Journal of Psychosomatic Research 188 (2025) 111983

Contents lists available at ScienceDirect

Journal of Psychosomatic Research

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

The factor structure of the Patient Health Questionnaire-9 in stroke: A comparison with a non-stroke population

J.J. Blake [a,*], T. Munyombwe [b], F. Fischer [c,q], T.J. Quinn [d], C.M. Van der Feltz-Cornelis [e,p], J.M. De Man-van Ginkel [f,1], I.S. Santos [g], Hong Jin Jeon [h], S. Köhler [i], M.T. Schram [i], J.L. Wang [j], H.F. Levin-Aspenson [k], M.A. Whooley [l], S.E. Hobfoll [m,2], S.B. Patten [n], A. Simning [o], F. Gracey [a], N.M. Broomfield [a]

[a] Department of Clinical Psychology and Psychological Therapies, University of East Anglia, Norwich NR4 7TJ, UK
[b] School of Medicine, Worsley Building, University of Leeds, Woodhouse, Leeds LS2 9JT, UK
[c] Center for Patient-Centered Outcomes Research, Department of Psychosomatic Medicine, Charité –Universitätsmedizin Berlin, Luisenstraße 2, 10117 Berlin, Germany
[d] School of Cardiovascular and Metabolic Health, 126 University Pl, University of Glasgow, Glasgow G12 8TA, UK
[e] Department of Health Sciences, Alcuin Research Resource Centre, Heslington, York YO10 5DD, UK
[f] Nursing Science, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG Utrecht, the Netherlands
[g] Post-graduation Program in Epidemiology, Faculty of Medicine, Federal University of Pelotas, Av. Duque de Caxias, 250 - Fragata, Pelotas, RS 96030-000, Brazil
[h] Department of Psychiatry, Samsung Medical Center, Sungkyunkwan University School of Medicine, (06351) 81 Irwon-Ro Gangnam-gu, Seoul, South Korea
[i] CARIM School for Cardiovascular Diseases, Maastricht University, Universiteitssingel, 506229 ER Maastricht, the Netherlands
[j] Department of Community Health & Epidemiology, Faculty of Medicine, Dalhousie University, Halifax, Nova Scotia B3H 4R2, Canada
[k] Department of Psychology, University of North Texas, Terrill Hall, Denton, TX 76201, USA
[l] Departments of Medicine, Epidemiology & Biostatistics, 550 16th Street, Second Floor, San Francisco, CA 94158, USA
[m] Department of Behavioral Sciences, Rush University Medical Center, 1645 W. Jackson Blvd., Chicago, IL 60612, USA
[n] Departments of Community Health Sciences and Psychiatry, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada
[o] Department of Psychiatry, University of Rochester, 601 Elmwood Avenue, Rochester, NY 14642, USA
[p] Institute of Health Informatics, University College London, London, UK
[q] Corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany

ARTICLE INFO

Keywords:
Depression
Stroke
Confirmatory factor analysis
Dimensionality
PHQ-9
Self-report

ABSTRACT

Background: It is unclear if certain post-stroke somatic symptoms load onto items of the Patient Health Questionnaire-9 (PHQ-9), a self-report depression questionnaire. We investigated these concerns in a stroke sample using factor analysis, benchmarked against a non-stroke comparison group.

Methods: The secondary dataset constituted 787 stroke and 12,016 non-stroke participants. A subsample of 1574 comparison participants was selected via propensity score matching. Dimensionality was assessed by comparing fit statistics of one-factor, two-factor, and bi-factor models. Between-group differences in factor structure were explored using measurement invariance.

Results: A two-factor model, consisting of somatic and cognitive-affective factors, showed better fit than the unidimensional model (CFI = 0.984 versus CFI = 0.974, $p < .001$), but the high correlation between the factors indicated unidimensionality ($r = 0.866$). Configural invariance between stroke and non-stroke was supported (CFI = 0.983, RMSEA = 0.080), as were invariant thresholds ($p = .092$) and loadings ($p = .103$). Strong invariance was violated ($p < .001$, ΔCFI = −0.003), stemming from differences in the tiredness and appetite intercepts. These differences resulted in a moderate overestimation of depression in stroke when using a summed score approach, relative to the comparison sample (Cohen's $d = 0.434$).

Conclusions: The findings suggest that the PHQ-9 measures a single factor in stroke. Because stroke patients may report higher tiredness on item 4, caution is advisable when classifying patients as depressed if they are near the cut-off and have significant post-stroke fatigue. Caution is also advised when comparing total scores between stroke and other populations.

* Corresponding author.
E-mail address: Joshua.blake@uea.ac.uk (J.J. Blake).
[1] Permanent address: Nursing Science, Department of Gerontology and Geriatrics, Leiden University Medical Centre, Albinusdreef 2, 2333 ZA Leiden, Netherlands
[2] Permanent address: STAR: Stress, Anxiety, and Resilience Consultants, Sandy, Utah, USA

https://doi.org/10.1016/j.jpsychores.2024.111983
Received 29 August 2024; Received in revised form 22 October 2024; Accepted 11 November 2024
Available online 16 November 2024

1. Introduction

The Patient Health Questionnaire-9 (PHQ-9) is one of the most widely used depression screening tools in stroke and has demonstrated acceptable reliability, validity and classification accuracy in this population [1–3]. Despite these advantages, concerns about the applicability of the PHQ-9 in stroke remain [4]. Several items contained within the PHQ-9, such as those relating to tiredness and concentration, may also capture experiences caused by other common complications in stroke recovery, such as post-stroke fatigue [5]. Even though there are often associations between depression and, for example, post-stroke fatigue or cognitive impairment, there is also evidence that these sequelae are clinically distinct with significant variance not explained by depression [6,7]. Thus, loading of this unshared variance could result in unintended multidimensionality [8,9]. Multidimensionality, if not accounted for, can obfuscate clinical interpretation of scores and add noise to optimal cut-off estimations if there is a general pattern of score inflation due to background physical comorbidity. This, in turn, could cause misclassification of individuals at increased risk of depression, potentially impacting treatment received [10].

Two recent publications have provided partial support for the unidimensionality of the PHQ-9 in stroke patients and robustness to extraneous sequelae, with a preliminary indication that the PHQ-9 may trend towards insufficient unidimensionality with increased time since stroke [4,11]. However, measurement accuracy can be compromised in more subtle ways, such as affecting item category thresholds or intercepts [10]. Differences in these parameters between populations could invalidate statistical comparisons. For example, unequal intercepts between populations can bias depression estimates using the traditional summed score approach, potentially increasing the likelihood of type I and II errors [10].

It is therefore important that the dimensionality of the PHQ-9 in stroke is further appraised and that the factor structure of the PHQ-9 in stroke is compared with other populations to understand population-level effects. We address the following aims in this study:

1) To assess the dimensionality of the PHQ-9 in a stroke population
2) To identify differences in factor structure that might be attributable to stroke via measurement invariance comparison with a non-stroke sample
3) To explore the significance of any non-invariance of item thresholds, loadings, intercepts, or residual variances between groups

2. Methods

2.1. Design

We conducted a cross-sectional secondary data analysis. Participating studies were identified by contacting authors that contributed to a previous individual participant meta-analysis [3]. Studies spanned multiple nations and languages (Table 1).

2.2. Participants

Five stroke datasets were included [12–16], covering inpatient and community settings. Individual stroke characteristic data, such as time since stroke and type of stroke, were not available. General post-stroke time points, where available, are provided in Table 1. The comparison group was sampled from eight studies [17–24] and recruited from a variety of settings. Most sampled the general population [19–21,24], but three of the smaller samples were recruited from specific groups to reflect a range of experiences and contexts [17,22,23]. A small sample of

Table 1
Demographic data, stratified by cluster.

Study	n	% female	Age (SD)	Country	Language	Population description/setting	Mean PHQ-9 (SD)	Approximate time since stroke, in months
	Stroke							
De Man-Van Ginkel (2012)	382	45.8	69.2 (14.5)	Netherlands	Dutch	Acute stroke patients	6.6 (5.6)	0.3–2.6 (M: 1.58)
Prisnie (2016)	114	56.1	59.6 (15.5)	Canada	English	Outpatient community stroke	4.8 (5.2)	2.2–10.6 (M: 3.6)
Quinn (unpublished)	135	47.7	68.4 (12.7)	UK	English	Inpatient acute stroke	6.7 (6.3	0–0.45
Simning (2018)	21	47.6	70.0 (6.5)	US	English	Older adults in public housing	4.7 (4.4)	Unavailable
Thombs (2008)	144	16.7	69.7 (10.1)	US	English	People with Coronary Artery Disease in the community	5.8 (5.6)	Unavailable
Total	796	42.3	67.8 (13.9)				6.1 (5.6)	
	Non-stroke comparison samples							
Kim (2017)	3071	56.6	38.8 (12.2)	South Korea	Korean	Randomly selected adults, via S. Korean census	2.2 (2.2)	
Liu (2015)	4182	55.0	44.7 (10.0)	Canada	English	Working population	3.2 (4.0)	
Janssen (2016)	3502	55.3	59.5 (8.6)	Netherlands	Dutch	Data from The Maastricht Study: a population-based cohort study	2.7 (3.2)	
Levin-Aspenson (2018)	408	68.6	45.0 (13.4)	US	English	General population community-based adults	6.5 (6.6)	
Santos (2013)	447	57.3	43.8 (15.1)	Brazil	Portuguese	General population via random household sampling	5.0 (5.1)	
Simning (2012)	169	59.2	67.3 (6.6)	US	English	Older adults in public housing	5.1 (4.3)	
Volker (2016)	93	50.5	46.4 (10.9)	Netherlands	Dutch	Employees on sickness leave in an occupational health setting	8.5 (7.6)	
Hobfoll (2011)	144	57.7	41.6 (15.2)	Israel	Hebrew and Arabic	Jewish and Palestinian residents of Jerusalem exposed to war	5.9 (5.9)	
Total	12,016	56.0	47.8 (13.5)				3.1 (4.0)	

people exposed to war was included because people with such experiences are commonly underrepresented in research [25].

Exclusion and inclusion criteria are specified in the source studies. We required that stroke source studies confirmed the presence of a non-transient stroke. Comparison group studies were excluded if they focused on any specific long-term health condition, though the existence of such conditions among individual participants was permitted. Stroke status was not recorded in most comparison group samples, meaning stroke-free status could not be guaranteed; however, this was only anticipated to make up a minority of cases based on published incident rates [26].

2.3. Measure

Each PHQ-9 item corresponds to one of the nine DSM-IV depression criteria [27] and is scored on a four-category scale, relating to the frequency of the symptom experienced in the past two weeks. Items are equally weighted, with a maximum total score of 27. An optimal cut-off of $\geq$10 is generally suggested in the literature for stroke and the general population [1,3].

2.4. Analysis

Propensity score matching was used to select a sample from the comparison dataset that was more demographically aligned to the stroke sample, using the MatchIt package in R [28]. The groups were matched by age, sex, country, and PHQ-9 total score. PHQ-9 total score was matched as an additional control for potential population differences. Missing data were excluded listwise.

2.4.1. Assessment of dimensionality

Confirmatory Factor Analysis (CFA) was modelled in R using Lavaan [29] and SEMTools [30]. A single-factor CFA model, representing general depression, was evaluated alongside two variations of a two-factor model and a bi-factor model. The two-factor models each consisted of a somatic and cognitive-affective factor but differed by the respective allocation of items seven, relating to trouble concentrating, and eight, relating to feeling slowed down. There are disagreements in the literature about the optimal specification of these two items so both were compared in our sample [8,9].

The bi-factor model consisted of a general factor and two specific factors, cognitive-affective and somatic [31]. The factor location of items seven and eight in the bifactor model was determined by examining which of the two-factor models had a superior fit. Bifactor models can indicate dimensionality through comparisons of correlations of bifactor global factor scores with one-factor model scores. If multidimensionality were present, the global depression factor in the bifactor model would lose significant variance to the specific uncorrelated factors because a large proportion of somatic item variance would not be due to depression and thus load to the specific somatic factor instead of the global depression factor. This would result in a comparatively weak correlation between bifactor global and one-factor model scores.

Each CFA model was fitted using Diagonally Weighted Least Squares (DWLS) estimation, which is suitable for ordinal data. Robust Root Mean Square Error of Approximation (RMSEA; values of $<$0.08 interpreted as acceptable fit) and Comparative Fit Index (CFI; $>$0.96 interpreted as acceptable fit) statistics were used to evaluate model fit [32]. Fit statistics were compared to identify whether one or two latent factors best described item responses in the stroke sample.

2.4.2. Measurement invariance

The stroke and comparison groups were assessed for measurement invariance, using established procedures for ordinal measures, to evaluate possible differences in factor structure [33]. This involves sequential equality testing of four parameters between groups: item thresholds, loadings, intercepts, and residual variances [34]. If the groups do not significantly differ in these parameters, the models are considered to be invariant and scores therefore comparable [10]. Invariance of thresholds is henceforth referred to as 'threshold invariance', invariance of loadings as 'metric invariance', invariance of item intercepts as 'scalar invariance', and invariance of item residuals as 'full invariance' [35]. In cases of non-invariance, a change in the direction of poorer fit would be expected because the model specifies equivalences that are not reflected in the data.

As per this methodology, a baseline multi-group CFA model, referred to as a configural model, was fit. Progressive equality restraints were sequentially applied and tested against the previous model. Changes to fit after each stage are typically examined via two methods: a one-way ANOVA significance test in chi-square fit statistics, using the Satorra (2000) method [36], and by inspecting changes to the CFI and RMSEA. However, caution is advised with interpreting changes to CFI and RMSEA when using ordered data and DWLS estimation [37].

Chi-square difference tests are sample-size dependent and find significant differences among small non-meaningful effects in large samples [34]. A pragmatic approach was, therefore, adopted; non-significant changes to chi-square values were assumed to be robust indicators of invariance, given the large samples. In cases where a significant *p*-value of chi-square difference was observed, a detailed exploration of invariance violation was explored to identify the meaningfulness of the observed differences.

2.4.3. Sample size

A sample size of 300–500 for CFA modelling has demonstrated robustness to low communalities and loadings, with a minimum of 200 [38]. Each stroke sample was insufficient in size to be modelled separately, so these clusters were combined. The combined samples in each group were therefore sufficient for robust parameter estimation. Statistically, accounting for clustering was impractical because of the number of clusters, the small size of each cluster, and the limitations of current software capabilities.

2.4.4. Ethics

This study was approved by the University of East Anglia Faculty of Medicine and Health Research Ethics Committee on 11th August 2021 (approval number: 2020/21–046). Participation consent had been provided to primary authors.

3. Results

3.1. Data processing and matching

Eight cases of missing data were removed from the stroke sample and five from the non-stroke sample before propensity matching, constituting 0.11 % of the original dataset. As suspected, substantial differences were found between stroke and non-stroke comparison samples in sex, nationality, language, age, and PHQ-9 total score (Table 2). Propensity matching was, therefore, necessary to reduce these differences. A 1:2 ratio sample was selected for analysis.

The final sample, after matching and removal of missing data, consisted of 787 stroke and 1574 comparison participants. Demographic details of the stroke group and propensity-matched comparison group, including significance tests of demographic differences, are summarised in Table 2. Compared with the pre-matched sample, the demographic differences in the matched sample were substantially reduced for each variable, with non-significant gender and PHQ-9 total score differences. Significant differences remained for nationality, language, and age, with medium-to-large effect sizes. Despite findings of significant differences in nationality, most participants in both groups were from western developed nations (94.9 % in non-stroke and 100 % in stroke) and may, therefore, represent similar cultural backgrounds.

Table 2
Demographic overview of samples before and after matching.

		Pre-matched Comparison Group ($n = 12{,}011$)	Stroke ($n = 787$)	Diff	Test statistic	p	Effect statistic	Effect size descriptor
% Female		56.1 %	42.3 %	13.8 %	56.92 (X^2)	<0.001	0.067 (ϕ)	Negligible
Age		47.8 (SD 13.5)	67.8 (SD 13.9)	−20.1	−40.26 (t)	< 0.001	1.46 (*d*)	Large
Country	UK	0.0 %	16.0 %	−16.0 %	2669.64 (X^2)	< 0.001	0.46 (*V*)	Large
	US	4.8 %	21.0 %	−16.2 %				
	Canada	34.8 %	14.5 %	20.3 %				
	Netherlands	29.9 %	48.5 %	−28.5 %				
	Brazil	3.7 %	0.0 %	3.7 %				
	Israel	1.2 %	0.0 %	1.2 %				
	South Korea	25.6 %	0.0 %	25.6 %				
	Total	100 %	100 %					
Language	English	39.6 %	51.5 %	−11.9 %	347.84 (X^2)	< 0.001	0.17 (*V*)	Small
	Dutch	29.9 %	48.5 %	−28.5 %				
	Portuguese	3.7 %	0.0 %	3.7 %				
	Hebrew	0.6 %	0.0 %	0.6 %				
	Arabic	0.6 %	0.0 %	0.6 %				
	Korean	25.6 %	0.0 %	25.6 %				
	Total	100 %	100 %					
PHQ-9 total		3.1 (4.0)	6.1 (5.6)	−3.0	−14.9 (t)	<0.001	0.62 (*d*)	Medium
		Matched Comparison Group ($n = 1574$)	**Stroke (n = 787)**	**Diff**	**Test statistic**	**p**	**Effect statistic**	**Effect size descriptor**
% Female		41.4 %	42.3 %	−0.9 %	0.17 (X^2)	0.679	0.01 (ϕ)	Non-significant
Age		61.8 (SD 11.3)	67.8 (SD 13.9)	−6.04	−10.56 (t)	< 0.001	0.48 (*d*)	Medium
Country	UK	0.0 %	16.0 %	−16.0 %	349.37 (X^2)	< 0.001	0.39 (*V*)	Medium
	US	17.1 %	21.0 %	−3.9 %				
	Canada	30.3 %	14.5 %	15.8 %				
	Netherlands	47.5 %	48.5 %	−1.0 %				
	Brazil	2.8 %	0.0 %	2.8 %				
	Israel	1.2 %	0.0 %	1.2 %				
	South Korea	1.1 %	0.0 %	1.1 %				
	Total	100 %	100 %					
Language	English	47.4 %	51.5 %	−4.1 %	42.41 (X^2)	< 0.001	0.13 (*V*)	Small
	Dutch	47.5 %	48.5 %	−1.0 %				
	Portuguese	2.8 %	0.0 %	2.8 %				
	Hebrew	0.7 %	0.0 %	0.7 %				
	Arabic	0.5 %	0.0 %	0.5 %				
	Korean	1.1 %	0.0 %	1.1 %				
	Total	100 %	100 %					
PHQ-9 total		5.9 (6.0)	6.1 (5.6)	0.24	−0.95 (t)	0.343	0.04 (*d*)	Non-significant

Table 3
Specification and fit statistics of stroke CFA models.

				Stroke group				Comparison group			
Model	Factors	Model specification (numbers denote items)	Model parameters	X2 (df)	Robust CFI	Robust RMSEA	Factor correlation	X2 (df)	Robust CFI	Robust RMSEA	Factor correlation
One-factor	1	All items onto a single factor	36	129.04 (27)	0.974	0.069	–	355.9 (27)	0.982	0.088	–
Two-factor A	2	Cognitive/ affective: 1,2,6 9 Somatic: 3, 4, 5, 7, 8	37	88.61 (26)	0.984	0.055	0.866	212.1 (26)	0.990	0.067	0.910
Two-factor B	2	Cognitive/ affective: 1,2,6, 7, 8, 9 Somatic: 3, 4, 5	37	108.27 (26)	0.979	0.063	0.865	252.1 (26)	0.988	0.074	0.908
Bifactor	3	Global depression: all items Cognitive/ affective: 1,2,6 9 Somatic: 3, 4, 5, 7, 8 Factor variances are freely estimated. Correlations between factors set to 0	45	39.36 (18)	0.995	0.039	Set to 0	121.8 (18)	0.994	0.061	Set to 0

X^2 = Chi-squared, df = degrees of freedom, CFI = Comparative Fit Index, RMSEA = Root Mean Square Error of Approximation. Item 1 = little interest, item 2 = down/ depressed/hopeless, item 3 = sleep problems, item 4 = tiredness, item 5 = appetite, item 6 = feeling bad about oneself, item 7 = trouble concentrating, item 8 = moving slowly, and item 9 = suicidality.

3.2. Dimensionality assessment

Fit statistics for each model are summarised in Table 3. All models had sufficient fit to the data, based on CFI values >0.96 in both groups. All stroke group models had a sufficient RMSEA fit of <0.08, but the non-stroke one-factor model did not.

Of the two alternate forms of the two-factor model, the model that specified problems with concentration and moving slowly in the somatic factor, two-factor A, had a superior fit in both groups. This suggests that the items measuring concentration and slowness covary more strongly with the somatic items, relating to sleep, tiredness, and appetite. The stroke two-factor A model had a superior fit to the stroke one-factor model, X^2 diff = 24.83, df diff = 1, $p < .001$. A high correlation ($r = 0.865$) was observed between factors in stroke, which is indicative of unidimensionality.

To further assess dimensionality in the stroke group, global depression factor scores were calculated from the bifactor model and plotted against latent depression scores from the one-factor model. The correlation between latent depression scores was 0.99 in the stroke group, indicating substantial shared variance in factor-derived scores and practical unidimensionality.

3.3. Measurement invariance

Measurement invariance findings are summarised in Table 4. The unidimensional configural model possessed sufficient fit (CFI > 0.96, RMSEA = 0.080). The constraining of item thresholds and loadings to be equal between groups did not significantly increase unstandardised chi-square statistics and only marginally affected CFI and RMSEA values, indicating equal thresholds and loadings. A significant reduction of model fit was observed after the constraint of item intercepts, as indicated by the *p*-value for the chi-square difference.

To identify the intercepts responsible for the violation of scalar invariance, nine partially invariant CFA models were specified, with item intercept constraints released one by one. The tiredness intercept was substantially greater in the stroke group, with a relative difference of 0.446, and the appetite intercept was significantly greater in the matched comparison group, with a relative difference of 0.346. The average absolute magnitude of intercept differences of the remaining items was 0.07. These findings indicate that tiredness scores are higher in the stroke group and appetite disruption scores are higher in the comparison group when controlling for latent depression severity.

A partially invariant model, which estimated the intercepts of items 4 and 5 freely while maintaining constraints for the remaining items was statistically compared to the constrained loadings model, with no significant reduction in model fit observed, $\Delta X^2 = 7.69$, $\Delta df = 6$, $p = .262$. This finding confirms the responsibility of items 4 and 5 for failed scalar invariance.

The intercept differences do not align with the original item scale and are not inherently meaningful. To identify the significance of the scalar invariance violation, the impact on depression estimates was explored. We compared the between-group differences using the standard summed score approach with those based on model-derived scores (see Table 5). The stroke and non-stroke groups did not significantly differ when using the traditional summed score approach, a consequence of the matching. However, model-implied depression scores indicated significantly greater latent depression in the non-stroke group. The effect size for the partially invariant model was 0.117 larger than for the fully invariant model, which suggests that falsely assumed scalar invariance results in moderate underestimation of between-groups differences.

The effect size derived from the partially invariant model scores was 0.434 larger than observed when using the summed score method, which is a difference that equates to a small-to-medium effect. Using the pooled SD observed in our sample, this effect approximates to an average difference of 2.3 points on the PHQ-9 total scale. Summed PHQ-9 scores may, therefore, modestly overestimate depression severity in patients with stroke.

4. Discussion

This study aimed to assess the dimensionality of the PHQ-9 in stroke and to identify possible differences in factor structure to those in the wider population. Despite the two-factor models demonstrating a better fit than the one-factor model, we found evidence of unidimensionality in stroke. This suggests that the PHQ-9 measures one unified construct of depression without the problematic loading of extraneous sequelae. This result builds on existing findings that the PHQ-9 possesses good psychometric performance and robustness to violations of unidimensionality in stroke [1,4,11] and in the general population [39].

The two groups were invariant in thresholds and factor loadings. This implies that the correlations of items with factors are broadly equivalent between groups and that the thresholds in which patients move to endorse a higher response category occur at approximately equal points on the item's latent continuum. Differences in intercepts were, however, observed; specifically, a large positive intercept of tiredness was observed in the stroke group and a large positive intercept of appetite disruption in the comparison group. This indicates that stroke patients are more likely to experience problems with tiredness, and non-stroke participants problems with appetite, when latent depression is held constant.

The violation of equal intercepts was clinically significant. A small-to-medium between-groups difference in latent depression was obscured by the inequalities in intercepts, where there was a net effect of inflated PHQ-9 total scores in stroke. Despite similar summed total scores between matched samples, latent depression estimates suggest that the stroke sample was, on average, less depressed. This confounds the between-group comparability of scores when using a summed score approach and could result in type I and II errors in research. Such bias from the tiredness item may have introduced noise and biased estimates of optimal cut-offs in stroke, leading to an overdetection of individuals at increased risk of depression in those with post-stroke fatigue and an underdetection in those without post-stroke fatigue [1]. Indeed, the potential bias of physical comorbidities and the effect on optimal cut-off points has been demonstrated in patients with diabetes [18,40].

A strength of this study is that it is one of the first to compare the factor structure of the PHQ-9 between stroke and non-stroke comparison groups. We have provided additional confirmation of the general

Table 4
Unstandardised X^2 fit statistics of multi-group CFA at each level of constraint.

Model	X^2 (df)	CFI scaled	RMSEA scaled	ΔX^2 (Δdf)	p	ΔCFI	$\Delta RMSEA$
Configural model	268.6 (54)	0.983	0.080				
Constrained thresholds	272.3 (63)	0.981	0.077	15.0 (9)	0.092	−0.001	−0.003
Constrained loadings	284.2 (71)	0.982	0.072	13.2 (8)	0.103	0.000	−0.005
Constrained intercepts	403.8 (79)	0.979	0.074	57.8 (8)	<0.001*	−0.003	0.001

X^2 = Chi-squared, df = degrees of freedom, CFI = Comparative Fit Index, RMSEA = Root Mean Square Error of Approximation, **Δ** = change in the associated fit statistic, the p-value corresponds to the significance of the change of chi-square of model fit.

Table 5
Between-groups differences in depression scores, derived from a sum score approach and model-derived latent factor estimations for fully and partially invariant models.

Scoring approach	Stroke M	Non-stroke M	t/z statistic	p	Effect size (Cohen's d)	Descriptor
Sum score	6.141	5.902	−0.94 (t)	0.255	−0.045	Non-significant
Fully invariant model	0	0.272	4.80 (z)	< 0.001	0.272	Small
Partially invariant model	0	0.389	5.91 (z)	<0.001	0.389	Small

Negative signs indicate higher estimated depression in the stroke group, and positive signs indicate higher depression in the non-stroke group.

robustness of the measure's overall dimensionality to somatic stroke sequelae [4,11].

Despite these strengths, several limitations were noted. First, it was difficult to statistically account for nested data because Lavaan has limited functionality for completing measurement invariance on nested data. Unaccounted clustering can lead to biased parameter estimations and standard errors because of the presence of item covariations within clusters that are not accounted for by the latent variable [41]. Second, the differences in demographics were too large for complete matching, resulting in a 6-year average difference in age, plus differences in country and language. Finally, a minority of the comparison sample may have suffered strokes, which may have marginally increased sample error and reduced the magnitude of observed differences.

Data on the amount of time elapsed since the index stroke event were not available. Similarly, it was not possible to explore the impact of community versus inpatient settings in the current study. As such, the stroke participants were likely to show substantial variance in their stage of stroke recovery. Time since the stroke event is important because a recent publication has demonstrated weak invariance as a factor of time since stroke [4] and because significant improvement in physical functioning can be observed up to one-year post-stroke, as well as cognitive decline because of emerging dementias [42].

The finding of unidimensionality suggests that clinicians can continue using the PHQ-9 in stroke practice. Though causes of fatigue cannot be easily separated for individual patients, clinicians should be mindful that stroke patients may report higher baseline tiredness on item 4 of the PHQ-9, and that those with significant post-stroke fatigue may have inflated total scores compared to those who do not. Caution in the interpretation of patients near the cut-off point is, therefore, advised until future research clarifies this concern. Interestingly, Simon and Von Korff (2006) found that somatic symptoms of depression, like fatigue, respond no worse to depression treatment in patients with physical health conditions compared to those without [43]. Researchers aiming to compare scores between stroke and non-stroke should consider estimating latent depression via modelling.

Several avenues for future research emerge from the current study. A mixed two-factor approach to measurement invariance methodology, whereby group-level and longitudinal-level measurement invariance are simultaneously investigated would promote a greater understanding of the longitudinal changes to factor structure associated with stroke recovery [4]. The causes of the differences in intercepts should be investigated in more detail. It is also important that measurement invariance, diagnostic accuracy, and differences in optimal cut-off points are assessed between people with post-stroke fatigue and those without, because of concerns about the tiredness item.

Funding statement

None received.

Data access

We are unable to share the data publicly.

CRediT authorship contribution statement

J.J. Blake: Writing – review & editing, Writing – original draft, Visualization, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **T. Munyombwe:** Writing – review & editing, Supervision, Methodology, Conceptualization. **F. Fischer:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization. **T.J. Quinn:** Writing – review & editing, Data curation, Conceptualization. **C.M. Van der Feltz-Cornelis:** Writing – review & editing, Data curation. **J.M. De Man-van Ginkel:** Writing – review & editing, Data curation. **I.S. Santos:** Writing – review & editing, Data curation. **Hong Jin Jeon:** Writing – original draft, Data curation. **S. Köhler:** Writing – review & editing, Data curation. **M.T. Schram:** Writing – review & editing, Data curation. **J.L. Wang:** Writing – review & editing, Data curation. **H.F. Levin-Aspenson:** Writing – review & editing, Data curation. **M.A. Whooley:** Writing – review & editing, Data curation. **S.E. Hobfoll:** Writing – review & editing, Data curation. **S.B. Patten:** Writing – review & editing, Data curation. **A. Simning:** Writing – review & editing, Data curation. **F. Gracey:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Conceptualization. **N.M. Broomfield:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We firmly thank Prof. Thombs, Dr. Levis, Dr. Benedetti, and Sheryl Sun of the DEPRESSD collaboration (McGill University) for their support in reviewing the study protocol and contacting primary authors to obtain the data used in this study.

References

[1] L.-J.J. Burton, S. Tyson, Screening for mood disorders after stroke: a systematic review of psychometric properties and clinical utility, Psychol. Med. 45 (2015) 29–49, https://doi.org/10.1017/S0033291714000336.
[2] K. Kroenke, R.L. Spitzer, J.B.W. Williams, The PHQ-9: validity of a brief depression severity measure, J. Gen. Intern. Med. 16 (2001) 606–613, https://doi.org/10.1046/j.1525-1497.2001.016009606.x.
[3] Z.F. Negeri, B. Levis, Y. Sun, C. He, A. Krishnan, Y. Wu, P.M. Bhandari, D. Neupane, E. Brehaut, A. Benedetti, B.D. Thombs, Accuracy of the Patient Health Questionnaire-9 for screening to detect major depression: updated systematic review and individual participant data meta-analysis, BMJ 375 (2021), https://doi.org/10.1136/BMJ.N2183.
[4] L. Dong, L.S. Williams, E. Briceno, L.B. Morgenstern, L.D. Lisabeth, Longitudinal assessment of depression during the first year after stroke: dimensionality and measurement invariance, J. Psychosom. Res. 153 (2022) 110689, https://doi.org/10.1016/J.JPSYCHORES.2021.110689.
[5] P. Langhorne, D.J. Stott, L. Robertson, J. MacDonald, L. Jones, C. McAlpine, F. Dick, G.S. Taylor, G. Murray, Medical complications after stroke, Stroke 31 (2000) 1223–1229, https://doi.org/10.1161/01.STR.31.6.1223.
[6] S.P. Van Der Werf, H.L.P. Van Den Broek, H.W.M. Anten, G. Bleijenberg, Experience of severe fatigue long after stroke and its relation to depressive

symptoms and disease characteristics, Eur. Neurol. 45 (2001) 28–33, https://doi.org/10.1159/000052085.
[7] J.M. de Man-van Ginkel, T.B. Hafsteinsdóttir, E. Lindeman, M.I. Geerlings, D. E. Grobbee, M.J. Schuurmans, Clinical manifestation of depression after stroke: is it different from depression in other patient populations? PLoS One 10 (2015) https://doi.org/10.1371/journal.pone.0144450.
[8] J. Chilcot, L. Rayner, W. Lee, A. Price, L. Goodwin, B. Monroe, N. Sykes, P. Hansford, M. Hotopf, The factor structure of the PHQ-9 in palliative care, J. Psychosom. Res. 75 (2013) 60–64, https://doi.org/10.1016/j.jpsychores.2012.12.012.
[9] J.S. Krause, K.S. Reed, J.J. McArdle, 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 (2010) 1218–1224, https://doi.org/10.1016/j.apmr.2010.04.015.
[10] E.S. Kim, M. Yoon, Testing measurement invariance: a comparison of multiple-group categorical CFA and IRT 18 (2011) 212–228, https://doi.org/10.1080/10705511.2011.557337.
[11] I.L. Katzan, B. Lapin, S. Griffith, L. Jehi, H. Fernandez, E. Pioro, S. Tepper, P. K. Crane, Somatic symptoms have negligible impact on Patient Health Questionnaire-9 depression scale scores in neurological patients, Eur. J. Neurol. 28 (2021) 1812–1819, 10.0.4.87/ene.14822.
[12] J.M. de Man-van Ginkel, T. Hafsteinsdóttir, E. Lindeman, H. Burger, D. Grobbee, M. Schuurmans, An efficient way to detect poststroke depression by subsequent administration of a 9-item and a 2-item Patient Health Questionnaire, Stroke 43 (00392499) (2012) 854–856, https://doi.org/10.1161/STROKEAHA.111.640276.
[13] J.C. Prisnie, K.M. Fiest, S.B. Coutts, S.B. Patten, C.A.M. Atta, L. Blaikie, A.G. M. Bulloch, A. Demchuk, M.D. Hill, E.E. Smith, N. Jetté, Validating screening tools for depression in stroke and transient ischemic attack patients, Int. J. Psychiatry Med. 51 (2016) 262–277, https://doi.org/10.1177/0091217416652616.
[14] A. Simning, C.L. Seplaki, Y. Conwell, The association of a heart attack or stroke with depressive symptoms stratified by the presence of a close social contact: findings from the National Health and Aging Trends Study Cohort, Int. J. Geriatr. Psychiatry 33 (2018) 96–103, https://doi.org/10.1002/gps.4684.
[15] B.D. Thombs, R.C. Ziegelstein, M.A. Whooley, Optimizing detection of major depression among patients with coronary artery disease using the patient health questionnaire: data from the heart and soul study, J. Gen. Intern. Med. 23 (2008) 2014–2017, https://doi.org/10.1007/S11606-008-0802-Y.
[16] T.J. Quinn, M. Taylor-Rowan, E. Elliott, B. Drozdowska, D. McMahon, N. M. Broomfield, M. Barber, M.J. MacLeod, V. Cvoro, A. Byrne, S. Ross, J. Crow, P. Slade, J. Dawson, P. Langhorne, Research protocol – Assessing Post-Stroke Psychology Longitudinal Evaluation (APPLE) study: a prospective cohort study in stroke, Cereb. Circ. - Cogn. Behav. 3 (2022) 100042, https://doi.org/10.1016/J.CCCB.2022.100042.
[17] S.E. Hobfoll, D. Canetti, B.J. Hall, D. Brom, P.A. Palmieri, R.J. Johnson, R. Pat-Horenczyk, S. Galea, Are community studies of psychological Trauma's impact accurate? A study among Jews and Palestinians, Psychol. Assess. 23 (2011) 599–605, https://doi.org/10.1037/A0022817.
[18] E.P.C.J. Janssen, S. Köhler, C.D.A. Stehouwer, N.C. Schaper, P.C. Dagnelie, S.J. S. Sep, R.M.A. Henry, C.J.H. van der Kallen, F.R. Verhey, M.T. Schram, 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 (2016) e201–e206, https://doi.org/10.1111/JGS.14388.
[19] H.F. Levin-Aspenson, D. Watson, Mode of administration effects in psychopathology assessment: analyses of gender, age, and education differences in self-rated versus interview-based depression, Psychol. Assess. 30 (2018) 287–295, https://doi.org/10.1037/PAS0000474.
[20] Y. Liu, J.L. Wang, Validity of the Patient Health Questionnaire-9 for DSM-IV major depressive disorder in a sample of Canadian working population, J. Affect. Disord. 187 (2015) 122–126, https://doi.org/10.1016/J.JAD.2015.07.044.
[21] I.S. Santos, B.F. Tavares, T.N. Munhoz, L.S.P. de Almeida, N.T.B. da Silva, B. D. Tams, A.M. Patella, A. Matijasevich, Sensibilidade e especificidade do Patient Health Questionnaire-9 (PHQ-9) entre adultos da população geral, Cad. Saude Publica 29 (2013) 1533–1543, https://doi.org/10.1590/0102-311X00144612.
[22] A. Simning, E. Van Wijngaarden, S.G. Fisher, T.M. Richardson, Y. Conwell, Mental healthcare need and service utilization in older adults living in public housing, Am. J. Geriatr. Psychiatry 20 (2012) 441–451, https://doi.org/10.1097/JGP.0B013E31822003A7.
[23] D. Volker, M.C. Zijlstra-Vlasveld, E.P.M. Brouwers, W.A. Homans, W.H.M. Emons, C.M. van der Feltz-Cornelis, Validation of the Patient Health Questionnaire-9 for major depressive disorder in the occupational health setting, J. Occup. Rehabil. 26 (2016) 237–244, https://doi.org/10.1007/S10926-015-9607-0.
[24] D.J. Kim, K. Kim, H.W. Lee, J.P. Hong, M.J. Cho, M. Fava, D. Mischoulon, J.Y. Heo, H.J. Jeon, Internet game addiction, depression, and escape from negative emotions in adulthood: a nationwide community sample of Korea, J. Nerv. Ment. Dis. 205 (2017) 568–573, https://doi.org/10.1097/NMD.0000000000000698.
[25] V. Bekteshi, M. Sifat, D.E. Kendzor, Reaching the unheard: overcoming challenges in health research with hard-to-reach populations, Int. J. Equity Health 23 (2024) 1–12, https://doi.org/10.1186/S12939-024-02145-Z/FIGURES/2.
[26] M. Hollander, P.J. Koudstaal, M.L. Bots, D.E. Grobbee, A. Hofman, M.M.B. Breteler, Incidence, risk, and case fatality of first ever stroke in the elderly population. The Rotterdam Study, J. Neurol. Neurosurg. Psychiatry 74 (2003) 317–321, https://doi.org/10.1136/JNNP.74.3.317.
[27] C.C. Bell, DSM-IV: diagnostic and statistical manual of mental disorders, JAMA J. Am. Med. Assoc. 272 (1994) 828, https://doi.org/10.1001/jama.1994.03520100096046.
[28] D.E. Ho, K. Imai, G. King, E.A. Stuart, MatchIt: nonparametric preprocessing for parametric causal inference, J. Stat. Softw. 42 (2011) 1–28, https://doi.org/10.18637/JSS.V042.I08.
[29] Y. Rosseel, Lavaan: an R package for structural equation modeling, J. Stat. Softw. 48 (2012), https://doi.org/10.18637/JSS.V048.I02.
[30] semTools Contributors, semTools: Useful tools for structural equation modeling. R package version 0.4–14. https://cran.r-project.org/package=semTools, 2016.
[31] F. Fischer, B. Levis, C. Falk, Y. Sun, J.P.A. Ioannidis, P. Cuijpers, I. Shrier, A. Benedetti, B.D. Thombs, Comparison of different scoring methods based on latent variable models of the PHQ-9: an individual participant data meta-analysis, Psychol. Med. (2021) 1–12, https://doi.org/10.1017/s0033291721000131.
[32] C.H. Li, The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables, Psychol. Methods 21 (2016) 369–387, https://doi.org/10.1037/MET0000093.
[33] H. Wu, R. Estabrook, Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes, Psychometrika 81 (2016) 1014, https://doi.org/10.1007/S11336-016-9506-0.
[34] F. Fischer, C. Gibbons, J. Coste, J.M. Valderas, M. Rose, A. Leplège, Measurement invariance and general population reference values of the PROMIS Profile 29 in the UK, France, and Germany, Qual. Life Res. 27 (2018) 999–1014, https://doi.org/10.1007/s11136-018-1785-8.
[35] D.L. Putnick, M.H. Bornstein, Measurement invariance conventions and reporting: the state of the art and future directions for psychological research, Dev. Rev. 41 (2016) 71, https://doi.org/10.1016/J.DR.2016.06.004.
[36] A. Satorra, Chapter 17: Scaled and adjusted restricted tests in multi-sample analysis of moment structures, in: R.D.H. Heijmans, A.S.D.S.G. Pollock (Eds.), Innov. Multivar. Stat. Anal, Springer, New York, NY, 2000, pp. 233–247, https://doi.org/10.1007/978-1-4615-4603-0_17.
[37] D.A. Sass, T.A. Schmitt, H.W. Marsh, Evaluating model fit with ordered categorical data within a measurement invariance framework: a comparison of estimators, Struct. Equ. Model. 21 (2014) 167–180, https://doi.org/10.1080/10705511.2014.882658/SUPPL_FILE/HSEM_A_882658_SM7145.DOCX.
[38] R.C. MacCallum, K.F. Widaman, S. Zhang, S. Hong, Sample size in factor analysis, Psychol. Methods 4 (1999) 84–99, https://doi.org/10.1037/1082-989X.4.1.84.
[39] R. Bianchi, J. Verkuilen, S. Toker, I.S. Schonfeld, M. Gerber, E. Brähler, K. Kroenke, Is the PHQ-9 a unidimensional measure of depression? A 58,272-participant study, Psychol. Assess. 34 (2022) 595–603, https://doi.org/10.1037/pas0001124.
[40] K.M. Van Steenbergen-Weijenburg, L. De Vroege, R.R. Ploeger, J.W. Brals, M. G. Vloedbeld, T.F. Veneman, L. Hakkaart-Van Roijen, F.F. Rutten, A.T. Beekman, C. M. Van Der Feltz-Cornelis, Validation of the PHQ-9 as a screening instrument for depression in diabetes patients in specialized outpatient clinics, BMC Health Serv. Res. 10 (2010), https://doi.org/10.1186/1472-6963-10-235.
[41] N.G. Dyer, P.J. Hanges, R.J. Hall, Applying multilevel confirmatory factor analysis techniques to the study of leadership, Leadersh. Q. 16 (2005) 149–167, https://doi.org/10.1016/J.LEAQUA.2004.09.009.
[42] C.A. McHutchison, V. Cvoro, S. Makin, F.M. Chappell, K. Shuler, J.M. Wardlaw, Functional, cognitive and physical outcomes 3 years after minor lacunar or cortical ischaemic stroke, J. Neurol. Neurosurg. Psychiatry 90 (2019) 436–443, https://doi.org/10.1136/jnnp-2018-319134.
[43] G.E. Simon, M. Von Korff, Medical co-morbidity and validity of DSM-IV depression criteria, Psychol. Med. 36 (2006) 27–36, https://doi.org/10.1017/S0033291705006136.