Research paper

Measurement invariance testing of the patient health questionnaire-9 (PHQ-9) across people with and without diabetes mellitus from the NHANES, EMHS and UK Biobank datasets

Arie Nouwen [a,b,1], Sonya S. Deschênes [c,1,*], Zhanna Balkhiyarova [d,e,f], Juan R. Albertorio-Díaz [g], Inga Prokopenko [d,h,i], Norbert Schmitz [j,k]

[a] *Middlesex University, London, UK*
[b] *The Open University, Milton Keynes, UK*
[c] *University College Dublin, Ireland*
[d] *University of Surrey, Guildford, UK*
[e] *Imperial College London, London, UK*
[f] *Bashkir State Medical University, Ufa, Russia*
[g] *Coalition to End Social Isolation and Loneliness DC, USA*
[h] *University of Lille, Lille, France*
[i] *Russian Academy of Sciences, Ufa, Russia*
[j] *McGill University, Montreal, Canada*
[k] *University of Tübingen, Tübingen, Germany*

ARTICLE INFO

ABSTRACT

Background: The prevalence of depression is higher among those with diabetes than in the general population. The Patient Health Questionnaire (PHQ-9) is commonly used to assess depression in people with diabetes, but measurement invariance of the PHQ-9 across groups of people with and without diabetes has not yet been investigated.

Methods: Data from three independent cohorts from the USA (n=1,886 with diabetes, n=4,153 without diabetes), Quebec, Canada (n= 800 with diabetes, n= 2,411 without diabetes), and the UK (n=4,981 with diabetes, n=145,570 without diabetes), were used to examine measurement invariance between adults with and without diabetes. A series of multiple group confirmatory factor analyses were performed, with increasingly stringent model constraints applied to assess configural, equal thresholds, and equal thresholds and loadings invariance, respectively. One-factor and two-factor (somatic and cognitive-affective items) models were examined.

Results: Results demonstrated that the most stringent models, testing equal loadings and thresholds, had satisfactory model fit in the three cohorts for one-factor models (RMSEA = .063 or below and CFI = .978 or above) and two-factor models (RMSEA = .042 or below and CFI = .989 or above).

Limitations: Data were from Western countries only and we could not distinguish between type of diabetes.

Conclusions: Results provide support for measurement invariance between groups of people with and without diabetes, using either a one-factor or a two-factor model. While the two-factor solution has a slightly better fit, the one-factor solution is more parsimonious. Depending on research or clinical needs, both factor structures can be used.

1. Introduction

People with diabetes have an increased risk of developing clinical and sub-clinical depression (Nouwen et al., 2010; Albertorio et al., 2017). Because depression in diabetes has been associated with adverse consequences, including a higher risk of diabetes complications (Nouwen et al., 2019), mortality (van Dooren et al., 2013), and cognitive decline (Schmitz et al., 2018), the regular screening of depressive

---

symptoms among people with diabetes has been suggested to clinicians (Young-Hyman et al., 2016). Originally developed in the US to screen for depressive disorders in primary care (Spitzer et al., 2001), the Patient Health Questionnaire-9 (PHQ-9) has become a widely used self-report depression screening tool for clinicians and researchers across the world. The scale is based on DSM-IV diagnostic criteria (Diagnostic and Statistical Manual of Mental Disorders, fourth edition; DSMIV; American Psychiatric Association, 1994) and assesses the affective, cognitive, and somatic symptoms of depression.

Despite the PHQ-9 having been developed and validated for use in primary care (Kroenke et al., 2010; Sung et al., 2013), the PHQ-9 is also commonly used in samples of people with diabetes. Some studies have provided psychometric support for use of the PHQ-9 in people with type 1 and type 2 diabetes. For instance, the PHQ-9 has been shown to have good internal consistency and convergent validity in people with diabetes (Janssen et al., 2016). However, questions remain regarding measurement invariance and the PHQ-9's underlying factor structure, using one or two factors, in diabetes samples (Boothroyd et al, 2019). In primary care samples, studies relying on exploratory factor analysis tend to report a single factor solution (i.e., all items loading on the same single factor) (e.g. Cameron et al., 2008; Dum, et al., 2008, Hanssen et al., 2009; Huang et al., 2006), while studies using confirmatory factor analysis have found that a two-factor solution, with one factor reflecting the somatic items of the PHQ-9 and the other factor the cognitive-affective items, has better model fit (Beard et al., 2016; Chil-cot et al., 2013; Petersen et al., 2015). The only study examining the PHQ-9 factor structure in people with diabetes, to our knowledge, found a better fit for a two-factor solution than the one-factor solution both in people with and without diabetes (Janssen et al., 2016).

Measurement invariance can be used to compare the stability of a scale's factor structure between groups (van de Schoot et al., 2015). A number of studies have shown that the PHQ-9 is scale or measurement invariant, that is, that the PHQ-9 items measure the same underlying factor structure across different groups and settings for both the one- and two-factor models. Specifically, invariance has been shown for different demographic (age, sex, education level and marital status; González-Blanch et al., 2018; Leung et al., 2020; Harry et al., 2019; Petersen et al., 2015; Villarreal-Zegarra et al., 2019), cross-cultural (Keum et al., 2018; Miranda and Scopetta, 2018), US student populations; Galenkamp et al, 2017), Netherlands Dutch and ethnic minorities (Arthurs et al., 2012; Merz et al., 2011), and patient groups (e.g. primary and secondary care) in the US and elsewhere. However, only one study reported assessing whether the PHQ-9 is scale invariant between people with and without diabetes, but no detailed statistical information on measurement invariance testing was provided. Research demonstrating measurement invariance is needed, as it currently remains unclear whether the meaning of the PHQ-9 items is similar for people with and without diabetes. This is particularly relevant for a scale assessing depressive symptoms in people with diabetes, given the overlap between the somatic symptoms of depression (e.g. trouble falling/staying asleep, sleeping too much; feeling tired or having little energy; poor appetite or overeating) with those of diabetes (Harding et al., 2019; McDade & Watson, 2011; Roy et al., 2012).

Taken together, depression is a common comorbidity of diabetes that is often assessed in research and clinical practice using the PHQ-9. However, the PHQ-9 has not yet been validated in terms of measurement invariance in this population. Therefore, the aim of this study was to first determine whether the PHQ-9 is measurement invariant between people with and without diabetes, and second, whether a one-factor or a two-factor was a better fit for diabetes samples, using three large data sets, namely the National Health and Nutrition Examination Survey (NHANES; USA), the Emotional Well-Being, Metabolic Factors and Health Status (EMHS) study and the Health, Inflammation, and Depression (HID) studies (Quebec, Canada), and the UK-Biobank (UK).

2. Methods

2.1. Samples

Participants were from three data banks, namely the US NHANES 2009-2014, the EMHS-HID from Canada, and the UK-Biobank.
NHANES-USA

2.1.1. Study design

Data source—The National Health and Nutrition Examination Survey (NHANES) is a multipurpose cross-sectional health survey that measures the health and nutritional status of the civilian noninstitutionalized U.S. population. NHANES uses a complex survey stratified multistage, probability cluster design to select a representative sample of the population. In the U.S., the NHANES is the only national health survey that contains a dual protocol for the collection of self-report health information and clinical, physical examination, and laboratory. The data collection is systematically achieved, allowing the identification of both diagnosed and undiagnosed health conditions (Boltri et al., 2005).

NHANES data collection is completed in two phases. The first phase is a face-to-face interview in the participant's household. The second phase consists of a series of private interviews, and physical and laboratory examinations held in a mobile examination center (MEC). Both the household and MEC interviews are performed using a standard protocol with trained staff and recorded using computer-assisted personal interviewing (CAPI) protocols. NHANES data collection protocol has been approved by the National Center for Health Statistics Research Ethics Review Board. All the participants provided written informed consent. More information about NHANES collection protocols can be found elsewhere (National Health and Nutrition Examination Survey, 2013, 2014).

Analyses were performed using NHANES public-use data files. NHANES data files are based on an independent two-year cycle. Each data cycle contains its respective complex survey elements such as weight, primary sample unit, and strata to conduct extrapolation to the general non-institutionalized adult population. For the development of an analytical study, we combined three NHANES cycles to increase sample size, subsequently increasing the statistical power, and precision of the subdomain of diabetes (National Health and Nutrition Examination Survey, 2011). Following this strategy, three NHANES data cycles were merged for this study (i.e., 2009–2010, 2011–2012, and 2013-2014 (National Health and Nutrition Examination Survey 2013)).

For the present study, a total of 8,221 eligible adults ages 20 and older who visited the MEC and provided complete case data on diabetes and depression were selected. The PHQ-9 was administered by trained interviewers. Of those eligible, 1,886 participants self-reported having a diagnosis of diabetes and 4,153 participants self-reported never having received a diagnosis of diabetes. Participants who had diabetes as confirmed by a positive lab test (glucose and HA1c) but were unaware of having the condition (n = 2,182) were excluded from the analysis.

Diabetes status in NHANES was based on self-reported physician diagnosis. Physician-diagnosed diabetes was obtained during the household interview. Women who reported having diabetes only during pregnancy were excluded from the sample. For the current study, sample weighting was not included in the invariance analysis since we aimed to study the measurement invariance and not derive non-institutionalized population estimates.
EMHS-HID - Canada

The EMHS and HID cohorts were combined for use in the present study to examine people with (HID) and without (EMHS) self-reported diabetes. Both samples were recruited from the original CARTaGENE study (www.cartagene.qc.ca), a large health survey of 20,004 French- and English-speaking residents of the Canadian province of Quebec between the ages of 40 and 69 years in 2009-2010 (Awadalla et al., 2013). The principal aim of EMHS study was to examine the combined role of metabolic dysregulations and depressive symptoms in the risk of

type 2 diabetes (Schmitz et al., 2016). The principal aim of the HID study was to examine the role systemic inflammation in depression among people with diabetes. The EMHS and HID study procedures were approved by the Douglas Mental Health University Institute Ethics Board and the St Justine Hospital Research Ethics Board.

A total of 2525 participants without diabetes at baseline from the CARTaGENE study participated in the EMHS follow-up in 2014-2015, and a total of 719 participants with diabetes from the CARTaGENE study participated in the HID follow-up in 2017. There were 87 participants that did not have diabetes at the CARTaGENE baseline assessment but developed diabetes by the EMHS assessment, and they were thus included the diabetes sample for the present analyses. For the EMHS study, individuals with depressive symptoms and metabolic factors were oversampled. For the HID study, individuals with depressive symptoms were oversampled. The original CARTaGENE study collected survey information on depressive symptoms, lifestyle, and demographic information, as well as health-related information, in addition to the collection of blood samples. CARTaGENE participants who participated in EMHS and HID completed a phone interview 5 years and 7 years, respectively, following the initial CARTaGENE assessment. These follow-up studies included the administration of the PHQ-9, used in the present study's analyses. Data were collected by CAPI.

Diabetes was assessed by self-report based on a diagnosis of diabetes made by a physician (EMHS) or by either a self-reported diagnosis of diabetes or HbA1c levels equal to or above 6.5 during the CARTaGENE baseline assessment (HID). Demographic variables were assessed by self-report.

For the present study, a total of 2,411 EMHS participants without diabetes and 800 HID participants with diabetes were included (total n = 3,211), based on having complete available data on all PHQ-9 items. French- and English-speaking participants were included, though the majority (n = 2,913; 91%) were French-speakers.

UK Biobank

The UK Biobank includes a population-based cohort consisting of 501,726 individuals aged between 40-69 years in 2006-2010, who were recruited through direct mailing invitations to 9.2 million National Health Service contact details living in reasonable proximity to one of 22 assessment centres throughout the United Kingdom. Assessment included demographic, socio-economic, psychosocial, and environmental factors, health status, and a range of physical measures (Sudlow et al., 2015). Participants entered their responses using a touch screen. Support was available at all times and a mouse and keyboard were available for those not comfortable with touchscreen computers (see https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/Touchscreen.pdf).

In 2016, an online link for a mental health questionnaire including the PHQ-9 was sent to all UK Biobank participants with an email address (n = 339,092) and 46% (n = 157,367) had responded (31% of total cohort) (Davis & Hotopf, 2019). The UK Biobank and subsequent amendments including the added mental health questionnaire were approved by the North West Multi-centre Research Ethics Committee.

For this study, we used a data set that was created to examine the genetical aspects of depression in people with diabetes and only included participants with white European heritage (n=487,320). Of those, 25,474 participants self-reported having diabetes mellitus, while 459,727 participants self-reported never having received a diagnosis of diabetes.

PHQ-9 data was available for 5,122 participants with diabetes and for 148,311 participants without diabetes. Among those, 4,981 participants with diabetes (97%) and 145,570 controls without diabetes (98%) had complete PHQ-9 data sets and were included in the current study (total n = 150,551).

2.2. Measures

Depressive symptoms. In all cohorts, the 9-item Patient Health Questionnaire (PHQ-9; Kroenke et al., 2001) was used to assess depressive symptoms. Item names, along with their frequencies across the three cohorts, are presented in Supplementary Table 1. Participants scored the items on 4-point scales ranging from 0 ("not at all"), 1 ("several days"), 2 ("more than half of the days"), and 3 ("nearly every day"). The total score ranged from 0 to 27. Because in the UK Biobank the PHQ-9 response options ranged from 1-4, these were recoded into 0-3.

Demographics included age, sex of birth, education level, and ethnic

Table 1

Demographic variables for participants with and without diabetes by the three datasets (NHANES, EMHS-HID, UK Biobank).

Variables	Diabetes	Control
NHANES		
	N=1,886	N=4,153
Age, mean (sd)	61.6 (13.1)	42.1 (17.7)
Sex, men n (%)	939 (49.7)	1,808 (43.5)
Qualifications, n (%)		
University/College	646 (34.3)	1,087 (18.3)
Some College	418 (22.2)	767 (18.5)
High school diploma or GED	521 (27.6)	1,202 (28.9)
Less than high school diploma	288 (15.3)	1,087 (26.2)
Ethnicity, n (%)		
White	678 (35.9)	1,840 (44.3)
Black	530 (28.1)	826 (19.9)
Hispanic	498 (26.4)	1,005 (24.2
Other	180 (9.5)	482 (11.6)
PHQ-9 levels of depression severity, n (%)		
Minimal (0-4)	1,214 (64.4)	3,187 (76.7)
Mild (5-9)	376 (19.9)	641 (15.4)
Moderate (10-14)	165 (8.7)	209 (5.0)
Moderately severe (15-19)	95 (5.0)	83 (2.0)
Severe (20-27)	36 (1.9)	33 (0.8)
Mean (sd; range)	4.4 (4.1; 0 to 25)	3.0 (4.1; 0 to 27)
EMHS-HID		
	N=800	N=2,411
Age, mean (sd)	63.8 (7.5)	58.4 (7.5)
Sex, men, n (%)	446 (62.8)	971 (42.0)
Qualifications, n (%)		
University/College	515 (73.5)	1,935 (81)
High school diploma	166 (23.7)	426 (17.9)
Less than high school diploma	20 (2.9)	22 (0.9)
Ethnicity White European, %	88%	93%
PHQ-9 levels of depression severity		
Minimal (0-4)	469 (58.6)	1,477 (61.3)
Mild (5-9)	202 (25.3)	578 (24.0)
Moderate (10-14)	79 (9.9)	207 (8.6)
Moderately severe (15-19)	36 (4.5)	112 (4.6)
Severe (20-27)	14 (1.8)	37 (1.5)
Mean (sd; range)	5.0 (5.1; 0 to 25)	4.7 (4.9; 0 to 27)
UK Biobank		
	N=4,981	N=145,570
Age, mean (sd)	59.06 (7.0)	56.3 (7.8)
Sex, men, n (%)	3,055 (61.3)	62,665 (43.0)
Qualifications, n (%)		
University/College degree	1,926 (39)	66,701 (46)
College (A-levels, professional, or equivalent)	1,305 (26)	34,198 (23)
Secondary education (O-levels, CSE or equivalent)	1,185 (24)	28,194 (19)
None of the above/prefer not to say	557 (11)	10,190 (7)
Missing	8 (0.2)	287 (0.2)
Ethnicity White European, n (%)	4,981 (100)	145,570 (100)
PHQ-9 levels of depression severity, n (%)		
Minimal (0-4)	3,572 (71.7)	116,205 (79.8)
Mild (5-9)	899 (18.0)	21,271 (14.6)
Moderate (10-14)	317 (6.4)	5,281 (3.6)
Moderately severe (15-19)	126 (2.5)	1,878 (1.3)
Severe (20-27)	67 (1.3)	935 (0.6)
Mean (sd; range)	3.6 (4.5; 0 to 27)	2.7 (3.7; 0 to 27)

Note: Unweighted percentages for NHANES education level do not add to 100% due to missing data

background (for details see Table 1).

2.3. Statistical analysis

A confirmatory factor analysis was first conducted to determine the fit of the PHQ-9 one-factor and two-factor solutions in each of the three cohorts for participants with and without diabetes. For the two-factor structure, the first factor included items reflecting a cognitive-affective domain of depression (items 1,2,6,9) and the second factor included items reflecting a somatic domain of depression (items 3,4,5,7,8) (Galenkamp et al., 2017). To examine PHQ-9 measurement invariance across groups, we conducted a series of measurement invariance tests using multigroup confirmatory factor analysis (CFA) with the weighted least squares mean and variance (WLSMV) estimator to account for the ordinal nature of the data (Svetina, et al., 2020). Measurement invariance tests are performed in a hierarchical manner, with increasingly stringent constraints applied to the parameters to determine if they are equal across groups. Measurement invariance support in each hierarchal level would lend strong support for the notion that the instrument is indeed measuring the same construct in a similar way across diabetes status groups. For the present study, we tested whether the one-factor and the two-factor structures of the PHQ-9 were invariant across groups of adults with diabetes and without diabetes.

For each test, PHQ-9 items were modelled as categorical variables. The recommended guidelines outlined by Wu and Estabrook (2016) and Svetina et al. (2020) for invariance testing of ordinal data were followed. This approach differs from the traditional approach for measurement invariance using continuous outcome variables (e.g., Vandenberg and Lance, 2000), though is better suited for ordinal variables such as the PHQ-9. Three increasingly restrictive models were hierarchically examined to determine the degree of measurement invariance between participants with and without diabetes within each cohort. Configural, threshold, and threshold and loading invariance were tested using the WLSMV estimator for ordinal data. In the first step, configural invariance was tested by examining whether the underlying factor structures are measured by the same items across those with and without diabetes, with an equal number and pattern of parameters across groups. In the second step, threshold invariance was tested by imposing an additional constraint on item thresholds, where these are constrained to be equal among diabetes groups. In a final step, threshold and loading invariance was tested by additionally constraining factor loadings to be equal among diabetes groups. An additional exploratory test was carried out to examine measurement invariance on the PHQ-9 across the three cohorts included in the present study (NHANES, EMHS-HID, and the UK Biobank). Analyses were conducted using MPlus version 7.4 (Muthén and Muthén, 2018).

Goodness of fit statistics were estimated for each CFA and measurement invariance models and were based on standard evaluative criteria (Chen, 2007). While a non-statistically significant chi-square value is often used to indicate model fit, this value is sensitive to sample size. Given the large sample sizes of the present study, we focused on indicators of model fit that are less sensitive to sample size (Chen, 2007), although we report chi-square statistics for completeness. A Root Mean Square Error of Approximation (RMSEA) equal to or below .08 and a Comparative Fit Index (CFI) value equal to or above .90 are indicative of acceptable model fit. For measurement invariance testing, an additional set of criteria were examined to determine measurement invariance based on changes in CFI and RMSEA values in the increasingly stringent measurement invariance models. Changes in RMSEA <0.015 and changes in CFI <0.01 indicate support for that level of measurement invariance (Chen, 2007). A chi-square difference test was also conducted, comparing the more stringent models with the model from the previous step (i.e., the thresholds only model was compared to the configural invariance model, and the thresholds and loadings model was compared to the thresholds only model), with a non-statistically significant chi square difference test indicating measurement invariance.

However, the chi-square difference test is also sensitive to sample size and therefore the focus for determining measurement invariance remained on changes in RMSEA and CFI.

3. Results

Sample characteristics are presented in Table 1. The frequency of responses on all PHQ-9 items, stratified by cohort, are presented in Supplementary Table 1. Overall, individuals with diabetes in all cohorts tended to report a greater severity of depressive symptoms compared to those without diabetes.

Results of the CFA tests demonstrated good model fit for both the one-factor and the two-factor solutions (Table 2) across each cohort. However, there was stronger support for a two-factor structure in all cohorts, and particularly for the UK biobank according to the RMSEA value. In the one-factor solution, RMSEA was .076 for those with diabetes and was .074 for those without diabetes, whereas in the two-factor solution, RMSEA was .052 for those with diabetes and was .049 for those without diabetes.

Table 3 presents the results of the measurement invariance tests per cohort for the one-factor solution models. We found that applying increasingly stringent constraints across the three levels of invariance testing did not significantly reduce model fit (all ΔCFI values were below 0.01 and all ΔRMSEA values were below 0.015) and model fit was acceptable for the configural, thresholds only, and loadings and thresholds models in all three cohorts. The most stringent model, testing equal loadings and thresholds, demonstrated satisfactory model fit in the three cohorts (RMSEA = 0.063 or below and CFI = .978 or above), providing support for measurement invariance between groups of people with and without diabetes using a one-factor solution.

Table 4 presents the results of the measurement invariance tests per cohort for the two-factor solution models. Similarly, applying increasingly stringent constrains did not significantly reduce model fit according to the critical values (all ΔCFI values were below 0.01 and all ΔRMSEA values were below 0.015), and model fit for the configural, thresholds only, and loadings and thresholds models in all three cohorts was acceptable. We also found that the most stringent model, testing equal loadings and thresholds, was satisfactory across the three cohorts (RMSEA = 0.042 or below and CFI = .990 or above). Model fit indices and chi-square difference tests provided greater support in the models testing a two-factor solution compared to the models testing a one-factor solution.

Finally, results indicated support for measurement invariance in the models comparing the PHQ-9 items in participants with diabetes across the three cohorts included in the present study (Table 5). Although the chi-square difference tests for the more stringent models were statistically significant, this was not surprising given the large sample sizes. Configural, equal thresholds, and equal thresholds and loadings measurement invariance was supported by examining the changes in RMSEA and CFI values across the increasingly stringent models (all ΔCFI values were below 0.01 and all ΔRMSEA values were below 0.015), and model fit was acceptable for the one-factor and the two-factor solutions, though with better fitting models in with a two-factor solution (Table 5).

4. Discussion

The present study examined whether the PHQ-9 was measurement invariant across people with and without diabetes in three large independent datasets from the USA (NHANES), Quebec, Canada (EMHS-HID), and the UK (UK-Biobank). The results showed that the PHQ-9 was measurement invariant across people with and without diabetes in each of the datasets. Measurement invariance was found for both the one-factor and the two-factor solutions of the PHQ-9. The findings suggest that the meaning of the PHQ-9 items is similar for people with and without diabetes. Thus, despite differences in the level of depression between people with diabetes and without diabetes, and the potential

Table 2
Confirmatory Factor Analysis Results

Model	N	Free parameters	χ^2 (df)	χ^2 p value	RMSEA	CFI
NHANES						
1-factor model						
Diabetes group	1,886	36	235.47(27)	<.00001	0.06	0.98
Control group	4,153	36	398.43(27)	<.00001	0.06	0.98
2-factor model						
Diabetes group	1,886	37	158.61(26)	<.00001	0.05	0.99
Control group	4,153	37	224.51(26)	<.00001	0.04	0.99
EMHS-HID						
1-factor model						
Diabetes group	800	36	96.85 (27)	<.00001	0.06	0.98
Control group	2,411	36	315.95(27)	<.00001	0.07	0.98
2-factor model						
Diabetes group	800	37	55.02(26)	0.0008	0.04	0.99
Control group	2,411	37	151.41(26)	<.00001	0.05	0.99
UK Biobank						
1-factor model						
Diabetes group	4,981	36	798.02 (27)	<.00001	0.08	0.98
Control group	145,570	36	21751.03 (27)	<.00001	0.07	0.98
2-factor model						
Diabetes group	4,981	37	379.12 (26)	<.00001	0.05	0.99
Control group	145,570	37	9174.46 (26)	<.00001	0.05	0.99

Note. The two-factor models including a somatic symptom cluster (items 3,4,5,7,8) and a cognitive-affective symptom cluster (items 1,2,6,9).

Table 3
Results of 1-factor invariance test

Model	Free parameters	χ^2 total(df)	χ^2 pvalue	χ^2 diabetes	χ^2 control	RMSEA	CFI	Reference model	ΔRMSEA	ΔCFI	Difftestχ^2 (df)	Difftestχ^2 p value
NHANES												
1.Configural	72	633.46 (54)	<.00001	235.43	398.03	0.06	0.98	-	-	-	-	-
2.Equal thresholds	63	678.60 (63)	<.00001	254.32	424.28	0.06	0.98	1	-0.003	-0.001	13.80(9)	0.1296
3.Equal loadings and thresholds	55	644.05 (71)	<.00001	248.16	395.89	0.05	0.98	2	-0.005	0.002	13.67(8)	0.0909
EMHS-HID												
1.Configural	72	408.03 (54)	<.00001	97.50	310.54	0.06	0.98	-	-	-	-	-
2.Equal thresholds	63	436.92 (63)	<.00001	106.83	330.08	0.06	0.98	1	-0.003	-0.001	9.51 (9)	0.3913
3.Equal loadings and thresholds	55	427.76 (71)	<.00001	124.77	302.99	0.06	0.98	2	-0.005	0.001	24.33 (8)	0.0020
UK Biobank												
1.Configural	72	23029.77 (54)	<.00001	776.91	22252.76	0.08	0.98	-	-	-	-	-
2.Equal thresholds	63	24182.33 (63)	<.00001	827.64	23354.69	0.07	0.98	1	-0.004	-0.001	21.36 (9)	0.0112
3.Equal loadings and thresholds	55	21416.51 (71)	<.00001	853.29	21061.00	0.06	0.98	2	-0.008	0.003	36.08(8)	<0.0001

overlap of some depression symptoms (e.g. trouble falling asleep, sleeping too much; feeling tired or having little energy; poor appetite or overeating) with symptoms of diabetes, these differences do not seem to be attributable to differences in interpretation of the PHQ-9 items.

Comparing the diabetes samples across the three datasets also showed measurement invariance of the PHQ-9 despite differences in ethnic make-up with NHANES being a multi-ethnic cohort, the EMHS-HIS cohort of predominantly composed of French Canadians, and the UK-Biobank cohort in this study entirely composed of white Europeans. Our results extend those of earlier studies showing PHQ-9 measurement invariance across various demographic, cultural and ethnic, and patient groups and settings (Merz et al., 2011; Arthurs et al., 2012; Petersen et al., 2015; Galenkamp et al., 2017; González-Blanch et al., 2018; Keum et al., 2018; Miranda and Scopetta, 2018; Harry et al, 2019; Leung et al., 2020; Villarreal-Zegarra et al. 2019) and demonstrate the robustness of the PHQ-9 across groups and settings.

It is important to note that despite measurement invariance indicating that any differences between cohorts cannot be attributed to

differences in interpretation of the PHQ-9 items, gender, socio-economic and cultural differences may exist between countries affecting the way depressive symptoms are perceived and expressed. Further research is needed to address this issue.

We also examined whether the factor structure of the PHQ-9 was better explained by a one-factor solution or a two-factor solution with one factor reflecting the somatic items and the other factor the cognitive-affective items. The results demonstrated that both factor solutions had good fit to the data, with the two-factor solution showing a somewhat better fit. Our results are consistent with Janssen et al. (2016) who also found support for both a one-factor and two-factor solutions in a sample of people with diabetes and with research in non-diabetes samples (e.g. Chilcot et al., 2013; Petersen et al., 2015; Beard et al., 2016). While the two-factor solution seems to better fit PHQ-9 data across populations, the one-factor solution is more parsimonious. The implication for using the PHQ-9 in people with diabetes is that both factor structures can be used for measuring depressive symptoms, and factor solution can be selected according to the research or clinical

Table 4
Results of 2-factor invariance test

Model	Free para-meters	χ² total(df)	χ² p value	χ² diabetes	χ² control	RMSEA	CFI	Reference model	ΔRMSEA	ΔCFI	Difftestχ² (df)	Difftestχ² p value
NHANES												
1.Configural	74	383.45 (52)	<.00001	157.37	226.08	0.046	0.99	-	-	-	-	-
2.Equal thresholds	65	414.71 (61)	<.00001	172.10	242.62	0.044	0.99	1	-0.003	-0.007	13.80(9)	0.1295
3.Equal loadings and thresholds	58	401.85 (68)	<.00001	171.83	230.02	0.040	0.99	2	-0.004	0.00	11.61(7)	0.1142
EMHS-HID												
1.Configural	74	205.57 (52)	<.00001	54.52	151.05	0.043	0.99	-	-	-	-	-
2.Equal thresholds	65	223.48 (61)	<.00001	61.66	161.83	0.041	0.99	1	-0.002	0.000	9.50(9)	0.3923
3.Equal loadings and thresholds	58	240.25 (68)	<.00001	84.67	155.58	0.040	0.99	2	-0.001	-.001	23.75(7)	0.0013
UK Biobank												
1.Configural	74	9729.75 (52)	<.00001	369.37	9360.38	0.050	0.99	-	-	-	-	-
2.Equal thresholds	65	10214.52 (61)	<.00001	400.22	9814.30	0.047	0.99	1	-0.00	-0.01	21.56(9)	0.0104
3.Equal loadings and thresholds	58	9284.75 (68)	<.00001	417.95	8866.80	0.042	0.99	2	-0.005	0.001	37.48(7)	<0.0001

Table 5
Results of invariance tests comparing diabetes groups from EMHS, NHANES and UK Biobank Follow-up data (N = 7,667, EMHS = 800, NHANES = 1,886, UK Biobank = 4,981).

Model	Free para-meters	χ² total (df)	χ² p value	χ²EMHS	χ² NHANES	χ²UKB	RMSEA	CFI	Reference model	ΔRMSEA	ΔCFI	Difftestχ² (df)	Difftestχ² p value
1-factor model													
1.Configural	108	1085.69 (81)	<.0001	100.60	235.56	749.52	0.07	0.98	-	-	-	-	-
2.Equal thresholds	90	1194.56 (99)	<.0001	111.50	270.29	812.77	0.07	0.98	1	-0.004	-0.001	50.46(18)	<0.0001
3.Equal loadings and thresholds	74	1143.99 (115)	<.0001	131.44	272.24	740.31	0.06	0.98	2	-0.007	0.001	61.85(16)	<0.0001
2-factor model													
1.Configural	111	570.82 (78)	<.00001	56.57	160.62	353.63	0.05	0.99	-	-	-	-	-
2.Equal thresholds	93	642.94 (96)	<.00001	64.34	190.70	387.89	0.05	0.99	1	-0.003	-0.001	50.08(18)	0.0001
3.Equal loadings and thresholds	79	662.93 (110)	<.00001	90.40	202.91	369.62	0.04	0.99	2	-0.003	0.00	58.93(14)	<0.0001

Indications for change in goodness of fit statistics: ΔCFI <0.01 and ΔRMSEA <0.015 indicating support for that level of invariance.

needs. The two-factor solution can provide additional information on the cognitive and somatic domains of depression in people with diabetes, whereas the one-factor solution can be used to screen for depression, with recommended cut-off scores that can be applied to the total scale summary score to indicate potentially high levels of depression (Spitzer et al., 2001).

The results of the current study should be interpreted against its limitations. First, the three datasets did not include information to distinguish between types of diabetes, although the cohorts were likely predominantly comprised of people with type 2 diabetes. For example, using treatment-based algorithms Mosslemi et al. (2020) showed that 94% of NHANES participants had type 2 diabetes, while Thomas et al. (Thomas et al., 2018) estimated that 4% of white Europeans with diabetes in the UK-Biobank had type 1 diabetes based on genetic risk scores. Further studies are needed to establish whether the factor structure of the PHQ-9 and measurement invariance differs between the two types of diabetes. Second, the study was based on data from three culturally closely linked western countries (USA; Quebec, Canada; UK) and, despite the availability of the PHQ-9 in many countries and languages across the globe, the results may not be generalisable to other countries,

cultures, and languages. While previous research has shown that the PHQ-9 is measurement invariant between different ethnic groups in the Netherlands (Baas et al., 2011; Galenkamp et al., 2017), Germans and native Russians living in Germany (Hirsch et al., 2013) and across sex, race/ethnicity and education within the NHANES data set 2015-2016 (Patel et al., 2019), only partial measurement invariance was found between Chinese and German student samples (Zhou et al., 2020). Further research is needed to examine whether the results can be extended to other cultures and languages. Third, two of the study samples (EMHS-HID and the UK Biobank) had a limited age range and consisted only of middle-aged participants. However, despite the differences in age range between these two cohorts and the NHANES cohort, the PHQ-9 showed invariance across the three cohorts. Moreover, our findings are in line with those of González-Blanch et al. (2018) who found measurement invariance of the PHQ-9 across younger (20-39 years old) and older (40-65 years old) adults in primary care patients in Spain. Fourth, the NHANES and UK Biobank assessed diabetes status by self-report, whereas the EMHS-HID cohort included a combination of self-report and HbA1c in the diabetes assessment. However, given that measurement invariance was found in all three cohorts, the nature of the

diabetes assessment, in this study, did not seem to impact the overall findings. Fifth, there were differences in the way the PHQ-9 was administered between the cohorts. In NHANES and EMHS-HID, trained interviewers were used while in the UK-Biobank participants entered their responses via a touchscreen. However, despite these differences, the PHQ-9 was measurement invariant between the cohorts for people with diabetes. Finally, participants in the EMHS-HID with high depressive symptoms were oversampled resulting in higher mean levels of depression compared to the NHANES and UK Biobank cohorts.

It can be concluded that the meaning of the PHQ-9 items is similar for people with and without diabetes. Both a two-factor solution and the one-factor solutions show good fit to the data suggesting that both factor structures can be used for measuring depressive symptoms among people with diabetes, and factor solution can be selected based on research or clinical need. Comparisons of PHQ-9 scores between these two groups are likely to be meaningful, whether presented as a total score or as sub-scores of somatic and/or cognitive-affective items.

Author statement

AN and SD conceptualized the research question, conducted the statistical analyses, and wrote the original draft of the manuscript. NS, ZB, IP, and JAD provided assistance with methodology. All authors reviewed and edited drafts of the manuscript and approved the final manuscript.

Role of the Funding Source

Declaration of Competing Interest

All authors declare that they have no conflicts of interest.

Acknowledgments

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jad.2021.05.031.

References

Albertorio-Diaz, J.R., Eberhardt, M.S., Oquendo, M., Mesa-Frias, M., Yulei He, Y., Jonas, B., Kang, K., 2017. Depressive states among adults with diabetes: Findings from the National Health and Nutrition Examination Survey, 2007-2012. Diabetes Res. Clin. Pract. 127, 80–88. https://doi.org/10.1016/j.diabres.2017.02.031.

American Psychiatric Association, 1994. Diagnostic and statistical manual of mental disorders. American Psychiatric Association, Washington, DC.

Arthurs, E., Steele, R.J., Hudson, M., Baron, M., Thombs, B.D., 2012. Are scores on English and French versions of the PHQ-9 comparable? An assessment of differential item functioning. PLoS One 7 (12), e52028. https://doi.org/10.1371/journal.pone.0052028.

Baas, K.D., Cramer, A.O., Koeter, M.W., 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 (1–3), 229–235. https://doi.org/10.1016/j.jad.2010.08.026.

Beard, C., Hsu, K.J., Rifkin, L.S., Busch, A.B., Björgvinsson, T., 2016. Validation of the PHQ-9 in a psychiatric sample. J. Affect. Disord. 193, 267–273. https://doi.org/10.1016/j.jad.2015.12.075.

Boltri, J.M., Okosun, I.S., Davis-Smith, M., Vogel, R.L., 2005. Hemoglobin A1c levels in diagnosed and undiagnosed black, Hispanic, and white persons with diabetes: results from NHANES 1999-2000. Ethn. Dis. 15 (4), 562–567.

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

Cameron, I.M., Crawford, J.R., Lawton, K., Reid, I.C., 2008. Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. Br. J. Gen. Pract. 58 (546), 32–36. https://doi.org/10.3399/bjgp08X263794.

Chen, F.F., 2007. Sensitivity of goodness of fit indexes to lack of measurement invariance. Struct. Equat. Model. 14 (3), 464–504. https://doi.org/10.1080/10705510701301834.

Chilcot, J., Rayner, L., Lee, W., Price, A., Goodwin, L., Monroe, B., Hotopf, M., 2013. The factor structure of the PHQ-9 in palliative care. J. Psychosom. Res. 75 (1), 60–64. https://doi.org/10.1016/j.jpsychores.2012.12.012.

Dum, M., Pickren, J., Sobell, L.C., Sobell, M.B., 2008. Comparing the BDI-II and the PHQ-9 with outpatient substance abusers. Addict. Behav. 33 (2), 381–387. https://doi.org/10.1016/j.addbeh.2007.09.017.

Galenkamp, H., Stronks, K., Snijder, M.B., Derks, E.M., 2017. Measurement invariance testing of the PHQ-9 in a multi-ethnic population in Europe: the HELIUS study. BMC Psychiatry 17. https://doi.org/10.1186/s12888-017-1506-9.

González-Blanch, C., Medrano, L.A., Muñoz-Navarro, R., Ruíz-Rodríguez, P., Moriana, J. A., Limonero, J.T., Cano-Vindel, A., 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. https://doi.org/10.1371/journal.pone.0193356.

Hansson, M., Chotai, J., Nordström, A., Bodlund, O., 2009. Comparison of two self-rating scales to detect depression: HADS and PHQ-9. Br. J. Gen. Pract. 59 (566), e283–e288. https://doi.org/10.3399/bjgp09X454070.

Harding, K.A., Pushpanathan, M.E., Whitworth, S.R., Nanthakumar, S., Bucks, R.S., Skinner, T.C., 2019. Depression prevalence in Type 2 diabetes is not related to diabetes–depression symptom overlap but is related to symptom dimensions within patient self-report measures: a meta-analysis. Diabet. Med. 36 (12), 1600–1611.

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. https://doi.org/10.1016/j.jad.2019.05.017.

Hirsch, O., Donner-Banzhoff, N., Bachmann, V., 2013. Measurement equivalence of four psychological questionnaires in native-born Germans, Russian-speaking immigrants, and native-born Russians. J. Transcult. Nurs. 24 (3), 225–235. https://doi.org/10.1177/1043659613482003.

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. https://doi.org/10.1111/j.1525-1497.2006.00409.x.

Janssen, E.P., Köhler, S., Stehouwer, C.D., Schaper, N.C., Dagnelie, P.C., Sep, S.J., Schram, M.T., 2016. The Patient Health Questionnaire-9 as a Screening Tool for Depression in Individuals with Type 2 Diabetes Mellitus: the Maastricht Study. J. Am. Geriatr. Soc. 64 (11), e201–e206. https://doi.org/10.1111/jgs.14388.

Keum, B.T., Miller, M.J., Inkelas, K.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. https://doi.org/10.1037/pas0000550.

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. https://doi.org/10.1046/j.1525-1497.2001.016009606.x.

Kroenke, K., Spitzer, R.L., Williams, J.B., Löwe, B., 2010. The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review. Gen. Hosp. Psychiatry 32 (4), 345–359. https://doi.org/10.1016/j.genhosppsych.2010.03.006.

Leung, D.Y.P., Mak, Y.W., Leung, S.F., Chiang, V.C.L., Loke, A.Y., 2020. Measurement invariances of the PHQ-9 across gender and age groups in Chinese adolescents. Asia Pac. Psychiatry 12 (3), e12381. https://doi.org/10.1111/appy.12381.

McDade, E.A., Watson, D., 2011. Examining the potential influence of diabetes on depression and anxiety symptoms via multiple sample confirmatory factor analysis. Ann. Behav. Med. 42 (3), 341–351.

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. Cultur. Divers. Ethnic Minor Psychol. 17 (3), 309–316. https://doi.org/10.1037/a0023883.

Miranda, C.A.C., Scoppetta, O., 2018. Factorial structure of the Patient Health Questionnaire-9 as a depression screening instrument for university students in Cartagena, Colombia. Psychiatry Res. 269, 425–429. https://doi.org/10.1016/j.psychres.2018.08.071.

Mosslemi, M., Park, H.L., McLaren, C.E., Wong, N.D., 2020. A treatment-based algorithm for identification of diabetes type in the National Health and Nutrition Examination Survey. Cardiovasc. Endocrinol. Metab. 9 (1), 9–16. https://doi.org/10.1097/XCE.0000000000000189.

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

National Health and Nutrition Examination Survey: Plan and Operations, 1999–2010 Program and Collection Procedures [article online], 2013. Available from http://www.cdc.gov/nchs/data/series/sr_01/sr01_056.pdf. 2014.

National Health and Nutrition Examination Survey: Analytic Guidelines, 2011-2012 [article online], 2013. Available from http://www.cdc.gov/nchs/data/nhanes/analytic_guidelines_11_12.pdf. Accessed March 20 2014.

National Health and Nutrition Examination Survey: Analytic Note Regarding 2007-2010 Survey Design Changes and Combining Data Across other Survey Cycles. [article online], 2011. Available from http://www.cdc.gov/nchs/data/nhanes/analyticnote_2007-2010.pdf.

Nouwen, A., Adriaanse, M.C., van Dam, K., Iversen, M.M., Viechtbauer, W., Peyrot, M., Pouwer, F., 2019. Longitudinal associations between depression and diabetes complications: a systematic review and meta-analysis. Diabet. Med. 36 (12), 1562–1572. https://doi.org/10.1111/dme.14054.

Nouwen, A., Winkley, K., Twisk, J., Lloyd, C.E., Peyrot, M., Ismail, K., Pouwer, F., 2010. Type 2 diabetes mellitus as a risk factor for the onset of depression: a systematic review and meta-analysis. Diabetologia 53 (12), 2480–2486. https://doi.org/10.1007/s00125-010-1874-x.

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 (9), 813–823. https://doi.org/10.1002/da.22940.

Petersen, J.J., Paulitsch, M.A., Hartig, J., Mergenthal, K., Gerlach, F.M., Gensichen, J., 2015. Factor structure and measurement invariance of the Patient Health Questionnaire-9 for female and male primary care patients with major depression in Germany. J. Affect. Disord. 170, 138–142. https://doi.org/10.1016/j.jad.2014.08.053.

Roy, T, Lloyd, C.E., Pouwer, F., Holt, R.I.G., Sartorius, N, 2012. Screening tools used for measuring depression among people with Type 1 and Type 2 diabetes: a systematic review. Diabet. Med. 29 (2), 164–175.

Schmitz, N., Deschênes, S.S., Burns, R.J., Danna, S.M., Franco, O.H., Ikram, M.A., Tiemeier, H., 2018. Cardiometabolic dysregulation and cognitive decline: potential role of depressive symptoms. Br. J. Psychiatry 212 (2), 96–102. https://doi.org/10.1192/bjp.2017.26.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Collins, R., 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 12 (3), e1001779 https://doi.org/10.1371/journal.pmed.1001779.

Sung, S.C., Low, C.C., Fung, D.S., Chan, Y.H., 2013. Screening for major and minor depression in a multiethnic sample of Asian primary care patients: a comparison of the nine-item Patient Health Questionnaire (PHQ-9) and the 16-item Quick Inventory of Depressive Symptomatology - Self-Report (QIDS-SR16). Asia Pac. Psychiatry 5 (4), 249–258. https://doi.org/10.1111/appy.12101.

Svetina, D., Rutkowski, L., Rutkowski, D., 2020. Multiple-Group Invariance with Categorical Outcomes Using Updated Guidelines: an Illustration Using Mplus and the lavaan/semTools Packages. Struct. Equat. Model. 27 (1), 111–130. https://doi.org/10.1080/10705511.2019.1602776.

Thomas, N.J., Jones, S.E., Weedon, M.N., Shields, B.M., Oram, R.A., Hattersley, A.T., 2018. Frequency and phenotype of type 1 diabetes in the first six decades of life: a cross-sectional, genetically stratified survival analysis from UK Biobank. Lancet Diabet. Endocrinol. 6 (2), 122–129. https://doi.org/10.1016/s2213-8587(17)30362-5.

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., Zondervan-Zwijnenburg, M., 2015. Editorial: Measurement Invariance. Front. Psychol. 6 (1064) https://doi.org/10.3389/fpsyg.2015.01064.

van Dooren, F.E., Nefs, G., Schram, M.T., Verhey, F.R., Denollet, J., Pouwer, F., 2013. Depression and risk of mortality in people with diabetes mellitus: a systematic review and meta-analysis. PLoS One 8 (3), e57058. https://doi.org/10.1371/journal.pone.0057058.

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 (1), 4–70. https://doi.org/10.1177/109442810031002.

Villarreal-Zegarra, D., Copez-Lonzoy, A., Bernabé-Ortiz, A., Melendez-Torres, G.J., Bazo-Alvarez, 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. https://doi.org/10.1371/journal.pone.0221717.

Wu, H., Estabrook, R., 2016. Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. Psychometrika 81 (4), 1014–1045. https://doi.org/10.1007/s11336-016-9506-0.

Young-Hyman, D., de Groot, M., Hill-Briggs, F., Gonzalez, J.S., Hood, K., Peyrot, M., 2016. Psychosocial care for people with diabetes: a position statement of the American diabetes association. Diabetes Care. 39 (12), 2126–2140. https://doi.org/10.2337/dc16-2053.

Zhou, Y., Xu, J., Rief, W., 2020. Are comparisons of mental disorders between Chinese and German students possible? An examination of measurement invariance for the PHQ-15, PHQ-9 and GAD-7. BMC Psychiatry 20 (1), 480. https://doi.org/10.1186/s12888-020-02859-8.