Journal of Affective Disorders Reports 4 (2021) 100121

Contents lists available at ScienceDirect

Journal of Affective Disorders Reports

journal homepage: www.elsevier.com/locate/jadr

Research Paper

Evaluating the cross-cultural measurement invariance of the PHQ-9 between American Indian/Alaska Native adults and diverse racial and ethnic groups

Melissa L. Harry[a],[*], R. Yates Coley[b], Stephen C. Waring[a], Gregory E. Simon[b]

[a] Essentia Health, Essentia Institute of Rural Health, 502 East Second Street, Duluth, MN 55805, United States
[b] Kaiser Permanente Washington Health Research Institute, 1730 Minor Ave, Suite 1600 Seattle, WA 98101-1466, United States

ARTICLE INFO

Keywords:
American Indian/Alaska Native
Cross-cultural
Health equity
Measurement equivalence
Measurement invariance
Patient Health Questionnaire-9

ABSTRACT

Background: The Patient Health Questionnaire-9 (PHQ-9), a self-reported depression screening instrument for measurement-based care (MBC), may have cross-cultural measurement invariance (MI) with a regional group of American Indian/Alaska Native (AI/AN) and non-Hispanic White adults. However, to ensure health equity, research was needed on the cross-cultural MI of the PHQ-9 between other groups of AI/AN peoples and diverse populations.

Methods: We assessed the MI of the one-factor PHQ-9 model and five previously identified two-factor models between non-Hispanic AI/AN adults (ages 18–64) from healthcare systems A (n = 1759) and B (n = 2701) using secondary data and robust maximum likelihood estimation. We then tested models for full or partial MI between either combined or separate AI/AN groups, respectively, and Hispanic (n = 7974), White (n = 7,974), Asian (n = 6988), Black (n = 6213), and Native Hawaiian/Pacific Islander (n = 1370) adults from healthcare system B. All had mental health or substance use disorder diagnoses and were seen in behavioral health or primary care settings from 1/1/2009 to 9/30/2017.

Results: The one-factor PHQ-9 model was partially invariant, with two-factor models partially, or in one case fully, invariant between AI/AN groups. The one-factor model and three two-factor models were partially invariant between all seven groups, while a two-factor model was fully invariant and another partially invariant between a combined AI/AN group and other racial and ethnic groups.

Conclusions: Achieving health equity in MBC requires ensuring the cross-cultural validity of measurement tools. Before comparing mean scores, PHQ-9 models should be assessed for individual racial and ethnic group fit for adults with mental health or substance use disorders.

1. Introduction

Measurement-based care (MBC) is important when evaluating the impact of mental health clinical care with patients over time in evidence-based practice, particularly when using standardized instruments (Coley et al., 2020; Scott and Lewis, 2015). One such standardized depression screener is the Patient Health Questionnaire-9 (PHQ-9) (Kroenke et al., 2001). The PHQ-9 is widely used by healthcare systems and plans in assessing levels of depression symptomology in patient populations (Kroenke et al., 2001), evaluating the effectiveness of interventions (National Committee for Quality Assurance, 2020), and gauging treatment response and depression remission as part of MBC for patients with depression diagnoses (Coley et al., 2020). PHQ-9 item 9, regarding thoughts of death or self-harm, is also recommended as a tool to identify suicidal ideation or suicide risk (Coleman et al., 2018; Simon et al., 2013, 2016). Research has shown that health care is commonly utilized before suicide (Ahmedani et al., 2014, 2019). This suggests that healthcare systems offer an opportunity for intervention, including depression screening, suicide risk assessment, referrals, and treatment.

As noted in prior work (Harry and Waring, 2019), depression, which is a leading cause of disability worldwide (GBD 2017 Disease and Injury Incidence and Prevalence Collaborators, 2018), is also a suicide risk factor (Chesney et al., 2014). Some populations may be disproportionally affected by suicide. In the United States, the suicide rate adjusted for age increased 33% from 1999 to 2017 (Hedegaard et al., 2018). While suicide rates increased significantly for most racial and ethnic groups, American Indian (or Native American) and Alaska Native (AI/AN) peoples showed the greatest percent changes over time,

[*] Corresponding author.
E-mail address: Melissa.Harry@EssentiaHealth.org (M.L. Harry).

https://doi.org/10.1016/j.jadr.2021.100121
Received 16 November 2020; Received in revised form 16 February 2021; Accepted 20 February 2021
Available online 22 February 2021

with increases in suicide rates for females (139%) outpacing those of males (71%) (Curtin and Hedegaard, 2019).

Depression prevalence rates may differ by group of AI/AN peoples (Asdigian et al., 2018; Bowen et al., 2020). Measures of depression may also lack cross-cultural validity testing with Indigenous peoples (Kisely et al., 2017). Furthermore, AI/AN peoples still face considerable health inequities (e.g., Carron, 2020; Garrett et al., 2015; Indian Health Services, 2019), and have experienced health disparities since colonization (Jones, 2006). This may be due to disparities in the distribution of wealth and power in the Americas (Jones, 2006) and historical trauma (e.g., Brave Heart and DeBruyn, 1998; Carron, 2020).

The COVID-19 pandemic has exposed and amplified the structural health inequities experienced by persons of color (e.g., Williams and Cooper, 2020). The pandemic also provides a reminder that employing culturally valid screening instruments within healthcare systems is a necessary component in achieving health equity, including for AI/AN peoples. Screening measures developed with primarily non-Hispanic White people and designed for the general population, such as the PHQ-9, should be assessed for cross-cultural measurement invariance (MI), or equivalence, before comparisons are made using these measures between different racial and ethnic groups (Cleary, 2013; Sass, 2011; Tran et al., 2017). When a scale is found to be cross-culturally measurement invariant, this means that different cultures define the construct measured by the instrument and the items it contains in a similar way, allowing comparisons to be made between different cultural groups (Tran et al., 2017).

Previous research has found mixed results for the standard one-factor PHQ-9 model with a range of racial and ethnic groups in the United States (Harry and Waring, 2019). While the one-factor model was shown to be a good fit for English and Spanish-speaking Latina women (Merz et al., 2011), two-factor models were a better fit for a group of Hispanic, Asian American, Black, and non-Hispanic White college students (Keum et al., 2018). Two-factor models, which typically separate PHQ-9 items into different forms of somatic or non-somatic (including those described as general, cognitive, and affective) latent factors, have generally displayed improved fits over the standard one-factor model in racially and ethnically diverse U.S. groups (Granillo, 2012; Harry and Waring, 2019; Keum et al., 2018; Patel et al., 2019). Specifically, while both one- and two-factor models were found to have good fits for non-Hispanic White, non-Hispanic Black, non-Hispanic Asian, Mexican American, and Hispanic adults in a study by Patel et al. (2019), the best-fitting two-factor model originally tested by Krause et al. (2008) was cross-culturally equivalent between groups. Granillo (2012) found that a seven-item, two-factor PHQ-9 model presented the best fit for Latina and non-Latina college students. Another study found a good fit for the one-factor model for Black, Chinese American, Hispanic, and non-Hispanic White adults separately, although some items displayed differential item functioning for Chinese American and Hispanic adults (Huang et al., 2006). In summary, no single factor model has been found to be a good fit between all racial and ethnic groups previously studied.

Harry and Waring (2019) recently reported that the PHQ-9 had an excellent fit and displayed cross-cultural MI for a number of models previously identified in the literature between AI/AN and non-Hispanic White general patient population adults seen at an integrated healthcare system with facilities located in the Upper Midwest. Tested models included the standard one-factor PHQ-9 model (Kroenke et al., 2001) and five two-factor models initially identified with other populations based on race, ethnicity, gender, or diagnosis (de Jonge et al., 2007; Granillo, 2012; Krause et al., 2008, 2010; Richardson and Richards, 2008). Many of these models have been tested together in other populations (e.g., Elhai et al., 2012; Keum et al., 2018). The work by Harry and Waring appears to be the first study focusing on how the PHQ-9 functions cross-culturally with Indigenous North American peoples. However, whether these results generalized to other AI/AN populations or if differences existed between AI/AN peoples and other racial and ethnic groups were unknown at the time of our study. While many groups of AI/AN peoples have the shared experience of intergenerational and historical trauma (Brave Heart and DeBruyn, 1998), AI/AN peoples are culturally, linguistically, spiritually, and geographically diverse (Livingston et al., 2019; Walls et al., 2017). There are over 570 federally recognized sovereign tribal nations in the United States (Indian Affairs Bureau, 2020), along with many other non-recognized bands and tribes. This vast diversity, along with the complex interplay of sociohistorical (e.g., historical trauma) and current contexts (e.g., health disparities, resilience), should be taken into consideration in regard to measurement with AI/AN peoples (Walls et al., 2017). Although there is a need for common measurement scales, finding that a scale works well with one tribal or AI/AN cultural group may not mean that the scale will work the same with another group of AI/AN people (Walls et al., 2017). In the present study, our primary goal was to examine the cross-cultural MI of the PHQ-9 between two populations of non-Hispanic AI/AN adults served by two geographically distant U.S. healthcare systems, as well as between non-Hispanic AI/AN adults and other racial and ethnic groups.

2. Guiding hypotheses

In this study, we had two aims: 1) assess whether any previously identified one- and two-factor PHQ-9 models were a good fit and displayed cross-cultural MI between two groups of non-Hispanic AI/AN adult peoples; and 2) if any models exhibited full or partial MI between the two groups of AI/AN adults, evaluate whether those models were also cross-culturally invariant between either a combined group (full MI) or separate groups (partial MI) of the two non-Hispanic AI/AN adults and a sample of other diverse racial and ethnic groups. We tested the following three progressively nested hypotheses in assessing cross-cultural MI for both study aims: 1) Equal factor structures between groups; 2) Equal factor structures and loadings between groups; and 3) Equal factor structures, loadings, and intercepts between groups (Brown, 2015).

3. Methods

3.1. Participants and procedures

The sample for this study came from two healthcare systems. Healthcare system A is a nonprofit serving patients in predominantly rural locations in the Upper Midwest. Healthcare system B serves patients in the Pacific Northwest. Eligibility for this study included adult patients ages 18 through 64 with at least one complete PHQ-9 in behavioral health or primary care (including family practice and internal medicine) visits where patients had a mental health or substance use disorder diagnosis from 1/1/2009 to 9/30/2017. Patients included non-Hispanic AI/AN adults at healthcare system A, and non-Hispanic White, Hispanic, non-Hispanic Asian, non-Hispanic Black, non-Hispanic AI/AN, and non-Hispanic Native Hawaiian/Pacific Islander (NH/PI) adults at healthcare system B. Data came from two secondary de-identified datasets: 1) a sample of AI/AN adults (Harry and Waring, 2019); and 2) another secondary dataset from healthcare system B. These datasets were originally generated from data extracted from electronic health records (EHR). Due to the extent of missing PHQ-9 item data within the EHR, and consequently both datasets, only the first complete PHQ-9 for individuals within the two datasets were analyzed in this study. Because the sample of non-Hispanic White adults with a complete PHQ-9 (n = 117,238) exceeded any other racial or ethnic group, a random sample matching the total number of the largest minority group, Hispanic adults (n = 7974) was drawn as a comparison group. The total study sample equaled 34,979. This study was reviewed and approved by the Institutional Review Boards for both healthcare systems.

3.2. Measures

The PHQ-9 was developed with primary care patients to assess levels of depression (Kroenke et al., 2001). PHQ-9 items were originally part of the Patient Health Questionnaire (PHQ), which is a patient self-reported version of the PRIME-MD (Spitzer et al., 1999). PHQ-9 questions are scored from 0 to 3, representing how often each item was experienced in the last two weeks (Kroenke et al., 2001). Increasing frequency corresponds with higher item scores and greater depression (Kroenke et al., 2001). Summing the scores for the nine items gives a total score, where 0–4 signifies 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 is also used internationally (e.g., Baas et al., 2011; Dreher et al., 2017; González-Blanch et al., 2018; Kim and Lee, 2019; Subotić et al., 2015; Villarreal-Zegarra et al., 2019). Previously reported levels of internal consistency reliability with Cronbach's alpha (α) were good at 0.86 and 0.89 (Kroenke et al., 2001), as were pooled sensitivity (81.3%) and specificity (85.3%) in a recent meta-analysis (Mitchell et al., 2016, as cited in Harry and Waring, 2019). Ordinal α also showed an excellent fit with AI/AN adults (0.94) (Harry and Waring, 2019).

3.3. Data analyses

Descriptive statistics, confirmatory factor analysis (CFA), and multiple group CFA (MGCFA) MI analyses were performed in Mplus version 8.3 (Muthén and Muthén, 2019). Internal consistency reliability analyses were conducted in R version 3.6.1 (R Core Team, 2019) with "psych" package version 1.9.12 (Revelle, 2019) for producing Cronbach's α (Cronbach, 1951), ordinal α (Zumbo et al., 2007), and item-total (or item-rest) correlations based on an underlying polychoric correlation matrix. We used robust maximum likelihood (MLR) estimation in CFA and MGCFA analyses. Weighted least squares means and variance adjusted (WLSMV) estimation was designed for categorical items with underlying normally distributed latent variables. However, MLR can handle moderately skewed observed categorical variables (Li, 2016), as was the case with the data for PHQ-9 items reported here aside from item 9 (Supplementary Material, Table S1). Item 9 would be expected to be more skewed since self-report of suicidal ideation is less common.

Like in prior work (Harry and Waring, 2019), we first identified best-fitting baseline CFA models for each racial and ethnic group for the standard one-factor PHQ-9 model and the five previously identified two-factor models (2A-2E) (Table 1) (Byrne, 2012). Models were considered as having a good fit if root mean square error of approximation (RMSEA) was < 0.05 or at most < 0.08, standardized root mean square residual (SRMR) was < 0.08 or at most < 0.10, and confirmatory fit index (CFI) and Tucker-Lewis index (TLI) were > 0.90 and ideally near 1.00 (Brown, 2015; Hu and Bentler, 1999; Vandenberg and Lance, 2000). If showing a good fit, these indexes can be used to assess the overall fit of a model (Tran et al., 2017). While a small and non-significant ($p < 0.05$) chi-square (χ^2) is also preferred, it is broadly understood that sensitivity to sample size makes χ^2 less than optimal for assessing model fit (Byrne, 2012).

MI is tested through progressively restrictive nested models commonly referred to as configural (equal factor structures), metric/weak (equal factor structures and loadings), scalar/strong (equal factor structures, loadings, and item intercepts or thresholds for ordinal data), and strict invariance (equal factor structures, loadings, intercepts or thresholds, and item residuals) (Meredith, 1993; Brown, 2015). Configural invariance represents that a model's structure, or the underlying latent factors and indicators, are equivalent between groups (Brown, 2015). Configural invariance is typically the first MGCFA model assessed (Brown, 2015). If configural noninvariance is found, one may pursue partial invariance by allowing item residuals representing the greatest drop in χ^2 to correlate separately by group where substantively appropriate (Byrne, 2012; Byrne et al., 1989). Partial invariance allows researchers to identify differences in specific loadings and/or intercepts (or thresholds) between groups (van de Schoot et al., 2012). Correlating item residuals for underlying psychological latent constructs to improve model fit is often required (Byrne et al., 1989). When a model displays metric invariance, the scale's metrics are equal between differing groups and meaningful comparisons can be made between groups on observed scale items and latent factors (Milfont and Fischer, 2010; Steenkamp and Baumgartner, 1998). However, metric noninvariance precludes moving on to scalar invariance testing; instead, researchers can revise the model's loading constraints, remove invariant loadings, or determine that model loadings vary between tested groups (Putnick and Bornstein, 2016). Findings of scalar invariance allow for comparisons of latent factor mean scores between groups (Brown, 2015; Steenkamp and Baumgartner, 1998; Steinmetz, 2013). Researchers who identify scalar noninvariance can release or add intercept (or threshold) constraints (Sass, 2011), remove items with invariant intercepts, or, like with metric invariance, determine that intercepts (or thresholds) vary between groups and halt further testing (Putnick and Bornstein, 2016). Strict invariance adds the additional constraint of equal item residuals between groups (Meredith, 1993; Brown, 2015). Findings of metric and scalar invariance are of prime importance in MGCFA MI testing (Muthén and Asparouhov, 2002). Like in similar work (Harry and Waring, 2019), in this paper we tested the most commonly reported configural, metric, and scalar invariance (Brown, 2015; Putnick and Bornstein, 2016). Putnick and Bornstein (2016) found that larger samples were associated with findings of strict noninvariance, making the assessment of residual invariance inappropriate with the present study's large sample.

In Mplus, nested MGCFA MI models are tested and compared sequentially using the MODEL = CONFIGURAL METRIC SCALAR command (Muthén and Muthén, 1998–2017). This command includes change in χ^2 ($\Delta\chi^2$) in MLR estimation, a log likelihood-based test with scaling correction factors (Muthén and Muthén, 1998–2017). We followed the advice that researchers calculate and report Satorra and Bentler's (2001) scaled χ^2 (S-Bχ^2) for estimating change (ΔS-Bχ^2) (Sass, 2011; Statmodel.com, n.d.). $\Delta\chi^2$ may over reject MGCFA models with large sample sizes (Putnick and Bornstein, 2016). ΔS-Bχ^2 appears to perform better with large samples (around 1000) than with small samples (Satorra and Bentler, 2001). However, the large sample in the present study could lead to χ^2 tests identifying small, spurious differences (Steinmetz, 2013). Other changes in goodness-of-fit criteria are also used when comparing nested MI models, such as ΔCFI, ΔRMSEA, and ΔSRMR, particularly with large samples when differences or changes in χ^2 are statistically significant (Putnick and Bornstein, 2016). Researchers are encouraged to use more than χ^2 in comparing model fit (Schermelleh-Engel and Moosbrugger, 2003). In the present study, we followed Cheung and Rensvold (2002) and based MI results on ΔCFI −0.01.

4. Results

Demographics for the sample, including separate groups of AI/AN adults at healthcare systems A and B, are presented in Table 2. In general, racial and ethnic group frequencies were similar, although more non-Hispanic AI/AN adults from healthcare system A were seen in primary care settings (85.6%) than any healthcare system B group. Also, more non-Hispanic AI/AN adults were aged 45–64 at healthcare system B than at healthcare system A. The majority of adults in the sample were female (69.6%) and seen in primary care (65.0%) versus behavioral health (35%) settings.

4.1. PHQ-9 item frequencies and internal consistency reliability

Table 3 presents item and total score means and standard deviations by racial and ethnic group. Medians and interquartile ranges for PHQ-9 items are presented in Table S2 (Supplementary Material). Mean total PHQ-9 scores were between 10 and 15, representing moderate levels of depression (Kroenke et al., 2001). Table 4 illustrates PHQ-9 item-total correlations and measures of internal consistency reliability for individual racial and ethnic groups. Some differences were seen in item-total correlations between the two groups of non-Hispanic AI/AN adults, as

Table 1
Previously Identified PHQ-9 Two-Factor Models and Underlying Latent Constructs.

PHQ-9 Items[a]	2A[b]	2B[c]	2C[d]	2D[e]	2E[f]
1 Little interest or pleasure in doing things	Non-Somatic	Non-Somatic	Somatic	Non-Somatic	Non-Somatic
2 Feeling down, depressed, or hopeless	Non-Somatic	Non-Somatic	Non-Somatic	Non-Somatic	Non-Somatic
3 Trouble falling or staying asleep, or sleeping too much	Somatic	Somatic	Somatic	Somatic	Somatic
4 Feeling tired or having little energy	Somatic	Somatic	Somatic	Somatic	Somatic
5 Poor appetite or overeating	Somatic	Somatic	Somatic	Somatic	Somatic
6 Feeling bad about yourself – or that you are a failure or have let yourself or your family down	Non-Somatic	Non-Somatic	Non-Somatic	Non-Somatic	Non-Somatic
7 Trouble concentrating on things, such as reading the newspaper or watching television	Somatic	Non-Somatic	Somatic	Non-Somatic	–
8 Moving or speaking so slowly that other people could have noticed? Or the opposite - being so fidgety or restless that you have been moving around a lot more than usual	Somatic	Non-Somatic	Somatic	Somatic	–
9 Thoughts that you would be better off dead or of hurting yourself in some way	Non-Somatic	Non-Somatic	Non-Somatic	Non-Somatic	Non-Somatic

Note. Adapted from Patel et al. (2019). The naming conventions of factors in prior two-factor PHQ-9 solutions included somatic and non-somatic (alternatively referred to as affective, cognitive, or general). [b,c,e,f]Non-somatic, affective, cognitive, and general factors have been labeled non-somatic here. Questions included within factors do overlap; those deemed somatic by one set of authors could be grouped with those grouped as non-somatic by other authors and vice versa. These labels may lack utility for clinicians and researchers due to this overlap.

[a] Kroenke et al. (2001).

[b] Richardson and Richards (2008).

[c] Krause et al. (2008).

[d] Krause et al. (2010).

[e] de Jonge et al. (2007).

[f] Granillo (2012).

well as other racial and ethnic groups, suggesting cross-cultural noninvariance (Tran et al., 2017). However, ordinal α showed an excellent level of internal consistency reliability (> 0.90) for the standard one-factor model. Cronbach's α was good (> 0.80) for all groups aside from non-Hispanic AI/AN adults from healthcare system A, where it was excellent (0.90).

4.2. Baseline CFA models

The one-factor model and all two-factor models had poor fits based on RMSEA or upper confidence interval (CI) RMSEA $\geq$ 0.080 (Supplementary Material, Table S3). Exceptions included model 2B for non-Hispanic Asian adults (RMSEA = 0.073, 90% CI 0.070–0.077), and model 2E for all groups aside for non-Hispanic NH/PI adults (RMSEA = 0.069, 90% CI 0.057–0.083) (Supplementary Material, Table S3). We progressively added correlating residuals based on modification indexes showing the greatest drop in χ^2, which improved RMSEA to < 0.080 for all other groups and models (Supplementary Material, Table S4). All correlations reflected items that could be associated with one another. For example, item 3 relates to sleep and item 4 relates to feeling tired.

Standardized loadings for initial (Supplementary Material, Table S5) and final (Supplementary Material, Table S6) baseline models show that most items loaded higher than 0.500 on all factors. The exception was item 9, which consistently had the lowest loading of any item regardless of model or group. Some loadings differed by more than 0.05 between both non-Hispanic AI/AN groups, as well as between other racial and ethnic groups, suggesting cross-cultural noninvariance (Tran et al., 2017). Most factor correlations in two-factor models were > 0.85. Factor correlations were highest for the group of non-Hispanic AI/AN adults from healthcare system A, where most exceeded 0.90.

4.3. Multiple group CFA

4.3.1. Separate non-hispanic AI/AN groups by healthcare system

Full MI was pursued for models 2B and 2E, and partial MI for the one-factor model and models 2A, 2C, and 2D between the two non-Hispanic AI/AN healthcare system groups. For all nested models, ΔS-Bχ^2 was statistically significant between the two groups of non-Hispanic AI/AN adults (Table 5). ΔCFI was −0.011 between metric and scalar models for the one-factor model (χ^2 = 677.58 [df = 64], p < 0.001, ΔS-Bχ^2 = 154.99 [df = 8], p < 0.001, RMSEA = 0.066 [90% CI 0.061–0.070], CFI = 0.957, ΔCFI = −0.011, TLI = 0.951, SRMR = 0.036) and model 2A (χ^2 = 736.78 [df = 64], p < 0.001, ΔS-Bχ^2 = 149.48 [df = 7], p < 0.001, RMSEA = 0.069 [90% CI 0.064–0.073], CFI = 0.953, ΔCFI = −0.011, TLI = 0.947, SRMR = 0.037) (not shown in Table 5). ΔCFI was equal to −0.007 for model 2B, −0.009 for model 2C, −0.010 for model 2D, and −0.008 for model 2E. These results suggest scalar noninvariance for the one-factor model and model 2A, full MI for models 2B and 2E, and partial MI for models 2C and 2D for the two groups of non-Hispanic AI/AN adults.

Due to the poor fit of the original scalar models for the one-factor model and model 2A, we pursued partial scalar MI for these two models. We relaxed the intercept for item 5 for the group of non-Hispanic AI/AN adults from healthcare system B in the one-factor model. Item 5 represented the greatest drop in χ^2 based on intercept modification indexes. Doing so improved ΔCFI to −0.008 (Table 5). We also relaxed the intercept for item 6 for the group of non-Hispanic AI/AN adults from healthcare system B in model 2A, which improved ΔCFI to −0.007 (Table 5). These findings suggest that the one-factor model and model 2A are partially scalar invariant between the two groups of non-Hispanic AI/AN adults

Table 2
Sample Demographic Information, including by Racial and Ethnic Group.

Demographics	Non-Hispanic AI/AN						Non-Hispanic									
	Healthcare System A (*n* = 1759)		Healthcare System B (*n* = 2701)		Hispanic (*n* = 7974)		White (*n* = 7974)		Asian (*n* = 6988)		Black (*n* = 6213)		NH/PI (*n* = 1370)		Full Sample (*n* = 34,979)	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Age group:																
18–29	578	32.9	758	28.1	2692	33.8	2065	25.9	2344	33.5	2103	33.8	530	38.7	11,070	31.6
30–44	611	34.7	863	32.0	2679	33.6	2449	30.7	2279	32.6	1975	31.8	515	37.6	11,371	32.5
45–64	570	32.4	1080	40.0	2603	32.6	3460	43.4	2365	33.8	2135	34.4	325	23.7	12,538	35.8
Sex:																
Male	494	28.1	757	28.0	2334	29.3	2633	33.0	2035	29.1	1909	30.7	467	34.1	10,629	30.4
Female	1265	71.9	1944	72.0	5640	70.7	5341	67.0	4953	70.9	4304	69.3	903	65.9	24,350	69.6
Department:																
Primary care[a]	1506	85.6	1759	65.1	5107	64.0	5010	62.8	4562	65.3	3922	63.1	884	64.5	22,750	65.0
Behavioral health	253	14.4	942	34.9	2867	36.0	2964	37.2	2426	34.7	2291	36.9	486	35.5	12,229	35.0

Note. AI/AN = American Indian/Alaska Native. NH/PI = Native Hawaiian/Pacific Islander.

[a] Includes family practice and internal medicine.

Table 3
PHQ-9 Descriptive Item Data by Racial and Ethnic Group.

PHQ-9 Items	Non-Hispanic AI/AN						Non-Hispanic							
	Healthcare System A (*n* = 1759)		Healthcare System B (*n* = 2701)		Hispanic (*n* = 7974)		White (*n* = 7974)		Asian (*n* = 6988)		Black (*n* = 6213)		NH/PI (*n* = 1370)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	1.62	1.19	1.45	1.09	1.41	1.07	1.30	1.08	1.42	1.06	1.50	1.09	1.53	1.07
2	1.64	1.20	1.54	1.06	1.52	1.05	1.39	1.05	1.53	1.04	1.59	1.07	1.61	1.06
3	1.94	1.20	1.96	1.10	1.87	1.11	1.75	1.13	1.78	1.13	1.98	1.12	1.94	1.11
4	1.89	1.18	1.92	1.05	1.85	1.05	1.77	1.06	1.83	1.04	1.88	1.06	1.96	1.02
5	1.41	1.26	1.57	1.16	1.52	1.16	1.31	1.15	1.34	1.14	1.61	1.16	1.68	1.15
6	1.37	1.25	1.47	1.16	1.45	1.15	1.32	1.13	1.44	1.12	1.49	1.16	1.57	1.14
7	1.48	1.27	1.36	1.16	1.38	1.15	1.20	1.12	1.31	1.13	1.36	1.15	1.43	1.14
8	1.09	1.21	0.90	1.07	0.89	1.07	0.71	0.99	0.85	1.05	0.88	1.07	1.04	1.11
9	0.41	0.86	0.38	0.76	0.39	0.77	0.31	0.68	0.43	0.78	0.41	0.79	0.51	0.87
Total score	12.84	7.96	12.56	6.89	12.28	6.82	11.06	6.69	11.92	6.69	12.69	6.83	13.27	6.93

Note. AI/AN = American Indian/Alaska Native. *M* = Mean. NH/PI = Native Hawaiian/Pacific Islander. *SD* = Standard deviation.

Table 4
PHQ-9 Item-Total Correlations and Internal Consistency Reliability Statistics by Racial and Ethnic Group.

	Non-Hispanic AI/AN			Non-Hispanic			
PHQ-9 Items	Healthcare System A (n = 1759)	Healthcare System B (n = 2701)	Hispanic (n = 7974)	White (n = 7974)	Asian (n = 6988)	Black (n = 6213)	NH/PI (n = 1370)
1	0.81	0.77	0.76	0.78	0.72	0.74	0.74
2	0.84	0.79	0.79	0.81	0.78	0.79	0.82
3	0.74	0.67	0.64	0.65	0.63	0.67	0.67
4	0.78	0.71	0.71	0.71	0.71	0.72	0.72
5	0.74	0.68	0.70	0.67	0.67	0.68	0.67
6	0.83	0.75	0.75	0.74	0.73	0.73	0.77
7	0.76	0.69	0.67	0.67	0.67	0.66	0.69
8	0.72	0.64	0.61	0.61	0.62	0.62	0.63
9	0.63	0.60	0.59	0.61	0.57	0.58	0.60
Ordinal α	0.94	0.91	0.91	0.91	0.91	0.91	0.91
Cronbach's α	0.90	0.88	0.88	0.88	0.87	0.87	0.88

Note. AI/AN = American Indian/Alaska Native. NH/PI = Native Hawaiian/Pacific Islander. Aside from Cronbach's α, results presented are based on polychoric correlation matrices.

Table 5
MGCFA between Non-Hispanic AI/AN Adults at Healthcare Systems A (n = 1759) and B (n = 2701).

Models		χ^2 (*df*)	ΔS-Bχ^2 (*df*)	RMSEA (90% CI)	CFI	ΔCFI	TLI	SRMR
1[a]	Configural	483.22 (48)***		0.064 (0.059–0.069)	0.969		0.954	0.030
	Metric	513.69 (56)***	38.17 (8)***	0.061 (0.056–0.065)	0.968	–0.001	0.958	0.031
	Scalar	629.77 (63)***	111.10 (7)***	0.064 (0.059–0.068)	0.960	–0.008	0.960	0.954
2A[b]	Configural	520.90 (50)***		0.065 (0.060–0.070)	0.967		0.952	0.031
	Metric	567.55 (57)***	33.83 (7)***	0.063 (0.059–0.068)	0.964	–0.003	0.954	0.033
	Scalar	668.71 (63)***	100.77 (6)***	0.066 (0.061–0.070)	0.957	–0.007	0.951	0.035
2B[c]	Configural	492.65 (50)***		0.063 (0.058–0.068)	0.969		0.955	0.031
	Metric	518.41 (57)***	32.38 (7)***	0.060 (0.056–0.065)	0.967	–0.002	0.959	0.032
	Scalar	633.55 (64)***	110.33 (7)***	0.063 (0.059–0.068)	0.960	–0.007	0.955	0.035
2C[d]	Configural	431.64 (48)***		0.060 (0.055–0.065)	0.973		0.959	0.028
	Metric	462.74 (55)***	35.91 (7)***	0.058 (0.053–0.063)	0.971	–0.002	0.962	0.030
	Scalar	624.57 (62)***	151.77 (7)***	0.064 (0.059–0.068)	0.960	–0.009	0.954	0.033
2D[e]	Configural	526.52 (49)***		0.066 (0.061–0.071)	0.966		0.950	0.031
	Metric	549.78 (56)***	30.82 (7)***	0.063 (0.058–0.068)	0.965	–0.001	0.955	0.031
	Scalar	698.87 (63)***	142.33 (7)***	0.067 (0.063–0.072)	0.955	–0.010	0.949	0.036
2E[f]	Configural	295.20 (26)***		0.068 (0.061–0.075)	0.975		0.959	0.027
	Metric	309.57 (31)***	19.01 (7)**	0.063 (0.057–0.070)	0.974	–0.001	0.965	0.027
	Scalar	403.96 (36)***	90.59 (5)***	0.068 (0.062–0.074)	0.966	–0.008	0.960	0.031

Note. ***p < 0.001. **p < 0.01. AI/AN = American Indian/Alaska Native. Preferred fits: non-significant χ^2 and ΔS-Bχ^2, RMSEA < 0.05 or at most < 0.08, CFI and TLI > 0.95 or near 1.00, ΔCFI no more than –0.01, and SRMR < 0.08 (Hu and Bentler, 1999; Satorra and Bentler, 2001; Vandenberg and Lance, 2000).

[a] Kroenke et al. (2001). Correlating item 7–8 residuals for both groups; Healthcare System A: 3–4; Healthcare System B: 2–5, 3–4, 6–9. Intercept for item 5 relaxed for Healthcare System B.

[b] Richardson and Richards (2008). Correlating item residuals: Healthcare System A: 3–4; Healthcare System B: 7–8. Intercept for item 6 relaxed for Healthcare System B.

[c] Krause et al. (2008). Correlating item 7–8 residuals for both groups.

[d] Krause et al. (2010). Correlating item 7–8 residuals for both groups; Healthcare System A: 3–4; Healthcare System B: 1–2.

[e] de Jonge et al. (2007). Correlating item 7–8 residuals for both groups; Healthcare System B: 4~8.

[f] Granillo (2012).

4.3.2. Non-Hispanic AI/AN adults compared to all other racial and ethnic groups

We pursued partial MI for the one-factor model and models 2A, 2C, and 2D between separate AI/AN groups and other racial and ethnic groups (Table 6). This included relaxing the intercept in the scalar model for item 5 in the AI/AN group from healthcare system B for the one-factor model and item 6 for the same group in model 2A. We pursued full MI for model 2B and partial MI for model 2E, both of which included a combined group of AI/AN adults compared to other racial and ethnic groups (Table 6). Based on ΔCFI, the one-factor model and models 2A, 2C, and 2D were partially cross-culturally invariant between separate AI/AN groups and other racial and ethnic groups. Model 2B was fully invariant and model 2E was partially invariant between the combined group of AI/AN adults and other racial and ethnic groups. ΔS-Bχ^2 was statistically significant between all nested models. However, ΔCFI was less than –0.01 in all cases, suggesting full MI for model 2B and partial MI for model 2E and the one-factor model. Other measures of fit were good (CFI > 0.94, SRMR < 0.04) or acceptable (RMSEA < 0.08) for configural, metric, and scalar models.

5. Discussion

In this study, we assessed the cross-cultural MI of previously identified one- and two-factor PHQ-9 models between two geographically distant groups of non-Hispanic AI/AN peoples and other racially and ethnically diverse adults. All had a mental health or substance use disorder diagnosis and were seen in either primary care or behavioral health settings within two healthcare systems. Results showed that internal consistency reliability was good to excellent for the standard one-factor model. Also, factor correlations were high (> 0.85) for all two-factor models, which could represent multicollinearity (Brown, 2015), and may support a one-factor PHQ-9 model (Harry and Waring, 2019). Furthermore, factors loaded well together regardless of structure, al-

Table 6
MGCFA MI between Non-Hispanic AI/AN Adults and Other Racial and Ethnic Groups at Healthcare System B.

Models	χ^2 (*df*)	ΔS-Bχ^2 (*df*)	RMSEA (90% CI)	CFI	ΔCFI	TLI	SRMR
Separate AI/AN Groups:							
1[a]							
Configural	4446.69 (169)***		0.071 (0.069–0.073)	0.959		0.938	0.032
Metric	4970.87 (217)***	629.37 (48)***	0.066 (0.065–0.068)	0.954	−0.005	0.947	0.036
Scalar	5747.09 (264)***	818.11 (47)***	0.064 (0.063–0.066)	0.947	−0.007	0.949	0.038
2A[b]							
Configural	4081.93 (176)***		0.067 (0.065–0.068)	0.962		0.946	0.031
Metric	4547.01 (218)***	496.95 (42)***	0.063 (0.061–0.065)	0.958	−0.004	0.952	0.035
Scalar	5224.08 (259)***	668.74 (41)***	0.062 (0.060–0.063)	0.952	−0.006	0.953	0.037
2C[c]							
Configural	3921.17 (172)***		0.066 (0.064–0.068)	0.964		0.947	0.029
Metric	4313.99 (214)***	469.63 (42)***	0.062 (0.060–0.064)	0.960	−0.004	0.953	0.033
Scalar	4994.69 (256)***	699.15 (42)***	0.061 (0.059–0.062)	0.954	−0.006	0.955	0.035
2D[d]							
Configural	4270.93 (174)***		0.069 (0.067–0.070)	0.960		0.943	0.032
Metric	4658.43 (216)***	469.89 (42)***	0.064 (0.063–0.066)	0.957	−0.003	0.950	0.035
Scalar	5327.66 (258)***	694.37 (42)***	0.063 (0.061–0.064)	0.951	−0.006	0.052	0.036
Combined AI/AN Groups:							
2B[e]							
Configural	3848.45 (150)***		0.065 (0.063–0.067)	0.964		0.949	0.031
Metric	4200.18 (185)***	222.49 (35)***	0.061 (0.059–0.063)	0.961	−0.003	0.955	0.034
Scalar	4720.53 (220)***	491.19 (35)***	0.059 (0.058–0.061)	0.957	−0.002	0.958	0.035
2E[f]							
Configural	2107.68 (77)***		0.067 (0.065–0.070)	0.974		0.957	0.025
Metric	2379.82 (102)***	306.69 (25)***	0.062 (0.060–0.064)	0.970	−0.004	0.963	0.027
Scalar	2755.51 (127)***	383.67 (25)***	0.060 (0.058–0.062)	0.966	−0.004	0.966	0.029

Note. ***$p < 0.001$. AI/AN = American Indian/Alaska Native. NH/PI = Native Hawaiian/Pacific Islander. AI/AN combined $n = 4460$, Healthcare System A $n = 1759$, Healthcare System B $n = 2701$; White $n = 7974$; Hispanic $n = 7974$; Asian $n = 6988$; Black $n = 6213$; NH/PI $n = 1370$. Preferred fits: non-significant χ^2 and ΔS-Bχ^2, RMSEA < 0.05 or at most < 0.08, CFI and TLI > 0.95 or near 1.00, ΔCFI no more than −0.01, and SRMR < 0.08 (Hu and Bentler, 1999; Satorra and Bentler, 2001; Vandenberg and Lance, 2000).

[a] Kroenke et al. (2001). Correlating item residuals: 7–8 for all groups; AI/AN Healthcare System A: 3–4, 7–8, Healthcare System B: 2–5, 3–4, 6–9, 7–8; White: 1–2, 3–4, 7–8; Hispanic: 2–5, 3–4, 7–8; Asian: 3–4, 7–8; Black: 1–2, 2–6, 7–8; NH/PI: 1–2, 2–6, 7–8. Relaxed intercept for item 5 for the AI/AN group from Healthcare System B in the scalar model.

[b] Richardson and Richards (2008). Correlating item residuals: AI/AN Healthcare System A: 3–4, Healthcare System B: 7–8; White: 7–8; Hispanic: 7–8; Black: 7–8; NH/PI: 7–8. Intercept for item 6 relaxed for the AI/AN group from Healthcare System B.

[c] Krause et al. (2010). Correlating item residuals: AI/AN Healthcare System A: 3–4, 7–8, Healthcare System B: 1–2, 7–8; White: 1–2, 7–8; Hispanic: 1–2; Asian: 7–8; Black: 7–8; NH/PI: 7–8.

[d] de Jonge et al. (2007). Correlating item residuals: All groups: 7–8; AI/AN Healthcare System B: 4–8.

[e] Krause et al. (2008). Correlating item residuals: All groups: 7–8.

[f] Granillo (2012). Correlating item residuals: NH/PI: 1–4.

though item 9 consistently loaded lower than other items, potentially signifying the uniqueness of this item. However, most models were a poor fit without modification based on RMSEA exceeding 0.10 for most groups and models (Browne and Cudeck, 1993; Browne and Mels, 1990; Steiger, 1989, as cited by Hu and Bentler, 1999). RMSEA is sensitive to misspecification of model factor loadings (Hu and Bentler, 1999; Vandenberg and Lance, 2000), although the acceptable loadings reported here argue against misspecification. Consequently, we pursued partial measurement invariance for most models by correlating residuals for items representing the greatest drop in χ^2 for individual racial and ethnic groups (Byrne, 2012; Byrne et al., 1989). We found that the one-factor model and two-factor model 2A exhibited scalar noninvariance between the two groups of non-Hispanic AI/AN adults ΔCFI exceeded −0.01, suggesting differences in intercepts. However, the one-factor model and model 2A both appeared partially scalar invariant between the two AI/AN groups when relaxing an item intercept. Models 2C and 2D also appeared to be partially invariant and models 2B and 2E fully invariant between the two groups of non-Hispanic AI/AN adults. We combined AI/AN groups for further cross-cultural assessment in models 2B and 2E due to these models having full MI between the two groups of AI/AN peoples, and left the two AI/AN groups separate when testing partial MI for other models. Based on ΔCFI −0.01, we found partial MI for the one-factor model and models 2A, 2C, and 2D when including separate AI/AN groups and all other racial and ethnic groups in our analyses. We also found full (model 2B) and partial (model 2E) MI between the combined AI/AN group and all other racial and ethnic groups.

Research has shown that two-factor PHQ-9 models can have better fits than the one-factor model for patients with depression or those with physical health problems (González-Blanch et al., 2018). Both González-Blanch and colleagues (2018) and Richardson and Richards (2008) note it is unclear whether two-factor models identify separate underlying latent constructs related to depression or alternatively identify somatic symptomology from physical health problems; perhaps it is a combination of both. However, research has shown that two-factor models tend to have improved fits over the one-factor model in cross-cultural MI testing between different racial and ethnic groups (Granillo, 2012; Harry and Waring, 2019; Keum et al., 2018; Patel et al., 2019). Our achievement of partial MI with the one-factor model between separate AI/AN groups and also across all other groups in the present study suggests that the one-factor model may require modification before mean scores can be compared between culturally diverse groups. This is particularly important for healthcare systems to consider as they focus on issues of health equity with the populations they serve, such as identifying and addressing social determinants of health.

When making cross-cultural comparisons in PHQ-9 scores, clinicians and researchers should first assess models to find the best fitting factor model(s) for separate groups (Byrne, 2012; Byrne et al., 1989), test those models for cross-cultural MI, then compare summed, or composite, mean PHQ-9 scores for models that are fully scalar invariant (Steinmetz, 2013). Most models in the present study required some degree of modification, such as correlating like-item residuals or relaxing an intercept. However, other research has found unmodified one-factor (Harry and Waring, 2019; Keum et al., 2018; Merz et al., 2011; Patel et al., 2019) and two-factor models (Granillo, 2012; Harry and Waring, 2019; Keum et al., 2018; Patel et al., 2019) to be cross-culturally measurement invariant between U.S. racial and ethnic groups using various estimators and changes in fit indexes. When considering AI/AN peoples, Harry and Waring (2019) evaluated the MI of the PHQ-9 between a general patient population of AI/AN and non-Hispanic White adults with WLSMV estimation in R version 3.4.4 (R Core Team, 2018) using the Lavaan package version 0.6–2 (Rosseel, 2018). They showed that the χ^2 difference test used, specifically Satorra and Bentler's (2001) scaled ΔS-Bχ^2 (the Mplus DIFFTEST was not available in Lavaan), was statistically significant in all cases (Harry and Waring, 2019). However, nested models had other good to excellent goodness-of-fit indexes (e.g., CFI > 0.99, RMSEA < 0.06), with ΔCFI < −0.01 and ΔRMSEA < 0.05 between configural and metric models and < 0.01 between metric and scalar models (Harry and Waring, 2019; Rutkowski and Svetina, 2017). These results support the cross-cultural MI of the PHQ-9 between the two groups in that study, including the standard one-factor model (Harry and Waring, 2019). Our findings of partial MI for the one-factor PHQ-9 model in the present study suggest that more research is needed on the differential item functioning of the PHQ-9 with AI/AN groups and other racial and ethnic groups to further assess if and how item responses differ.

Regarding clinical use of the PHQ-9, aside from high RMSEA values, the one-factor model showed good fits for all study groups individually. Together with the high factor correlations in two-factor models (Boothroyd et al., 2019; Brown, 2015; Harry and Waring, 2019), this suggests that clinicians and healthcare systems and plans can still employ the one-factor model used in standard practice when calculating PHQ-9 scores to examine depression treatment response and depression remission within individuals over time. Indeed, Boothroyd et al. (2019) recommends use of a one-factor model with the PHQ-9 due to high factor correlations and "that separately assessing factors will not provide any useful information for the majority of patients" (p. 533).

In this study, we were unable to state whether the PHQ-9 captured all aspects of depression for AI/AN peoples. As noted in prior research (Harry and Crea, 2018; Harry and Waring, 2019), some groups of AI/AN peoples may experience loneliness as a symptom of depression (Armenta et al., 2014; Beals et al., 2003; O'Nell, 2004), which the PHQ-9 does not specifically assess. Future community- and tribal-based participatory research is needed (Skewes et al., 2020; Walls et al., 2017). This research could employ the framework for guiding measurement with AI/AN populations described by Walls et al. (2017) in determining if the PHQ-9 is an appropriate common measure for use in screening for depression with AI/AN adults or if more tailored instrument(s) would best capture depression symptomology for AI/AN communities and specific tribal and cultural groups.

5.1. *Limitations*

This retrospective study was limited in that data were included from two secondary datasets extracted from EHR data for other studies at two healthcare systems. The sample included those who were either seeking treatment for mental health diagnoses or substance use disorders or recognized by a primary care provider as having a mental health condition or substance use disorder. Consequently, our results might not apply to people with unrecognized mental health or substance use problems or people who do not seek treatment. Furthermore, the presence of a mental health or substance use disorder diagnosis in the medical record may vary based on race and ethnicity due to differing rates of healthcare utilization for mental health or substance use-related needs. We were also unable to differentiate mental health or substance use disorder diagnosis type for all study groups. Results may differ between diagnosis groups, as mental health and substance use disorder diagnoses are broad, and depression is not a universal symptom. However, the PHQ-9 is regularly applied in the general population for depression screening, where not all patients have a depression diagnosis.

Differently sized samples may have adversely affected study results. We also only included non-Hispanic AI/AN adults from healthcare system A. Moreover, grouping individuals by self-identified race does not automatically make individuals within these racial groupings equivalent or homogenous (Han et al., 2019). Until healthcare systems adopt broader options for the self-identification of race and ethnicity among patient populations, we will be unable to make finer cultural distinctions between different populations of a single racial or ethnic group. This includes the diverse and culturally rich AI/AN peoples. Healthcare systems do often allow for the selection of multiple racial and ethnic groups, which healthcare system researchers can use to identify multiracial groups. Unfortunately, this does little to aid in identifying cultural within-group differences. More prospective research is needed in this area.

Additionally, we were unable to recode item 9 as a dichotomous variable in our analyses. This is because Mplus multiple-indicator, multiple-causes mixture models do not compare all parameters tested in MGCFA. Furthermore, we did not specifically assess how item 9, which has been shown to predict suicide risk (Coleman et al., 2018; Simon et al., 2013, 2016), functioned between study groups. Future research could assess this. We found that most items loaded well on one factor, aside from item 9, similar to Harry and Waring (2019). Further research is needed on the differential item functioning of individual items between diverse groups. Understanding differential item functioning would assist researchers and clinicians in adapting models to fit the groups they want to compare cross-culturally. Similarly, due to space constraints, we did not explore other factor models between study groups. We also did not compare means for summed PHQ-9 total scores between groups due to the one-factor model exhibiting only partial scalar invariance between the two AI/AN groups (Steinmetz, 2013). MI analyses are also unable to ascertain forms of response bias either between or within groups (Steinmetz, 2013), another potential area for future research.

6. Conclusion

To ensure health equity in MBC, standardized instruments like the PHQ-9 should function similarly with diverse cultural groups. In this study, we compared the cross-cultural MI of the PHQ-9 depression screener between two geographically distant groups of non-Hispanic AI/AN adults, as well as other diverse racial and ethnic groups, all of whom had mental health or substance use disorder diagnoses. We found support for the partial cross-cultural MI of the one-factor PHQ-9 model and partial or full MI for five two-factor models. However, more research is needed on how the PHQ-9 functions with diverse populations. This includes evaluating the differential item functioning of PHQ-9 items, as well as assessing differences in item 9 and suicide risk. Other instruments or modified versions of the PHQ-9 may also be a better fit for screening for depression in different groups of AI/AN peoples and other racial and cultural groups, which future community-based participatory research should assess (Skewes et al., 2020).

Contributors

MLH conceptualized the study, drafted the manuscript, and conducted the data analyses. RYC, SCW, and GES contributed to the conceptualization of the study and development of the manuscript. All authors approved the final manuscript.

Declaration of Competing Interest

All authors declare that they have no conflicts of interest.

Funding statement

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

Acknowledgments

Essentia Institute of Rural Health and Kaiser Permanente Washington Health Research Institute.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jadr.2021.100121.

References

Ahmedani, B.K., Simon, G.E., Stewart, C., Beck, A., Waitzfelder, B.E., Rossom, R., Lynch, F., Owen-Smith, A., Hunkeler, E.M., Whiteside, U., Operskalski, B.H., Coffey, M.J., Solberg, L.I., 2014. Health care contacts in the year before suicide death. J. Gen. Intern. Med. 29 (6), 870–877. doi:10.1007/s11606-014-2767-3.

Ahmedani, B.K., Westphal, J., Autio, K., Elsiss, F., Peterson, E.L., Beck, A., Waitzfelder, B.E., Rossom, R.C., Owen-Smith, A.A., Lynch, F., Lu, C.Y., Frank, C., Prabhakar, D., Braciszewski, J.M., Miller-Matero, L.R., Yeh, H.H., Hu, Y., Doshi, R., Waring, S.C., Simon, G.E., 2019. Variation in patterns of health care before suicide: a population case-control study. Prev. Med. 127, 105796. doi:10.1016/j.ypmed.2019.105796.

Armenta, B.E., Sittner Hartshorn, K.J., Whitbeck, L.B., Crawford, D.M., Hoyt, D.R., 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.

Asdigian, N.L., Running Bear, U., Beals, J., Manson, S.M., Kaufman, C.E., 2018. Mental health burden in a national sample of American Indian and Alaska Native adults: differences between multiple-race and single-race subgroups. Soc. Psychiatry Psychiatr. Epidemiol. 53, 521–530. doi:10.1007/s00127-018-1494-1.

Baas, K.D., Cramer, A.O.J., Koeter, M.W.J., van de Lisdonk, E.H., van Weert, H.C., Schene, A.H., 2011. Measurement invariance with respect to ethnicity of the Patient Health Questionnaire-9 (PHQ-9. J. Affect Disord. 129, 229–235. doi:10.1016/j.jad.2010.08.026.

Beals, J., Manson, S.M., Mitchell, C.M., Spicer, P., Team AI-SUPERPFP, 2003. Cultural specificity and comparison in psychiatric epidemiology: walking the tightrope in American Indian research. Cult. Med. Psychiatry 27 (3), 259–289. doi:10.1023/a:1025347130953.

Boothroyd, L., Dagnan, D., Muncer, S., 2019. PHQ-9: one factor or two? Psychiatry Res. 271, 532–534. doi:10.1016/j.psychres.2018.12.048.

Bowen, D.J., Powers, D.M., Russo, J., Arao, R., LePoire, E., Sutherland, E., Ratzliff, A.D.H., 2020. Implementing collaborative care to reduce depression for rural native American/Alaska native people. BMC Health Serv. Res. 20 (1), 34. doi:10.1186/s12913-019-4875-6.

Brave Heart, M.Y.H., DeBruyn, L.M., 1998. The American Indian Holocaust: healing historical unresolved grief. Am. Indian Alsk. Native Ment. Health Res. 8 (2), 56–78.

Brown, T.A., 2015. Confirmatory Factor Analysis for Applied Research, second ed. Guilford Press, New York.

Browne, M.W., Cudeck, R., 1993. Alternative ways of assessing model fit. In: K.A., Bollen, J.S., Long (Eds.), Testing structural equation models. Sage, Newbury Park, CA, pp. 136–162.

Browne, M.W., Mels, G., 1990. RAMONA user's guide. Unpublished report, Department of Psychology, The Ohio State University, Columbu.

Byrne, B.M., 2012. Structural Equation Modeling With Mplus: Basic Concepts, Applications, and Programming, first ed. Taylor & Francis Group, LLC, New York.

Byrne, B.M., Shavelson, R.J., Muthén, B., 1989. Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. Psychol. Bull. 105 (3), 456–466.

Carron, R., 2020. Health disparities in American Indians/Alaska Natives: implications for nurse practitioners. Nurse Pract. 45 (6), 26–32. doi:10.1097/01.NPR.0000666188.79797.a7.

Chesney, E., Goodwin, G.M., Fazel, S., 2014. Risks of all-cause and suicide mortality in mental disorders: a meta-review. World Psychiatry 13, 153–160.

Cheung, G.W., Rensvold, R.B., 2002. Evaluating goodness-of-fit indexes for testing measurement invariance. Struct. Eq. Model. 9, 233–255. doi:10.1207/S15328007SEM0902_5.

Cleary, L.M., 2013. Cross-Cultural Research with Integrity: Collected Wisdom from Researchers in Social Settings. Palgrave Macmillan, New York.

Coleman, K.J., Johnson, E., Ahmedani, B.K., Beck, A., Rossom, R.C., Shortreed, S.M., Simon, G.E., 2018. Predicting suicide attempts for racial and ethnic groups of patients during routine clinical care. Suicide Life Threat. Behav. 49 (3), 724–734. doi:10.1111/sltb.12454.

Coley, R.Y., Boggs, J.M., Beck, A., Hartzler, A.L., Simon, G.E., 2020. Defining success in measurement-based care for depression: a comparison of common metrics. Psychiatr. Serv. 71 (4), 312–318. doi:10.1176/appi.ps.201900295.

Cronbach, L.J., 1951. Coefficient alpha and the internal structure of tests. Psychometrika 16, 297–334.

Curtin, S.C., Hedegaard, H., 2019. Suicide rates for females and males by race and ethnicity: United States, 1999 and 2017. https://www.cdc.gov/nchs/data/hestat/suicide/rates_1999_2017.pdf (Accessed 2 September 2020).

de Jonge, P., Mangano, D., Whooley, M.A., 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.

Dreher, A., Hahn, E., Diefenbacher, A., Nguyen, M.H., Böge, K., Burian, H., Dettling, M., Burian, R., Tam Ta, T.M., 2017. Cultural differences in symptom representation for depression and somatization measured by the PHQ between Vietnamese and German psychiatric outpatients. J. Psychosom. Res. 102, 71–77. doi:10.1016/j.jpsychores.2017.09.010.

Elhai, J.D., Contractor, A.A., Tamburrino, M., Fine, T.H., Prescott, M.R., Shirley, E., Chan, P.K., Slembarski, R., Liberzon, I., Galea, S., Calabrese, J.R., 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. doi:10.1016/j.psychres.

Garrett, M.D., Baldridge, 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.

GBD 2017 Disease and Injury Incidence and Prevalence Collaborators, 2018. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 392, 1789–1858. doi:10.1016/S0140-6736(18)32279-7.

González-Blanch, C., Medrano, L.A., Muñoz-Navarro, R., Ruíz-Rodíguez, P., Moriana, J.A., Limonero, J.T., Schmitz, F., Cano-Vindel, A.on behalf of the 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.

Granillo, T.M., 2012. Structure and function of the Patient Health Questionnaire-9 among Latina and Non-Latina White female college students. J. Soc. Soc. Work Res. 3 (2), 80–93. doi:10.5243/jsswr.2012.6.

Han, K., Colarelli, S.M., Weed, N.C., 2019. Methodological and statistical advances in the consideration of cultural diversity in assessment: a critical review of group classification and measurement invariance testing. Psychol. Assess. 31 (12), 1481–1496. doi:10.13140/RG.2.2.16638.23368.

Harry, M.L., Crea, T.M., 2018. Examining the measurement invariance of a Modified CES-D for American Indian and Non-Hispanic White adolescents and young adults. Psychol. Assess. 30 (8), 1107–1120. doi:10.1037/pas0000553.

Harry, M.L., Waring, S.C., 2019. The measurement invariance of the Patient Health Questionnaire-9 for American Indian adults. J. Affect. Disord. 254, 59–68. doi:10.1016/j.jad.2019.05.017.

Hedegaard, H., Curtin, S.C., Warner, M., 2018. Suicide Mortality in the United States, 1999–2017. NCHS Data Brief, no. 330. National Center for Health Statistics, Hyattsville, MD https://www.cdc.gov/nchs/products/databriefs/db330.htm.

Huang, F.Y., Chung, H., Kroenke, K., Delucchi, K.L., Spitzer, R.L., 2006. Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. J. Gen. Intern. Med. 21 (6), 547–552. doi:10.1111/j.1525-1497.2006.00409.x.

Hu, L.-t., Bentler, P.M., 1999. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct. Eq. Model. 6 (1), 1–55. doi:10.1080/10705519909540118.

Indian Affairs Bureau, 2020. Indian entities recognized by and eligible to receive services from the United States Bureau of Indian Affairs. https://www.federalregister.gov/documents/2020/01/30/2020-01707/indian-entities-recognized-by-and-eligible-to-receive-services-from-the-united-states-bureau-of (Accessed 2 September 2020).

Indian Health Services, 2019. Disparities. https://www.ihs.gov/newsroom/factsheets/disparities/ (Accessed 29 December 2020).

Jones, D.S., 2006. The persistence of American Indian health disparities. Am. J. Public Health. 96 (12), 2122–2134. doi:10.2105/AJPH.2004.054262.

Keum, B., Miller, M.J., 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. doi:10.1037/pas0000550.

Kim, Y.E., Lee, B., 2019. The psychometric properties of the Patient Health Questionnaire-9 in a sample of Korean university students. Psychiatry Investig. 16 (12), 904–910. doi:10.30773/pi.2019.0226.

Kisely, S., Katarzyna Alichniewicz, K., Black, E.B., 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.

Krause, J.S., Bombardier, C., Carter, R.E., 2008. Assessment of depressive symptoms during inpatient rehabilitation for spinal cord injury: is there an underlying somatic factor when using the PHQ? Rehabil. Psychol. 53 (4), 513–620. doi:10.1037/a0013354.

Krause, J.S., Reed, K.S., McArdle, J.J., 2010. Factor structure and predictive validity of somatic and nonsomatic symptoms from the Patient Health Questionnaire-9: a longitudinal study after spinal cord injury. Arch. Phys. Med. Rehabil. 91, 1218–1224.

Kroenke, K., Spitzer, R.L., Williams, J.B., 2001. The PHQ-9: validity of a brief depression severity measure. J. Gen. Intern. Med. 16 (9), 606–613.

Li, C.H., 2016. Confirmatory factor analysis with ordinal data: comparing robust maximum likelihood and diagonally weighted least squares. Behav. Res. Methods 48, 936–949. doi:10.3758/s13428-015-0619-7.

Livingston, R., Daily, R.S., Guerrero, A.P.S., Walkup, J.T., Novins, D.K., 2019. No Indians to spare: depression and suicide in Indigenous American children and youth. Child Adolesc. Psychiatr. Clin. N. Am. 28 (3), 497–507. doi:10.1016/j.chc.2019.02.015.

Meredith, W., 1993. Measurement invariance, factor analysis and factorial invariance. Psychometrika 58, 525–543. doi:10.1007/BF02294825.

Merz, E.L., Malcarne, V.L., Roesch, S.C., Riley, N., Sadler, G.R., 2011. A multigroup confirmatory factor analysis of the Patient Health Questionnaire-9 among English and Spanish-speaking Latinas. Cult. Divers. Ethnic Minor. Psychol. 17 (3), 309–316. doi:10.1037/a0023883.

Milfont, T.L., Fischer, R., 2010. Testing measurement invariance across groups: applications in cross-cultural research. Int. J. Psychol. Res. 3, 111–121. doi:10.21500/20112084.857.

Mitchell, A.J., 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. doi:10.1192/bjpo.bp.115.001685.

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 2 September 2020.

Muthén, L.K., Muthén, B.O., 1998-2017. Mplus User's Guide, eighth ed. Muthén & Muthén, Los Angeles, CA.

Muthén, L.K., Muthén, B.O., 2019. Mplus version 8.3. Muthén & Muthén, Los Angeles, CA. [Software].

National Committee for Quality Assurance, 2020. HEDIS Depression measures specified for electronic clinical data systems. https://www.ncqa.org/hedis/the-future-of-hedis/hedis-depression-measures-specified-for-electronic-clinical-data/ (Accessed 2 September 2020).

O'Nell, T., 2004. Culture and pathology: flathead loneliness revisited. Cult. Med. Psychiatry 28, 221–230.

Patel, J.S., Oh, Y., Rand, K.L., Wu, W., Cyders, M.A., Kroenke, K., Stewart, J.C., 2019. 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. Depress. Anxiety 36, 813–823. doi:10.1002/da.22940.

Putnick, D.L., Bornstein, M.H., 2016. Measurement invariance conventions and reporting: the state of the art and future directions for psychological research. Dev. Rev. 41, 71–90. doi:10.1016/j.dr.2016.06.004, https://doi.org/:.

R Core Team, 2018. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria. [Software].

R Core Team, 2019. R: a Language and Environment for Statistical Computing. R Foundation For Statistical Computing Vienna, Austria. [Software].

Revelle, W., 2019. psych: Procedures for Personality and Psychological Research. Northwestern University, Evanston, IL Version = 1.9.12. [Software].

Richardson, E.J., Richards, J.S., 2008. Factor structure of the PHQ-9 screen for depression across time since injury among persons with spinal cord injury. Rehabili. Psychol. 53 (2), 243–249. doi:10.1037/0090-5550.53.2.243.

Rosseel, Y., 2018. Lavaan = 0.6.2 [Software].

Rutkowski, L., Svetina, D., 2017. Measurement invariance in international surveys: categorical indicators and fit measure performance. Appl. Meas. Educ. 30 (1), 39–51. doi:10.1080/08957347.2016.1243540.

Sass, D.A., 2011. Testing measurement invariance and comparing latent means within a confirmatory factor analysis framework. J. Psychoeduc. Assess. 29, 347–363.

Satorra, A., Bentler, P.M., 2001. A scaled difference chi-square test statistic for moment structure analysis. Psychometrika 66 (4), 507–514.

Schermelleh-Engel, K., Moosbrugger, E.L., 2003. Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. Methods Psychol. Res. Online 8, 23–74.

Scott, K., Lewis, C.C., 2015. Using measurement-based care to enhance any treatment. Cognit. Behav. Pract. 22 (1), 49–59. doi:10.1016/j.cbpra.2014.01.010.

Simon, G.E., Coleman, K.J., Rossom, R.C., Beck, A., Oliver, M., Johnson, E., Whiteside, U., Operskalski, B., Penfold, R.B., Shortreed, S.M., 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.

Simon, G.E., Rutter, C.M., Peterson, D., Oliver, M., Whiteside, U., Operskalski, B., Ludman, E.J., 2013. Does response on the PHQ-9 Depression Questionnaire predict subsequent suicide attempt or suicide death? Psychiatr. Serv. 64 (12), 1195–1202. doi:10.1176/appi.ps.201200587.

Skewes, M.C., Gonzalez, V.M., Gameon, J.A., FireMoon, P., Salois, E., Rasmus, S.M., Lewis, J.P., Gardner, S.A., Ricker, A., Reum, M., 2020. Health disparities research with American Indian communities: the importance of trust and transparency. Am. J. Community Psychol. 66 (3–4), 302–313. doi:10.1002/ajcp.12445.

Spitzer, R.L., Kroenke, K., Williams, J.B.W., 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.

Statmodel.com., n.d. Mplus: chi-square difference testing using the Satorra-Bentler scaled chi-square. https://www.statmodel.com/chidiff.shtml (Accessed 2 September 2020).

Steiger, J.H. (1989). EzPATH: A supplementary module for SYSTAT and SYGRAPH. Evanston, IL: SYSTAT.

Steenkamp, J.-B.E.M., Baumgartner, H., 1998. Assessing measurement invariance in cross-national consumer research. J. Consum. Res. 25, 78–107. doi:10.1086/209528.

Steinmetz, H., 2013. Analyzing observed composite differences across groups: is partial measurement invariance enough? Methodology 9 (1), 1–12.

Subotić, S., Knežević, I., Dimitrijević, S., Miholjčić, D., Šmit, S., Karać, M., Mijatović, J., 2015. The factor structure of the Patient Health Questionnaire (PHQ-9) in a non-clinical sample. J. Soc. Technol. Dev. 2015, 20–28.

Tran, T.V., Nguyen, T.H., Chan, K.T., 2017. Developing Cross-Cultural Measurement in Social Work Research and Evaluation, second ed. Oxford University Press, New York.

Vandenberg, R.J., Lance, C.E., 2000. A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. Organ. Res. Methods 3, 4–70. doi:10.1177/109442810031002.

van de Schoot, R., Lugtig, P., Hox, J., 2012. A checklist for testing measurement invariance. Eur. J. Dev. Psychol. 9 (4), 486–492. doi:10.1080/17405629.2012.686740.

Villarreal-Zegarra, D., Copez-Lonzoy, A., Bernabé-Ortiz, A., Melendez-Torres, G.J., Melendez-Torres, J.C., 2019. Valid group comparisons can be made with the Patient Health Questionnaire (PHQ-9): a measurement invariance study across groups by demographic characteristics. PLoS One 14 (9), e0221717. doi:10.1371/journal.pone.0221717.

Walls, M.L., Rumbaugh Whitesell, N., Barlow, A., Sarche, M., 2017. Research with American Indian and Alaska Native populations: measurement matters. J. Ethn. Subst. Abuse 18 (1), 129–149. doi:10.1080/15332640.2017.1310640.

Williams, D.R., Cooper, L.A., 2020. COVID-19 and health equity – a new kind of “herd immunity. JAMA 323 (24), 2478–2480. doi:10.1001/jama.2020.8051.

Zumbo, B.D., Gadermann, A.M., Zeisser, C., 2007. Ordinal versions of coefficients alpha and theta for likert rating scales. J. Modern Appl. Stat. Method 6 (1). doi:10.22237/jmasm/1177992180.