Asian Nursing Research 19 (2025) 1–12

Contents lists available at ScienceDirect

Asian Nursing Research

journal homepage: www.asian-nursingresearch.com

Review Article

Internal Structure of the Patient Health Questionnaire-9: A Systematic Review and Meta-analysis

Duckhee Chae,[1] Jiyeon Lee,[2] Eun-Hyun Lee[3,*]

[1] College of Nursing, Chonnam National University, Republic of Korea
[2] College of Nursing and Mo-Im Kim Nursing Research Institute, Yonsei University, Republic of Korea
[3] Graduate School of Public Health, Ajou University, Republic of Korea

ARTICLE INFO

Article history:
Received 3 April 2024
Received in revised form
13 December 2024
Accepted 17 December 2024

Keywords:
depression
questionnaire
reproducibility of results
systematic review

SUMMARY

Purpose: This review aimed to evaluate the internal structure (structural validity, internal consistency, and measurement invariance) of the Patient Health Questionnaire-9 (PHQ-9), which is one of the most widely used self-administered instruments for assessing and screening depression.

Methods: The updated COnsensus-based Standards for the selection of health Measurement Instruments methodology for a systematic review of self-reported instruments was used. PubMed, Embase, CINAHL, PsycINFO, and the Cochrane Library databases were searched from their inception up to February 28, 2023.

Results: This study reviewed 98 psychometric studies reported on in 90 reports conducted in 40 countries. Various versions of the PHQ-9 were identified: one-factor structures (8 types), two-factor structures (10 types), bifactor structures (4 types), three-factor structure (1 type), and second-order three-factor structure (1 type). There was sufficient high-quality evidence for structural validity of the one-factor structure with nine items scored using a four-point Likert scale based on confirmatory factor analysis, for internal consistency with a quantitatively pooled Cronbach α of .85, and for measurement invariance across sex, age, education level, marital status, and income groups. There was sufficient high-quality evidence for structural validity, internal consistency (Cronbach's $\alpha = .76–.92$, $\omega = 0.83–.92$), and measurement invariance across sex for the PHQ-8 (which excluded item 9: "suicidality or self-harm thoughts").

Conclusion: The one-factor PHQ-9 and PHQ-8 (excluding item 9) scored using a four-point Likert scale have the best internal structure based on the current evidence. The one-factor PHQ-9 and PHQ-8 justify the use of aggregated total scores in both practice and research. The total score of the PHQ-9 using a four-point Likert scale can be used to compare depression levels across sex, age, education level, marital status, and income groups due to the availability of sufficient evidence for measurement invariance across these demographic groups.

Introduction

Depression is a common mental disorder, with approximately 5% of adults estimated to be suffering from it globally [1]. Depression can adversely affect the ability to perform the activities of daily living as well as normal tasks at work/school and in the community, which leads to a poor quality of life or possibly even suicide [1]. This situation indicates the importance of the early detection and prompt management of depression, which requires an instrument for assessing depression that is brief, easy to apply, and accurate for measurements and utilization.

The Patient Health Questionnaire (PHQ)-9 is one of the most widely used self-administered instruments developed to assess and screen depression according to criteria in the fourth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) [2] along with other leading major depressive disorder symptoms [3]. The PHQ-9 comprises nine items scored using a four-point Likert scale with response options ranging "not at all" (score of 0) to "nearly every day" (score of 3). Its total score is obtained by

Duckhee Chae: https://orcid.org/0000-0003-3259-7385; Jiyeon Lee: https://orcid.org/0000-0001-6413-329X; Eun-Hyun Lee: https://orcid.org/0000-0001-7188-3857

* Correspondence to: Eun-Hyun Lee, Graduate School of Public Health, Ajou University, Suwon, Republic of Korea.
E-mail address: ehlee@ajou.ac.kr

https://doi.org/10.1016/j.anr.2024.12.005

summing the scores for all items, with higher scores indicating more severe depressive symptoms. PHQ-9 scores of 0–5, 6–10, 11–15, and 16–20 are taken to represent mild, moderate, moderately severe, and severe depression, respectively [3]. The psychometric properties of internal consistency, test–retest reliability, convergent validity, and criteria validity of the PHQ-9 were satisfied in the original study involving 6000 patients aged ≥18 years in 8 primary-care and 7 obstetrics clinics in the USA [3].

The PHQ-9 has been translated into various languages and psychometrically tested in many countries in diverse populations with different conditions. However, considerable problems have found regarding its internal structure, comprising structural validity, internal consistency, and measurement invariance [4]. The results for the structural validity of the PHQ-9 have been inconsistent (e.g., for one, two, and three factors), and the underlying structure of the PHQ-9 is criticized as still being inconclusive [4]. The internal consistency of a scale is dependent on its structure [5]; for example, if a result for structural validity yields three factors, the internal consistency (e.g., Cronbach α) for each subscale needs to be calculated rather than that for the total scale. In other words, structural validity is a prerequisite for evaluating internal consistency. When the PHQ-9 produces heterogeneous structural results, it is questionable how items should be clustered for ensuring internal consistency. The measurement invariance of an instrument refers to how its items perform across groups and hence is an important property when comparing differences in scores between groups [6]. In other words, the group difference is not due to the true nature of the construct to be measured if measurement noninvariance is present. Evaluations of the measurement invariance of the PHQ-9 across demographic groups (e.g., age and sex) have produced inconsistent findings [7–11]. Nevertheless, studies have found differences in PHQ-9 scores across demographic and various medical condition groups, with it being unclear whether these score differences are due to differences in how items function across the groups or to differences in the true nature of depression [7,9].

The PHQ-9 has been psychometrically applied in diverse populations, and various PHQ-9 versions and inconsistent psychometric properties have been reported [8]. In such cases, a systematic review of the measurement properties of an instrument can be used to identify all existing instruments (types) measuring a concept of interest and provide psychometric information to determine which one is the best [6]. Several types of systematic review have been applied to the PHQ-9, but most of them were narrative and summarized measurement properties rather than performing quality assessments or data syntheses, or they were disease-specific systematic reviews [4,8,12–14]. Thus, a systematic review of the PHQ-9 needs to be conducted using an internationally acceptable standard guideline for the systematic review of measurement properties without a population restriction. The most common topic addressed by the systematic reviews was the screening of depression, focusing on the criterion validity using the gold standard of a structural clinical interview by a trained health professional, and identifying a cutoff value for detecting major depressive disorder [14–19]. Therefore, the aim of this study was to conduct a systematic review of the internal structure (structural validity, internal consistency, and measurement invariance) of the PHQ-9 without a population restriction.

Methods

Study design and searching strategy

The aim of this study was to conduct a systematic review of depression (construct) without a population restriction (population) measured using the PHQ-9 (instrument) about structural validity, internal consistency, and measurement invariance (measurement properties of interest). This systematic review adhered to the updated COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) methodology for a systematic review of self-reported instruments [20,21]. PubMed, Embase, CINAHL, PsycINFO, and the Cochrane Library databases were searched from their inception up to February 28, 2023. According to the COSMIN, the databases were searched by combining the following key elements using AND and NOT Boolean operators [22]: instrument names (PHQ-9 and its short forms), measurement properties, and an exclusion filter reported by Terwee et al. [22] (detailed filters are presented in Supplementary files).

Eligibility criteria

Studies eligible for inclusion in this study were peer-reviewed, full-text psychometric studies reported on in English that examined the internal structure of the PHQ-9 and its short forms. The studies that exclusively employed exploratory factor analysis for structural validity were excluded because this is a preliminary method that provides potential hypotheses for the factor structure, such as providing empirical evidence for the structure for use in confirmatory factor analysis (CFA) [23]. Eligible studies included not only healthy general populations but also patients with diseases as long as the involved individuals were ≥18 years of age since the PHQ-9 was originally developed for this age group. Additionally, studies that used the PHQ-9 and its short forms for validating other instruments were also excluded.

Selection of studies

The processes to select psychometric studies are presented in Figure 1, showing an updated Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) flow diagram for systematic reviews [24]. The records identified in the searched databases were exported to EndNote (Thomson Reuters, New York, NY, USA) to remove duplicates. Two authors (D.C. and J.L.) independently screened records based on their titles and abstracts. The full texts of the screened records were then assessed for eligibility. The reference lists of the identified records were also checked manually to identify any other relevant reports. In cases of any disagreement, consensus about inclusion was obtained by consulting with a third reviewer (E.-H.L.).

Data extraction

The following data were extracted from each included study: the characteristics of instruments (name of PHQ-9 versions, numbers of subscales and items, administration mode, and language), the characteristics of studies (sample size, age, sex, marital status, education level, study population, study setting, and country), and the results of measurement property evaluations (structural validity, internal consistency, and cross-cultural/measurement invariance). Data extraction was performed by the three authors independently (D.C., J.L., and E.-H.L.), with any disagreement resolved through discussion.

Assessment of methodological quality and measurement property results

The methodological quality of each property (structural validity, internal consistency, and cross-cultural validity/measurement invariance) for each study was assessed using the COSMIN Risk of Bias checklist [20]. The risk of bias was rated on a four-point scale of “very good,” “adequate,” “doubtful,” and “inadequate” using the

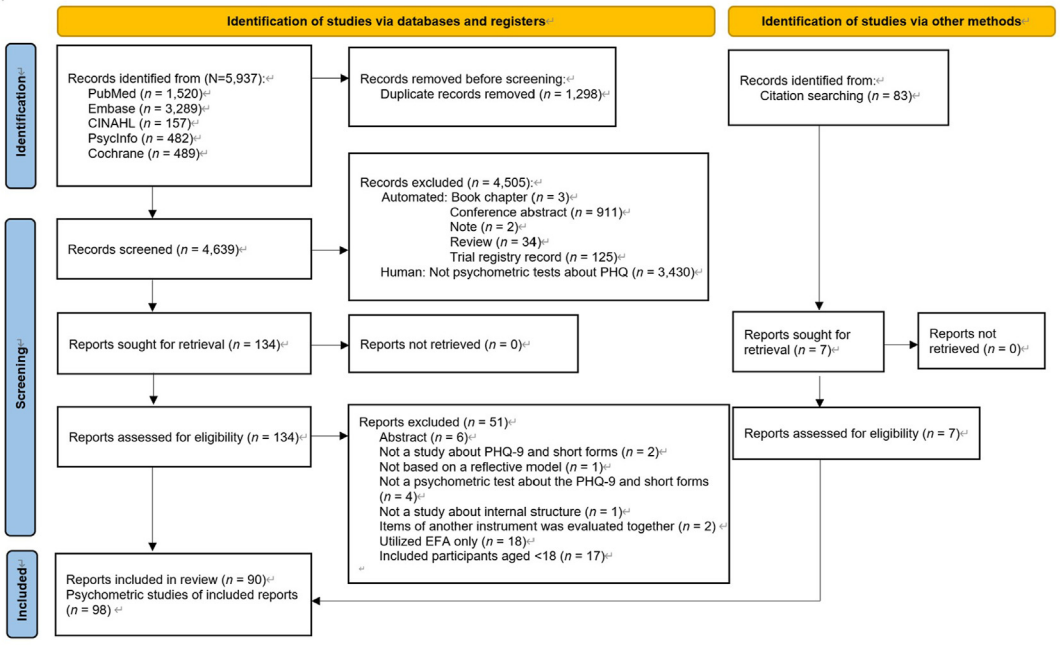

Figure. 1. PRISMA flow diagram.

worst score counts principle. For the methodological assessment of structural validity, the reporting of a particular estimation method being used was additionally defined as an important requirement for the CFA in this study: the risk of bias was rated as "doubtful" if the estimation method was not described, while the flaw was rated as "inadequate" if an inappropriate estimation was used.

The results of individual studies for the internal structure were rated against the updated criteria for good measurement properties as sufficient (+), insufficient (−), or indeterminate (?) [21]. According to the updated criteria, only Cronbach α ($\geq$ .70) can be used as a rating indicator for internal consistency. Other indicators have frequently been used recently, such as McDonald's omega (ω) [25]. Therefore, the additional criteria suggested by Lee et al. [26] and Perreira et al. [27] were also applied for the quality rating of internal consistency: sufficient (+) for $\omega \geq .70$ (hierarchical omega: $\omega_H \geq .50$), item/person reliability of $\geq .70$, and person/item separation of $\geq$1.50 for each unidimensional scale or subscale; insufficient (−) for $\omega < .70$ ($\omega_H < .50$), item/person reliability of < .70, and person/item separation of <1.50; and indeterminate (?) if the values were not reported. Regarding the quality of cross-cultural validity/measurement invariance, a sufficient (+) rating was assigned in this study if configural (equivalence of model form), metric (equivalence of item loadings on factors), and scalar (equivalence of item intercepts of metric invariant items) invariances across groups were satisfied, because the error-variance invariance (equivalence of item residuals of metric and scalar invariant items) is excessively stringent to achieve in the application for multiple-group confirmatory factor analysis in practice [28]. These assessments were performed independently by all three authors (D.C., J.L., and E.-H.L.), with any disagreement resolved through discussion.

Summary of evidence and grading the quality of evidence

To draw an overall conclusion about the quality of an instrument, the results of studies for each measurement property were first summarized or quantitatively pooled. For internal consistency, a pooled estimate of Cronbach α values was obtained using the statistical package Meta in R software (version 4.3.2, R Core Team) [29]. The summarized or pooled result was compared with the criteria for good measurement properties to determine an overall rating for each measurement property: sufficient (+), insufficient (−), inconsistent (±), or indeterminate (?). The quality of the evidence was then graded according to the modified Grading of Recommendations Assessment, Development, and Evaluation approach, taking into account four factors (risk of bias, inconsistency, imprecision, and indirectness) and grading the quality of evidence as high, moderate, low, or very low [21]. The overall rating and its quality of evidence were not assessed if only a single study had analyzed each of structural validity, internal consistency, and measurement invariance, in order to avoid overweighting by that single study [30]. The three authors (D.C., J.L., and E.-H.L.) independently performed the data syntheses and then together came to the final resolution through discussion.

Results

Identified studies

As shown in Figure 1, the database search identified 5,937 records. After removing 1,298 duplicates, a further 4,505 records were excluded after screening the titles and abstracts. Among the remaining 134 full-text reports retrieved, 51 failed to meet the inclusion criteria and thus were also excluded. An additional 7 reports were identified from the reference lists in these 83 reports. Eight of these 90 reports reported on 2 psychometric studies that examined the measurement properties using different samples or different structures of the instrument and hence 98 psychometric studies reported on in 90 reports were finally included in this systematic review.

Characteristics of instruments

The 98 studies included in this study were classified based on the underlying structures (one-factor, two-factor, bifactor, three-factor, and second-order three-factor structures). Each

structure was then sorted by the used item responses (binary, three-point Likert, and four-point Likert scales) and clustered items (parceled items into a factor or factors). There were 24 different versions included (Table 1). Regarding the administration modes, paper-and-pencil and interview (face-to-face or phone) modes were the most frequently used, followed by an online (computer) mode (Supplementary Table 1).

Characteristics of the included studies

The PHQ-9 and derived versions thereof were applied to general populations (48 studies), patients with disease (44 studies), or both general and patient populations (6 studies). The studies were conducted in 40 countries, with the largest proportion in the USA (27 studies), followed by Germany (11 studies) and the UK (7 studies). The most frequently utilized setting in which data were collected was the community (46 studies), followed by clinics and then both communities and clinics. The original English version had been translated into 28 languages: Amharic, Brazilian, Chinese, Danish, Dutch, French, German, Hebrew, Indian (local languages), Indonesian, Malayalam, Japanese, Korean, Liberian English, Lithuanian, Norwegian, Persian, Filipino, Portuguese, Russian, Sesotho, Spanish, Swahili, Swedish, Thai, Turkish, Twi, and Vietnamese (Supplementary Table 1).

Synthesized evidence

The summarized and pooled results for the structural validity, internal consistency, and cross-cultural/measurement invariance of each version of the PHQ-9 are presented in Supplementary Table 2. The overall rating and quality of evidence for measurement properties are presented in Table 2.

The one-factor PHQ-9 with 9 items scored using a 4-point Likert scale was the most frequently evaluated (48 results) using CFA (42 results) and item response theory (IRT)/Rasch analysis (6 results) (Supplementary Table 2). There was insufficient (−) high-quality evidence for this version of the PHQ-9 since 72.9% out of 48 results supported the structure (which is below the COSMIN criterion of >75%) (Table 2). When this evidence was divided into using the CFA and IRT/Rasch approaches, sufficient (+) high-quality evidence (81.0% of the results) was found in the CFA studies, while insufficient (−) high-quality evidence (16.7% of the results) was produced

Table 1 *PHQ-9 and Derived Versions.*

Factor	Number of items	Item response	Factors and clustered items
One factor (8 versions)	9	4-point Likert	One: 1–9
	9	3-point Likert	One: 1–9
	9	Binary responses	One: 1–9
	8	4-point Likert	One: 1–8
	8	4-point Likert	One: 1–6, 7, 9
	7	4-point Likert	One: 3–9
	5	3-point Likert	One: 1, 2, 4, 6, 9
	2	4-point Likert	One: 1, 2
Two factors (10 versions)	9	4-point Likert	Somatic: 3–5, 7, 8 Nonsomatic (cognitive/affective): 1, 2, 6, 9
	9	4-point Likert	Somatic: 3–5 Cognitive/affective: 1, 2, 6–9
	9	4-point Likert	Somatic: 3, 5, 7, 8 Affective: 1, 2, 4, 6, 9
	9	4-point Likert	Somatic: 1–5 Cognitive/affective: 6–9
	9	4-point Likert	Somatic: 3–5, 8 Cognitive/affective: 1, 2, 6, 7, 9
	9	4-point Likert	Somatic: 1, 3–5, 8 Nonsomatic: 2, 6, 7, 9
	8	4-point Likert	Somatic: 3–5, 8 Cognitive/affective: 1, 2, 6, 7
	8	4-point Likert	Somatic: 3–5, 7, 8 Nonsomatic: 1, 2, 6
	7	4-point Likert	Somatic: 3–5 Nonsomatic: 1, 2, 6, 9
	6	4-point Likert	Somatic: 3–5 Nonsomatic: 2, 6, 9
Bifactor (4 versions)	9	4-point Likert	General: 1–9 Somatic: 3–5, 7, 8 Cognitive/affective: 1, 2, 6, 9
	9	4-point Likert	General: 1–9 Somatic: 3–5 Cognitive/affective: 1, 2, 6–9
	9	4-point Likert	General: 1–9 Somatic: No information Cognitive/affective: No information
	8	4-point Likert	General: 1–8 Somatic: 3–5, 8 Cognitive/affective: 1, 2, 6, 7
Three factors (1 version)	9	4-point Likert	Somatic: 7, 8 Cognitive: 1, 2, 6, 9 Pregnancy symptoms: 3–5
Second-order three factors (1 version)	9	4-point Likert	Somatic: 3–5 Affective: 1, 2, 6, 9 Cognitive: 7, 8

PHQ-9 = Patient Health Questionnaire-9.

Table 2 *Overall Rating and Quality of Evidence for Measurement Properties of the Internal Structure of the PHQ-9 Versions.*

Versions	Structural validity	Internal consistency	Cross-cultural/measurement invariance
	Overall rating/quality of evidence	Overall rating/quality of evidence	Overall rating/quality of evidence
• **Development study of the PHQ-9: Unclear information about factor(s)**			
PHQ-9 [3]		No rating/grading	
• **One factor**			
One-factor PHQ-9 scored using a 4-point Likert scale (items 1–9) [7,10,11,32,39–43,47–80]	Insufficient (−)/high <Based on the analysis method> CFA: Sufficient (+)/high IRT/Rasch: Insufficient (−)/high	Sufficient (+)/high	Sex: Sufficient (+)/high Age: Sufficient (+)/high Education level: Sufficient (+)/high Marital status: Sufficient (+)/high Income: Sufficient (+)/high Language/ethnicity: Insufficient (−)/high Over time: Insufficient (−)/high Eating disorder: Insufficient (−)/high Occupation[a] University type[a] Domestic violence[a] Neurological/community groups[a] Administration mode[a] Alcohol consumption[a] GOLD severity[a] Residence area[a] Symptom burden[a]
One-factor PHQ-9 scored using a 3-point Likert scale (items 1–9) [81–87]	Insufficient (−)/high	Sufficient (+)/high	Sex: Insufficient (−)/high Age: Sufficient (−)/high Education level: Sufficient (+)/high Employment[a] Duration of visual impairment[a] Systemic/ocular comorbidity[a] Visual impairment[a] Implantation[a] Heart failure severity[a]
One-factor PHQ-9 scored using binary responses (items 1–9) [88]	[a]		Symptoms[a] Chemotherapy[a]
One-factor PHQ-8 scored using a 4-point Likert scale (items 1–8) [35,54,77,89–91]	Sufficient (+)/high	Sufficient (+)/high	Sex: Sufficient (+)/high Ethnicity[a] Age[a] Education level[a]
One-factor PHQ-8 scored using a 4-point Likert scale (items 1–7, 9) [92]	Insufficient (−)/high	Sufficient (+)/high	Sex[a] Age: Not available Multimorbidity[a]
One-factor PHQ-7 scored using a 4-point Likert scale (items 3–9) [93]	[a]	[a]	
One-factor PHQ-5 scored using a 3-point Likert scale (items 1, 2, 4, 6, 9) [94]	[a]	[a]	
One-factor PHQ-2 scored using a 4-point Likert scale (items 1, 2) [51,93,95]	Insufficient (−)/high	Insufficient (−)/high	Over time[a]
• **Two factors**			
Two-factor PHQ-9 scored using a 4-point Likert scale: Somatic (items 3–5, 7, 8) nonsomatic (cognitive/affective) (items 1, 2, 6, 9) [96–102]	Sufficient (+)/high		Sex: Insufficient (−)/high Diabetes: Sufficient (+)/high Country[a]
Two-factor PHQ-9 scored using a 4-point Likert scale: Somatic (items 3–5) cognitive/affective (nonsomatic) (items 1, 2, 6–9) [9,38,103–106]	Sufficient (+)/high	Sufficient (+)/high	Ethnicity: Sufficient (+)/high Over time: Sufficient (+)/high Sex: Sufficient (+)/high Education level: Sufficient (+)/high
Two-factor PHQ-9 scored using a 4-point Likert scale: Somatic (items 3, 5, 7, 8) affective (items 1, 2, 4, 6, 9) [107]	[a]		Over time[a] Treatment groups across the following time periods[a]
Two-factor PHQ-9 scored using a 4-point Likert scale: Somatic (items 1–5) cognitive/affective (items 6–9) [108,109]	Insufficient (−)/high[a]	[a]	
Two-factor PHQ-9 scored using a 4-point Likert scale: Somatic (items 3–5, 8) cognitive/affective (items 1, 2, 6, 7, 9) [110]	[a]	[a]	Sex[a] Region[a]
Two-factor PHQ-9 scored using a 4-point Likert scale: Somatic (items 1, 3–5, 8) nonsomatic (items 2, 6, 7, 9) [111]	[a]		

(*continued on next page*)

Table 2 (*continued*)

Versions	Structural validity	Internal consistency	Cross-cultural/measurement invariance
	Overall rating/quality of evidence	Overall rating/quality of evidence	Overall rating/quality of evidence
Two-factor PHQ-8 scored using a 4-point Likert scale: Somatic (items 3–5, 8) cognitive/affective (items 1, 2, 6, 7) [112]			
Two-factor PHQ-8 scored using a 4-point Likert scale: Somatic (items 3–5, 7, 8) Nonsomatic (items 1, 2, 6) [113,114]	Sufficient (+)/high		Over time[a] Intervention/control groups[a]
Two-factor PHQ-7 scored using a 4-point Likert scale: Somatic (items 3–5), nonsomatic (items 1, 2, 6, 9) [115]	[a]		Ethnicity[a]
Two-factor PHQ-6 scored using a 4-point Likert scale: Somatic (items 3–5), nonsomatic (items 2, 6, 9) [116]	[a]		
• **Bifactors**			
Bifactor PHQ-9 scored using a 4-point Likert scale: General (items 1–9), Somatic (items 3–5, 7, 8) cognitive/affective (items 1, 2, 6, 9) [31]	[a]		Sex[a] Diagnostic group[a] Comorbidity[a]
Bifactor PHQ-9 scored using a 4-point Likert scale: General (items 1–9) Somatic (items 3–5), cognitive/affective (items 1, 2, 6–9) [45,46]	Sufficient (+)/high	Insufficient (−)/high	Nonclinical and MDD groups[a] MDD and MDD with anxiety disorder groups[a] Age[a]
Bifactor PHQ-9 scored using a 4-point Likert scale: General (items 1–9) Somatic (no information) cognitive/affective (no information) [44]	Insufficient (−)/moderate	[a]	Sex[a] Age[a] Language[a] Country[a]
Bifactor PHQ-8 scored using a 4-point Likert scale: General (items 1–8) Somatic (items 3–5, 8) cognitive/affective (items 1, 2, 6, 7) [117]	[a]	[a]	
• **Three factors**			
Three-factor PHQ-9 scored using a 4-point Likert scale: Cognitive/affective (items 1, 2, 6, 9) Pregnancy symptoms (items 3–5) Somatic (items 7, 8) [118]	[a]	[a]	
• **Second-order three factors**			
Second-order three-factor PHQ-9 scored using a 4-point Likert scale: Somatic (items 3–5) affective (items 1, 2, 6, 9) Cognitive (items 7, 8) [119]	[a]	[a]	

MDD = major depressive disorder; PHQ-9 = Patient Health Questionnaire-9.

[a] No rating/grading: Overall rating and its quality of evidence was not assessed if there existed a single study for the structural validity, internal consistency, and measurement properties in order to avoid overweighting by that single study.

in the IRT/Rasch studies. There was sufficient (+) high-quality evidence for the internal consistency of the one-factor PHQ-9 scored using a four-point Likert scale. The pooled Cronbach α was .85 (95% confidence interval = .83–.86, I^2 = 98%) (Figure 2). The summarized Cronbach α values ranged from .65 to .92, and other indicators for the internal consistency are presented in Supplementary Table 2. Regarding the cross-cultural validity/measurement invariance of this one-factor PHQ-9, 17 different types of groups were assessed. Among them, there was sufficient (+) high-quality evidence for measurement invariance for the sex, age, education level, marital status, and income groups (Table 2).

Ancillary paper-and-pencil, interview, and online modes were also applied to the 1-factor PHQ-9 scored using a 4-point Likert scale in 14, 12, and 8 studies, respectively (studies using more than one mode type were not considered). There was sufficient (+) high-quality evidence for structural validity of the one-factor PHQ-9 in the studies using the paper-and-pencil and online modes, while insufficient (−) high-quality evidence was found in the studies that administered the interview mode (Table 3).

Among the one-factor structures, the second most frequently reported structure was the PHQ-9 scored using a three-point Likert scale and the PHQ-8 (items 1–8) scored using a four-point Likert scale. There was insufficient (−) high-quality evidence for structural validity and sufficient (+) high-quality evidence for internal consistency of the PHQ-9 scored using a three-point Likert scale, with a Cronbach α of .82, a person separation reliability of .66– .82, and an item separation of .96. There was sufficient (+) high-quality evidence for measurement invariance across age and education level groups, but insufficient (−) high-quality evidence across sex groups. In the case of the PHQ-8 (excluding item 9), there was sufficient (+) high-quality evidence for structural validity, internal consistency (Cronbach α = .76-.92, ω = .83-.92), and

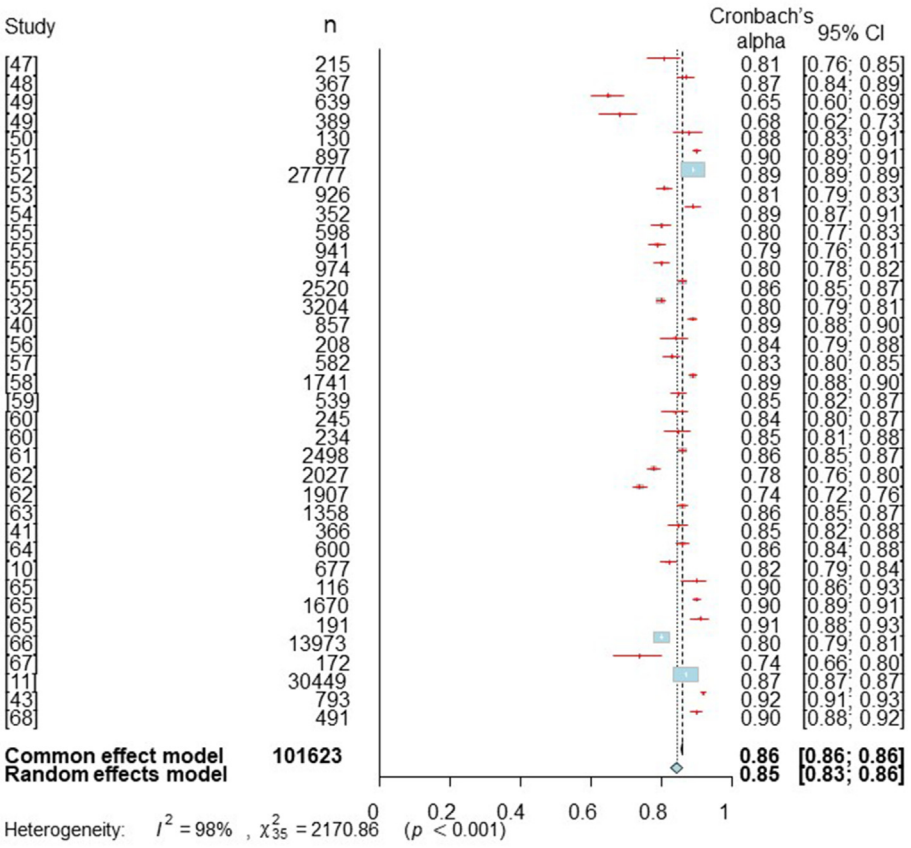

Figure. 2. The pooled Cronbach's alpha.

measurement invariance across sex groups (Table 2 and Supplementary Table 2).

Regarding two-factor structures, the best version was that of the PHQ-9 (scored using a four-point Likert scale) with a somatic factor (items 3–5) and a cognitive/affective (nonsomatic) factor (items 1, 2, 6–9). There was sufficient (+) high-quality evidence for the structural validity and internal consistency of this version of the PHQ-9, with Cronbach α values of .72– .76 and .69– .76 for the somatic and cognitive/affective factors, respectively. There was sufficient (+) high-quality evidence for cross-cultural/measurement invariance across each ethnicity, over time, sex, and education level groups. There was sufficient (+) high-quality evidence for structural validity of the two-factor PHQ-9 with different subscale configurations (somatic factor comprising items 3–5, 7, and 8; cognitive/affective factor comprising items 1, 2, 6, and 9), and for measurement invariance across disease status (diabetes), but there was insufficient (−) high-quality evidence for sex groups, and no evidence for its internal consistency (Table 2 and Supplementary Table 2).

Associated with bifactors, there was sufficient (+) high-quality evidence for structural validity of the PHQ-9 scored using a four-point Likert scale comprising a general factor with items 1–9, a somatic factor with items 3–5, and a cognitive/affective factor with items 1, 2, and 6–9. However, there was insufficient (−) high-quality evidence for internal consistency, in particular with low ω_H values of .47, .44, and .26 for the general, somatic, and cognitive/affective factors, respectively (Table 2 and Supplementary Table 2).

The remaining versions of the three-factor PHQ-9 and the second-order three-factor PHQ-9 were assessed only once, and so the overall rating and quality of the evidence could not be determined (Table 2 and Supplementary Table 2).

Discussion

Principal findings

This study reviewed 98 studies reported on in 90 reports on the internal structure (structural validity, internal consistency, and

Table 3 *Overall Rating and Quality of Evidence for the Internal Structure of the One-Factor PHQ-9 Scored Using a Four-Point Likert Scale According to Administration Modes.*

	Structural validity	Internal consistency	Cross-cultural/measurement invariance
Administration mode	Overall rating/quality of evidence	Overall rating/quality of evidence	Overall rating/quality of evidence
Paper and pencil	Sufficient (+)/high	Sufficient (+)/high	Sex: Sufficient (+)/high Language: Insufficient (−)/high
Interview (face-to-face or phone)	Insufficient (−)/high	Sufficient (+)/high	Sex: Sufficient (+)/high Age: Sufficient (+)/high Education level: Sufficient (+)/high
Online	Sufficient (+)/high	Sufficient (+)/high	Sex: Sufficient (+)/high Education level: Insufficient (−)/high

PHQ-9 = Patient Health Questionnaire-9.

cross-cultural/measurement invariance) of the PHQ-9 and derived versions, which had been conducted in a wide variety of countries, languages, and settings in adult populations. The PHQ-9 was originally developed according to criteria in Diagnostic and Statistical Manual of Mental Disorders-IV without evaluations of the construct domains of its items, which has resulted in subsequent efforts to identify the underlying structure, but for which the findings remain inconsistent [10,31]. The present systematic review has identified the following versions: one-factor structures (8 types), two-factor structures (10 types), bifactor structures (4 types), three-factor structure (1 type), and second-order three-factor structure (1 type).

The one-factor PHQ-9 scored using a four-point Likert scale was the most frequently assessed in the analyzed studies, and there was insufficient high-quality evidence for its structural validity. Dividing this evidence into the studies that applied CFA and IRT/Rasch analysis for structural validity yielded different results: sufficient high-quality evidence in the CFA studies and insufficient high-quality evidence in the IRT/Rasch studies. According to the COSMIN, the quality of structural validity is rated as sufficient (+) in a study using IRT/Rasch analysis when there are no violations of assumptions (unidimensionality, local independence, and monotonicity) and the model fit is satisfactory. All but one of the reports on IRT/Rasch analyses of the one-factor PHQ-9 [32] provided no information about the assumptions. In other words, it is unclear for these studies whether the assumptions were satisfied but not reported, violated and so not reported, or not assessed. No information about the assumptions was the main reason contributing to the insufficient rating for the quality of structural validity in the IRT/Rasch studies. For this reason, Lee et al. [26] emphasized that researchers need to report the assumption results for psychometric studies applying IRT/Rach analysis. The internal consistency of the one-factor PHQ-9 scored using a four-point Likert scale showed sufficient high-quality evidence utilizing both summarized and quantitatively pooled (meta-analysis) results. To the best of our knowledge, this is the first report about pooled Cronbach α values of the PHQ-9 obtained in a meta-analysis. Cronbach α has been a dominant indicator for assessing internal consistency in multiple-item measurement scales. However, this indicator is criticized as performing poorly because of the violation of tau-equivalence [33], which has resulted in a recent shift to reporting McDonald's ω instead of Cronbach α. The ω values of the PHQ-9 were reported in only a few articles [10,11,72,77] and so the present study did not conduct quantitative pooling. It is recommended that a future systematic review conducts a meta-analysis of the ω for the PHQ-9 once sufficient values have been accumulated. In this study, the measurement invariance of the one-factor PHQ-9 scored using a four-point Likert scale yielded sufficient high-quality evidence for the general characteristics of sex, age, education level, marital status, and income. In other words, the difference in the scores measured by the instrument represents a true difference rather than bias due to different perceptions of depression and so score differences can be meaningfully compared across groups. However, the measurement invariance of this version was restricted to mainly general characteristics and so further research is needed to extend to other groups such as disease and country.

Self-reported questionnaires are traditionally mainly administered in a paper-and-pencil mode. When a paper-and-pencil questionnaire is transformed into another mode, such as an interview or computer-based mode, this new mode must be reassessed because the administration mode can affect the responses [34]. Paper-and-pencil and online modes give respondents more time to answer the questions, whereas in an interview mode, respondents may feel pressured to answer the interviewer's questions rapidly or may respond disingenuously to the interviewer due to factors such as the content of a question being socially undesirable. The PHQ-9 was originally developed for administering in a paper-and-pencil mode and includes a very sensitive question (item 9: suicidality and self-harm thoughts) to allow frank answers. This systematic review found that the evidence for structural validity of the one-factor PHQ-9 scored using a four-point Likert scale was better for the paper-and-pencil and online modes than for the interview mode. While further studies are needed to confirm these findings, the usability of the administration mode also has to be considered in both practice and research.

The PHQ-8 (items 1–8) version excluded item 9 ("suicidality or self-harm thoughts") for several reasons: (a) the depression severities measured by the PHQ-9 and PHQ-8 were empirically reported to be strongly correlated ($r > .99$), implying a trivial effect of item 9; (b) it is not possible to immediately provide an adequate intervention with an interview survey (e.g., by phone) when a participant answers that they have a suicidal intention; (c) suicidality and thoughts of self-harm are not common in a general population; (d) this sensitive question can increase the difficulty of obtaining permission from an institutional review board; and (e) many respondents will not be willing to answer this sensitive question [35–37]. Due to these reasons, the PHQ-8 was evaluated again to verify whether or not the remaining eight items were psychometrically satisfactory. This systematic review found that the one-factor PHQ-8 version exhibited almost the same evidence as that for the one-factor PHQ-9 scored using a four-point Likert scale.

Regarding the two-factor versions, there was sufficient high-quality evidence for structural validity, internal consistency, and measurement invariance (across sex, education level, ethnicity, and over time) of the PHQ-9 comprising a somatic factor (items 3–5) and a cognitive/affective factor (items 1, 2, 6–9). However, between-factor correlations also need to be considered [9,38]. Strong two-factor correlations (coefficients ranging from .85 to .90) were found in several other studies [7,39–43]. These studies found that the results for both one- and two-factor solutions of the PHQ-9 were statistically acceptable, but the researchers selected one factor as the final structure since the strongly correlated two-factor solution indicates a multicollinearity problem.

Bianchi et al. [44] suggested using a bifactor model when strong between-factor correlations are present. Four types of bifactor structures of the PHQ-9 and PHQ-8 have been reported. Of them, the bifactor PHQ-9 comprising a general factor (all nine items), a specific-somatic factor (items 3–5), and a specific-cognitive/affective factor (items 1, 2, 6–9) was the best, but there was sufficient high-quality evidence only for its structural validity. Moreover, two studies found that there was insufficient high-quality evidence for its internal consistency. The Cronbach α values for the general factor and the two-specific factors in a Japanese study exceeded .70 [45], but the ω_H values for both the specific-somatic and cognitive/affective factors were below the criterion threshold of .50 [27] in a Puerto Rican study [46]. The low ω_H values for these specific factors reflect the trivial proportion of the reliable variance in the subset scores for each specific factor while controlling for the general factor, which makes it difficult to interpret each factor score in both practice and research. Further study is recommended for the internal consistency of the bifactor PHQ-9 using ω_H.

Implications for practice and research

Overall, this study found that the internal structure (structural validity, internal consistency, and cross-cultural/measurement invariance) was best for the one-factor PHQ-9 and PHQ-8

(excluding item 9) scored using a four-point Likert scale. Since these versions supported unidimensionality, the use of an aggregated total score is justified. The invariance of the one-factor PHQ-9 across various demographic groups including sex, age, education level, marital status, and income was supported. Thus, the total scores of the PHQ-9 can be meaningfully compared between demographic groups in both practice and research. In particular, it is recommended to use the one-factor PHQ-8 (excluding item 9) when applying an interview survey (e.g., by phone) as the administration mode for a general population, where an immediate intervention is difficult to provide or the inclusion of a particularly sensitive question might interfere with the survey accuracy.

Strengths and limitations

The main strength of this study is that it is the first to have performed a systematic review of the PHQ-9 regarding its internal structure, in terms of the structural validity, internal consistency, and cross-cultural/measurement invariance. This is the first study to quantitatively pool Cronbach α values for the PHQ-9 in a meta-analysis. On the other hand, a limitation of this study is that it did not assess other psychometric properties such as convergent validity, discriminant validity, criterion validity, test–retest reliability, measurement error, or responsiveness. It is recommended that future systematic reviews assess these other psychometric properties reported for the one-factor PHQ-9 and PHQ-8 (excluding item 9) scored using a four-point Likert scale. Another limitation is this study only included peer-reviewed journal studies published in English, potentially leading to selection bias.

Conclusions

This study reviewed 98 studies reported on in 90 reports on the internal structure (structural validity, internal consistency, and cross-cultural/measurement invariance) of the PHQ-9 that were conducted in a wide variety of countries, languages, and settings in adult populations. The one-factor PHQ-9 and PHQ-8 (excluding item 9) scored using a four-point Likert scale have the best internal structures based on the current evidence. There was sufficient high-quality evidence for structural validity of the one-factor PHQ-9 based on CFA, internal consistency, and cross-cultural/measurement invariance across various demographic groups. It is recommended that the one-factor PHQ-8 (excluding item 9) is used in interview surveys (e.g., by phone) of the general population where it is difficult to provide an immediate intervention. Future systematic reviews should assess the other psychometric properties of the one-factor PHQ-9 and PHQ-8 scored using a four-point Likert scale, including their convergent validity, discriminant validity, criterion validity, test–retest reliability, measurement error, and responsiveness.

Author contributions

E.-H.L. conceived the study. All authors were involved in selecting articles, in assessing the methodological quality of each study and the quality of the measurement properties, in evaluating the available evidence. Two authors (D.C. and E.-H.L.) were involved in writing this manuscript. All authors approved the final submitted version of the manuscript.

Availability of data and materials

The data included in the review were retrieved from published studies.

Ethics approval and consent to participants

Not applicable.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

All authors have no competing interests.

Acknowledgments

We thank Nawon Kim (librarian at the Yonsei Medical Library, Yonsei University, Seoul, Republic of Korea) for supporting the data search and Hyun Young Lee (statistician at the Clinical Trial Center, Ajou University Hospital, Suwon, Republic of Korea) for supporting the statistical analyses.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.anr.2024.12.005.

References

1. World Health Organization. Depression. 2023. [cited 2023 December 11]. Available from https://www.who.int/health-topics/depression#tab=tab_1.
2. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM-IV-TR). 4th ed. Washington DC: American Psychiatric Association; 2000.
3. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16(9):606–13. https://doi.org/10.1046/j.1525-1497.2001.016009606.x
4. van Dijk SEM, Adriaanse MC, van der Zwaan L, Bosmans JE, van Marwijk HWJ, van Tulder MW, et al. Measurement properties of depression questionnaires in patients with diabetes: a systematic review. Qual Life Res. 2018;27:1415–30. https://doi.org/10.1007/s11136-018-1782-y
5. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res. 2012;21:651–7. https://doi.org/10.1007/s11136-011-9960-1
6. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: practical guides to biostatistics and epidemiology. Cambridge: Cambridge University Press; 2011.
7. González-Blanch C, Medrano LA, Muñoz-Navarro R, Ruíz-Rodríguez P, Moriana JA, Limonero JT, et al. Factor structure and measurement invariance across various demographic groups and over time for the PHQ-9 in primary care patients in Spain. PLoS One. 2018;13(2):e0193356. https://doi.org/10.1371/journal.pone.0193356
8. Lamela D, Soreira C, Matos P, Morais A. Systematic review of the factor structure and measurement invariance of the Patient Health Questionnaire-9 (PHQ-9) and validation of the Portuguese version in community settings. J Affect Disord. 2020;276:220–33. https://doi.org/10.1016/j.jad.2020.06.066
9. Patel JS, Oh Y, Rand KL, Wu W, Cyders MA, Kroenke K, et al. 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. 2019;36(9):813–23. https://doi.org/10.1002/da.22940
10. Rahman MA, Dhira TA, Sarker AR, Mehareen J. Validity and reliability of the Patient Health Questionnaire scale (PHQ-9) among university students of Bangladesh. PLoS One. 2022;17(6):e0269634. https://doi.org/10.1371/journal.pone.0269634
11. Villarreal-Zegarra D, Copez-Lonzoy A, Bernabé-Ortiz A, Melendez-Torres GJ, Bazo-Alvarez JC. Valid group comparisons can be made with the Patient Health Questionnaire (PHQ-9): a measurement invariance study across groups by demographic characteristics. PLoS One. 2019;14(9):e0221717. https://doi.org/10.1371/journal.pone.0221717
12. Carroll HA, Hook K, Perez OFR, Denckla C, Vince CC, Ghebrehiwet S, et al. Establishing reliability and validity for mental health screening instruments in resource-constrained settings: systematic review of the PHQ-9 and key recommendations. Psychiatr Res. 2020;291:113236. https://doi.org/10.1016/j.psychres.2020.113236
13. El-Den S, Chen TF, Gan YL, Wong E, O'Reilly CL. The psychometric properties of depression screening tools in primary healthcare settings: a systematic

review. J Affect Disord. 2018;225:503–22. https://doi.org/10.1016/j.jad.2017.08.060
14. Patrick S, Connick P. Psychometric properties of the PHQ-9 depression scale in people with multiple sclerosis: a systematic review. PLoS One. 2019;14(2): e0197943. https://doi.org/10.1371/journal.pone.0197943
15. Costantini L, Pasquarella C, Odone A, Colucci ME, Costanza A, Serafini G, et al. Screening for depression in primary care with Patient Health Questionnaire-9 (PHQ-9): a systematic review. J Affect Disord. 2021;279:473–83. https://doi.org/10.1016/j.jad.2020.09.131
16. Fekadu A, Demissie M, Birhane R, Medhin G, Bitew T, Hailemariam M, et al. Under detection of depression in primary care settings in low and middle-income countries: a systematic review and meta-analysis. Syst Rev. 2022;11:21. https://doi.org/10.1186/s13643-022-01893-9
17. Manea L, Gilbody S, McMillan D. A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. Gen Hosp Psychiatr. 2015;37(1):67–75. https://doi.org/10.1016/j.genhosppsych.2014.09.009
18. Negeri ZF, Levis B, Sun Y, He C, Krishnan A, Wu Y, et al. Accuracy of the Patient Health Questionnaire-9 for screening to detect major depression: updated systematic review and individual participant data meta-analysis. BMJ. 2021;375:n2183. https://doi.org/10.1136/bmj.n2183
19. Yin L, Teklu S, Pham H, Li R, Tahir P, Garcia ME. Validity of the Chinese language patient health questionnaire 2 and 9: a systematic review. Health Equity. 2022;6(1):574–94. https://doi.org/10.1089/heq.2022.0030
20. Mokkink LB, de Vet HCW, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, et al. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27:1171–9. https://doi.org/10.1007/s11136-017-1765-4
21. Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27:1147–57. https://doi.org/10.1007/s11136-018-1798-3
22. Terwee CB, Jansma EP, Riphagen II, de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual Life Res. 2009;18:1115–23. https://doi.org/10.1007/s11136-009-9528-5
23. Byrne BM. Factor analytic models: viewing the structure of an assessment instrument from three perspectives. J Pers Assess. 2005;85(1):17–32. https://doi.org/10.1207/s15327752jpa8501_02
24. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. https://doi.org/10.1136/bmj.n71
25. Hayes AF, Coutts JJ. Use omega rather than Cronbach's alpha for estimating reliability. But…. Commun Methods Meas. 2020;14(1):1–24. https://doi.org/10.1080/19312458.2020.1718629
26. Lee J, Lee EH, Chae D. eHealth literacy instruments: systematic review of measurement properties. J Med Internet Res. 2021;23(11):e30644. https://doi.org/10.2196/30644
27. Perreira TA, Morin AJ, Hebert M, Gillet N, Houle SA, Berta W. The short form of the Workplace Affective Commitment Multidimensional Questionnaire (WACMQ-S): a bifactor-ESEM approach among healthcare professionals. J Vocat Behav. 2018;106:62–83. https://doi.org/10.1016/j.jvb.2017.12.004
28. Chen YJ, Tang TLP. Attitude toward and propensity to engage in unethical behavior: measurement invariance across major among university students. J Bus Ethics. 2006;69:77–93. https://doi.org/10.1007/s10551-006-9069-6
29. R Core Team. R: a language and environment for statistical computing. 2010 [cited 2023 Dec 5]. Available from: http://www.R-project.orgversion 4.3.2
30. Lee J, Lee EH, Moon SH. Systematic review of the measurement properties of the Depression Anxiety Stress Scales-21 by applying updated COSMIN methodology. Qual Life Res. 2019;28:2325–39. https://doi.org/10.1007/s11136-019-02177-x
31. Brattmyr M, Lindberg MS, Solem S, Hjemdal O, Havnen A. Factor structure, measurement invariance, and concurrent validity of the Patient Health Questionnaire-9 and the Generalized Anxiety Disorder scale-7 in a Norwegian psychiatric outpatient sample. BMC Psychiatr. 2022;22:461. https://doi.org/10.1186/s12888-022-04101-z
32. Jiraniramai S, Wongpakaran T, Angkurawaranon C, Jiraporncharoen W, Wongpakaran N. Construct validity and differential item functioning of the PHQ-9 among health care workers: rasch analysis approach. Neuropsychiatric Dis Treat. 2021;17:1035–45. https://doi.org/10.2147/NDT.S271987
33. Hoekstra R, Vugteveen J, Warrens M, Kruyen P. An empirical analysis of alleged misunderstandings of coefficient alpha. Int J Soc Res Methodol. 2019;22(4):351–64. https://doi.org/10.1080/13645579.2018.1547523
34. Thorpe JM, Smith D, Kuzla N, Scott L, Ersek M. Does mode of survey administration matter? using measurement invariance to validate the mail and telephone versions of the bereaved family survey. J Pain Symptom Manag. 2016;51(3):546–56. https://doi.org/10.1016/j.jpainsymman.2015.11.006
35. Alpizar D, Laganá L, Plunkett SW, French BF. Evaluating the eight-item Patient Health Questionnaire's psychometric properties with Mexican and Central American descent university students. Psychol Assess. 2018;30(6):719–28. https://doi.org/10.1037/lat0000087
36. Kroenke K, Spitzer RL, Williams JB, Löwe B. The Patient Health Questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen Hosp Psychiatr. 2010;32(4):345–59. https://doi.org/10.1016/j.genhosppsych.2010.03.006
37. Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord. 2009;114(1-3):163–73. https://doi.org/10.1016/j.jad.2008.06.026
38. De Man J, Absetz P, Sathish T, Desloge A, Haregu T, Oldenburg B, et al. Are the PHQ-9 and GAD-7 suitable for use in India? a psychometric analysis. Front Psychol. 2021;12:676398. https://doi.org/10.3389/fpsyg.2021.676398
39. Harry ML, Waring SC. The measurement invariance of the Patient Health Questionnaire-9 for American Indian adults. J Affect Disord. 2019;254:59–68. https://doi.org/10.1016/j.jad.2019.05.017
40. Keum BT, Miller MJ, Inkelas KK. Testing the factor structure and measurement invariance of the PHQ-9 across racially diverse U.S. college students. Psychol Assess. 2018;30(8):1096–106. https://doi.org/10.1037/pas0000550
41. Quiñonez-Freire C, Vara MD, Tomás JM, Baños RM. Psychometric properties of the Spanish version of the Patient Health Questionnaire-9 in users of the Ecuadorian public health care system. Rev Latinoam Psicol. 2021;53:210–7. https://doi.org/10.14349/rlp.2021.v53.23
42. Schuler M, Strohmayer M, Mühlig S, Schwaighofer B, Wittmann M, Faller H, et al. Assessment of depression before and after inpatient rehabilitation in COPD patients: psychometric properties of the German version of the Patient Health Questionnaire (PHQ-9/PHQ-2). J Affect Disord. 2018;232:268–75. https://doi.org/10.1016/j.jad.2018.02.037
43. Wisting L, Johnson SU, Bulik CM, Andreassen OA, Rø Ø, Bang L. Psychometric properties of the Norwegian version of the Patient Health Questionnaire-9 (PHQ-9) in a large female sample of adults with and without eating disorders. BMC Psychiatr. 2021;21:6. https://doi.org/10.1186/s12888-020-03013-0
44. Bianchi R, Verkuilen J, Toker S, Schonfeld IS, Gerber M, Brähler E, et al. Is the PHQ-9 a unidimensional measure of depression? a 58,272-participant study. Psychol Assess. 2022;34(6):595–603. https://doi.org/10.1037/pas0001124
45. Doi S, Ito M, Takebayashi Y, Muramatsu K, Horikoshi M. Factorial validity and invariance of the Patient Health Questionnaire (PHQ)-9 among clinical and non-clinical populations. PLoS One. 2018;13(7):e0199235. https://doi.org/10.1371/journal.pone.0199235
46. Rosario-Hernández E, Rovira-Millán LV, Merino-Soto C, Angulo-Ramos M. Review of the psychometric properties of the Patient Health Questionnaire-9 (PHQ-9) Spanish version in a sample of Puerto Rican workers. Front Psychiatr. 2023;14:1024676. https://doi.org/10.3389/fpsyt.2023.1024676
47. Arrieta J, Aguerrebere M, Raviola G, Flores H, Elliott P, Espinosa A, et al. Validity and utility of the Patient Health Questionnaire (PHQ)-2 and PHQ-9 for screening and diagnosis of depression in rural Chiapas, Mexico: a cross-sectional study. J Clin Psychol. 2017;73(9):1076–90. https://doi.org/10.1002/jclp.22390
48. Baldellou Lopez M, Goldstein LH, Robinson EJ, Vitoratou S, Chalder T, Carson A, et al. Validation of the PHQ-9 in adults with dissociative seizures. J Psychosom Res. 2021;146:110487. https://doi.org/10.1016/j.jpsychores.2021.110487
49. Barthel D, Barkmann C, Ehrhardt S, Schoppen S, Bindt C, International CDS Study Group. Screening for depression in pregnant women from Côte d'Ivoire and Ghana: psychometric properties of the Patient Health Questionnaire-9. J Affect Disord. 2015;187:232–40. https://doi.org/10.1016/j.jad.2015.06.042
50. Dadfar M, Kalibatseva Z, Lester D. Reliability and validity of the Farsi version of the patient health questionnaire-9 (PHQ-9) with Iranian psychiatric outpatients. Trends Psychiatr Psychother. 2018;40(2):144–51. https://doi.org/10.1080/13674676.2019.1699042
51. Errazuriz A, Beltrán R, Torres R, Passi-Solar A. The validity and reliability of the PHQ-9 and PHQ-2 on screening for Major Depression in Spanish speaking immigrants in Chile: a cross-sectional study. Int J Environ Res Publ Health. 2022;19(21):13975. https://doi.org/10.3390/ijerph192113975
52. Familiar I, Ortiz-Panozo E, Hall B, Vieitez I, Romieu I, Lopez-Ridaura R, et al. Factor structure of the Spanish version of the patient health questionnaire-9 in Mexican women. Int J Methods Psychiatr Res. 2015;24(1):74–82. https://doi.org/10.1002/mpr.1461
53. Gelaye B, Williams MA, Lemma S, Deyessa N, Bahretibeb Y, Shibre T, et al. Validity of the patient health questionnaire-9 for depression screening and diagnosis in East Africa. Psychiatr Res. 2013;210(2):653–61. https://doi.org/10.1016/j.psychres.2013.07.015
54. González-Rivera JA. Validation and dimensionality pf patient health questionnaire for depression (PHQ-8 and PHQ-9) in Hispanic LGBT+ community. Int J Recent Sci Res. 2019;10(12):36670–6. https://doi.org/10.24327/ijrsr.2020.1012.4970
55. Huang FY, Chung H, Kroenke K, Delucchi KL, Spitzer RL. Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. J Gen Intern Med. 2006;21(6):547–52. https://doi.org/10.1111/j.1525-1497.2006.00409.x
56. Kigozi G. Confirmatory factor analysis of the Patient Health Questionnaire-9: a study amongst tuberculosis patients in the Free State province. South Afr Infect Dis. 2020;35(1):a242. https://doi.org/10.4102/sajid.v35i1.242
57. Kim YE, Lee B. The psychometric properties of the patient health questionnaire-9 in a sample of Korean university students. Psychiatry Investig. 2019;16(12):904–10. https://doi.org/10.30773/pi.2019.0226
58. Ma S, Yang J, Yang B, Kang L, Wang P, Zhang N, et al. The patient health questionnaire-9 vs. The Hamilton rating scale for depression in assessing major depressive disorder. Front Psychiatr. 2021;12:747139. https://doi.org/10.3389/fpsyt.2021.747139
59. Maroufizadeh S, Omani-Samani R, Almasi-Hashiani A, Amini P, Sepidarkish M. The reliability and validity of the Patient Health Questionnaire-9 (PHQ-9) and

PHQ-2 in patients with infertility. Reprod Health. 2019;16:137. https://doi.org/10.1186/s12978-019-0802-x
60. Merz EL, Malcarne VL, Roesch SC, Riley N, Sadler GR. A multigroup confirmatory factor analysis of the Patient Health Questionnaire-9 among English- and Spanish-speaking Latinas. Cult Divers Ethnic Minor Psychol. 2011;17(3):309–16. https://doi.org/10.1037/a0023883
61. Nguyen TQ, Bandeen-Roche K, Bass JK, German D, Nguyen NT, Knowlton AR. A tool for sexual minority mental health research: the Patient Health Questionnaire (PHQ-9) as a depressive symptom severity measure for sexual minority women in Viet Nam. J Gay Lesb Ment Health. 2016;20(2):173–91. https://doi.org/10.1080/19359705.2015.1080204
62. Odero SA, Mwangi P, Odhiambo R, Nzioka BM, Shumba C, Ndirangu-Mugo E, et al. Psychometric evaluation of PHQ-9 and GAD-7 among community health volunteers and nurses/midwives in Kenya following a nation-wide telephonic survey. Front Psychiatr. 2023;14:1123839. https://doi.org/10.3389/fpsyt.2023.1123839
63. Pranckeviciene A, Saudargiene A, Gecaite-Stonciene J, Liaugaudaite V, Griskova-Bulanova I, Simkute D, et al. Validation of the patient health questionnaire-9 and the generalized anxiety disorder-7 in Lithuanian student sample. PLoS One. 2022;17(1):e0263027. https://doi.org/10.1371/journal.pone.0263027
64. Rafiey H, Alipour F, LeBeau R, Salimi Y, Ahmadi S. Factor structure of Persian translation of the Patient Health Questionnaire in Iranian earthquake survivors. Iran J Psychiatry Behav Sci. 2018;12(4):e59416. https://doi.org/10.5812/ijpbs.59416
65. Reich H, Rief W, Brähler E, Mewes R. Cross-cultural validation of the German and Turkish versions of the PHQ-9: an IRT approach. BMC Psychol. 2018;6:26. https://doi.org/10.1186/s40359-018-0238-z
66. Tibubos AN, Beutel ME, Schulz A, Klein EM, Brähler E, Michal M, et al. Is assessment of depression equivalent for migrants of different cultural backgrounds? results from the German population-based Gutenberg Health Study (GHS). Depress Anxiety. 2018;35(12):1178–89. https://doi.org/10.1186/s12888-021-03234-x
67. Titov N, Dear BF, McMillan D, Anderson T, Zou J, Sunderland M. Psychometric comparison of the PHQ-9 and BDI-II for measuring response during treatment of depression. Cognit Behav Ther. 2011;40(2):126–36. https://doi.org/10.1080/16506073.2010.550059
68. Xiong N, Fritzsche K, Wei J, Hong X, Leonhart R, Zhao X, et al. Validation of Patient Health Questionnaire (PHQ) for major depression in Chinese outpatients with multiple somatic symptoms: a multicenter cross-sectional study. J Affect Disord. 2015;174:636–43. https://doi.org/10.1016/j.jad.2014.12.042
69. Amtmann D, Kim J, Chung H, Bamer AM, Askew RL, Wu S, et al. Comparing CESD-10, PHQ-9, and PROMIS depression instruments in individuals with multiple sclerosis. Rehabil Psychol. 2014;59(2):220–9. https://doi.org/10.1037/a0035919
70. Baas KD, Cramer AO, Koeter MW, van de Lisdonk EH, van Weert HC, Schene AH. Measurement invariance with respect to ethnicity of the patient health questionnaire-9 (PHQ-9). J Affect Disord. 2011;129(1-3):229–35. https://doi.org/10.1016/j.jad.2010.08.026
71. Barroso SM, Melo APS, Silva MAd, Guimarães MDC. Evaluation of the Brazilian version of patient health questionnaire (PHQ-9) in quilombola population using the item response theory. Salud Ment. 2019;42(1):43–50. https://doi.org/10.17711/SM.0185-3325.2019.006
72. Bélanger E, Thomas KS, Jones RN, Epstein-Lubow G, Mor V. Measurement validity of the Patient-Health Questionnaire-9 in US nursing home residents. Int J Geriatr Psychiatr. 2019;34(5):700–8. https://doi.org/10.1002/gps.5074
73. Chung H, Kim J, Askew RL, Jones SM, Cook KF, Amtmann D. Assessing measurement invariance of three depression scales between neurologic samples and community samples. Qual Life Res. 2015;24:1829–34. https://doi.org/10.1007/s11136-015-0927-5
74. Donlan W, Lee J. Screening for depression among indigenous Mexican migrant farmworkers using the Patient Health Questionnaire-9. Psychol Rep. 2010;106(2):419–32. https://doi.org/10.2466/pr0.106.2.419-432
75. Fischer HF, Tritt K, Klapp BF, Fliege H. How to compare scores from different depression scales: equating the patient health questionnaire (PHQ) and the ICD-10-symptom rating (ISR) using item response theory. Int J Methods Psychiatr Res. 2011;20(4):203–14. https://doi.org/10.1002/mpr.350
76. Galenkamp H, Stronks K, Snijder MB, Derks EM. Measurement invariance testing of the PHQ-9 in a multi-ethnic population in Europe: the HELIUS study. BMC Psychiatr. 2017;17:349. https://doi.org/10.1186/s12888-017-1506-9
77. Gómez-Gómez I, Benítez I, Bellón J, Moreno-Peral P, Oliván-Blázquez B, Clavería A, et al. Utility of PHQ-2, PHQ-8 and PHQ-9 for detecting major depression in primary health care: a validation study in Spain. Psychol Med. 2023;53:5625–35. https://doi.org/10.1017/S0033291722002835
78. Hirsch O, Donner-Banzhoff N, Bachmann V. Measurement equivalence of four psychological questionnaires in native-born Germans, Russian-speaking immigrants, and native-born Russians. J Transcult Nurs. 2013;24(3):225–35. https://doi.org/10.1177/1043659613482003
79. Moreno-Agostino D, Chua KC, Peters TJ, Scazufca M, Araya R. Psychometric properties of the PHQ-9 measure of depression among Brazilian older adults. Aging Ment Health. 2022;26(11):2285–90. https://doi.org/10.1080/13607863.2021.1963951
80. Walker ER, Engelhard Jr G, Thompson NJ. Using Rasch measurement theory to assess three depression scales among adults with epilepsy. Seizure. 2012;21(6):437–43. https://doi.org/10.1016/j.seizure.2012.04.009
81. Christensen KS, Oernboel E, Zatzick D, Russo J. Screening for depression: Rasch analysis of the structural validity of the PHQ-9 in acutely injured trauma survivors. J Psychosom Res. 2017;97:18–22. https://doi.org/10.1016/j.jpsychores.2017.03.117
82. Dyer JR, Williams R, Bombardier CH, Vannoy S, Fann JR. Evaluating the psychometric properties of 3 depression measures in a sample of persons with traumatic brain injury and major depressive disorder. J Head Trauma Rehabil. 2016;31(3):225–32. https://doi.org/10.1097/HTR.0000000000000177
83. Gothwal VK, Bagga DK, Sumalini R. Rasch validation of the PHQ-9 in people with visual impairment in South India. J Affect Disord. 2014;167:171–7. https://doi.org/10.1016/j.jad.2014.06.019
84. Lamoureux EL, Tee HW, Pesudovs K, Pallant JF, Keeffe JE, Rees G. Can clinicians use the PHQ-9 to assess depression in people with vision loss? Optom Vis Sci. 2009;86(2):139–45. https://doi.org/10.1097/OPX.0b013e318194eb47
85. Pedersen SS, Mathiasen K, Christensen KB, Makransky G. Psychometric analysis of the Patient Health Questionnaire in Danish patients with an implantable cardioverter defibrillator (The DEFIB-WOMEN study). J Psychosom Res. 2016;90:105–12. https://doi.org/10.1016/j.jpsychores.2016.09.010
86. Williams RT, Heinemann AW, Bode RK, Wilson CS, Fann JR, Tate DG. Improving measurement properties of the Patient Health Questionnaire-9 with rating scale analysis. Rehabil Psychol. 2009;54(2):198–203. https://doi.org/10.1037/a0015529
87. Zhong Q, Gelaye B, Fann JR, Sanchez SE, Williams MA. Cross-cultural validity of the Spanish version of PHQ-9 among pregnant Peruvian women: a Rasch item response theory analysis. J Affect Disord. 2014;158:148–53. https://doi.org/10.1016/j.jad.2014.02.012
88. Jones SM, Ludman EJ, McCorkle R, Reid R, Bowles EJ, Penfold R, et al. A differential item function analysis of somatic symptoms of depression in people with cancer. J Affect Disord. 2015;170:131–7. https://doi.org/10.1016/j.jad.2014.09.002
89. Alpizar D, Plunkett SW, Whaling K. Reliability and validity of the 8-item Patient Health Questionnaire for measuring depressive symptoms of Latino emerging adults. J Lat Psychol. 2018;6(2):115–30. https://doi.org/10.1037/lat0000087
90. Fabian KE, Fannoh J, Washington GG, Geninyan Weetol WB, Nyachienga B, Cyrus G, et al. Psychometric properties of two mental health screening tools in southeast Liberia: the Liberian Distress Screener and Patient Health Questionnaire. Transcult Psychiatr. 2022;59(4):425–37. https://doi.org/10.1177/13634615221107201
91. Pagán-Torres OM, González-Rivera JA, Rosario-Hernández E. Psychometric analysis and factor structure of the Spanish version of the Eight-Item Patient Health Questionnaire in a general sample of Puerto Rican adults. Hisp J Behav Sci. 2020;42(3):401–15. https://doi.org/10.1177/0739986320926524
92. Forkmann T, Gauggel S, Spangenberg L, Brähler E, Glaesmer H. Dimensional assessment of depressive severity in the elderly general population: psychometric evaluation of the PHQ-9 using Rasch Analysis. J Affect Disord. 2013;148(2-3):323–30. https://doi.org/10.1016/j.jad.2012.12.019
93. Horton M, Perry AE. Screening for depression in primary care: a Rasch analysis of the PHQ-9. BJPsych Bull. 2016;40(5):237–43. https://doi.org/10.1192/pb.bp.114.050294
94. Gothwal VK, Bagga DK, Bharani S, Sumalini R, Reddy SP. The Patient Health Questionnaire-9: validation among patients with glaucoma. PLoS One. 2014;9(7):e101295. https://doi.org/10.1371/journal.pone.0101295
95. Downey L, Hayduk LA, Curtis JR, Engelberg RA. Measuring depression-severity in critically ill patients' families with the Patient Health Questionnaire (PHQ): tests for unidimensionality and longitudinal measurement invariance, with implications for CONSORT. J Pain Symptom Manag. 2016;51(5):938–46. https://doi.org/10.1016/j.jpainsymman.2015.12.303
96. Beard C, Hsu KJ, Rifkin LS, Busch AB, Björgvinsson T. Validation of the PHQ-9 in a psychiatric sample. J Affect Disord. 2016;193:267–73. https://doi.org/10.1016/j.jad.2015.12.075
97. Chen IP, Liu SI, Huang HC, Sun FJ, Huang CR, Sung MR, et al. Validation of the Patient Health Questionnaire for depression screening among the elderly patients in Taiwan. Int J Gerontol. 2016;10(4):193–7. https://doi.org/10.1016/j.ijge.2016.05.002
98. Hinz A, Mehnert A, Kocalevent RD, Brähler E, Forkmann T, Singer S, et al. Assessment of depression severity with the PHQ-9 in cancer patients and in the general population. BMC Psychiatr. 2016;16:22. https://doi.org/10.1186/s12888-016-0728-6
99. Janssen EP, Köhler S, Stehouwer CD, Schaper NC, Dagnelie PC, Sep SJ, et al. 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. 2016;64(11):e201–6. https://doi.org/10.1111/jgs.14388
100. Miranda CAC, Scoppetta O. Factorial structure of the Patient Health Questionnaire-9 as a depression screening instrument for university students in Cartagena, Colombia. Psychiatr Res. 2018;269:425–9. https://doi.org/10.1016/j.psychres.2018.08.071
101. Nouwen A, Deschênes SS, Balkhiyarova Z, Albertorio-Díaz JR, Prokopenko I, Schmitz N. 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. J Affect Disord. 2021;292:311–8. https://doi.org/10.1016/j.jad.2021.05.031
102. Petersen JJ, Paulitsch MA, Hartig J, Mergenthal K, Gerlach FM, Gensichen J. 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. 2015;170:138–42. https://doi.org/10.1016/j.jad.2014.08.053
103. Appiah R, Schutte L, Wilson Fadiji A, Wissing MP, Cromhout A. Factorial validity of the Twi versions of five measures of mental health and well-being in Ghana. PLoS One. 2020;15(8):e0236707. https://doi.org/10.1371/journal.pone.0236707
104. Chilcot J, Rayner L, Lee W, Price A, Goodwin L, Monroe B, et al. The factor structure of the PHQ-9 in palliative care. J Psychosom Res. 2013;75(1):60–4. https://doi.org/10.1016/j.jpsychores.2012.12.012
105. Hall BJ, Patel A, Lao L, Liem A, Mayawati EH, Tjipto S. Structural validation of the Patient Health Questionnaire-9 (PHQ-9) among Filipina and Indonesian female migrant domestic workers in Macao: structural validation of PHQ-9. Psychiatr Res. 2021;295:113575. https://doi.org/10.1016/j.psychres.2020.113575
106. Lee B. Factor structure and validity of the Korean version of the Patient Health Questionnaire-9 among early childhood teachers. Open Psychol J. 2021;14:69–75. https://doi.org/10.2174/1874350102114010069
107. Guo B, Kaylor-Hughes C, Garland A, Nixon N, Sweeney T, Simpson S, et al. Factor structure and longitudinal measurement invariance of PHQ-9 for specialist mental health care patients with persistent major depressive disorder: exploratory structural equation modelling. J Affect Disord. 2017;219:1–8. https://doi.org/10.1016/j.jad.2017.05.020
108. Shin C, Ko YH, An H, Yoon HK, Han C. Normative data and psychometric properties of the Patient Health Questionnaire-9 in a nationally representative Korean population. BMC Psychiatr. 2020;20:194. https://doi.org/10.1186/s12888-020-02613-0
109. Vu LG, Le LK, Dam AVT, Nguyen SH, Vu TTM, Trinh TTH, et al. Factor structures of Patient Health Questionnaire-9 instruments in exploring depressive symptoms of suburban population. Front Psychiatr. 2022;13:838747. https://doi.org/10.3389/fpsyt.2022.838747
110. Tibubos AN, Otten D, Zöller D, Binder H, Wild PS, Fleischer T, et al. Bidimensional structure and measurement equivalence of the Patient Health Questionnaire-9: sex-sensitive assessment of depressive symptoms in three representative German cohort studies. BMC Psychiatr. 2021;21:238. https://doi.org/10.1002/da.22831
111. Zhong Q, Gelaye B, Rondon M, Sánchez SE, García PJ, Sánchez E, et al. Comparative performance of patient health questionnaire-9 and Edinburgh postnatal depression scale for screening antepartum depression. J Affect Disord. 2014;162:1–7. https://doi.org/10.1016/j.jad.2014.03.028
112. Mattsson M, Sandqvist G, Hesselstrand R, Nordin A, Boström C. Validity and reliability of the Patient Health Questionnaire-8 in Swedish for individuals with systemic sclerosis. Rheumatol Int. 2020;40:1675–87. https://doi.org/10.1007/s00296-020-04641-1
113. Moehring A, Guertler D, Krause K, Bischof G, Rumpf HJ, Batra A, et al. Longitudinal measurement invariance of the patient health questionnaire in a German sample. BMC Psychiatr. 2021;21:386. https://doi.org/10.1186/s12888-021-03390-0
114. Pressler SJ, Subramanian U, Perkins SM, Gradus-Pizlo I, Kareken D, Kim J, et al. Measuring depressive symptoms in heart failure: validity and reliability of the Patient Health Questionnaire-8. Am J Crit Care. 2011;20(2):146–52. https://doi.org/10.4037/ajcc2010931
115. Granillo MT. Structure and function of the Patient Health Questionnaire-9 among Latina and non-Latina White female college students. J Soc Soc Work Res. 2012;3(2):80–93. https://doi.org/10.5243/jsswr.2012.6
116. Krause JS, Saunders LL, Bombardier C, Kalpakjian C. Confirmatory factor analysis of the Patient Health Questionnaire-9: a study of the participants from the spinal cord injury model systems. Pharm Manag PM R. 2011;3(6):533–40. https://doi.org/10.1016/j.pmrj.2011.03.003
117. Pavlov C, Egan K, Limbers C. Reliability and validity of the PHQ-8 in first-time mothers who used assisted reproductive technology. Hum Reprod Open. 2022;2022(2):hoac019. https://doi.org/10.1093/hropen/hoac019
118. Marcos-Nájera R, Le HN, Rodríguez-Muñoz MF, Olivares Crespo ME, Izquierdo Mendez N. The structure of the Patient Health Questionnaire-9 in pregnant women in Spain. Midwifery. 2018;62:36–41. https://doi.org/10.1016/j.midw.2018.03.011
119. Monteiro S, Bártolo A, Torres A, Pereira A, Albuquerque E. Examining the construct validity of the Portuguese version of the Patient Health Questionnaire-9 among college students. Psicologia. 2019;33(2):1–8. https://doi.org/10.17575/rpsicol.v33i2.1421