models identified in the literature (note: this is not an exhaustive list; other studies could exist).

## Methods

### Study Population

The study population included 4,443 American Indian adults (18 years of age and over) and a random sample of 4,443 Caucasian American adults seeking care from a large, integrated, upper Midwestern healthcare system with locations in northern Minnesota, North Dakota, and Wisconsin. Inclusion criteria for either group included having at least one PHQ-9 total score in the healthcare system's electronic health record (EHR) from 12/1/2005 to 12/31/2017. The PHQ-9 was inconsistently utilized in the American Indian population; prior to this date range, PHQ-9 scores appeared only sporadically in the EHR.

### Instrument

The PHQ-9 contains a subset of nine depression-related questions originally included in the self-reported Patient Health Questionnaire (PHQ) (Kroenke et al., 2001), which was designed for use in primary care settings (Spitzer et al., 1999). The PHQ-9 was previously employed in diagnosing major depressive disorder or other depression with the DSM-IV

(Kroenke et al., 2001), as well as major depressive disorder with the DSM-5 (American Psychological Association, 2013).

In the PHQ-9, individuals are asked to self-rate how often they experienced nine depression-related “problems” in the past two weeks on a four-point scale: “not at all = 0,” “several days = 1,” “more than half the days = 2,” and “nearly every day = 3” (Kroenke et al., 2001). The PHQ-9 total score equals the sum of the 9 item scores and ranges from 0 to 27. Scores on the PHQ-9 represent varied levels of depression symptomology: 0-4 = minimal depression; 5-9 = mild depression; 10-14 = moderate depression; 15-19 = moderately severe depression; and 20-27 = severe depression (Kroenke et al., 2001). Kroenke et al. (2001) reported good levels of internal reliability for the PHQ-9 based on two samples (Cronbach’s alpha of 0.86 and 0.89). A recent pooled meta-analysis showed acceptable sensitivity (81.3%) and specificity (85.3%) across 16 studies assessing the linear (summed score) PHQ-9 for total scores  $\geq 10$ , with a ROC of 97.5 (Mitchell et al., 2016).

## Procedures

In this retrospective study, we analyzed a dataset composed of existing healthcare system EHR data. We excluded individuals who opted out of research at the healthcare system. Because we conducted a retrospective, EHR data-only study, we requested and received a waiver of informed consent from the healthcare system Institutional Review Board that approved this study.

The sample in this study included patients with self-reported EHR race data of either American Indian or Caucasian American. Hispanic ethnicity documentation includes “Yes”, “No”, or “Unknown”. Race is documented in the EHR as either White, American Indian, Black, Asian, Hispanic, or unknown. Patients can identify with multiple races; however, Caucasian Americans in our sample only identified as White and non-Hispanic. Other descriptive data elements extracted from the EHR included: sex; age; health insurance type; clinic Rural-Urban Commuting Area (RUCA) codes (United States Department of Agriculture, 2016); and depression diagnosis codes (ICD-9: 296.0-296.9, 300.4, 309.0, 309.1, 311; ICD-10: F31, F31.0-F31.9, F32, F32.0-F34.9, F39) (National Center for Health Statistics, 2010; World Health Organization, 2012). We also generated two binary (Yes/No) indicator variables for PHQ-9 total scores 10-27 (moderate or greater depression) and 15-27 (moderately-severe or greater depression, the diagnostic cut off for major depressive disorder). We used these binary indicators as dependent variables in multivariate logistic regression controlling for American Indian or Caucasian American race, mean-centered age, sex, having a major depression diagnosis in the past year, and clinic RUCA code. Due to the potential for differences in PHQ-9 response based on age, we also conducted subgroup analyses for those ages 65 and over in our sample. In this study, we report on only the sample’s first PHQ-9 scores in the EHR during the eligibility period (12/1/2005-12/31/2017).

## Data Analysis

We conducted descriptive statistics, as well as bivariate analyses (chi-square [ $\chi^2$ ] cross tabulations for nominal data and Mann Whitney U for nonparametric skewed continuous

data) and multivariate logistic regression to assess differences between groups in IBM® SPSS® Statistics Version 23.0 (IBM Corp, 2015). We performed two-tailed analyses with a significance level of .05. We also utilized R version 3.4.4 (R Core team, 2018) for polychoric correlations, internal consistency reliability, confirmatory factor analysis (CFA), and multigroup measurement invariance testing.

Due to the ordinal nature of PHQ-9 items, we employed weighted least squares means and variance adjusted (WLSMV) estimation for categorical data in both CFA and measurement invariance testing (Muthén and Muthén, 1998-2015). Brown (2015) reported that Mplus was the best software available for CFA modeling with categorical data. R's "lavaan" package 0.6-2 allows users to "mimic" Mplus in WLSMV estimation (Rosseel, 2018), which we utilized in our confirmatory factor and measurement invariance analyses. Regarding missing data, PHQ-9 total score data were complete. However, some PHQ-9 item scores were marked as "NA" at random in some EHR charts, representing a lack of item data. We assessed cases with "NA" for missing data patterns, recoded "NA" as blank for analysis, then dropped the associated cases from the confirmatory factor and measurement invariance analyses using listwise deletion as required with WLSMV estimation in R.

**Internal Consistency Reliability.**—Due to the categorical nature of PHQ-9 items, we present ordinal alpha ( $\alpha$ ) for internal consistency reliability analysis of all PHQ-9 factors tested. While Cronbach's  $\alpha$  (Cronbach, 1951) is more widely reported, it is based on Pearson covariance matrices for continuous data (Zumbo et al., 2007). Ordinal  $\alpha$  provides more accurate estimates for ordinal-level items by basing  $\alpha$  on polychoric correlation matrices (Zumbo et al., 2007). We also present item-rest correlations, or the correlation between an item and other items in a scale or factor excluding that item, based on ordinal  $\alpha$  (Revelle, 2018). In this study, we calculated ordinal  $\alpha$  and item-rest correlations using R's "psych" package 1.8.10 (Revelle, 2018).

**Confirmatory Factor Analysis.**—We evaluated the one- and two-factor PHQ-9 models presented in Figure 1 for goodness of fit using CFA separately in both racial groups. We selected the two-factor models based on goodness of fit shown in prior testing with other populations (e.g., de Jonge et al., 2007; Granillo, 2012; Krause et al., 2008; Krause et al., 2010; Kroenke et al., 2001; Richardson and Richards, 2008). Note that Richardson and Richards (2008) originally identified the two-factor model 2A for the 1-year post-spinal cord injury group. Also, the two-factor model 2E has only seven items, as Granillo (2012) dropped items 7 and 8 when they cross-loaded on more than one factor in exploratory factor analysis.

Regarding assessing CFA model fit, while a nonsignificant chi-square ( $\chi^2$ ) is preferred,  $\chi^2$  is sensitive to sample size. As such, researchers commonly use other goodness of fit measures (Brown, 2015; Tran et al., 2017). These include comparative fit index (CFI) and Tucker-Lewis Index (TLI) > 0.90 and preferably near 1.00, root mean square error of approximation (RMSEA) < 0.05 or at most < 0.08, and standardized root mean square residual (SRMR) < 0.08 or at most < 0.10 (Steenkamp and Baumgartner, 1998; Vandenberg and Lance, 2000). We also report factor correlations for two-factor models, where correlations > 0.85 suggest multicollinearity (Brown, 2015).

**Measurement Invariance Testing.**—We assessed the cross-cultural measurement invariance of the one- and two-factor PHQ-9 models that showed a good fit in CFA by testing three progressively constrained, or nested, models (Brown, 2015). We did so by first adding a constraint for equal latent factor structures (or patterns) between groups (configural invariance); otherwise, we freely estimated factor models between groups. A finding of configural invariance means the latent factor structure (i.e., number of latent factors and number of items composing each latent factor) is the same between groups. Next, we added a constraint for equal factor loadings to that of equal latent factor structures between groups (metric or weak invariance). When a scale exhibits metric invariance, this suggests that the meaning of latent factors is the same between groups. Finally, due to the ordinal nature of PHQ-9 items, we added a constraint for equal item thresholds along with the constraints for equal factor loadings and latent factor structures between groups (scalar or strong invariance). Findings of scalar invariance allow for comparing mean scores between different groups (Brown, 2015). Metric and scalar invariance are necessary to support multi-group measurement invariance (Muthén & Asparouhov, 2002). Due to the restrictive nature in real world practice of adding a fourth common constraint to these nested models, equal item residuals (error terms) between groups (strict invariance) (Brown, 2015), we did not test this constraint in our study. Findings of configural, metric, and scalar invariance are sufficient for determining if a factor model is invariant between groups (Brown, 2015).

Measures of goodness of fit between nested measurement invariance models employed in this study include the scaled  $\chi^2$  model difference test ( $\chi^2_{\text{diff}}$ ) (Satorra and Bentler, 2001), as differences in  $\chi^2$  are not distributed as  $\chi^2$  in WLSMV estimation (Muthén and Muthén, 1998-2015). Statistically nonsignificant ( $p > 0.05$ )  $\chi^2_{\text{diff}}$  between nested models supports measurement invariance. However, like other  $\chi^2$  tests,  $\chi^2_{\text{diff}}$  is sensitive to sample size; even small differences between models may be significant due to large samples (Schermelleh-Engel and Moosbrugger, 2003). Consequently, we followed Cheung and Rensvold (2002) where a change of  $-0.01$  in CFI (CFI) represents a difference between nested models. RMSEA, CFI, TLI, and SRMR values in the same ranges as in CFA are also preferred. Lastly, although weighted root mean square residual (WRMR) is experimental and not a replacement for SRMR (Muthén, 2016), Yu (2002) suggested that WRMR around 1.00 shows a good fit.

## Results

As shown in Table 1, females composed the largest share of the sample for both groups. Age ranged from 18 to 96 for the American Indian group and 18 to 98 for the Caucasian American group. Of those ages 65 and over, 924 were Caucasian American and 365 were American Indian ( $p < 0.001$ ). Approximately 2% of American Indian adults also identified as Hispanic. Significantly more American Indian adults received care in clinics within rural areas and fewer in micropolitan and urban clinics than Caucasian American adults ( $p < 0.001$ ). Most members of the sample also had some form of health insurance with small, yet significant differences between racial groups. Regarding depression diagnoses, 41% of American Indian and 38% of Caucasian American adults had at least one depression diagnosis in the year prior to and including the index date (the date of the first PHQ-9 in the EHR during the eligibility period). Some had multiple depression diagnoses. Significantly

more American Indian adults (12%) had a major depressive disorder diagnosis as of the index date compared to Caucasian Americans (10%):  $\chi^2 = 11.48$  ( $df = 1$ ),  $p = 0.001$ . American Indian adults (14%) also had higher rates of being diagnosed with a major depressive disorder over the past year compared to Caucasian American adults (11%):  $\chi^2 = 17.73$  ( $df = 1$ ),  $p < 0.001$ . American Indian adults ( $Mdn = 11$ ) did have significantly higher median PHQ-9 total scores compared to the Caucasian American group ( $Mdn = 7$ ) ( $p < 0.001$ ). American Indian adults (3%) also had significantly more bipolar disorder diagnoses in the past year than Caucasian American adults (2%):  $\chi^2 = 10.31$  ( $df = 1$ ),  $p = 0.001$ . Results for adults ages 65 and over ( $n = 1,289$ ) are presented in Table S2 in the supplementary materials.

Polychoric correlations were high between PHQ-9 items 1 and 2 and items 6 and 2 for both American Indian (0.80) and Caucasian American (0.87 and 0.86, respectively) adults ages 18 and over (Supplementary materials, Table S3). Item 9 had the lowest correlations over all, with the lowest correlation between items 9 and 4 in both groups (0.36 for American Indians and 0.42 for Caucasian Americans). Other item correlations ranged from 0.38 to 0.74 for American Indian adults and 0.45 and 0.78 for Caucasian American adults. Similar results were also found for adults ages 65 and over (Supplementary materials, Table S3).

### Internal Consistency Reliability

Table 2 shows PHQ-9 item-rest correlations based on ordinal  $\alpha$  for the standard one-factor model by racial group for adults ages 18 and over. Item-rest correlations for items 1-8 were above the preferred 0.70 for both American Indian and Caucasian American adults, while item 9 was 0.63 and 0.66 for each group, respectively. Item-rest correlations were quite similar between groups, suggesting cross-cultural equivalency (Tran et al., 2015). Median scores for each item, along with 25<sup>th</sup> and 75<sup>th</sup> percentiles, are also presented in Table 2. Significant differences were seen between both racial groups on all items. American Indian adults did have higher median scores on items 1, 2, 5, 6, and 7. Table S4 in the supplementary materials illustrates item internal consistency reliability and median scores for the subgroup ages 65 and over; of note, item 9 was the only item without a significant difference between racial groups.

Table 3 presents means, and standard errors, and ordinal  $\alpha$  for all tested latent factors by racial group for those ages 18 and over. The standard one-factor model had the highest ordinal  $\alpha$  for both groups (American Indian  $\alpha = 0.94$ , Caucasian American  $\alpha = 0.95$ ), internal consistency reliability estimates that were slightly better than the levels of Cronbach's  $\alpha$  (0.89 and 0.86) reported for two samples by Kroenke et al. (2001). All other factors showed good ( $> 0.80$ ) to excellent ( $> 0.90$ ) levels of ordinal  $\alpha$ . Compared to Caucasian American adults, American Indian adults had higher mean scores on each factor, comparable standard errors, and slightly lower ordinal  $\alpha$ . In the subgroup analysis of those ages 65 and over (Supplementary material, Table S5), mean factor scores were still higher for American Indian older adults. Ordinal  $\alpha$ , still ranging from good ( $> 0.80$ ) to excellent ( $> 0.90$ ), was also higher for this group compared to Caucasian Americans for some factors.



## Confirmatory Factor Analysis

Based on CFA goodness of fit indexes, all tested models had good fits for both racial groups in those ages 18 and over (Table 4) and the subgroup of adults ages 65 and over (Supplementary materials, Table S6). For the 18 and over group, the two-factor model 2E with seven items and Affective and Somatic factors presented the best fit, followed by the two-factor model 2B with nine items and Non-Somatic and Somatic factors, although the standard one-factor model showed a good fit as well. Similar results were seen for the subgroup ages 65 and over; however, while all models showed excellent fits, after model 2E, model 2B had the best fit for American Indian adults, and model 2A for Caucasian American adults in this subgroup.

Regardless of factor model or group in those 18 and over, standardized item loadings were all  $> 0.700$  for items 1 through 7, with item 8 having loadings between 0.600 and 0.700, and item 9 having loadings between 0.400 and 0.500 (Supplementary materials, Table S7). Correlations between the two-factor models were all high ( $> 0.85$ ), ranging from 0.920 to 0.985 for Caucasian American adults and 0.904 to 0.943 for American Indian adults (Supplementary materials, Table S7). Similar good fits were seen for those 65 and over, along with high factor correlations ( $> 0.85$ ) and some slightly lower standardized item loadings for all two-factor models (Supplementary materials, Table S7).

## Measurement Invariance Testing

All tested models displayed configural, metric, and scalar invariance between American Indian and Caucasian American adults ages 18 and over (Table 5). While no model had a nonsignificant  $\chi^2_{diff}$ , changes in  $\chi^2$  between nested models were small, suggesting sample size affected the significance of  $\chi^2_{diff}$ . Furthermore, CFI did not exceed  $-0.01$  and all CFI values were  $> 0.99$ . Model 2E fit best, followed by 2B, 2A, 2D, 2C, and the standard one-factor model. All tested models were also measurement invariant in subgroup analyses of those ages 65 and over, with 2E showing the best fit, followed by 2B, 2A, 2C, 2D, and the standard one-factor model (Supplementary materials, Table S8). However, given the high levels ( $> 0.85$ ) of factor correlations for two-factor models, the one-factor model appears to fit the data best for this sample of American Indian adults.

## Multivariate Logistic Regression

Multivariate logistic regression models showed that American Indian adults 18 and over had significantly higher odds of having a PHQ-9 total score  $\geq 10$  (OR = 1.53, 95% CI = 1.40-1.68,  $p < .001$ ) and  $\geq 15$  (OR = 1.64, 95% CI = 1.48-1.82,  $p < .001$ ) than Caucasian American adults (see Table 1). When looking at the subgroup ages 65 and over, American Indian older adults also had significantly higher odds of PHQ-9 total scores  $\geq 10$  (OR = 1.41, 95% CI = 1.07-1.88,  $p = .016$ ) and  $\geq 15$  (OR = 1.54, 95% CI = 1.04-2.27,  $p = .030$ ) (Supplementary materials, Table S2).

## Discussion

Culturally sensitive screening tools are necessary for assessing depression across diverse populations. In this study, we addressed a gap in the literature on the cross-cultural

measurement invariance of the PHQ-9 depression scale with American Indian adults. We did so by testing the standard one-factor model and five previously identified two-factor models, finding that the PHQ-9 performed similarly between American Indian and Caucasian American adults seen in an upper Midwestern healthcare system. All tested models displayed good to excellent levels of internal consistency reliability based on ordinal  $\alpha$ . The models also had good CFA fits and were all cross-culturally measurement invariant for the full adult sample ages 18 and over, as well as for adults ages 65 and over. This included the standard one-factor model typically used in clinical practice to calculate the total PHQ-9 score (Kroenke et al., 2001). This suggests that PHQ-9 total scores can be meaningfully compared between American Indian and Caucasian American adults in populations like our study. As such, we also found significant differences related to PHQ-9 total scores between American Indian and Caucasian American adults in multivariate logistic regression models, with American Indian adults having significantly greater odds of higher total scores regardless of age group. This suggests an opportunity for enhancing depression clinical care for this population.

While all models were cross-culturally invariant between racial groups, findings that two-factor models fit the sample better than the standard one-factor model support previous research on two-factor models (e.g., Beard et al., 2016; Chilcot et al., 2013; Elhai et al., 2012; Krause et al., 2008; Krause et al., 2010; Morehead, 2012; Petersen et al., 2014; Richardson and Richards, 2008). This suggests that the PHQ-9 may be assessing more than one latent factor, such as those related to somatic and affective depression symptomology. However, little difference was seen between fit indexes for competing two-factor models, supporting their equivalence for this sample. Also, the high factor correlations seen in all tested two-factor models suggest that factors could be combined (Brown, 2015), as they are in the standard one-factor PHQ-9. In addition to multicollinearity, high factor correlations may also suggest the presence of an over-arching latent bifactor (Holzinger & Swineford, 1937). Future research could assess the usefulness of a bifactor PHQ-9 model.

Nevertheless, the one-factor model used in clinical care for calculating the PHQ-9 total score was cross-culturally measurement invariant and showed a good fit between groups. This finding supports the use of the standard summed score one-factor PHQ-9 in assessing levels of depression for American Indian adults similar to our sample. Our findings of cross-cultural measurement invariance of the one-factor PHQ-9 model also allows us to make meaningful comparisons in PHQ-9 total scores for the sample in our study. We found that American Indian adults had significantly higher median and mean PHQ-9 total scores and were more likely to have a score  $\geq 10$  or  $\geq 15$ , signifying a higher level of depression symptomology than Caucasian American adults. Given the scarcity of epidemiological data on depression prevalence and incidence in the American Indian population (Garrett et al., 2015; Kisely et al., 2017), these findings are important to note as they may represent an opportunity to positively impact patient care by identifying American Indian adults displaying symptoms of depression and linking them with culturally competent treatment resources. Furthermore, strengths-focused perspectives, such as emphasizing positive mental health (Kading et al., 2015) and resilience (Goodkind et al., 2015) with American Indian people, may provide a counterpoint to pathology-focused assessments of depression symptomology.

Lastly, item 9 loaded lower than other PHQ-9 items regardless of the factor model or racial group. Recent research has shown the predictive utility of item 9 in assessing suicidality (Coleman et al., 2018; Simon et al., 2013; Simon et al., 2016). Future research could consider developing additional questions related to suicidality for the PHQ-9 and creating a separate factor that assesses suicide risk. Dube and colleagues (2010) did develop the P4, a four-item suicide screener that assesses: whether a patient made a past suicide attempt; has a plan; the probability they will complete suicide; as well as preventative factors. The P4 is triggered by a positive response to the PHQ-9 item 9 (Dube et al., 2010).

### Limitations

Data for this study came from a single healthcare system in the upper Midwest. We also could not conduct convergent validity analyses due to a lack of other depression scales employed in the healthcare system. However, other research assessing the convergent validity of the PHQ-9 with American Indian populations is available (e.g., Heck, 2018), a model future research could follow. Furthermore, due to journal word limits, low percentages of other minority racial and ethnic groups seen in the healthcare system, and previous research reporting on the measurement invariance of the PHQ-9 with other racial and ethnic groups in the U.S. (e.g., Granillo, 2012; Keum et al., 2018; Merz et al., 2011; Morehead, 2012), we limited our analyses to comparing American Indian adults with the dominant Caucasian American group. Moreover, while the PHQ-9 total score measures depression symptom severity (Kroenke et al., 2001), item 9 is predictive of suicidality (Coleman et al., 2018; Simon et al., 2013; Simon et al., 2016). However, only a small number of individuals in this study had evidence of attempted or completed suicide in the EHR (based on ICD-9 or ICD-10 diagnosis codes). Consequently, we were unable to assess and compare the sensitivity and specificity of item 9 as a screening measure for attempted or completed suicide for American Indian adults. Future research should explore the predictive utility of item 9 in assessing suicide likelihood.

### Conclusion

The PHQ-9 showed a good fit and cross-cultural equivalency in all tested models between American Indian and Caucasian American adults in this upper Midwestern population. The standard one-factor model used by clinicians in calculating PHQ-9 depression scores, assessing depression severity, and diagnosing major depressive disorder appears acceptable for use with American Indian adults. Our results also showed that American Indian adults had significantly higher odds of PHQ-9 total scores signifying moderate to moderately-severe depression compared to Caucasian American adults. Future research with larger samples could examine the predictive utility of question 9 in assessing suicide risk, as well as determine whether these findings generalize to other populations of American Indian people.

### Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements:

The authors thank Dr. Gregory Simon with Kaiser Permanente Washington and Dr. Brian Ahmedani with Henry Ford Health System for scientific review, and Essentia Institute of Rural Health Research Informatics Analysts Austin Land and Nicholas Cameron and Research Informatics Supervisor Paul Hitz for assistance with data collection.

Funding statement: This work was partially supported by Cooperative Agreement with the National Institute of Mental Health [grant number U19 MH092201].

## References

- American Psychological Association, 2013 Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychological Association, Washington, D.C.
- Armenia BE, Sittner Hartshorn KJ, Whitbeck LB, Crawford DM, Hoyt DR, 2014 A longitudinal examination of the measurement properties and predictive utility of the Center for Epidemiologic Studies Depression Scale among North American indigenous adolescents. *Psychol Assess* 26(4), 1347–1355. doi:10.1037/a0037608 [PubMed: 25181394]
- Baas KD, Cramer AOJ, Koeter MWJ, van de Lisdonk EH, van Weert HC, Schene AH, 2011 Measurement invariance with respect to ethnicity of the Patient Health Questionnaire-9 (PHQ-9). *J Affect Disord* 129, 229–235. [PubMed: 20888647]
- Beads J, Manson SM, Whitesell NR, Mitchell CM, Novins DK, Simpson S, Spicer P, 2005a Prevalence of major depressive episode in two American Indian reservation populations: unexpected findings with a structured interview. *Am. J. Psychiatry* 162, 1713e1722. [PubMed: 16135632]
- Beads J, Novins DK, Whitesell WR, Spicer P, Mitchell CM, Spiro MM, 2005b Prevalence of mental disorders and utilization of mental health services in two American Indian reservation populations: Mental health disparities in a national context. *Am J Psychiatry*. 162(9), 1723–1732. doi:10.1176/appi.ajp.162.9.1723 [PubMed: 16135633]
- Beard C, Hsu KJ, Rifkin LS, Busch AB, Bjorgvinsson T, 2016 Validation of the PHQ-9 in a psychiatric sample. *J Affect Disord* 193, 267–273. [PubMed: 26774513]
- Brave Heart MYH, DeBruyn LM, 1998 The American Indian Holocaust: Healing historical unresolved grief. *Am Indian Alsk Native Ment Health Res* 8(2), 56–78. [PubMed: 9842066]
- Brown TA, 2015 Confirmatory Factor Analysis for Applied Research, second ed Guilford Press, New York.
- Brown-Rice K, 2013 Examining the theory of historical trauma among Native Americans. *The Professional Counselor*. 3(3), 117–130.
- C' de Baca I, Lapham SC, Skipper BJ, Hunt WC, 2004 Psychiatric disorders of convicted DWI offenders: A comparison among Hispanics, American Indians and non-Hispanic Whites. *J Stud Alcohol*. 65, 419–427. [PubMed: 15376815]
- Cheung GW, Rensvold RB, 2002 Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*. 9:233–255. doi: 10.1207/S15328007SEM0902\_5
- Chilcot J, Rayner L, Lee W, Price A, Goodwin L, Monroe B, Sykes N, Hansford P, Hotopf M, 2013 The factor structure of the PHQ-9 in palliative care. *J Psychosom Res* 75(1), 60–64. doi:10.1016/j.jpsychores.2012.12.012 [PubMed: 23751240]
- Chung H, Kim J, Askew RL, Jones SMW, Cook KF, Amtmann D, 2015 Assessing measurement invariance of three depression scales between neurologic samples and community samples. *Qual Life Res* 24, 1829–1834. doi:10.1007/s11136-015-0927-5 [PubMed: 25627670]
- Cleary LM, 2013 Cross-Cultural Research with Integrity: Collected Wisdom from Researchers in Social Settings. Palgrave Macmillan, New York.
- Coleman KJ, Johnson E, Ahmedani BK, Beck A, Rossom RC, Shortreed SM, Simon GE, 2018 Predicting suicide attempts for racial and ethnic groups of patients during routine clinical care. *Suicide Life Threat Behav* 10.1111/sltb.12454
- Costello EJ, Farmer EMZ, Angold A, Burns BJ, Erkanli A, 1997 Psychiatric disorders among American Indian and White youth in Appalachia: The Great Smoky Mountains Study. *Am J Public Health*. 87(5):827–32. [PubMed: 9184514]

- Crockett LJ, Randall BA, Shen YS, Russell ST, Driscoll AK, 2005 Measurement equivalence of the Center for Epidemiological Studies Depression Scale for Latino and Anglo Adolescents: A national study. *J Consult Clin Psychol* 73(1), 47–58. doi: 10.1037/0022-006X.73.1.47 [PubMed: 15709831]
- Cronbach LJ (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334. doi: 10.1007/BF02310555
- 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. *Psychosom Med* 69, 735–739. [PubMed: 17942844]
- Dube P, Kroenke K, Bair MJ, Theobald D, Williams LS, 2010 The P4 screener: evaluation of a brief measure for assessing potential suicidal risk in 2 randomized effectiveness trials of primary care and oncology patients. *Prim Care Companion J Clin Psychiatry*. 12(6), e1–e8.
- Dum M, Pickren J, Sobell LC, Sobell MB, 2008 Comparing the BDI-II and the PHQ-9 with outpatient substance abusers. *Addictive Behaviors*. 33, 381–387. 10.1016/j.addbeh.2007.09.017 [PubMed: 17964079]
- Elhai JD, Contractor AA, Tamburrino M, Fine TH, Prescott MR, Shirley E, Chan PK, Slembariski R, Liberzon I, Galea S, Calabrese JR, 2012 The factor structure of major depression symptoms: A test of four competing models using the Patient Health Questionnaire-9. *Psychiatry Res* 199, 169–173. 10.1016/j.psychres [PubMed: 22698261]
- 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*. 17:349. [PubMed: 29065874]
- Garrett MD, Baldrige D, Benson W, Crowder J, Aldrich N, 2015 Mental health disorders among an invisible minority: Depression and dementia among American Indian and Alaska Native elders. *Gerontologist*. 55(2), 227–236. doi:10.1093/geront/gnu181 [PubMed: 26035598]
- González-Blanch C, Medrano LA, Muñoz-Navarro R, Ruíz-Rodríguez P, Moriana JA, Limonero JT, Schmitz F, Cano-Vindel A, PsicAP Research Group, 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. doi: 10.1371/journal.pone.0193356. [PubMed: 29474410]
- Goodkind JR, Gorman B, Hess JM, 2015 Reconsidering culturally competent approaches to American Indian healing and well-being. *Qual Health Res* 25(4), 486–99. doi: 10.1177/1049732314551056. [PubMed: 25239567]
- Granillo TM, 2012 Structure and function of the Patient Health Questionnaire-9 among Latina and Non-Latina white female college students. *J Soc Social Work Res* 3(2), 80–93. doi: 10.5243/jsswr.2012.6
- Guo B, Kaylor-Hughes C, Garland A, Nixon N, Sweeney T, Simpsonc S, Dagleish T, Ramana R, Yang M, Morrissa R, 2017 Factor structure and longitudinal measurement invariance of PHQ-9 for specialist mental health care patients with persistent major depressive disorder: Exploratory structural equation modelling. *J Affect Disord* 219, 1–8. 10.1016/j.jad.2017.05.020 [PubMed: 28501679]
- Heck JL, 2018 Screening for postpartum depression in American Indian/Alaskan Native women: A comparison of two instruments. *Am Indian Alsk Native Ment Health Res* 25(2), 74–102. doi: 10.5820/aian.2502.2018.74 [PubMed: 29889949]
- Hinz A, Mehnert A, Kocalevent RD, Braehler E, Forkmann T, Singer S, Schulte T, 2016 Assessment of depression severity with the PHQ-9 in cancer patients and in the general population. *BMC Psychiatry*. 2, 16:22. doi:10.1186/s12888-016-0728-6.
- Holzinger KJ, Swineford F, 1937 The bi-factor method. *Psychometrika*. 2, 41–54. doi: 10.1007/bf02287965
- Huang B, Grant BF, Dawson DA, Stinson FS, Chou SP, Saha TD, Goldstein RB, Smith SM, Ruan WJ, Pickering RP, 2006 Race-ethnicity and the prevalence and co-occurrence of Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, alcohol and drug use disorders and Axis I and II disorders: United States, 2001 to 2002. *Compr Psychiatry*. 47(4):252–257. [PubMed: 16769298]

- Huang FY, Chung H, Kroenke K, Delucchi KF, Spitzer RF, 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. doi: 10.1111/j.1525-1497.2006.00409.x [PubMed: 16808734]
- IBM Corp. 2015 IBM SPSS Statistics for Windows, Version 23.0. [Software.]
- Kading MF, Hautala DS, Palombi FC, Aronson BD, Smith RC, Walls MF, 2015 Flourishing: American Indian positive mental health. *Soc Ment Health*. 5(3), 203–2017. doi: 10.1177/2156869315570480 [PubMed: 28966866]
- Kalpakjian CZ, Toussaint FF, 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. *J Spinal Cord Med* 32(2), 147–156. [PubMed: 19569462]
- Keum B, Miller MJ, Kurotsuchi Inkelas K, 2018 Testing the factor structure and measurement invariance of the PHQ-9 across racially diverse U.S. college students. *Psychol Assess* 30(8): 1096–1106. 10.1037/pas0000550 [PubMed: 29565614]
- Kisely S, Katarzyna Alichniewicz K, Black EB, Siskind D, Spurling G, & Toombs M, 2017 The prevalence of depression and anxiety disorders in indigenous people of the Americas: A systematic review and meta-analysis. *J Psychiatr Res* 84, 137–152. [PubMed: 27741502]
- 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? *Rehabil Psychol* 53(4), 513–620. doi:10.1037/a0013354
- 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. *Arch Phys Med Rehabil* 91, 1218–1224. [PubMed: 20684902]
- 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, Spitzer RL, Williams JB, Löwe B, 2010 The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: A systematic review. *General Hospital Psychiatry*, 32, 345–359. 10.1016/j.genhosppsy.2010.03.006 [PubMed: 20633738]
- Leavitt RA, Ertl A, Sheats K, Petrosky E, Ivey-Stephenson A, Fowler KA, 2018 Suicides among American Indian/Alaska Natives — National Violent Death Reporting System, 18 states, 2003–2014. *Morbidity and Mortality Weekly Report*. 67(8), 237–242. <https://www.cdc.gov/mmwr/volumes/67/wr/pdfs/mm6708a1-H.pdf> (accessed 11 March 2019). [PubMed: 29494572]
- Li C, Ford ES, Strine TW, Mokdad AH, 2008 Prevalence of depression among U.S. adults with diabetes: findings from the 2006 behavioral risk factor surveillance system. *Diabetes Care*. 31: 105–107. [PubMed: 17934145]
- Melville JL, Gavin A, Guo Y, Fan M-Y, Katon WJ, 2010 Depressive disorders during pregnancy: Prevalence and risk factors in a large urban sample. *Obstet Gynecol* 116(5): 1064–1070. doi: 10.1097/AOG.0b013e3181f60b0a. [PubMed: 20966690]
- Merz EL, Malcarne VL, Roesch SC, Riley N, Robins Sadler G, 2011 A multigroup confirmatory factor analysis of the Patient Health Questionnaire-9 among English- and Spanish-speaking Latinas. *Cultur Divers Ethnic Minor Psychol* 17(3): 309–316. doi:10.1037/a0023883 [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. *BJPsych Open*.2(2), 127–138. [PubMed: 27703765]
- Morehead DL, 2012 Length of time in the United States, adiposity, and blood pressure as predictors of depression in an ethnically diverse sample. [Dissertation] Howard University, Washington, DC.
- Muthén B, & Asparouhov T, 2002 Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. *Mplus Web Notes*, 4, 1–22. <https://www.statmodel.com/download/webnotes/CatMGLong.pdf> (accessed 8 March 2019)
- Muthén LK, 1 June 2016 – 6:10 a.m. Re: Model fit output for continues/binary outcomes. <http://www.statmodel.com/discussion/messages/11/22754.html#POST125108> (accessed 7 March 2019).
- Muthén LK, Muthén BO, 1998–2015 Mplus user’s guide (7th ed.). Muthén & Muthén, Los Angeles, CA.

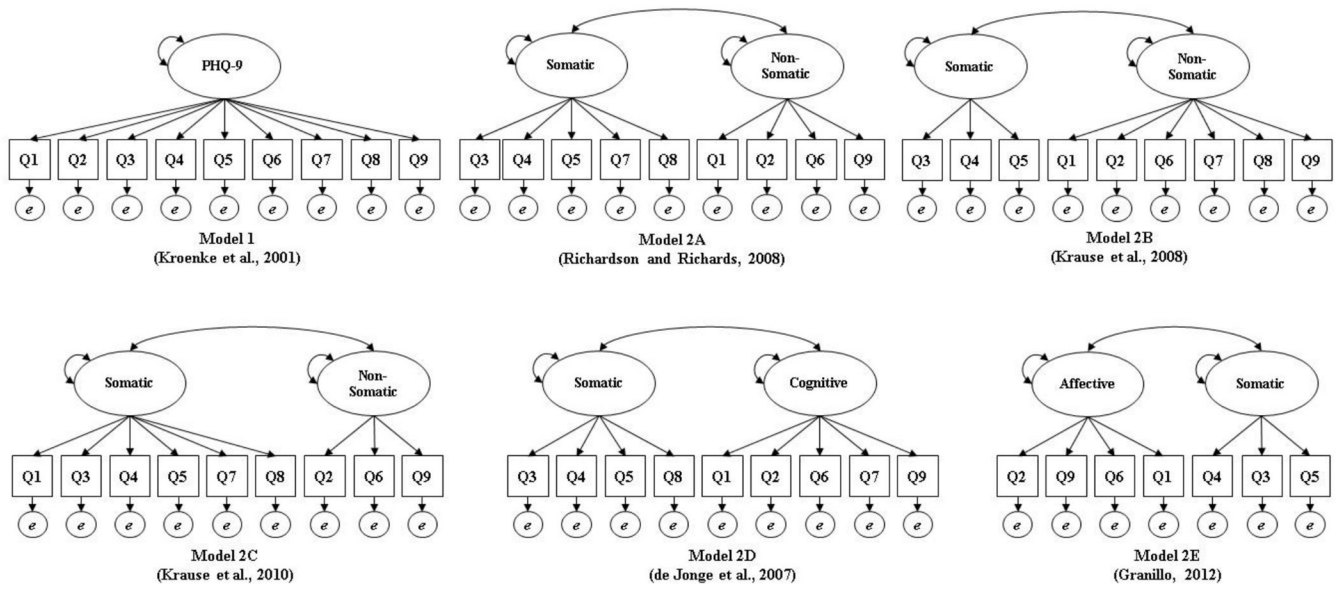
- National Center for Health Statistics, 2010 International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9). Atlanta, GA: Centers for Disease Control and Prevention.
- National Committee for Quality Assurance, 2018 HEDIS and Performance Measurement. <https://www.ncqa.org/hedis/> (accessed 11 March 2019)
- National Conference of State Legislatures, 2018 Federal and state recognized tribes. <http://www.ncsl.org/research/state-tribal-institute/list-of-federal-and-state-recognized-tribes.aspx> (accessed 11 March 2019).
- O'Neil T, 2004 Culture and pathology: Flathead loneliness revisited. The 2001 Roger Allan Moore Lecture. *Cult Med Psychiatry*. 28, 221–230. [PubMed: 15470949]
- Patel JS, 2017 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. [Thesis] Purdue University, Indianapolis, IN.
- Petersen JJ, Paulitsch MA, Hartig J, Mergenthal K, Gerlach FM, Gensichen J, 2014 Factor structure and measurement invariance of the Patient Health Questionnaire-9 for female and male primary care patients with major depression in Germany. *J Affect Disord* 170, 138–142. [PubMed: 25240840]
- R Core Team, 2018 R: A language and environment for statistical computing. Vienna, Austria: Foundation for Statistical Computing, [software]
- Revelle W, 2018 psych 1.8.10: Procedures for personality and psychological research. [Software.]
- 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, 243–249.
- Rossee Y, 2018 lavaan 0.6-2. [Software.]
- Satorra A, Bentler PM, 2001 A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*. 66(4), 507–514.
- Schermelleh-Engel K, Moosbrugger EL, 2003 Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*. 8, 23–74.
- Simon GE, Coleman KJ, Rossom RC, Beck A, Oliver M, Johnson E, Whiteside U, Operskalski B, Penfold RB, Shortreed SM, Rutter C, 2016 Risk of suicide attempt and suicide death following completion of the Patient Health Questionnaire depression module in community practice. *J Clin Psychiatry*. 77(2), 221–227. doi: 10.4088/JCP.15m09776 [PubMed: 26930521]
- Simon GE, Rutter CM, Peterson D, Oliver M, Whiteside U, Operskalski B, Ludman EJ, 2013 Does response on the PHQ-9 Depression Questionnaire predict subsequent suicide attempt or suicide death? *Psychiatr Serv* 64(12), 1195–1202. [PubMed: 24036589]
- Smith SM, Stinson FS, Dawson DA, Goldstein R, Huang B, Grant BF, 2006 Race/ethnic differences in the prevalence and co-occurrence of substance use disorders and independent mood and anxiety disorders: Results from the National Epidemiologic Survey on Alcohol and Related Conditions. *Psychol. Med* 36, 987–998. [PubMed: 16650344]
- Spitzer RL, Kroenke K, Williams JBW, 1999 Patient Health Questionnaire study group: Validity and utility of a self-report version of PRIME-MD: the PHQ Primary Care Study. *JAMA*. 282, 1737–1744. [PubMed: 10568646]
- Steenkamp J-BEM, Baumgartner H, 1998 Assessing measurement invariance in cross national consumer research. *J Consum Res* 25, 78–107. 10.1086/209528
- Suicide Prevention Resource Center, 2013 Suicide among racial/ethnic populations in the U.S.: American Indians/Alaska Natives. Waltham, MA: Education Development Center, Inc. [http://www.sprc.org/sites/default/files/migrate/library/AI\\_AN%20Sheet%20Aug%2028%202013%20Final.pdf](http://www.sprc.org/sites/default/files/migrate/library/AI_AN%20Sheet%20Aug%2028%202013%20Final.pdf) (accessed 11 March 2019).
- Tran TV, Nguyen T, Chan K, 2017 Developing Cross-Cultural Measurement in Social Work Research and Evaluation. Oxford University Press, New York.
- Tucker RP, Wingate LR, O'Keefe VM, 2016 Historical loss thinking and symptoms of depression are influenced by ethnic experiences in American Indian college students. *Cultur Divers Ethnic Minor Psychol* 22(3), 350–358. doi:10.1037/cdp0000055 [PubMed: 26371791]
- United States Department of Agriculture, 2016 Rural-Urban Commuting Area Codes. <https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/> (accessed 11 March 2019).

- Vandenberg RI, Lance CE, 2000 A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organ Res Methods*. 3, 4–70. 10.1177/109442810031002
- Whitbeck LB, McMorris BJ, Hoyt DR, Stubben JD, Lafromboise T, 2002 Perceived discrimination, traditional practices, and depressive symptoms among American Indians in the upper midwest. *J Health Soc Behav* 43, 400–418. 10.2307/3090234 [PubMed: 12664673]
- Whitbeck LB, Walls ML, Johnson KD, Morrisseau AD, McDougall CM, 2009 Depressed affect and historical loss among North American Indigenous adolescents. *Am Indian Alsk Native Ment Health Res* 16 (3), 16–41. 10.5820/aian.1603.2009.16 [PubMed: 20052631]
- World Health Organization, 2017 Depression and other common mental disorders: Global Health Estimates. <http://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf;jsessionid=9C80428261103FA6A1C28F6D43C3562F?sequence=1> (accessed 11 March 2019)
- World Health Organization, 2012 International Classification of Diseases, Tenth Revision (ICD-10). Herndon, VA: Stylus Publishing, LLC.
- Yu CY, 2002 Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes (Doctoral dissertation). <https://www.statmodel.com/download/Yudissertation.pdf> (assessed 11 March 2019)
- Yu X, Tam WW, Wong PT, Earn TH, Stewart SM, 2012 The Patient Health Questionnaire-9 for measuring depressive symptoms among the general population in Hong Kong. *Compr Psychiatry*. 53, 95–102. 10.1016/j.comppsy.2010.11.002 [PubMed: 21193179]
- Zumbo BD, Gadermann AM, Zeisser C, 2007 Ordinal versions of coefficients alpha and theta for likert rating scales. *Journal of Modern Applied Statistical Methods*.6:1. doi: 10.22237/jmasm/1177992180



**Highlights**

- The PHQ-9 was invariant for American Indian and Caucasian American adults.
- 1- and 2-factor models tested were equivalent between groups.
- Due to high factor correlations for 2-factor models, the 1-factor model may be best.
- The PHQ-9 is a suitable depression scale for American Indian adults in clinical care.



**Figure 1.**  
Tested one- and two-factor PHQ-9 models.

**Table 1.**

American Indian and Caucasian American adult demographics.

Demographics	American Indian ( <i>n</i> = 4,443)	Caucasian American ( <i>n</i> = 4,443)	<i>p</i>
	Count (%)	Count (%)	
Age <i>Mdn</i> ( <i>M</i> , <i>SD</i> , range) <sup><i>a</i></sup>	38 (40.41, 15.88, 18-96)	47 (47.25, 19.07, 18-98)	< 0.001
Age >=65 ( <i>n</i> = 1,289) <sup><i>a</i></sup>	365 (28%)	924 (72%)	< 0.001
Clinic RUCA code <sup><i>b</i></sup>			
Rural (1-3)	2,386 (54%)	1,295 (29%)	< 0.001
Micropolitan (4-6)	668 (15%)	1,030 (23%)	< 0.001
Urban (7-10)	1,389 (31%)	2,118 (48%)	< 0.001
Diagnosis at index date (first PHQ-9 administration) <sup><i>b,c</i></sup>			
Major depressive disorder	528 (12%)	429 (10%)	0.001
Major depressive disorder in remission	17 (<1%)	17 (<1%)	1.00
Depression diagnoses during the past year <sup><i>b,c</i></sup>			
Bipolar disorder	123 (3%)	78 (2%)	0.001
Bipolar disorder in remission	83 (2%)	58 (1%)	0.034
Major depressive disorder	639 (14%)	506 (11%)	< 0.001
Major depressive disorder in remission	27 (<1%)	30 (<1%)	0.690
Other depression diagnosis	1,043 (24%)	1,085 (24%)	0.296
Other mood disorder	165 (4%)	145 (4%)	0.569
Female <sup><i>b</i></sup>	3,056 (69%)	2,923 (66%)	0.003
Health Insurance Type <sup><i>b</i></sup>			
Insurance	3,965 (89%)	4,088 (92%)	0.001
Self-pay	329 (7%)	256 (6%)	0.001
Unknown	149 (3%)	99 (2%)	0.001
Hispanic	82 (2%)	0 (0%)	n/a
Refused	10 (<1%)	0 (0%)	n/a
Unknown	3 (<1%)	0 (0%)	n/a
Total PHQ-9 Score <i>Mdn</i> ( <i>M</i> , <i>SD</i> , range) <sup><i>a</i></sup>	11 (10.96, 7.14, 0-27)	7 (8.29, 6.74, 0-27)	< 0.001
Total PHQ-9 Score >=10 <sup><i>b</i></sup>	2,408 (54%)	1,750 (39%)	< 0.001
Total PHQ-9 Score >=15 <sup><i>b</i></sup>	1,476 (33%)	895 (20%)	< 0.001

**Logistic Regression Results (Ages 18 and Over, *n* = 8,886)<sup>*d*</sup>**

Independent Variables	PHQ-9 >=10			PHQ-9 >=15		
	OR	CI	<i>p</i>	OR	CI	<i>p</i>
American Indian	1.53	1.40-1.68	<0.001	1.64	1.48-1.82	<0.001
Female	0.91	0.83-0.99	0.040	0.96	0.86-1.06	0.421

Demographics	American Indian ( <i>n</i> = 4,443)	Caucasian American ( <i>n</i> = 4,443)	<i>p</i>			
	Count (%)	Count (%)				
Mean-centered age	0.98	0.98-0.99	<0.001	0.98	0.98-0.99	<0.001
Major depression diagnosis in the past year	3.50	3.04-4.03	<0.001	3.03	2.66-3.45	<0.001
Urban RUCA code	0.78	0.71-0.87	<0.001	0.76	0.68-0.84	<0.001
Micropolitan RUCA code	0.77	0.68-0.87	<0.001	0.73	0.63-0.84	<0.001
Model Fit						
$\chi^2$ ( <i>df</i> )	763.17 (6)		<0.001	650.45 (6)		<0.001
Cox-Snell R <sup>2</sup>	.08			.07		
Nagelkerke R <sup>2</sup>	.11			.10		
Hosmer-Lemeshow Test $\chi^2$ ( <i>df</i> )	17.33 (8)		0.027	27.17 (8)		0.001

Note. CI = Confidence interval. *df* = Degrees of freedom. *M* = Mean. *Mdn* = Median. n/a = Not applicable. OR = Odds ratio. *SD* = Standard deviation.  $\chi^2$  = Chi-square.

<sup>a</sup>Wilcoxon Rank Sums.

<sup>b</sup>Chi-square cross tabulation.

<sup>c</sup>Individuals could have more than one depression diagnosis.

<sup>d</sup>Comparison is a male Caucasian American of average age who did not have a major depression diagnosis in the past year, and the PHQ-9 clinic location had a rural RUCA code.

**Table 2.**

PHQ-9 descriptive statistics for American Indian and Caucasian American adults ages 18 and over.

PHQ-9 Items	American Indian					Caucasian American					<i>p</i> <sup>c</sup>
	Item-Rest Correlation <sup>a</sup>	Mdn	Percentiles		<i>n</i>	Item-Rest Correlation <sup>b</sup>	Mdn	Percentiles		<i>n</i>	
			25 <sup>th</sup>	75 <sup>th</sup>				25 <sup>th</sup>	75 <sup>th</sup>		
1. Little interested or pleasure in doing things	0.80	2	0	3	3,282	0.87	0	0	2	3,296	< 0.001
2. Feeling down, depressed, or hopeless	0.86	2	0	3	3,210	0.90	0	0	2	3,180	< 0.001
3. Trouble falling or staying asleep, or sleeping too much	0.73	2	1	3	3,454	0.76	2	0	3	3,428	< 0.001
4. Feeling tired or having little energy	0.73	2	1	3	3,256	0.75	2	0	3	3,225	< 0.001
5. Poor appetite or overeating	0.74	1	0	3	3,407	0.78	0	0	2	3,495	< 0.001
6. Feeling bad about yourself – or that you are a failure or have let yourself or your family down	0.85	1	0	3	3,378	0.85	0	0	2	3,450	< 0.001
7. Trouble concentrating on things, such as reading the newspaper or watching television	0.79	1	0	3	3,460	0.82	0	0	2	3,520	< 0.001
8. Moving or speaking so slowly that other people could have noticed	0.71	0	0	2	3,500	0.74	0	0	0	3,681	< 0.001
9. Thoughts that you would be better off dead or of hurting yourself in some way	0.60	0	0	0	3,657	0.66	0	0	0	3,830	< 0.001

Note. Mdn = Median. Items 1-9 ranged from 0-3. Some item data were missing. Medians are presented rather than means due to the skewed nature of PHQ-9 item data.

<sup>a</sup> *n* = 1,806.

<sup>b</sup> *n* = 1,811.

<sup>c</sup> Wilcoxon Rank Sums.

**Table 3.**

Factor scores, standard errors, and ordinal alpha for tested models for American Indian and Caucasian American adults ages 18 and over.

Tested Models	American Indian			Caucasian American		
	Mean (SE)	$\alpha$	<i>n</i>	Mean (SE)	$\alpha$	<i>n</i>
1) One-factor PHQ-9 <sup>a</sup>	11.41 (0.19)	0.94	1,806	7.93 (0.18)	0.95	1,811
2A) Two-factor PHQ-9 <sup>b</sup>						
Factor 1	7.40 (0.10)	0.89	2,282	5.31 (0.10)	0.90	2,223
Factor 2	4.23 (0.08)	0.90	2,411	2.85 (0.07)	0.92	2,482
2B) Two-factor PHQ-9 <sup>c</sup>						
Factor 1	5.32 (0.06)	0.86	2,686	4.21 (0.06)	0.88	2,588
Factor 2	6.11 (0.12)	0.92	2,149	3.85 (0.10)	0.94	2,245
2C) Two-factor PHQ-9 <sup>d</sup>						
Factor 1	8.88 (0.13)	0.91	2,123	6.27 (0.13)	0.92	2,050
Factor 2	2.78 (0.05)	0.87	2,622	1.88 (0.05)	0.90	2,722
2D) Two-factor PHQ-9 <sup>e</sup>						
Factor 1	6.15 (0.08)	0.86	2,443	4.59 (0.08)	0.87	2,386
Factor 2	5.38 (0.10)	0.91	2,260	3.51 (0.09)	0.93	2,322
2E) Two-factor, seven-item PFIQ-9 <sup>f</sup>						
Factor 1	4.23 (0.08)	0.90	2,411	2.85 (0.07)	0.92	2,482
Factor 2	5.32 (0.06)	0.86	2,686	4.21 (0.06)	0.88	2,588

Notes. *SE* = Standard error.  $\alpha$  = Ordinal alpha (Zumbo et al., 2007). Count data unless otherwise specified. Factor means, standard errors, and ordinal alpha were calculated only with complete data on all included PHQ-9 questions. Rounded to nearest hundredth place.

<sup>a</sup>Kroenke et al. (2001).

<sup>b</sup>Richardson and Richards (2008).

<sup>c</sup>Krause et al. (2008).

<sup>d</sup>Krause et al. (2010).

<sup>e</sup>de Jonge et al. (2007).

<sup>f</sup>Granillo (2012).

**Table 4.**

Confirmatory factor analysis results for tested one- and two-factor PHQ-9 models by American Indian and Caucasian American group (ages 18 and over).

Models	<i>n</i>	$\chi^2$ ( <i>df</i> )	CFI	TLI	RMSEA (CI)	SRMR
1) One-Factor <sup>a</sup>						
American Indian	1,806	133.29 (27) ***	0.995	0.993	0.047 (0.039-0.055)	0.039
Caucasian American	1,811	101.38 (27) ***	0.996	0.995	0.039 (0.031-0.047)	0.040
2A) Two-Factor <sup>b</sup>						
American Indian	1,806	108.56 (26) ***	0.996	0.994	0.042 (0.034-0.050)	0.036
Caucasian American	1,811	78.18 (26) ***	0.997	0.996	0.033 (0.025-0.042)	0.036
2B) Two-Factor <sup>c</sup>						
American Indian	1,806	95.83 (26) ***	0.997	0.995	0.039 (0.030-0.047)	0.034
Caucasian American	1,811	70.27 (26) ***	0.998	0.997	0.031 (0.022-0.039)	0.035
2C) Two-Factor <sup>d</sup>						
American Indian	1,806	121.76 (26) ***	0.995	0.994	0.045 (0.037-0.053)	0.037
Caucasian American	1,811	90.73 (26) ***	0.997	0.995	0.037 (0.029-0.046)	0.038
2D) Two-Factor <sup>e</sup>						
American Indian	1,806	118.77 (26) ***	0.995	0.994	0.044 (0.037-0.053)	0.037
Caucasian American	1,811	83.42 (26) ***	0.997	0.996	0.035 (0.027-0.043)	0.038
2E) Two-Factor <sup>f</sup>						
American Indian	1,947	48.09 (13) ***	0.997	0.996	0.037 (0.026-0.049)	0.030
Caucasian American	1,931	29.85 (13) ***	0.999	0.998	0.026 (0.014-0.038)	0.027

Note. CI = Confidence interval. *df* = Degrees of freedom. Acceptable fits include non-significant  $\chi^2$ , RMSEA < 0.05 or at most < 0.08, CFI and TLI > 0.90 and preferably > 0.95, and SRMR < 0.05 or at most < 0.10 (Steenkamp and Baumgartner, 1998; Vandenberg and Lance, 2000).

\*\*\*  
*p* < 0.001.

<sup>a</sup>Kroenke et al. (2001).

<sup>b</sup>Richardson and Richards (2008).

<sup>c</sup>Krause et al. (2008).

<sup>d</sup>Krause et al. (2010).

<sup>e</sup>de Jonge et al. (2007).

<sup>f</sup>Granillo (2012).

Table 5.

Measurement invariance testing for one- and two-factor PHQ-9 models between American Indian and Caucasian American adults ages 18 and over.

Models	$\chi^2$ (df)	$\chi^2_{diff}$ (df) <sup>a</sup>	CFI	CFI	TLI	RMSEA (CI)	SRMR	WRMR
1) One-Factor <sup>b</sup>								
Configural	234.75 (54)***		0.995		0.994	0.043 (0.037-0.049)	0.036	2.09
Metric	366.17 (62)***	70.32 (8)***	0.992	-0.003	0.991	0.052 (0.047-0.057)	0.043	2.60
Scalar	381.50 (70)***	30.43 (8)***	0.992	0.000	0.992	0.050 (0.045-0.055)	0.044	2.66
2A) Two-Factor <sup>b</sup>								
Configural	186.80 (52)***		0.997		0.995	0.038 (0.032-0.044)	0.033	1.86
Metric	318.75 (59)***	70.22 (7)***	0.993	-0.004	0.992	0.049 (0.044-0.055)	0.040	2.43
Scalar	333.92 (66)***	31.12 (7)***	0.993	0.000	0.993	0.047 (0.042-0.052)	0.041	2.49
2B) Two-Factor <sup>b</sup>								
Configural	166.15 (52)***		0.997		0.996	0.035 (0.029-0.041)	0.031	1.75
Metric	255.08 (59)***	47.89 (7)***	0.995	-0.002	0.994	0.043 (0.038-0.048)	0.037	2.17
Scalar	269.51 (66)***	29.94 (7)***	0.995	0.000	0.994	0.041 (0.036-0.046)	0.038	2.23
2C) Two-Factor <sup>b</sup>								
Configural	212.57 (52)***		0.996		0.994	0.041 (0.036-0.047)	0.034	1.98
Metric	337.33 (59)***	66.01 (7)***	0.993	-0.003	0.991	0.051 (0.046-0.056)	0.041	2.50
Scalar	351.75 (66)***	29.12 (7)***	0.993	0.000	0.992	0.049 (0.044-0.054)	0.042	2.55
2D) Two-Factor <sup>b</sup>								
Configural	202.26 (52)***		0.996		0.995	0.040 (0.034-0.046)	0.035	1.94
Metric	330.23 (59)***	68.00 (7)***	0.993	-0.003	0.992	0.050 (0.045-0.056)	0.042	2.47
Scalar	345.73 (66)***	31.57 (7)***	0.993	0.000	0.992	0.048 (0.043-0.054)	0.043	2.53
2E) Two-Factor <sup>c</sup>								
Configural	77.96 (26)***		0.998		0.997	0.032 (0.024-0.040)	0.025	1.49
Metric	101.28 (31)***	15.42 (5)**	0.998	0.000	0.997	0.034 (0.027-0.042)	0.028	1.70
Scalar	112.35 (36)***	23.80 (5)***	0.997	-0.001	0.997	0.033 (0.026-0.040)	0.030	1.79

Note. CI = Confidence interval. *df* = Degrees of freedom.  $\Delta$  = Change between nested models. Preferred fit indexes: nonsignificant  $\chi^2$  and  $\chi^2_{diff}$ ;

CFI no more than -0.01; CFI and TLI at least > 0.90 and preferably > 0.95; RMSEA < 0.05 or at most < 0.08; SRMR < 0.08 or at most < 0.10; WRMR around 1.00 (Bentler, 1990; Cheung and Rensvold, 2002; Steenkamp and Baumgartner, 1998; Vandenberg and Lance, 2000; Yu, 2002).

\*\*  
*p* < 0.01.

\*\*\*  
*p* < 0.001.

<sup>a</sup>Satorra and Bentler (2001) scaled  $\chi^2_{diff}$  test.

<sup>b</sup>American Indian *n* = 1,806; Caucasian American *n* = 1,811.



<sup>c</sup>American Indian  $n = 1,947$ ; Caucasian American  $n = 1,931$ .

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript