Manuscript Details

Manuscript number	JAD_2019_2432_R1
Title	Systematic Reviews of the Factor Structure and Measurement Invariance of the Patient Health Questionnaire-9 (PHQ-9) and Validation of the European Portuguese Version in Community Settings
Article type	Research Paper

Abstract

Background: This research sought to review studies that examined the factor structure of the PHQ-9 using a confirmatory factor analysis approach (Study 1); to review studies that tested the measurement invariance of the PHQ-9 (Study 2); to examine the psychometric properties of the European Portuguese version in the general population (Study 3). Methods: Using PRISMA guidelines, a search was performed on Web of Science, PsycINFO, and Scopus from 2001 to August 2019. Assessment of eligibility criteria and data extraction were conducted by two independent researchers (Studies 1 and 2). In Study 3, data were collected from 1479 Portuguese adults, using a cross-sectional design. The BDI-II and the GDS-15 were administered to examine convergent validity. Results: The systematic review identified four-factor models of the PHQ-9 (Study 1). Nineteen studies supported a one-factor model, whereas 12 found evidence for a two-factor model. Both models were supported in general, clinical, psychiatric, and international samples. Study 2 identified ten studies that examined PHQ-9 measurement invariance across 18 groups. The PHQ-9 measurement invariance was fully supported across studies. Study 3 revealed that a two-factor model showed a close fit to data in the European Portuguese version of the PHQ-9. Measurement invariance, reliability, and convergent and divergent validity were also established. Limitations: Study 3 did not include a gold standard measure of depression to evaluate PHQ-9 diagnostic properties. Conclusions: Conceptual implications of the findings are discussed, and recommendations for using the Portuguese version of the PHQ-9 as a screening measure in community settings are also highlighted.

Keywords	PHQ-9; factor structure; measurement invariance; multigroup confirmatory factor analysis; depression; systematic review

Corresponding Author	Diogo Lamela
Corresponding Author's Institution	Lusófona University of Porto
Order of Authors	Diogo Lamela, Cátia Soreira, Paula Matos, Ana Morais

Suggested reviewers	Rui Campos, Henrike Galenkamp, César González-Blanch, Carlos Arturo Cassiani Miranda

Submission Files Included in this PDF

File Name [File Type]

JAD_2019 cover letter R1.pdf [Cover Letter]

JAD_2019_2432 RESPONSE LETTER R1.pdf [Response to Reviewers]

Highlights PHQ-9 R1.docx [Highlights]

Abstract PHQ-9 JAD R1.docx [Abstract]

Title page PHQ-9 Portuguese version R1.docx [Title Page (with Author Details)]

PHQ-9 JAD R1 19may.docx [Manuscript File]

Conflict of Interest PHQ-9.docx [Conflict of Interest]

Author statement PHQ-9.docx [Author Statement]

Supplementary Material PHQ-9 R1.doc [Supplementary Material]

To view all the submission files, including those not included in the PDF, click on the manuscript title on your EVISE Homepage, then click 'Download zip file'.

Highlights

- We conducted the first systematic reviews addressing the factor structure and measurement invariance of the PHQ-9
- Stronger evidence was found for the one- and two-factor models of the PHQ-9
- Strong evidence was found that PHQ-9 is equivalent across groups
- A two-factor model of the Portuguese PHQ-9 showed the best fit to data
- The measurement invariance of the Portuguese version of the PHQ-9 was established

Abstract

Background: This research sought to review studies that examined the factor structure of the PHQ-9 using a confirmatory factor analysis approach (Study 1); to review studies that tested the measurement invariance of the PHQ-9 (Study 2); to examine the psychometric properties of the European Portuguese version in the general population (Study 3).

Methods: Using PRISMA guidelines, a search was performed on Web of Science, PsycINFO, and Scopus from 2001 to August 2019. Assessment of eligibility criteria and data extraction were conducted by two independent researchers (Studies 1 and 2). In Study 3, data were collected from 1479 Portuguese adults, using a cross-sectional design. The BDI-II and the GDS-15 were administered to examine convergent validity.

Results: The systematic review identified four-factor models of the PHQ-9 (Study 1). Nineteen studies supported a one-factor model, whereas 12 found evidence for a two-factor model. Both models were supported in general, clinical, psychiatric, and international samples. Study 2 identified ten studies that examined PHQ-9 measurement invariance across 18 groups. The PHQ-9 measurement invariance was fully supported across studies. Study 3 revealed that a two-factor model showed a close fit to data in the European Portuguese version of the PHQ-9. Measurement invariance, reliability, and convergent and divergent validity were also established.

Limitations: Study 3 did not include a gold standard measure of depression to evaluate PHQ-9 diagnostic properties.

Conclusions: Conceptual implications of the findings are discussed, and recommendations for using the Portuguese version of the PHQ-9 as a screening measure in community settings are also highlighted.

Keywords: PHQ-9; factor structure; measurement invariance; multigroup confirmatory factor analysis; depression; systematic review

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

Diogo Lamela, Cátia Soreira, Paula Matos and Ana Morais

Lusófona University of Porto

Author Note

Diogo Lamela, Cátia Soreira, Paula Matos and Ana Morais, Digital Human-Environment Interaction Lab, Lusófona University of Porto.

Correspondence concerning this article should be addressed to Diogo Lamela, Faculty of Psychology, Education, and Sports, Lusófona University of Porto, Rua Augusto Rosa 24, 4000-098, Porto, Portugal. E-mail: lamela@ulp.pt

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

Depression is a major public health problem in Western societies (Cassano and Fava, 2002) with epidemiologic research reporting it is one of the mental health conditions with the highest rates of incidence and prevalence (Kessler and Bromet, 2013; Thornicroft et al., 2017). Clinical levels of depression are associated with a higher risk of physical health problems, labor absenteeism, poverty, low-quality family relationships, and lower life expectancy (Evans-Lacko and Knapp, 2016; Lamela et al., 2017; Laursen et al., 2016; Whooley and Wong, 2013).

National health systems have been implementing universal preventive programs to reduce the incidence and prevalence of depression and the personal and societal costs of depression-related impairments (Cuijpers et al., 2007; Hegerl et al., 2008). Portugal has one of the highest estimated annual prevalence rates of depression among Western countries (World Health Organization, 2017). As only 10% of the patients in the primary health care system are diagnosed with a depressive disorder, these epidemiological data suggest a chronic underdiagnosis of this mental health condition in Portugal (Direção-Geral da Saúde, 2017). A top priority of the Portuguese National Plan for Mental Health is to increase the detection rates of depression by primary health care providers (Direção-Geral da Saúde, 2017). To achieve this priority, the Portugal National Plan for Mental Health aims to expand the psychiatric treatment of depression in primary care and promote the use of well-validated screening measures in both community and primary care settings (Direção-Geral da Saúde, 2017).

Among such screening measures, the Patient Health Questionnaire-9 (PHQ-9) is one of the most widely used self-reported measures in research and health care settings worldwide (Mitchell et al., 2016). The PHQ-9 was developed to be administered in primary care settings as a screener of the depressive symptoms specified by the *Diagnostic and Statistical Manual of Mental Disorders IV-TR*: anhedonia, depressed mood, sleep disturbance, fatigue, appetite

changes, low self-esteem, concentration problems, psychomotor disturbances, and suicidal ideation (Kroenke et al., 2001). The PHQ-9 can be scored using an algorithm method to diagnose major depressive disorder (MDD) or a severity-based method to classify different clinical levels of depressive symptoms, from minimal to severe depression (Kroenke et al., 2001). Besides, PHQ-9 has become one of the most commonly used depression screening measures in primary care and other clinical settings worldwide due to its ease of use combined with good accuracy and sensitivity (Levis et al., 2019). As DSM-TR-IV-driven measure, not examined the factor structure of the PHQ-9. The factor structure of the PHQ-9 has been extensively examined, and several alternative structures have received some empirical support. Due to these inconsistent results across studies, there is a need to systematically review the international literature regarding the factor structure of the PHQ-9 and to identify possible methodological and clinical sources of such variability. The fully understanding whether in measurement invariance is a question that also remains unanswered.

The PHQ-9 is available free of charge for non-commercial purposes in 49 languages and 32 additional cultural adaptations (https://www.phqscreeners.com). However, only a small portion of these international versions of the PHQ-9 has received psychometric validation (Barthel et al., 2015; Galenkamp et al., 2017; Nguyen et al., 2016). Two versions in the Portuguese language were developed to accommodate semantic specificities of Portuguese as spoken in Portugal (European Portuguese) and Brazil (Brazilian Portuguese). However, no previous research has examined the factor structure and measurement invariance of either of these Portuguese versions, constraining the use of the PHQ-9 in a universe of 250 million native Portuguese speakers.

The current paper is comprised of three studies conducted to address these gaps. In Study 1, we sought to review empirical studies that examined the factor structure of the PHQ-9 using a CFA approach. In Study 2, we reviewed the empirical studies that tested the measurement invariance of the PHQ-9. In Study 3, we sought to examine the factor structure,

measurement invariance, and convergent validity of the European Portuguese version in the general population of adult European Portuguese speakers.

2. Study 1: Systematic Review of the Factor Structure of the PHQ-9

The factor structure of the PHQ-9 has been examined in community, primary care, and clinical samples, across socially diverse populations in fifteen countries. To date, however, there has been little agreement on the optimal factor structure of the PHQ-9 (Krause et al., 2011). Individual studies have found support for several factor models of the PHQ-9, ranging from one-factor to three-factor structures (Bélanger et al., 2019; Marcos-Nájera et al., 2018). Such heterogeneity in factor structures has been justified by the criteria-driven nature of the measure or by the distinct statistical approaches used to test the PHQ-9 factor structure, for example, exploratory factor analysis (EFA) vs. confirmatory factor analysis (CFA) (Petersen et al., 2015).

So far, there has been little understanding about the extent of the inconsistency between factor structures and whether the proposed factor structures depend on participants' sociodemographic, cultural or clinical characteristics. The purpose of this review was to identify empirical articles that examined the factor structure of the PHQ-9 using a confirmatory factor analysis (CFA) approach.

2.1 Method

2.1.1 Strategy of searching

The review methods were informed by the PRISMA standards for reporting systematic reviews (Moher et al., 2009). We conducted a systematic search for peer-reviewed articles published in English, Spanish, Portuguese, or French through three electronic databases: Web of Science, Scopus, and PsycINFO (searches were queried on August 12, 2019). We used the following search terms and logic: TI=(Patient health questionnaire 9 OR patient health

questionnaire-9 OR PHQ-9) AND TS=(factor structure OR factor model OR factor OR factor solution OR confirmatory factor analysis OR confirmatory factor analyses OR CFA). Since the first paper regarding PHQ-9 was published in 2001, searches were limited to full-text articles published from 2001 and 2019. Further manual searching of reference lists from identified studies was also undertaken. These search parameters yielded the following number of hits in each database: Web of Science (88), Scopus (202), PsycINFO (50), and manual search (3).

2.1.2 Screening

First, initial search results were merged, and duplicates entries were removed. As a second step, we searched the abstract, title, and keyword fields. We excluded articles that: (1) did not examine the factor structure of PHQ-9 and (2) were entries of conference papers, study protocols, or dissertations, and theses. Subsequently, we downloaded full-text articles of the remaining records to assess the article's eligibility for the review. In this third step, we excluded articles that: (1) only examined the PHQ-9 factor structure via exclusively exploratory factor analysis, (2) examined the PHQ-9 factor structure in samples of adolescents, (3) combined adult and adolescent participants in the total sample, (4) examined the PHQ-9 factor structure in samples with less than 200 participants, and (5) were published in another language than English, Spanish, Portuguese or French. Thirty-three articles were included in the systematic review. The screening procedure is described in Figure 1.

2.1.3 Data extraction

Records were analyzed by one reviewer and checked for accuracy by a second reviewer. The extracted data included: author, year, sample size, participants' age range (or age *M* and *SD* if age range was not reported), participant characteristics, country, whether the preferred factor model was selected by comparison with competing models, and the selected factor structure (e.g., one-factor model, two-factor model). We also evaluated the goodness-of-

fit of the selected factor model, using the recommended guidelines for interpreting model fit measures (Hu & Bentler, 1999; Little, 2013; Patel et al., 2019): a) Root mean square error of approximation (RMSEA) and the standardized root mean squared residual (SRMR): exact fit = 0.00, close fit = 0.01–0.050, acceptable fit = 0.051–0.080, mediocre fit = 0.081–0.10, and poor fit ≥ than .010; b) Tucker–Lewis index (TLI) and comparative fit index (CFI): exact fit = 1.00, close fit = .95–.99, acceptable fit = .90–0.95, mediocre fit = .85–90, and poor fit ≤ .85.

2.2 Results and discussion

The summary of findings (Table A1) are presented in Supplemental Appendix A. We reported the sample size, age range (or *M* and *SD*), primary sample's characteristics, and country/region for each study. We also reported whether the preferred factor model was selected by comparison with competing models and the fit measures obtained by the preferred factor model.

Of the thirty-three studies that examined the factor structure of the PHQ-9 using a CFA approach, nineteen (57.6%) found support for a one-factor structure, twelve (36.4%) a two-factor structure, and two (6%) for bifactor or three-factor models. For those studies that confirmed a one-factor model, the sample sizes ranged from 202 to 1,986,783 participants (Bélanger et al., 2019; Williams et al., 2009). Ten studies (52.6%) tested the PHQ-9 factor structure in the general population, four (21.1%) in primary care patients, and five (26.3%) in clinical groups, including multiple sclerosis patients (Amtmann et al., 2014) and patients with HIV infection (Crane et al., 2010). A one-factor model was selected after comparison with competing two-factor models in ten of the CFA studies (52.6%). Despite reporting a better fit for the two-factor model, three of these studies selected the one-factor structure due to the high intercorrelation between the two factors (Boothroyd et al., 2019; González-Blanch et al., 2018; Keum et al., 2018). In terms of the model's goodness-of-fit (Hu and Bentler, 1999; Masyn,

2013), four studies (21.1%) reported a close model-data fit, thirteen an acceptable fit (68.4%), and two a mediocre fit (10.5%).

The sample size of the twelve studies that supported a two-factor structure ranged from 300 to 31,366 participants (Chilcot et al., 2013; Patel et al., 2019). Two-factor models were confirmed in both general (41.7%) and clinical (58.3%) populations, including autistic adults (Arnold et al., 2019), patients diagnosed with cancer (Hinz et al., 2016), and primary care patients with MDD (Petersen et al., 2015). Ten (83%) of these studies tested competing factor models before the selection of the two-factor structure. In terms of model's goodness-of-fit, close and acceptable model-data fit were demonstrated in five (41.7%) and seven (58.3%) studies, respectively.

The CFA studies that found evidence for two-factor models demonstrated two slightly different factor structures.[1] Three studies (25%) obtained evidence for a latent factor comprising six cognitive/affective symptoms (anhedonia, depressed mood, low self-esteem, concentration problems, psychomotor disturbances, and suicidal ideation) and a latent factor comprising three somatic symptoms (sleep disturbance, fatigue, appetite changes) (Arnold et al., 2019; Chilcot et al., 2013; Patel et al., 2019).[2] Seven studies (66.7%) supported a different pattern of items' distribution on each of the latent factors: one comprising four non-somatic symptoms (anhedonia, depressed mood, low self-esteem, and suicidal ideation) and one with five somatic symptoms (sleep disturbance, fatigue, appetite changes, concentration problems, and psychomotor disturbances) (Elhai et al., 2012; Hinz et al., 2016; Janssen et al., 2016; Krause et al., 2011; Miranda and Scoppetta, 2018; Petersen et al., 2015; Zhong et al., 2014).

[1] One study did not provide precise information about how items were distributed among the two factors (Beard, Hsu, Rifkin, Busch, & Björgvinsson, 2016).

[2] Granillo (2012) reported a factor structure comprising a somatic factor with three items (sleep disturbance, fatigue, appetite changes) and a cognitive/affective factor. However, the tested version had seven items, since two items (concentration problems and psychomotor disturbances) were removed after the initial exploratory factor analysis (Granillo, 2012).

Taken together, the systematic search of literature regarding the PHQ-9 factor structure revealed four patterns of findings. First, the one-factor structure received higher empirical support than the two-factor model. Second, a higher proportion of two-factor models demonstrated close fits to data than one-factor models (Table A1). Third, better fits to data were generally obtained in CFAs with general/non-clinical populations in both one- and two-factor models. Fourth, little support was found for the hypothesis that the PHQ-9 factor model might depend on specific characteristics of samples (clinical vs. non-clinical), as previously suggested (Petersen et al., 2015).

Our systematic review indicated that one- and two-factor models were both supported in general, clinical, psychiatric, and international samples. This suggests that the heterogeneity in the PHQ-9 factor structures might not be exclusively explained by samples' sociodemographic diversity but instead by the absence of a conceptual model of how depressive symptoms are interrelated (see general discussion). The heterogeneity in the factor structures of the PHQ-9, along with the inconsistencies in items included in different sets of symptoms across different two-factor models, suggested that these previous results need to be interpreted with caution and that a further empirical examination of the factor structure of the PHQ-9 should be conducted.

3. Study 2: Systematic Review of Measurement Invariance of the PHQ-9

Measurement invariance is a critical condition to ensure the validity of psychological assessment procedures (Putnick and Bornstein, 2016). Measurement invariance provides psychometric evidence regarding whether a set of items will represent a targeted latent construct similarly across groups. Without establishing measurement invariance, researchers and clinicians do not have psychometric guarantees that potential differences in a psychological construct between groups and subsequent associations with other variables

reflect the inferred underlying psychological processes rather than a lack of measurement accuracy (Adolf et al., 2014).

In depression screening measures, measurement invariance assumes particular methodological significance, since scores from these screeners are used to identify individuals or sociodemographic groups who are at risk or in higher need of care (Siu et al., 2016). Without empirical evidence for the equivalence in response patterns across comparable sociodemographic groups, screening measures are more likely to produce Type 1 or Type 2 errors in identifying individuals at high risk of depression. However, comparing with the extensive research about the factor structure and the sensitivity of the PHQ-9 to detect individuals with clinical levels of depression, there is less information about the measurement invariance of the PHQ-9 across sociodemographic and clinical groups. The purpose of this review was to identify empirical articles that examined the measurement invariance of the PHQ-9 using a multigroup confirmatory factor analysis approach.

3.1 Method

3.1.1 Strategy of searching

The review methods were informed by the PRISMA standards for reporting systematic reviews (Moher et al., 2009). We conducted a systematic search for peer-reviewed articles published in English, Spanish, Portuguese, or French through three electronic databases: Web of Science, Scopus, and PsycINFO (searches were queried on August 12, 2019). We used the following search terms and logic: TI=(Patient health questionnaire-9 OR patient health questionnaire 9 OR PHQ-9) AND TS=(measurement invariance OR invariance OR multigroup confirmatory factor analysis OR multigroup confirmatory factor analysis OR multiple-group confirmatory factor analysis OR multigroup OR multigroup OR multiple-group OR MG-CFA OR configural OR metric OR scalar OR strict OR strong OR weak). Since the first paper regarding PHQ-9 was published in 2001, searches were limited to full-text articles published from 2001 and 2019. Further manual searching of reference lists from identified studies was

also undertaken. These search parameters yielded the following number of hits in each database: Web of Science (88), Scopus (202), PsycINFO (50), and manual search (0).

3.1.2 Screening

Figure 2 describes the screening procedure. First, initial search results were merged, and duplicates entries were removed. As a second step, we searched the abstract, title, and keyword fields. We excluded articles that: (1) did not examine the measurement invariance of PHQ-9 and (2) were entries of conference papers, study protocols, or dissertations, and theses. Subsequently, we downloaded full-text articles of the remaining records to assess the article's eligibility for the review. In this third step, we excluded articles that: (1) examined the PHQ-9 measurement invariance via another approach rather than multigroup CFA (e.g., multiple-indicator multiple-cause model), (2) did not perform or report statistics for nested model comparisons, including difference values between at least one fit measure for the compared invariance models (ΔRMSEA or ΔSRMR or ΔCFI), (3) determined measurement invariance based exclusively on chi-square difference tests ($\Delta\chi 2$), (4) combined data from adults and adolescents, and (5) were published in a language other than English, Spanish, Portuguese, or French. Ten articles were included in the systematic review.

3.1.3 Data extraction

Records were analyzed by one reviewer and checked for accuracy by a second reviewer. The extracted data included: author, year, sample size, participants' age range (or age *M* and *SD* if age range was not reported), participant characteristics, country, and the selected factor structure (e.g., one-factor model, two-factor model). For each study, we also indicated (1) which groups were used to evaluate PHQ-9 measurement invariance and (2) results obtained in the tests of measurement invariance, as reported by the study's authors.

3.2 Results and discussion

For each study, we reported the sample size, age range (or *M* and *SD*), characteristics of the primary sample, country/region, target groups, and the findings obtained in the different steps of measurement invariance testing (Tables B1 and B2).

Based on the ten extracted studies, the measurement invariance of the PHQ-9 was tested across nineteen different groups (Table S3). Sex and race/ethnicity were the sociodemographic groups most often used to examine the PHQ-9 measurement invariance, including in general population (Patel et al., 2019), college students (Keum et al., 2018), and primary care patients (González-Blanch et al., 2018). Surprisingly, measurement invariance across age and marital status groups were evaluated in only one study (González-Blanch et al., 2018). Two studies addressed the measurement invariance across education level (González-Blanch et al., 2018; Patel et al., 2019), and three studies examined measurement invariance across clinical conditions (Chung et al., 2015; Doi et al., 2018; Schuler et al., 2018). Invariance across measurement occasions was tested in two studies involving patients with chronic obstructive pulmonary disease (Schuler et al., 2018) and primary care patients (González-Blanch et al., 2018). Finally, the measurement invariance of the one-factor model was tested in seven studies (Galenkamp et al., 2017; Merz et al., 2011), while two evaluated the measurement invariance of a two-factor model (Miranda and Scoppetta, 2018; Patel et al., 2019) and one a bifactor model (Doi et al., 2018).

Measurement invariance was supported across eighteen groups and one measurement occasion. Only the study of patients with chronic obstructive pulmonary disease found partial scalar invariance across sex and partial scalar and strict invariance across measurement occasions (Schuler et al., 2018). The findings of the studies established measurement invariance of the PHQ-9 across sociodemographic variables and clinical conditions, which suggests that the PHQ-9 scores can be meaningfully compared between different sociodemographic and clinical groups.

However, significant variability was observed among the previous studies in the number of tests used to establish measurement invariance (Table S3). In particular, five of the ten studies (50%) tested the four steps recommended (Putnick and Bornstein, 2016) to support measurement invariance: (1) configural, equivalence of model form; (2) metric, equivalence of factor loadings; (3) scalar, equivalence of item intercepts; and (4) strict, equivalence of items' residuals or unique variances. Besides, three studies evaluated additional steps of measurement invariance (Gregorich, 2006), including dimensional invariance and invariance in factor variance and covariances (Doi et al., 2018; Patel et al., 2019; Schuler et al., 2018). Four studies tested three steps of invariance, skipping either the scalar invariance test (Merz et al., 2011) or the strict invariance test (Chung et al., 2015; Harry and Waring, 2019; Keum et al., 2018).

4. Study 3: Validation of the European Portuguese version of the PHQ-9

Study 3 was designed to examine three primary goals. The first was to determine the factor structure of the European Portuguese version of the PHQ in the general adult population of Portuguese speakers. We tested and compared four competing factor models of the PHQ-9 identified by previous research using a CFA approach (Figure 3) in Study 1:[3] Model 1, a one-factor model, comprising the nine items of PHQ-9 (Baas et al., 2011; Bélanger et al., 2019); Model 2, a two-factor model, comprising a cognitive/affective factor with six items (anhedonia, depressed mood, low self-esteem, concentration problems, psychomotor disturbances, and suicidal ideation) and a somatic factor with three items (sleep disturbance, fatigue, appetite changes) (Patel et al., 2019); Model 3, two-factor model, comprising a cognitive/affective factor with four cognitive/affective items (anhedonia, depressed mood, low

[3] We did not test a fifth-factor structure (three-factor model) found in our systematic search of the literature since that model was specifically created for pregnant women (i.e., comprising a factor with pregnancy-related depression symptoms) (Marcos-Nájera et al., 2018).

self-esteem, and suicidal ideation) and a factor with five somatic symptoms (sleep disturbance, fatigue, appetite changes, concentration problems, and psychomotor disturbances) (Petersen et al., 2015); Model 4, a bifactor model, with a general factor added to the factors of Model 2 (Doi et al., 2018). This model specified that each of the nine items loads on a general PHQ-9 factor in parallel to their loading on their respective factor (somatic or cognitive-affective).

The comparison of competing models derived from previous empirical research is a common strategy to determine the factor structure of the PHQ-9 (Krause et al., 2011; Patel et al., 2019; Petersen et al., 2015). However, prior studies have indiscriminately compared factor models obtained via EFA and CFA approaches. As these two statistical approaches do not produce mathematically equivalent results, the merit of this strategy to select competing models of PHQ-9 is questionable. In order to avoid this methodological flaw, our study offers the first test of competing factor models of the PHQ-9 drawn exclusively from prior studies that used a CFA approach to examine its factor structure as reviewed in Study 1.

After selection of the factor structure that best fit depressive symptoms as measured by the European Portuguese version of the PHQ-9, our second goal was to examine the measurement invariance of the PHQ-9 across sex, age, marital status, education level, and administration format (pencil-and-paper vs. online). Finally, the third goal was to examine the reliability and convergent validity of the Portuguese PHQ-9. Convergent validity was tested with two well-established measures of depressive symptoms in Portuguese context: the Beck Depression Inventory-II (Beck et al., 1996) and the Geriatric Depression Scale-15 (Sheikh and Yesavage, 1986).

4.1 Method

4.1.1 Participants

The dataset for the current study combined data from three independent samples. The total sample comprised 1479 adults residing in Portugal (69.8% women). The average participants' average age was 42.24 years (*SD* = 19.47; range: 18–96 years). In terms of marital

status, 25.3% of the participants were single, 56.5% married or cohabitating, 8.4% divorced, and 9.8% widowed. The median education for the sample was a college degree, 33% had earned a high school diploma, and 24% of the participants had not earned a high school diploma or equivalent. The distribution of the total sample across the sociodemographic categories is displayed in Table 4. A detailed description of the sociodemographic variables for each of the three independent samples is presented in Table C1 of Supplemental Appendix C.

4.1.2 Measures

Patient Health Questionnaire-9 (Kroenke et al., 2001)**.** The PHQ-9 assesses depressive symptoms linked to the DSM-V criteria for major depressive disorder (MDD). Individuals are asked to self-rate on how many days during the past two weeks they experienced the following symptoms: anhedonia, depressed mood, sleep disturbance, fatigue, appetite changes, low self-esteem, concentration problems, psychomotor disturbances, and suicidal ideation. Items are answered on a 4-point scale: “not at all = 0,” “several days = 1,” “more than half the days = 2,” and “nearly every day = 3”. The PHQ-9 total score is calculated by summing the nine items' scores (range 0–27). Higher total scores are indicative of greater symptoms of depression. The PHQ-9 can be used as a continuous measure or as a diagnostic algorithm to make a probable diagnosis of MDD. The original version of the PHQ-9 demonstrated high internal consistency with good sensitivity and specificity for identifying cases of MDD (Kroenke et al., 2001; Levis et al., 2019; Mitchell et al., 2016).

In the current study, we used the European Portuguese version of the PHQ-9 available on the official PHQ screeners website (https://www.phqscreeners.com). In order to confirm the quality of this adaptation, the European Portuguese version was back-translated to English by a bilingual specialist with a Ph.D. in Psychology prior to data collection. Close similarity between the original and back-translated versions suggested semantic and content equivalence between the two versions. In addition, we submitted the European Portuguese version to a

"think-aloud" group of twelve native Portuguese-speaking adults to ensure face validity (Hambleton et al., 2005). After answering the PHQ-9, participants in this group were asked to discuss their opinion regarding the: (1) suitability and accessibility of language; (2) intelligibility of items, instructions and response scale; and (3) meaning and interpretation of items. Based on this discussion, evidence for face validity was obtained. Therefore, we adopted this version as a linguistically and culturally suitable adaptation of the PHQ-9 to the Portuguese context.

Beck Depression Inventory-II (BDI-II) (Beck et al., 1996) is a 21-item self-report measure that assesses symptoms corresponding to the diagnostic criteria for depressive disorders listed in the DSM-IV. Items are answered using a 4-point scale (from 0 to 3); higher scores reflect greater levels of depressive symptoms. The Portuguese version of the BDI-II possesses satisfactory psychometric properties (Campos and Gonçalves, 2011). In the current study, the BDI-II showed excellent internal consistency (Cronbach's $\alpha = 0.90$; $n = 127$).

Geriatric Depression Scale-15 (GDS-15) (Sheikh and Yesavage, 1986) is a widely used screening measure to assess depressive symptoms in elderly individuals. Each item is answered using a dichotomic scale (1 *yes* or 0 *no*); higher scores reflect higher levels of depressive symptoms (range from 0 to 15). The Portuguese version of the GDS-15 possesses satisfactory psychometric properties (Apóstolo et al., 2014). In the current study, Cronbach's alpha for the GDS-15 was .80 ($n = 127$).

4.1.3 Procedure

The dataset for the current study combined three independent samples to increase the heterogeneity of participants' sociodemographic characteristics. Participants in the current research sample must have been at least 18 years old and be residents in Portugal. The Institutional Review Board approved all research projects described in this work. All participants provided written informed consent and did not receive financial compensation for their participation.

The first sample was collected for a cross-sectional study on personal stigma against depression. The assessment protocol included a sociodemographic form, the PHQ-9, and measures related to the personal stigma against depression. We randomly included the BDI-II in 25% of the assessment protocols with the intent of testing convergent validity between the PHQ-9 and the BDI-II. Participants were college students recruited at two universities in Northern Portugal and adults recruited from students' social networks. For data collection, we first contacted department chairs to ask for permission to approach students enrolled in their departments. After obtaining permission, members of the research team contacted the students in classrooms and asked for voluntary and anonymous participation.

A total of 282 students voluntarily completed the survey administered during class (response rate = 92%). We also employed a snowball sampling procedure by asking for students' assistance in the recruitment of other potential participants among their acquaintances. To increase the sample's sociodemographic heterogeneity, we asked for potential referrals who were not students in higher education. The assessment protocols and informed consent form were provided in sealed envelopes to the students, who handed them to the potential volunteers. Completed assessment protocols were then returned to the research team by the students two weeks later. We received 252 of the 327 assessment protocols we delivered using the snowball sampling procedure (response rate = 77%). In total, we collected data from 534 participants. We removed 19 participants for failing to complete at least 70% of the assessment protocol (Funk and Rogge, 2007). Ultimately, the final sample included in the current study consisted of 514 participants.

The second sample was collected using a cross-sectional online survey designed to explore the relationship between mental health and family functioning in adulthood. The survey was available on a Portuguese website hosted on a university server for six months. Participants were recruited via online forums, social media websites, and e-mails to institutional public entities' web accounts. Prior to statistical analysis, the data were cleaned

(Funk and Rogge, 2007). First, 45 participants were removed for not meeting the inclusion criteria. Second, 32 of the remaining participants were excluded due to a lack of effort or attention, using the cut-score of the Directed Questions Scale (Maniaci and Rogge, 2014), designed to detect careless or inattentive responses in online surveys. Finally, 26 of the remaining participants were removed for failing to complete at least 70% of the assessment protocol. In total, the data-cleaning process removed 103 respondents (12.3%), leaving a final sample of 738 participants.

The third and final sample was collected for the current study. Participants were elderly individuals (≥ 65 years) receiving care in senior centers, day-care centers, or nursing homes in three metropolitan areas of Northern Portugal (Porto, Aveiro, and Viana do Castelo). The assessment protocol comprised a sociodemographic form, the PHQ-9, and the GDS-15 in order to test the convergent validity between the PHQ-9 and the GDS-15. After information regarding the study was disclosed in local organizations, 239 elderly individuals volunteered to participate. An additional exclusion criterion for this sample was severe physical, sensorial, or cognitive impairments. After providing a detailed description of the study, written informed consent was obtained from participants. Data collection was carried out in the organizations' facilities with the assistance of trained researchers. The assessment protocol was completed using a self-administration format. Trained researchers were available to identify potential difficulties in completing the self-administered questionnaires and address queries in items' comprehension by the participants. From the original volunteers, 12 were removed for failing to complete at least 70% of the assessment protocol, leaving a final sample of 227 participants.

4.1.4 Data analysis strategy

To test the factor structure of the Portuguese version of the PHQ-9, we conducted CFA using a full information maximum likelihood estimator. To determine the factor model, we performed four independent CFAs on the entire sample, testing the competing PHQ-9 factor solutions found in previous research (Figure 3). Though there are no universally accepted

cutoff values for approximate fit indices (Kline, 2015), the competing model fits were evaluated using the Tucker–Lewis index (TLI), comparative fit index (CFI), the root mean square error of approximation (RMSEA), and the standardized root mean squared residual (SRMR). We used established guidelines to evaluate model fit (Hu and Bentler, 1999; Little, 2013; Patel et al., 2019): RMSEA and SRMR, exact fit = 0.00, close fit = 0.01–0.050, acceptable fit = 0.051–0.080, mediocre fit = 0.081–0.10, and poor fit ≥ than .010. For TLI and CFI, exact fit = 1.00, close fit = .95–.99, acceptable fit = .90–0.95, mediocre fit = .85–90, and poor fit ≤ .85.

Using single and multiple-group CFAs, we implemented a five-step procedure to examine the measurement invariance of the factor structure that best fits depressive symptoms as measured by the Portuguese version of the PHQ-9 (Gregorich, 2006; Putnick and Bornstein, 2016). This five-step procedure comprises five tests of specific levels of measurement invariance, conducted sequentially from the least to the most restrictive level of invariance: dimensional, configural, metric, scalar, and strict invariance.

In the first step, we assess dimensional invariance, which requires that the number of latent factors is equivalent across groups. (Gregorich, 2006). We conducted a series of single-group CFA to separately examine the fit of the factor structure found with the entire sample across each group (sex, age, marital status, education level, and administration format). We used RMSEA, SRMR, TLI, and CFI as the criteria to evaluate model fit. Second, we assessed configural invariance, requiring each common latent factor to have the same pattern of free and fixed loadings across groups. Support for configural invariance was established by fitting the selected factor structure to the groups within each sociodemographic variable using a multiple-group procedure. A good model fit suggested equivalent factor structures across groups.

If configural invariance was supported, the third step was to examine for metric invariance, which requires the equivalence of factor loadings across groups. Metric invariance was tested by constraining factor loadings to be equivalent in the groups. If configural

invariance was supported, the fourth step was to evaluate scalar invariance, which required the equivalence of item intercepts in addition to invariant loadings and the same pattern of item loadings on latent factors. Scalar invariance was tested by constraining the item intercepts to be equivalent in the groups along with the constraints in factor loadings of the metric invariance model. If scalar invariance was supported, the final step was to test strict invariance, defined as the equivalence of item residuals of metric and scalar invariant items. Strict invariance was determined by imposing equal constrains on item residual variance of the scalar invariance model.

In each step of this procedure, the more restricted model was accepted if (1) the RMSEA and CFI values indicated close or acceptable model fit; and (2) values for the $\Delta CFI \leq .010$ and the $\Delta RMSEA \leq .015$ were found, these metrics are widely used as rules of thumb of model fit indexes to determine invariance (Chen, 2007; Cheung and Rensvold, 2002; Little, 2013). We also report the significance of the $\Delta\chi^2$; however, we do not focus on χ^2 tests for model evaluation because they are overly sensitive to sample size and minor misspecifications (Chen, 2007; Cheung and Rensvold, 2002).

We then conducted *t*-tests and ANOVAs to compare PHQ-9 scores on demographic variables used in measurement invariance analyses. We tested the reliability of the PHQ-9 by performing internal consistency analyses (Cronbach is α). Cronbach's values of $\leq.70$, $\leq.80$, $\leq.90$, and $\geq .90$ are indicative of questionable, acceptable, good, and excellent internal consistency, respectively (Nunnally and Bernstein, 1994). Finally, we tested convergent validity by performing Pearson correlations between PHQ-9 and both BDI-II and GDS-15. CFA and measurement invariance analyses were conducted using the lavaan package in R (Rosseel, 2012). Other statistical analyses (ANOVA, *t*-test, correlations, and Cronbach's alpha) were performed using IBM SPSS 25.

4.2 Results

4.2.1 PHQ-9 factor structure

To determine the factor structure of the PHQ-9, we performed four independent single-group CFAs to compare the four competing factor structures described in past research (Figure 3). Table 1 displays fit indices for the four models. All four models possessed acceptable-to-close data fits. However, Model 2 provided a better fit than the one-factor (Model 1), the other two-factor (Model 3), and the bifactor (Model 4) models. Based on the goodness-of-fit indices, Model 2 demonstrated a close model-data fit, as demonstrated by the RMSEA and SRMR values between 0.020–0.050 and the TLI and CFI values within the .95–.99 range. Similar to previous findings (Patel et al., 2019), the correlation between the cognitive/affective factor and the somatic factor was moderate (r = .57). All items exhibited substantial factor saturation, as shown by their high factor loading (all λ .58–.89; all $p < 0.001$). As such, Model 2, corresponding to the two-factor structure comprising a factor of cognitive/affective symptoms (6 items) and a factor of somatic symptoms (3 items) (Arnold et al., 2019; Chilcot et al., 2013; Patel et al., 2019), was identified as the best-fit factor structure of the European Portuguese version of the PHQ-9. All further analyses were conducted for Model 2.

4.2.2 Measurement invariance across sex, age, marital status, educational level, and administration format

Using a single-group CFA procedure, we independently fit Model 2 to each group to examine dimensional invariance (Patel et al., 2019). All models revealed close to acceptable fit (Table 2), suggesting that the number of latent factors was equivalent across groups. As dimensional invariance was supported, we then assessed configural invariance, by testing Model 2's fit for the groups within each sociodemographic variable (sex, age, marital status, educational level, and administration format). The four multiple-group CFAs indicated that all models for each sociodemographic variable provided acceptable fits (Table 3). In addition, the factor loadings across the groups were similar, higher than .40, and significant at .001 (Table 2). The pattern of factor loadings across groups was marginally more similar in the somatic

factor than in the cognitive/affective factor. This comparable pattern of loadings of items across groups suggests that the two-factor structure is supported in the sex, age, marital status, educational level, and administration format groups.

As configural invariance was supported, we evaluated metric invariance by constraining factor loadings to be equivalent in the groups. As shown in Table 3, all models for sex, age, marital status, educational level, and administration format provided acceptable fits (Table 3). The ΔCFI (≤ 0.10) and ΔRMSEA (≤ 0.15) criteria were met for all four models, indicating acceptable metric invariance. These results suggest that each item contributed to the latent factors to a similar degree across groups, meaning the item loadings of both somatic and cognitive/affective factors are equivalent across groups.

As metric invariance was supported, we then examined scalar invariance by constraining the item intercepts to be equivalent for all groups, along with the constraints in factor loadings of the metric invariance model. As presented in Table 3, all models for sex, age, marital status, educational level, and administration format provided acceptable fits (Table 3). Values of the measures to test invariance (ΔCFI ≤ 0.10; ΔRMSEA ≤ 0.15) suggested that scalar invariance was established. These results indicate that mean differences in the latent factors capture all mean differences in the items' shared variance.

As scalar invariance was supported, we evaluated strict invariance, by constraining item residual variance of the scalar invariance model. As presented in Table 3, all models for sex, age, marital status, educational level, and administration format provided acceptable fits (Table 3). Values of the measures to test invariance (ΔCFI ≤ 0.10; ΔRMSEA ≤ 0.15) indicated that strict invariance was established. These results reveal the equivalence of the sum of specific variance and error variance (measurement error) across groups.

4.2.3 PHQ-9 and sociodemographic variables

Means and standards deviations for the entire PHQ-9 and the cognitive/affective and somatic factors across groups (sex, age, marital status, educational level, and administration format) are presented in Table 4. We tested whether the PHQ-9 total score and cognitive/affective and somatic factors differed between sociodemographic groups. Detailed information regarding values of test differences, effect sizes, and posthoc comparisons for PHQ-9 total score and cognitive/affective and somatic factors across groups is presented in Table C2 of Supplemental Appendix C.

Overall, women reported higher scores in PHQ-9 total and cognitive/affective and somatic factors than men (all $p < .001$; d between .30 and .42). The age group 35–60 years displayed lower scores for the somatic factor than the other two age groups ($p < .001$; $\eta^2 = 0.013$). The age groups did not differ in PHQ-9 total score and cognitive/affective factor. Regarding marital status, single and divorced/widowed individuals exhibited greater total PHQ-9 scores ($p < .001$; $\eta^2 = 0.01$). Divorced/widowed individuals also had higher scores for the cognitive/affective and somatic factors than married individuals ($\eta^2 = 0.008$ and $\eta^2 = 0.010$, respectively). Groups based on education levels did not differ in PHQ-9 total score and the cognitive/affective factor. In the somatic factor, individuals with nine or fewer years of education reported higher scores than individuals with a master/doctoral degree ($p < .05$; $\eta^2 = 0.007$). Finally, PHQ-9 total scores and scores for both factors did not differ in terms of administration format (pencil-and-paper vs. internet).

4.2.4 Internal consistency and convergent validity

Internal consistency was assessed using Cronbach's α. For the total score, α = .86, the somatic factor, α = .82., and the cognitive factor for the entire sample, α = .90. The internal consistency of the total PHQ-9 score across sociodemographic groups ranged from good to excellent, with the exception of the age group over 61 years, which had acceptable internal consistency (α = .75) (Table 4). To examine convergent validity, we computed correlation

coefficients between the total score of the PHQ-9, BDI-II, and GDS-15 using two subsamples of the entire sample. As expected, we found a strong positive correlation between the PHQ-9 and the BDI-II (n = 145; r = .76, p < .001), and a moderate positive correlation between the PHQ-9 and the GDS-15 (n = 227; r = .64, p < .001). Both PHQ-9 somatic and cognitive/affective factors revealed a positive correlation with the BDI-II (r = .60, p < .001; r = .74, p < .001, respectively) and the GDS-15 (r = .52, p < .001; r = .62, p < .001, respectively).

4.3 Discussion

Study 3 examined the factor structure, measurement invariance, reliability, and convergent validity of the European Portuguese version of PHQ-9 in the general adult Portuguese-speaking population. After testing four competing factor models, Model 2 was identified as the model with the best fit to the data. It is important to note, however, that all the remaining models also reported an acceptable fit. Thus, our results suggested that no authoritative conclusions can be drawn regarding the factor structure, and future research should replicate and expand our findings. In particular, based on fit-values of Models 1 and 4, it is plausible to hypothesize that the Portuguese version of the PHQ-9 might also be considered as a unidimensional construct.

This pattern of findings (i.e., a two-factor model with close fit and a one-factor factor model with acceptable fit) was already documented in past empirical research (Arnold et al., 2019; Boothroyd et al., 2019; González-Blanch et al., 2018; Keum et al., 2018; Patel et al., 2019). Remarkably, the one-factor model was preferred over the two-factor model in three of those studies due to the strong intercorrelation between the somatic and cognitive-affective factors (.85 to .97), which precluded the statistical meaning and conceptual interpretability of the two-factor solution (Boothroyd et al., 2019; González-Blanch et al., 2018; Keum et al., 2018). As a result, the one-factor model was determined as the simplest structure of the PHQ-9 for screening utility since it was more parsimonious, easier to score and interpret, and apply the diagnosis algorithm to detect depression (Keum et al., 2018).

However, our findings revealed a moderate intercorrelation between the two factors, in line with other empirical work (Patel et al., 2019). This suggests that the European Portuguese version of the PHQ-9 may reflect two distinct but interdependent dimensions of depression that are conceptually interpretable (see general discussion). This two-factor structure of the PHQ-9 does not prevent the use of the total score for screening proposes. According to psychometric theory, a total score can still be computed in a multidimensional measure, when factors are significantly correlated, and the total score has acceptable properties (Furr and Bacharach, 2013). The appropriateness of the use of the PHQ-9 total score finds additional support in the acceptable fit of the bifactor model with high loading on an overall depression (Arnold et al., 2019).

Although the acceptable fit of the bifactor model (Model 4), we selected the two-model factor (Model 2) over the bifactor model because the former demonstrated a better data-model fit and was more theoretically defensible. Despite the potential statistical advantages over two second-order factor models, bifactor models still lack conceptual interpretability, and the clinical and research utility of such factor structures has not been fully clarified (Dere et al., 2015).

Our findings regarding the factor structure of the European Portuguese version of the PHQ-9 showed that Model 2 presented the best-fit data. However, the acceptable fit of the remaining models advises a further investigation of the factor structure of this version. Our results suggest that the total score of the PHQ-9 can be used for screening purposes, while both factors can be used in research to investigate depression clusters and their association with psychological and health outcomes.

We also established the measurement invariance of the European Portuguese version of the PHQ-9 across sex, age, marital status, education level, and administration format (paper-and-pencil vs. internet). Only two studies supported a two-factor model that previously evaluated the measurement invariance of this factor structure across groups (Miranda and

Scoppetta, 2018; Patel et al., 2019). Despite their contribution, the measurement invariance analyses in those studies were restricted to three sociodemographic groups (sex, race/ethnicity, and education level). In particular, Patel et al. (2019) demonstrated the PHQ-9 measurement invariance across sex, race/ethnicity, and Miranda and Scoppetta (2018) found support across sex. Our research replicated these findings for sex and education level. Only one previous study tested and found support for measurement invariance across age and marital status (González-Blanch et al., 2018). The results of the current research corroborated this initial evidence for measurement invariance across these main sociodemographic groups.

Beyond being the first study that tested measurement invariance across age and marital status in a two-factor model of the PHQ-9, our study brought higher specificity in categorizing the sociodemographic groups. While González-Blanch et al. (2018) tested measurement invariance across two age groups (young adults and adults) and two marital status groups (with or without an intimate relationship), we performed invariance testing across three age groups (young adults, middle-aged adults, and elderly) and three marital status groups (single, married/cohabiting, and divorced/widowed).

The evaluation of the measurement invariance across age, including a group of elderly participants, assumes critical importance since other depression measures have been shown to have a differential function in elderly individuals (Estabrook et al., 2015; Kim et al., 2002). Though previous research has tested the factor structure of the PHQ-9 among elderly individuals (Bélanger et al., 2019), no study to date has demonstrated the statistical equivalence of the factor structure of the PHQ-9 across age, having a group exclusively consisted of elderly individuals. Our results provide the first empirical support of the measurement invariance of the PHQ-9 across young adults, middle-aged adults, and elderly adults, suggesting that the PHQ-9 produces comparable response patterns across groups with different levels of depression. Thus, valid interpretations of total scores, along with explainable and meaningful comparisons between groups, can be performed. However, our results

regarding the measurement invariance across age groups should be read with caution, since the internal consistency values of the PHQ-9 were lower in the elderly group, and the PHQ-9 total score was only moderately correlated with the GDS-15 total score.

Our results regarding measurement invariance suggest that the two-factor Portuguese version of the PHQ-9 carries similar meaning across sex, age, marital status, education level, and administration format in Portuguese adults. These findings highlight that the PHQ-9 produces comparable response patterns across these major Portuguese sociodemographic groups. Thus, the PHQ-9 allows valid interpretations of total and factor scores, and explainable and meaningful comparisons between groups can be performed with minimal risk of bias.

As indicated by the internal consistency tests, the reliability of the total score of the PHQ-9 was good in the entire sample and among sociodemographic groups. These values were comparable to the two-factor structure reported, for example, in Familiar et al. (2014) and Janssen et al. (2016). As expected, the PHQ-9 revealed a strong correlation with the BDI-II and a moderate correlation with the GDS-15, supporting the convergent validity of the PHQ-9 (Schutt et al., 2016; Shin et al., 2019). In particular, this finding is consistent with prior research that has systematically documented strong associations between the PHQ-9 and the BDI-II, suggesting that both measures can be used interchangeably to assess depressive symptoms in clinical and community populations (Kung et al., 2013; Schutt et al., 2016).

5. General discussion

Based on our systematic search of previous studies that examined the factor structure of the PHQ-9 using a CFA approach (Study 1), we found substantial inconsistency in the number of factors and the item composition of factors across the proposed structures. In contrast with previous hypotheses that suggested the heterogeneity in factor models of the PHQ-9 could be accounted for by samples' sociodemographic, clinical, and cultural variations (Petersen et al., 2015), we did not find a consistent pattern of associations between participants' characteristics

and the selected factor model. Results from Study 1 indicated that both one- and two-factor models were supported in the community and clinical settings, as well as in both sexes, multiple age groups, and in different countries. This heterogeneity in factor structures might be due to the lack of a conceptual model of the PHQ-9 that predicts how depressive symptoms are interrelated. The PHQ-9 is a clinical-driven measure which translated the DSM-IV-TR diagnosis criteria into self-reported items. The original paper of PHQ-9 did not explore the measure's conceptual rationale nor conducted factor analyses (Kroenke et al., 2001). The first studies that tested the factor structure of PHQ-9 using CFA assumed a priori unidimensional structure without further examination of alternative factor structures (Crane et al., 2010; Williams et al., 2009). The acceptance of the one-factor model relied exclusively on instrumental reasons (e.g., easier scoring and interpretation) and statistical fit rather than in a conceptual model that explains the rationale (and implications) of a unidimensional perspective of depression (Williams et al., 2009).

In contrast, some authors have argued that a two-factor structure of the PHQ-9 might be more interpretable in light of conceptual models of depression (Patel et al., 2019; Vrany et al., 2016). A common assumption of psychological, social psychiatric, and neurobiological models of depression is that cognitive/affective and somatic symptoms are distinct sets of symptoms (that may or may not co-occur) and that the onset and maintenance of each of these two sets of symptoms reflect independent but intercorrelated underlying processes (Harshaw, 2015; Kendler et al., 2013; Penninx et al., 2013; Silverstein and Levin, 2014). A large body of previous empirical work has supported this assumption by demonstrating that individuals with high comorbidity of cognitive/affective and somatic symptoms of depression exhibited greater depression chronicity, and lower remission rates and response to treatment when compared to individuals with only cognitive/affective depressive symptoms (Bekhuis et al., 2016; Huijbregts et al., 2013; Stegenga et al., 2012). Person-centered research has also documented two clinically distinct profiles of depressive symptoms: one with high levels of

cognitive/affective symptoms and low levels of somatic symptoms, and other with high levels of both cognitive/affective symptoms and somatic symptoms (Baldassin et al., 2013; Illi et al., 2012; Lamela et al., 2017; Novick et al., 2013; Vrany et al., 2016). The latter typology is reported to be connected to worse health and psychosocial outcomes, including higher depression severity, higher risk for children's physical maltreatment, and higher severity in sickness behavior among patients diagnosed with cancer (Illi et al., 2012; Lamela et al., 2017; Novick et al., 2013). Empirical research using the PHQ-9 has also replicated these two depressive symptoms clusters and found the support of their differential impact on physical health, including inflammation, insulin resistance, and mortality rates in patients with heart failure (Case and Stewart, 2014; Hwang et al., 2015; Vrany et al., 2016). This conceptual and empirical work suggests that a two-factor structure might improve the research and clinical utility of the PHQ-9 in the screening of specific patterns of onset of the depressive symptoms and the prediction of subsequent prognosis pathways (Beard et al., 2016).

By identifying two alternative two-factor models of the PHQ-9, the results of study 1 highlighted the need for a careful theoretical consideration regarding the items' composition of each factor in the two-factor models. Three studies supported a somatic factor comprised of sleep disturbance, fatigue, and appetite changes (Model 2), while seven supported a somatic factor comprised of two additional items (concentration difficulties and psychomotor disturbances) (Model 3).

Presenting a theory-driven rationale for the selection of Model 2 over Model 3, Patel et al. (2019) suggested that concentration difficulties and psychomotor disturbances should be conceptualized as a cognitive and affective symptom of depression, respectively. They argued that concentration difficulties reflect depression-related impairs in attentional processes (a domain of the cognitive functioning) and thus a cognitive manifestation of depression (Duivis et al., 2013; Patel et al., 2019). It is also plausible to hypothesize that the significant associations between concentration difficulties and the somatic factor found in some studies

might be partially explained by the well-documented impairments caused by deprived sleep patterns in cognitive functioning, especially in memory and attentional processes (Harris et al., 2015; Vargas et al., 2017). Patel et al. (2019) suggested that psychomotor disturbances are theoretically well-supported as an affective symptom of depression since psychomotor retardation and anhedonia (an affective symptom) are thought to be caused by similar neurobiological substrates of altered reward processing system (Stein, 2008). However, these research-driven hypotheses should be consistently examined in future research to provide additional conceptual support to Model 2.

Despite the clinical utility of screening measures of depression, the implementation of a generalized screening strategy in both community and primary health care settings is a controversial topic in public health. Some researchers argue that routine screening might increase the population's mental health literacy, combat negative attitudes towards depression, and increase the professional seeking behaviors for mental health problems (Mojtabai et al., 2011; Siu et al., 2016). Conversely, other researchers pointed out that depression screening in community settings might lead to higher false-positive rates, increased risk of harm, self-treatment, and does not increase positive outcomes in treatment (Gilbody et al., 2006; Thombs et al., 2012). More recently, the routine use of screening tools in community and primary health settings care has been suggested as an effective strategy to identify individuals at high risk of or with clinical levels of depression only when adequate systems for formal assessment, treatment, and follow-up are available (Ferenchick et al., 2019). Therefore, the use of the PHQ-9 as a screening tool for depression in community settings in Portugal might only be recommended under a structured national policy that foresaw access to adequate psychiatric care.

Limitations

Several limitations of the present research warrant discussion. First, we only reviewed studies that examined the factor structure and measurement invariance of PHQ-9 using CFA

approaches (Studies 1 and 2). Although this methodological option excluded studies that employed EFA to test the PHQ-9 factor structure, other studies with other adequate statistical procedures were not included in the review (e.g., Item response theory). Second, in Study 3, caution should be exercised in generalizing these findings to the whole population of Portuguese adults. Despite the sociodemographic diversity of our sample, future research should replicate our findings with a nationally representative sample. Third, the examination of the effects of race/ethnicity and financial status was beyond the scope of Study 3. Future empirical research should address the measurement invariance of the European Portuguese version of the PHQ-9 across these two major sociodemographic groups. Fourth, no gold-standard measure was used to assess depressive symptoms in Study 3. Thus, diagnostic accuracy and sensitivity for the PHQ-9 were not examined in our study. Fifth, we did not formally assess cognitive and sensorial status in the subsample of elderly participants in Study 3; the definition of participants' cognitive and sensorial functioning relied solely on the clinical diagnosis reported by the staff in senior centers, day-care centers, or nursing homes. In addition, only 25% of participants completed both PQH-9 and GDS; we did not calculate the response rate since the number of potential participants was not collected. Finally, due to the cross-sectional design of Study 3, we did not examine the measurement invariance of the PHQ-9 over time. In contrast with previous studies that evaluated measurement invariance of the one-factor structure over time (González-Blanch et al., 2018; Schuler et al., 2018), no similar study has been conducted for the two-factor structure. A further empirical inquiry should address this limitation to expand the clinical utility of the two-factor structure of the PHQ-9.

Despite these limitations, our studies provided new light regarding the methodological and psychometric factor validity of the PHQ-9. In particular, Study 3 was one of the first examinations of the measurement invariance of a two-factor model of the PHQ-9 across main sociodemographic variables (sex, age, marital status, and education level). As the first research to address the factor structure and measurement invariance of a Portuguese version of the

PHQ-9, our findings suggest that valid interpretations and meaningful comparisons between groups can be performed using the PHQ-9 in the Portuguese context. This initial quantitative effort to examine the psychometric properties of the European Portuguese version of PHQ-9 highlights the potential clinical utility of the PHQ-9 for screening depressive symptoms in native Portuguese speakers.

References

Adolf, J., Schuurman, N.K., Borkenau, P., Borsboom, D., Dolan, C. V., 2014. Measurement invariance within and between individuals: a distinct problem in testing the equivalence of intra- and inter-individual model structures. Front. Psychol. 5, 883. https://doi.org/10.3389/fpsyg.2014.00883

Amtmann, D., Kim, J., Chung, H., Bamer, A.M., Askew, R.L., Wu, S., Cook, K.F., Johnson, K.L., 2014. Comparing CESD-10, PHQ-9, and PROMIS depression instruments in individuals with multiple sclerosis. Rehabil. Psychol. 59, 220–229. https://doi.org/10.1037/a0035919

Apóstolo, J., Loureiro, L., Reis, I., Silva, I., Cardoso, D., Sfetcu, R., 2014. Contribuição para a adaptação da Geriatric Depression Scale-15 para a língua portuguesa [Contribution to the adaptation of the Geriatric Depression Scale-15 into portuguese]. Rev. Enferm. Ref. IV Série, 65–73. https://doi.org/10.12707/RIV14033

Arnold, S., Uljarević, M., Hwang, Y.I., Richdale, A.L., Trollor, J.N., Lawson, L.P., 2019. Psychometric properties of the Patient Health Questionaire-9 (PHQ-9) in autistic adults. J. Autism Dev. Disord. 1–9. https://doi.org/10.1007/s10803-019-03947-9

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

Baldassin, S., Silva, N., De Toledo Ferraz Alves, T.C., Castaldelli-Maia, J.M., Bhugra, D.,

Nogueira-Martins, M.C.F., De Andrade, A.G., Nogueira-Martins, L.A., 2013. Depression in medical students: Cluster symptoms and management. J. Affect. Disord. 150, 110–114. https://doi.org/10.1016/j.jad.2012.11.050

Barthel, D., Barkmann, C., Ehrhardt, S., Schoppen, S., Bindt, C., 2015. Screening for depression in pregnant women from Côte d′Ivoire and Ghana: Psychometric properties of the Patient Health Questionnaire-9. J. Affect. Disord. 187, 232–240. https://doi.org/10.1016/J.JAD.2015.06.042

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

Beck, A., Steer, R., Brown, G., 1996. Manual for the Beck Depression Inventory-II. Psychological Corporation, San Antonio.

Bekhuis, E., Boschloo, L., Rosmalen, J.G.M., de Boer, M.K., Schoevers, R.A., 2016. The impact of somatic symptoms on the course of major depressive disorder. J. Affect. Disord. 205, 112–118. https://doi.org/10.1016/j.jad.2016.06.030

Bélanger, E., Thomas, K.S., Jones, R.N., Epstein-Lubow, G., Mor, V., 2019. Measurement validity of the Patient-Health Questionnaire-9 in US nursing home residents. Int. J. Geriatr. Psychiatry 34, 700–708. https://doi.org/10.1002/gps.5074

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

Campos, R.C., Gonçalves, B., 2011. The portuguese version of the Beck Depression Inventory-II (BDI-II): Preliminary psychometric data with two nonclinical samples. Eur. J. Psychol. Assess. 27, 258–264. https://doi.org/10.1027/1015-5759/a000072

Case, S.M., Stewart, J.C., 2014. Race/ethnicity moderates the relationship between depressive symptom severity and C-reactive protein: 2005–2010 NHANES data. Brain. Behav. Immun. 41, 101–108. https://doi.org/10.1016/J.BBI.2014.04.004

Cassano, P., Fava, M., 2002. Depression and public health: An overview. J. Psychosom. Res. 53, 849–857. https://doi.org/10.1016/S0022-3999(02)00304-5

Chen, F.F., 2007. Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. Struct. Equ. Model. A Multidiscip. J. 14, 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G.W., Rensvold, R.B., 2002. Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. Struct. Equ. Model. A Multidiscip. J. 9, 233–255. https://doi.org/10.1207/S15328007SEM0902_5

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

Chung, H., Kim, J., Askew, R.L., Jones, S.M.W., Cook, K.F., Amtmann, D., 2015. Assessing measurement invariance of three depression scales between neurologic samples and community samples. Qual. Life Res. 24, 1829–1834. https://doi.org/10.1007/s11136-015-0927-5

Crane, P.K., Gibbons, L.E., Willig, J.H., Mugavero, M.J., Lawrence, S.T., Schumacher, J.E., Saag, M.S., Kitahata, M.M., Crane, H.M., 2010. Measuring depression levels in HIV-infected patients as part of routine clinical care using the nine-item Patient Health Questionnaire (PHQ-9). AIDS Care 22, 874–885. https://doi.org/10.1080/09540120903483034

Cuijpers, P., Smit, F., Oostenbrink, J., de Graaf, R., ten Have, M., Beekman, A., 2007. Economic costs of minor depression: A population-based study. Acta Psychiatr. Scand. 115, 229–236. https://doi.org/10.1111/j.1600-0447.2006.00851.x

Dere, J., Watters, C.A., Yu, S.C.-M., Bagby, R.M., Ryder, A.G., Harkness, K.L., 2015. Cross-cultural examination of measurement invariance of the Beck Depression Inventory–II. Psychol. Assess. 27, 68–81. https://doi.org/10.1037/pas0000026

Direção-Geral da Saúde, 2017. Programa Nacional para a Saúde Mental 2017 [National Program for Mental Health 2017]. Lisbon, Portugal.

Doi, S., Ito, M., Takebayashi, Y., Muramatsu, K., Horikoshi, M., 2018. Factorial validity and invariance of the Patient Health Questionnaire (PHQ)-9 among clinical and non-clinical populations. PLoS One 13, e0199235. https://doi.org/10.1371/journal.pone.0199235

Duivis, H.E., Vogelzangs, N., Kupper, N., de Jonge, P., Penninx, B.W.J.H., 2013. Differential association of somatic and cognitive symptoms of depression and anxiety with inflammation: Findings from the Netherlands Study of Depression and Anxiety (NESDA). Psychoneuroendocrinology 38, 1573–1585. https://doi.org/10.1016/J.PSYNEUEN.2013.01.002

Elhai, J.D., Contractor, A.A., Tamburrino, M., Fine, T.H., Prescott, M.R., Shirley, E., Chan, P.K., Slembarski, R., Liberzon, I., Galea, S., Calabrese, J.R., 2012. The factor structure of major depression symptoms: A test of four competing models using the Patient Health Questionnaire-9. Psychiatry Res. 199, 169–173. https://doi.org/10.1016/J.PSYCHRES.2012.05.018

Estabrook, R., Sadler, M.E., McGue, M., Denmark, O.C., 2015. Differential item functioning in the Cambridge Mental Disorders in the Elderly (CAMDEX) Depression Scale across middle age and late life. Psychol. Assess. 27, 1219–33. https://doi.org/10.1037/pas0000114

Evans-Lacko, S., Knapp, M., 2016. Global patterns of workplace productivity for people with depression: Absenteeism and presenteeism costs across eight diverse countries. Soc. Psychiatry Psychiatr. Epidemiol. 51, 1525–1537. https://doi.org/10.1007/s00127-016-1278-4

Ferenchick, E.K., Ramanuj, P., Pincus, H.A., 2019. Depression in primary care: Part 1 - screening and diagnosis. BMJ. https://doi.org/10.1136/bmj.l794

Funk, J.L., Rogge, R.D., 2007. Testing the ruler with item response theory: Increasing

precision of measurement for relationship satisfaction with the Couples Satisfaction Index. J. Fam. Psychol. 21, 572–583. https://doi.org/10.1037/0893-3200.21.4.572

Furr, M., Bacharach, V., 2013. Psychometric: An Introduction. SAGE Publications, Thousand Oaks, CA.

Galenkamp, H., Stronks, K., Snijder, M.B., Derks, E.M., 2017. Measurement invariance testing of the PHQ-9 in a multi-ethnic population in Europe: the HELIUS study. BMC Psychiatry 17, 349. https://doi.org/10.1186/s12888-017-1506-9

Gilbody, S., Sheldon, T., Wessely, S., 2006. Should we screen for depression? BMJ 332, 1027–1030. https://doi.org/10.1136/BMJ.332.7548.1027

González-Blanch, C., Medrano, L.A., Muñoz-Navarro, R., Ruíz-Rodríguez, P., Moriana, J.A., Limonero, J.T., Schmitz, F., Cano-Vindel, A., Group, on behalf of the P.R., 2018. Factor structure and measurement invariance across various demographic groups and over time for the PHQ-9 in primary care patients in Spain. PLoS One 13, e0193356. https://doi.org/10.1371/journal.pone.0193356

Granillo, M.T., 2012. Structure and function of the Patient Health Questionnaire-9 among latina and non-Latina white female college students. J. Soc. Social Work Res. 3, 80–93. https://doi.org/10.5243/jsswr.2012.6

Gregorich, S.E., 2006. Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. Med. Care 44, S78-94. https://doi.org/10.1097/01.mlr.0000245454.12228.8f

Hambleton, R., Merenda, P., Spielberger, C., 2005. Adapting educational and psychological tests for cross-cultural assessment. Lawrence S. Erlbaum Publishers, Hillsdale, NJ.

Harris, K., Spiegelhalder, K., Espie, C.A., MacMahon, K.M.A., Woods, H.C., Kyle, S.D., 2015. Sleep-related attentional bias in insomnia: A state-of-the-science review. Clin. Psychol. Rev. 42, 16–27. https://doi.org/10.1016/J.CPR.2015.08.001

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

Harshaw, C., 2015. Interoceptive dysfunction: Toward an integrated framework for understanding somatic and affective disturbance in depression. Psychol. Bull. 141, 311–63. https://doi.org/10.1037/a0038101

Hegerl, U., Wittmann, M., Arensman, E., van Audenhove, C., Bouleau, J.-H., van der Feltz-Cornelis, C., Gusmao, R., Kopp, M., Löhr, C., Maxwell, M., Meise, U., Mirjanic, M., Oskarsson, H., Perez Sola, V., Pull, C., Pycha, R., Ricka, R., Tuulari, J., Värnik, A., Pfeiffer-Gerschel, T., 2008. The 'European Alliance Against Depression (EAAD)': A multifaceted, community-based action programme against depression and suicidality. World J. Biol. Psychiatry 9, 51–58. https://doi.org/10.1080/15622970701216681

Hinz, A., Mehnert, A., Kocalevent, R.-D., Brähler, E., Forkmann, T., Singer, S., Schulte, T., 2016. Assessment of depression severity with the PHQ-9 in cancer patients and in the general population. BMC Psychiatry 16, 22. https://doi.org/10.1186/s12888-016-0728-6

Hu, L., Bentler, P.M., 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Struct. Equ. Model. A Multidiscip. J. 6, 1–55. https://doi.org/10.1080/10705519909540118

Huijbregts, K.M.L., de Jong, F.J., van Marwijk, H.W.J., Beekman, A.T.F., Adèr, H.J., van der Feltz-Cornelis, C.M., 2013. A high physical symptom count reduces the effectiveness of treatment for depression, independently of chronic medical conditions. J. Psychosom. Res. 74, 179–185. https://doi.org/10.1016/j.jpsychores.2013.01.004

Hwang, B., Moser, D.K., Pelter, M.M., Nesbitt, T.S., Dracup, K., 2015. Changes in Depressive Symptoms and Mortality in Patients With Heart Failure. Psychosom. Med. 77, 798–807. https://doi.org/10.1097/PSY.0000000000000221

Illi, J., Miaskowski, C., Cooper, B., Levine, J.D., Dunn, L., West, C., Dodd, M., Dhruva, A.,

Paul, S.M., Baggott, C., Cataldo, J., Langford, D., Schmidt, B., Aouizerat, B.E., 2012. Association between pro- and anti-inflammatory cytokine genes and a symptom cluster of pain, fatigue, sleep disturbance, and depression. Cytokine 58, 437–447. https://doi.org/10.1016/j.cyto.2012.02.015

Janssen, E.P.C.J., Köhler, S., Stehouwer, C.D.A., Schaper, N.C., Dagnelie, P.C., Sep, S.J.S., Henry, R.M.A., van der Kallen, C.J.H., Verhey, F.R., Schram, M.T., 2016. The Patient Health Questionnaire-9 as a screening tool for depression in individuals with type 2 diabetes mellitus: The Maastricht study. J. Am. Geriatr. Soc. 64, e201–e206. https://doi.org/10.1111/jgs.14388

Kendler, K.S., Aggen, S.H., Neale, M.C., 2013. Evidence for multiple genetic factors underlying DSM-IV criteria for major depression. JAMA psychiatry 70, 599–607. https://doi.org/10.1001/jamapsychiatry.2013.751

Kessler, R.C., Bromet, E.J., 2013. The Epidemiology of Depression Across Cultures. Annu. Rev. Public Health 34, 119–138. https://doi.org/10.1146/annurev-publhealth-031912-114409

Keum, B.T., Miller, M.J., Inkelas, K.K., 2018. Testing the factor structure and measurement invariance of the PHQ-9 across racially diverse U.S. college students. Psychol. Assess. 30, 1096–1106. https://doi.org/10.1037/pas0000550

Kim, Y., Pilkonis, P.A., Frank, E., Thase, M.E., Reynolds, C.F., 2002. Differential functioning of the Beck Depression inventory in late-life patients: Use of item response theory. Psychol. Aging 17, 379–391. https://doi.org/10.1037/0882-7974.17.3.379

Kline, R., 2015. Principles and Practice of Structural Equation Modeling. Guilford, New York.

Krause, J.S., Saunders, L.L., Bombardier, C., Kalpakjian, C., 2011. Confirmatory factor analysis of the Patient Health Questionnaire-9: A study of the participants from the Spinal Cord Injury Model Systems. PM&R 3, 533–540. https://doi.org/10.1016/J.PMRJ.2011.03.003

Kroenke, K., Spitzer, R.L., Williams, J.B.W., 2001. The PHQ-9: Validity of a brief depression severity measure. J. Gen. Intern. Med. 16, 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kung, S., Alarcon, R.D., Williams, M.D., Poppe, K.A., Jo Moore, M., Frye, M.A., 2013. Comparing the Beck Depression Inventory-II (BDI-II) and Patient Health Questionnaire (PHQ-9) depression measures in an integrated mood disorders practice. J. Affect. Disord. 145, 341–343. https://doi.org/10.1016/J.JAD.2012.08.017

Lamela, D., Jongenelen, I., Morais, A., Figueiredo, B., 2017. Cognitive-affective depression and somatic symptoms clusters are differentially associated with maternal parenting and coparenting. J. Affect. Disord. 219. https://doi.org/10.1016/j.jad.2017.05.006

Laursen, T.M., Musliner, K.L., Benros, M.E., Vestergaard, M., Munk-Olsen, T., 2016. Mortality and life expectancy in persons with severe unipolar depression. J. Affect. Disord. 193, 203–207. https://doi.org/10.1016/J.JAD.2015.12.067

Levis, B., Benedetti, A., Thombs, B.D., 2019. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: Individual participant data meta-analysis. BMJ-British Med. J. 365, l1476. https://doi.org/10.1136/bmj.l1476

Little, T., 2013. Longitudinal structural equation modeling. Guilford, New York.

Maniaci, M.R., Rogge, R.D., 2014. Caring about carelessness: Participant inattention and its effects on research. J. Res. Pers. 48, 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

Marcos-Nájera, R., Le, H.-N., Rodríguez-Muñoz, M.F., Olivares Crespo, M.E., Izquierdo Mendez, N., 2018. The structure of the Patient Health Questionnaire-9 in pregnant women in Spain. Midwifery 62, 36–41. https://doi.org/10.1016/J.MIDW.2018.03.011

Masyn, K., 2013. Latent class analysis and finite mixture modeling, in: Little, T. (Ed.), The Oxford Handbook of Quantitative Methods in Psychology (Volume 2). Oxford University Press, New York, NY, pp. 551–611.

Merz, E.L., Malcarne, V.L., Roesch, S.C., Riley, N., Sadler, G.R., 2011. A multigroup

confirmatory factor analysis of the Patient Health Questionnaire-9 among English- and Spanish-speaking Latinas. Cult. Divers. Ethn. Minor. Psychol. 17, 309–316. https://doi.org/10.1037/a0023883

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

Mitchell, A.J., Yadegarfar, M., Gill, J., Stubbs, B., 2016. Case finding and screening clinical utility of the Patient Health Questionnaire (PHQ-9 and PHQ-2) for depression in primary care: A diagnostic meta-analysis of 40 studies. Br. J. Psychiatry Open 2, 127–138. https://doi.org/10.1192/bjpo.bp.115.001685

Mojtabai, R., Olfson, M., Sampson, N.A., Jin, R., Druss, B., Wang, P.S., Wells, K.B., Pincus, H.A., Kessler, R.C., 2011. Barriers to mental health treatment: results from the National Comorbidity Survey Replication. Psychol. Med. 41, 1751–1761. https://doi.org/10.1017/S0033291710002291

Nguyen, T.Q., Bandeen-Roche, K., Bass, J.K., German, D., Nguyen, N.T.T., Knowlton, A.R., 2016. 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 Lesbian Ment. Health 20, 173–191. https://doi.org/10.1080/19359705.2015.1080204

Novick, D., Montgomery, W., Aguado, J., Kadziola, Z., Peng, X., Brugnoli, R., Haro, J.M., 2013. Which somatic symptoms are associated with an unfavorable course in Asian patients with major depressive disorder? J. Affect. Disord. 149, 182–188. https://doi.org/10.1016/j.jad.2013.01.020

Nunnally, J., Bernstein, I., 1994. Psychometric Theory. McGraw-Hill, New York.

Patel, J.S., Oh, Y., Rand, K.L., Wu, W., Cyders, M.A., Kroenke, K., Stewart, J.C., 2019. Measurement invariance of the patient health questionnaire-9 (PHQ-9) depression

screener in U.S. adults across sex, race/ethnicity, and education level: NHANES 2005–2016. Depress. Anxiety da.22940. https://doi.org/10.1002/da.22940

Penninx, B.W., Milaneschi, Y., Lamers, F., Vogelzangs, N., 2013. Understanding the somatic consequences of depression: biological mechanisms and the role of depression symptom profile. BMC Med. 11, 129. https://doi.org/10.1186/1741-7015-11-129

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

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

Rosseel, Y., 2012. lavaan: An R Package for Structural Equation Modeling. J. Stat. Softw. 48, 1–36. https://doi.org/10.18637/jss.v048.i02

Schuler, M., Strohmayer, M., Mühlig, S., Schwaighofer, B., Wittmann, M., Faller, H., Schultz, K., 2018. 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. 232, 268–275. https://doi.org/10.1016/J.JAD.2018.02.037

Schutt, P.E., Kung, S., Clark, M.M., Koball, A.M., Grothe, K.B., 2016. Comparing the Beck Depression Inventory-II (BDI-II) and Patient Health Questionnaire (PHQ-9) Depression Measures in an Outpatient Bariatric Clinic. Obes. Surg. 26, 1274–1278. https://doi.org/10.1007/s11695-015-1877-2

Sheikh, J.I., Yesavage, J.A., 1986. Geriatric depression scale (GDS) recent evidence and development of a shorter version. Clin. Gerontol. 5, 165–173. https://doi.org/10.1300/J018v05n01_09

Shin, C., Park, M.H., Lee, S.-H., Ko, Y.-H., Kim, Y.-K., Han, K.-M., Jeong, H.-G., Han, C., 2019. Usefulness of the 15-item Geriatric Depression Scale (GDS-15) for classifying minor and major depressive disorders among community-dwelling elders. J. Affect. Disord. 259, 370–375. https://doi.org/10.1016/J.JAD.2019.08.053

Silverstein, B., Levin, E., 2014. Differences in the developmental patterns of depression with and without additional somatic symptoms. Psychiatry Res. 220, 254–257. https://doi.org/10.1016/j.psychres.2014.07.054

Siu, A.L., Bibbins-Domingo, K., Grossman, D.C., Baumann, L.C., Davidson, K.W., Ebell, M., García, F.A.R., Gillman, M., Herzstein, J., Kemper, A.R., Krist, A.H., Kurth, A.E., Owens, D.K., Phillips, W.R., Phipps, M.G., Pignone, M.P., 2016. Screening for depression in adults: US Preventive Services Task Force Recommendation Statement. JAMA 315, 380. https://doi.org/10.1001/jama.2015.18392

Stegenga, B.T., Kamphuis, M.H., King, M., Nazareth, I., Geerlings, M.I., 2012. The natural course and outcome of major depressive disorder in primary care: the PREDICT-NL study. Soc. Psychiatry Psychiatr. Epidemiol. 47, 87–95. https://doi.org/10.1007/s00127-010-0317-9

Stein, D.J., 2008. Depression, Anhedonia, and Psychomotor Symptoms: *The Role of Dopaminergic Neurocircuitry*. CNS Spectr. 13, 561–565. https://doi.org/10.1017/S1092852900016837

Thombs, B.D., Coyne, J.C., Cuijpers, P., Jonge, P. de, Gilbody, S., Ioannidis, J.P.A., Johnson, B.T., Patten, S.B., Turner, E.H., Ziegelstein, R.C., 2012. Rethinking recommendations for screening for depression in primary care. CMAJ 184, 413–418. https://doi.org/10.1503/CMAJ.111035

Thornicroft, G., Chatterji, S., Evans-Lacko, S., Gruber, M., Sampson, N., Aguilar-Gaxiola, S., Al-Hamzawi, A., Alonso, J., Andrade, L., Borges, G., Bruffaerts, R., Bunting, B., de Almeida, J.M.C., Florescu, S., de Girolamo, G., Gureje, O., Haro, J.M., He, Y., Hinkov,

H., Karam, E., Kawakami, N., Lee, S., Navarro-Mateu, F., Piazza, M., Posada-Villa, J., de Galvis, Y.T., Kessler, R.C., 2017. Undertreatment of people with major depressive disorder in 21 countries. Br. J. Psychiatry 210, 119–124. https://doi.org/10.1192/bjp.bp.116.188078

Vargas, I., Drake, C.L., Lopez-Duran, N.L., 2017. Insomnia Symptom Severity Modulates The Impact of Sleep Deprivation on Attentional Biases to Emotional Information. Cognit. Ther. Res. 41, 842–852. https://doi.org/10.1007/s10608-017-9859-4

Vrany, E.A., Berntson, J.M., Khambaty, T., Stewart, J.C., 2016. Depressive symptoms clusters and insulin resistance: Race/ethnicity as a moderator in 2005–2010 NHANES data. Ann. Behav. Med. 50, 1–11. https://doi.org/10.1007/s12160-015-9725-0

Whooley, M.A., Wong, J.M., 2013. Depression and Cardiovascular Disorders. Annu. Rev. Clin. Psychol. 9, 327–354. https://doi.org/10.1146/annurev-clinpsy-050212-185526

Williams, R.T., Heinemann, A.W., Bode, R.K., Wilson, C.S., Fann, J.R., Tate, D.G., 2009. Improving measurement properties of the Patient Health Questionnaire–9 with rating scale analysis. Rehabil. Psychol. 54, 198–203. https://doi.org/10.1037/a0015529

World Health Organization, 2017. Depression and other common mental disorders: Global health estimates. Geneva.

Zhong, Q., Gelaye, B., Rondon, M., E. Sánchez, S., J. García, P., Sánchez, E., V. Barrios, Y., E. Simon, G., C. Henderson, D., May Cripe, S., A. Williams, M., 2014. Comparative performance of Patient Health Questionnaire-9 and Edinburgh Postnatal Depression Scale for screening antepartum depression. J. Affect. Disord. 162, 1–7. https://doi.org/10.1016/J.JAD.2014.03.028

Figure 1.
Study selection flow diagram for the systematic review of the factor structure of the PHQ-9.

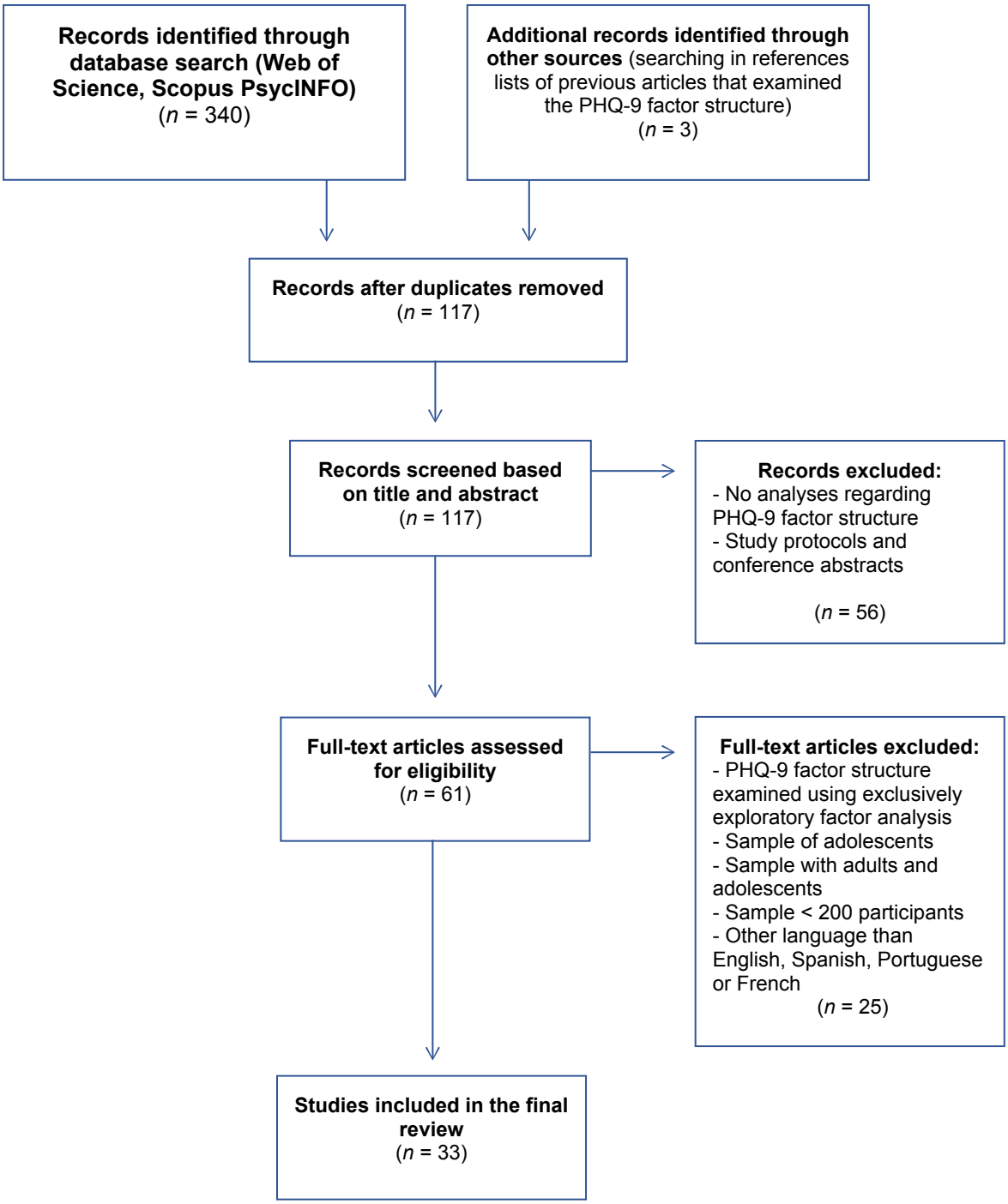
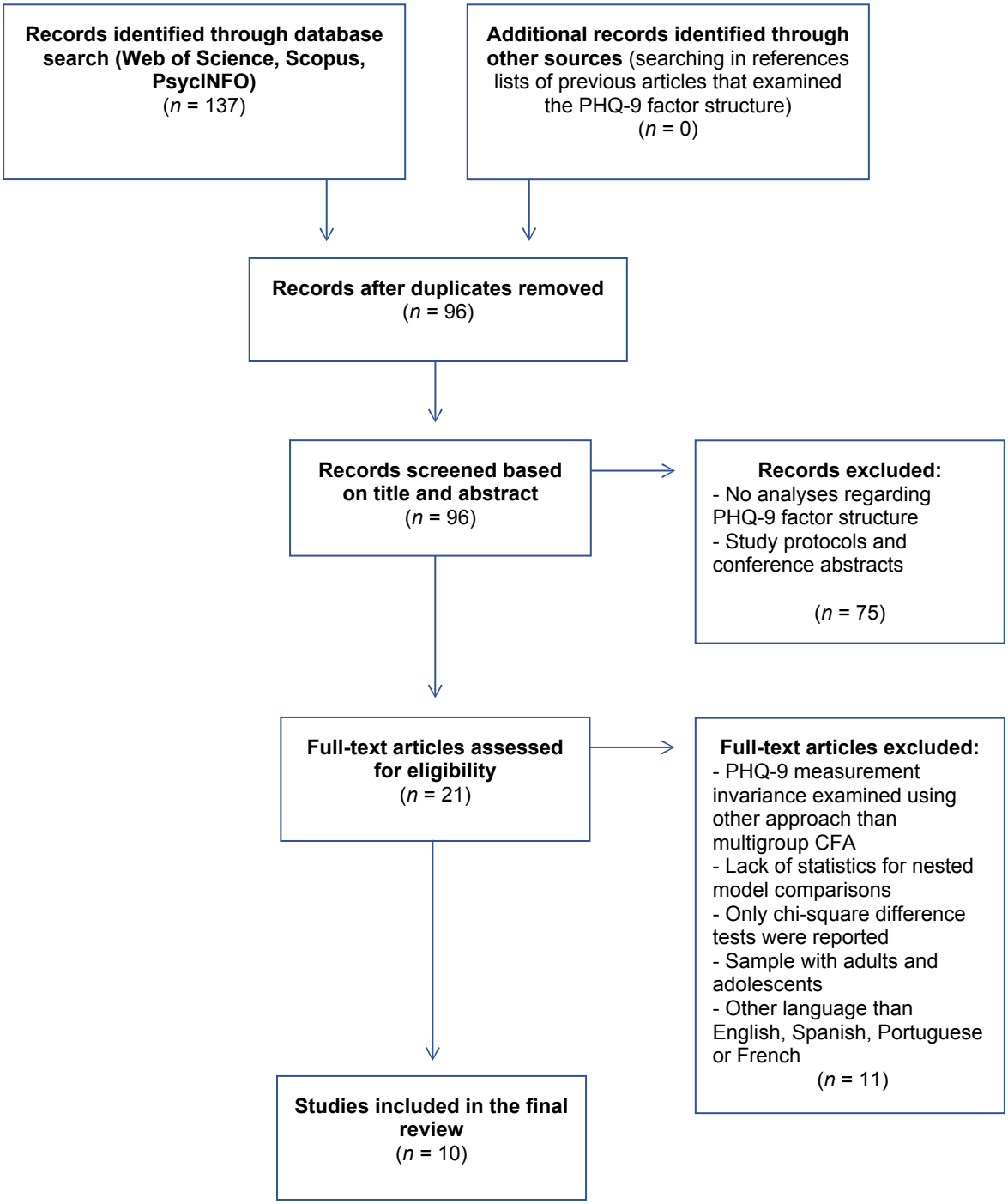


Figure 2
Study selection flow diagram for the systematic review of the measurement invariance of the PHQ-9.



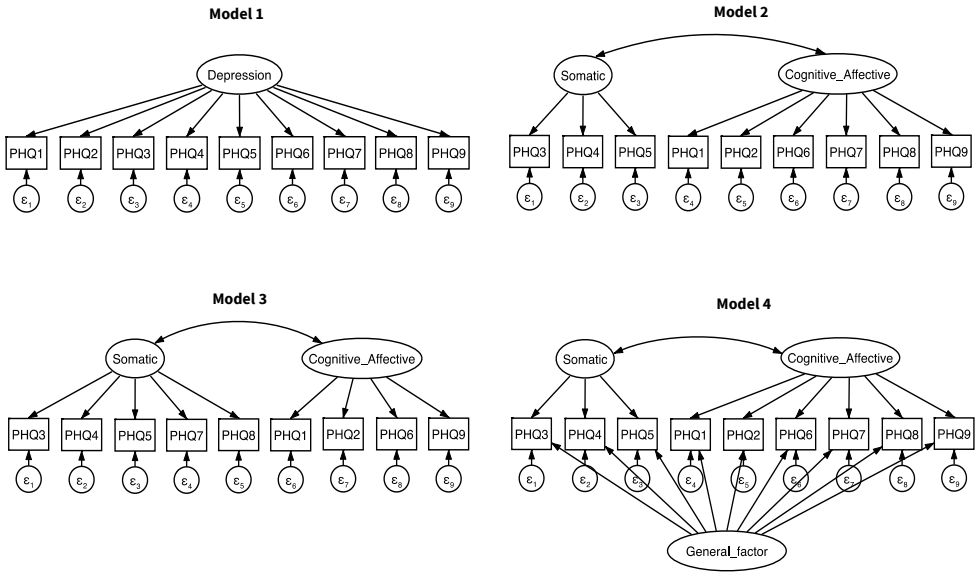


Figure 3. Evaluated PHQ-9 factor structures.
Note. PHQ1 = Anhedonia, PHQ2 = Depressed mood, PHQ3 = Sleep disturbance, PHQ4 = Fatigue, PHQ5 = Appetite changes, PH6 = Low Self-Esteem, PHQ7 = Concentration difficulties, PHQ8 = Psychomotor disturbances, PHQ9 = Suicidal ideation

Table 1

Fit Indices for the Competing Factor Models of the PHQ-9

	Model 1	Model 2	Model 3	Model 4
Chi-Square (χ2)	222.859	128.640	189.205	128.157
df	27	26	26	18
p value	<.001	<.001	<.001	< .001
TLI	0.916	0.979	0.927	0.929
CFI	0.937	0.985	0.948	0.965
RMSEA	0.076	0.049	0.071	0.070
RMSEA 90% CI	0.067-0.085	0.040-0.059	0.061-0.081	0.059-0.081
SRMR	0.040	0.023	0.037	0.029

Note. Model 1: one-factor model, comprising the nine items of PHQ-9; Model 2: two-factor model, comprising a cognitive/affective factor with six items (anhedonia, depressed mood, low self-esteem, concentration problems, psychomotor disturbances, and suicidal ideation) and a somatic factor with three items (sleep disturbance, fatigue, appetite changes); Model 3: two-factor model, comprising a cognitive/affective factor with four cognitive/affective items (anhedonia, depressed mood, low self-esteem, and suicidal ideation) and a factor with five somatic symptoms (sleep disturbance, fatigue, appetite changes, concentration problems, and psychomotor disturbances); Model 4: bifactor model, with a general factor added to the factors of Model 2.
df = degrees of freedom; TLI = Tucker-Lewis Index; CFI = comparative fit index; RMSEA = root-mean-square error of approximation; RMSEA 90% CI = 90% confidence interval for RMSEA; SRMR = standardized root-mean-square residual.

Table 2

Factor Loadings and Fit Indices for the Two-Factor Model (Model 2) of the PHQ-9 in the Total Sample and by Sociodemographic Group

	Factor Loadings of the PHQ-9 Items[a]									χ^2	*df*	*p*	TLI	CFI	RMSEA	RMSEA 90% CI	SRMR
	Cognitive/Affective Factor						Somatic Factor										
	PHQ 1	PHQ 2	PHQ 6	PHQ 7	PHQ 8	PQH 9	PHQ 3	PHQ 4	PHQ 5								
Total sample (*N* = 1479)	.58	.66	.82	.71	.71	.74	.85	.89	.88	128.64	26	< .001	.98	.99	0.049	0.041-0.059	0.023
Sex (*n* = 1479)																	
Women	.59	.66	.83	.72	.71	.74	.85	.89	.90	67.29	26	< .001	.99	.99	0.039	0.028-0.052	0.018
Men	.49	.65	.80	.63	.70	.74	.85	.89	.88	105.96	26	< .001	.93	.95	0.079	0.067-0.100	0.047
Age (*n* = 1479)																	
18-34 years	.62	.75	.80	.70	.69	.70	.85	.86	.88	82.77	26	< .001	.99	.98	0.065	0.050-0.082	0.031
35-60 years	.65	.67	.80	.69	.72	.68	.88	.93	.88	84.09	26	< .001	.98	.98	0.055	0.042-0.069	0.027
> 61 years	.40	.54	.75	.48	.50	.65	.75	.85	.95	75.09	26	< .001	.91	.93	0.079	0.066-0.113	0.056
Marital status (*n* = 1468)																	
Single	.54	.65	.81	.69	.73	.69	.84	.86	.88	48.79	26	< .001	.98	.99	0.050	0.028-0.070	0.034
Married/Cohabiting	.66	.71	.84	.73	.78	.78	.88	.90	.89	117.81	26	< .001	.97	.98	0.065	0.054-0.078	0.030
Divorced/Widowed	.49	.55	.82	.66	.49	.70	.78	.85	.95	70.42	26	< .001	.92	.94	0.079	0.058-0.103	0.055
Education level (*n* =1448)																	
≤ 9th grade	.61	.72	.81	.74	.69	.72	.81	.89	.88	85.73	26	< .001	.93	.95	0.080	0.063-0.100	0.041
High school graduate or equivalent	.63	.76	.77	.70	.66	.68	.86	.92	.88	47.40	26	< .01	.99	.99	0.042	0.022-0.061	0.028
College degree	.43	.44	.73	.52	.61	.64	.83	.87	.91	58.95	26	< .001	.97	.98	0.059	0.039-0.079	0.030
Master/Doctorate degree	.67	.65	.81	.67	.72	.63	.86	.88	.89	58.82	26	< .001	.96	.97	0.069	0.046-0.093	0.039
Administration format (*n* = 1479)																	
Pencil-and-paper	.51	.59	.82	.66	.64	.72	.84	.86	.90	118.86	26	< .001	.96	.97	0.069	0.057-0.080	0.031
Internet	.71	.74	.84	.75	.78	.77	.88	.90	.89	82.33	26	< .001	.98	.99	0.054	0.041-0.068	0.023

Note. df = degrees of freedom; TLI = Tucker-Lewis Index; CFI = comparative fit index; RMSEA = root-mean-square error of approximation; RMSEA 90% CI = 90% confidence interval for RMSEA; SRMR = standardized root-mean-square residual.

[a]In all cases, $p < .001$.

Table 3

Results of the Measurement Invariance Tests of the PHQ-9 Across the Sociodemographic Groups

		χ^2	*df*	$\Delta\chi^2$	*p*	CFI	RMSEA	ΔCFI	ΔRMSEA
Sex	Configural	173.58	52	–	–	.98	0.056	–	–
(*n* = 1479)	Metric	192.05	59	18.47	< .05	.98	0.055	0.002	0.001
	Scalar	221.01	66	28.95	< .001	.98	0.056	0.003	0.001
	Strict	257.10	68	54.10	< .001	.97	0.064	0.008	0.008
Age	Configural	274.59	78	–	–	.98	0.072	–	–
(*n* = 1479)	Metric	322.64	92	48.06	< .001	.97	0.071	0.005	0.000
	Scalar	419.74	106	97.10	< .001	.96	0.078	0.010	0.006
	Strict	441.43	110	21.69	< .001	.96	0.078	0.002	0.001
Marital status	Configural	256.64	78	–	–	.98	0.069	–	–
(*n* = 1468)	Metric	399.69	92	44.06	< .001	.97	0.068	0.004	0.000
	Scalar	337.98	106	37.29	< .001	.97	0.067	0.003	0.001
	Strict	351.69	110	13.71	< .001	.97	0.067	0.001	0.000
Education level	Configural	284.10	104	–	–	.98	0.069	–	–
(*n* =1448)	Metric	307.04	125	22.93	*ns*	.98	0.064	0.000	0.006
	Scalar	368.12	146	61.09	< .001	.97	0.065	0.006	0.001
	Strict	381.03	152	12.91	< .05	.97	0.065	0.001	0.000
Administration format	Configural	201.19	52	–	–	.98	0.062	–	–
(*n* = 1479)	Metric	215.42	59	14.29	< .05	.98	0.060	0.005	0.003
	Scalar	257.02	66	41.60	< .001	.97	0.063	0.005	0.003
	Strict	263.98	68	6.96	< .05	.97	0.063	0.001	0.000

Note. *df* = degrees of freedom; $\Delta\chi^2$ = change in χ^2; CFI = comparative fit index; RMSEA = root-mean-square error of approximation; ΔCFI = change in CFI; ΔRMSEA = change in RMSEA.

Table 4

Means (SD) and Internal Consistency Values for the Total Sample and Sociodemographic Groups

	n (%)	PHQ-9 total score		PHQ-9 cognitive /affective factor		PHQ-9 somatic factor	
		M (SD)	α	*M* (SD)	α	*M* (SD)	α
Total sample	1479 (100)	6.22 (5.03)	.86	2.98 (3.10)	.82	3.93 (3.06)	.90
Sex (*n* = 1479)							
Women	1032 (69.8)	6.80 (5.22)	.88	3.24 (3.22)	.85	4.30 (3.10)	.91
Men	447 (30.2)	4.88 (4.29)	.82	2.36 (2.72)	.82	3.09 (2.79)	.91
Age (*n* = 1479)							
18-34 years	513 (34.7)	6.30 (5.06)	.88	3.00 (2.98)	.86	3.95 (3.07)	.90
35-60 years	730 (49.4)	5.96 (5.09)	.90	2.97 (3.22)	.85	3.68 (2.98)	.92
> 61 years	236 (16.0)	6.22 (5.03)	.75	2.96 (3.00)	.76	4.67 (3.18)	.87
Marital status (*n* = 1468)							
Single	372 (25.3)	6.55 (4.90)	.87	3.16 (2.92)	.84	4.11 (3.03)	.90
Married/Cohabiting	829 (56.5)	5.77 (5.08)	.89	2.74 (3.12)	.88	3.58 (3.05)	.92
Divorced/Widowed	267 (18.2)	7.10 (4.99)	.90	3.46 (2.76)	.79	4.43 (3.07)	.92
Education level (*n* =1448)							
≤ 9th grade	345 (23.8)	6.48 (4.93)	.81	2.89 (2.93)	.77	4.32 (3.21)	.91
High school graduate or equivalent	470 (32.5)	6.46 (5.16)	.90	3.14 (3.04)	.86	4.00 (3.04)	.91
College degree	371 (25.6)	6.06 (5.11)	.89	2.93 (3.23)	.89	3.80 (3.03)	.92
Master or doctorate degree	262 (18.1)	5.73 (4.84)	.88	2.84 (3.10)	.89	3.55 (2.91)	.91
Administration format (*n* = 1479)							
Pencil-and-paper	741 (50.1)	6.20 (4.84)	.83	2.87 (2.99)	.80	4.00 (3.08)	.90
Internet	738 (49.9)	6.24 (5.22)	.90	3.09 (3.20)	.89	3.86 (3.04)	.93

Conflict of Interest

All authors have no conflict of interest

Credit authorship contribution statement

Diogo Lamela contributed to the study conceptual background, study design, supervision of data collection, data analysis and interpretation, writing the original draft and revising the final version of the manuscript. Cátia Soreira contributed to study conceptual background, study design, project management, data analysis and interpretation, writing the original draft and revising the final version of the manuscript. Paula Matos contributed to data collection, data analysis and interpretation, and writing the original draft and revising the final version of the manuscript. Ana Morais contributed to study conceptual background, study design, data collection, data analysis and interpretation, writing the original draft and revising the final version of the manuscript

Role of the Funding source

This research did not receive any funding or grant

Acknowledgements

None

Appendix A

Study 1: Summary of Findings in the Systematic Review of Factor Structure of the PHQ-9

Table A1
Summary of Studies Examining the Factor Structure of the PHQ-9 Using CFA

Study	*N*	Age range or *M (SD)*	Participant characteristics	Country	Competing models tested?	Selected model	Goodness-of-fit
Patel et al. (2019)	31366	18+	Representative general population	USA	Yes	Two-factor	Close
Harry and Waring (2019)	8886	18-98	American Indian & Caucasian American	USA	Yes	One-factor	Close
Bélanger et al. (2019)	1986783	65-85+	Nursing home residents	USA	Yes	One-factor	Close
Arnold et al. (2019)	581	15-85	Autistic adults	Australia	Yes	Two-factor	Acceptable
Saldivia et al. (2019)	1738	18-75	Primary care patients	Chile	No	One-factor	Acceptable
Boothroyd et al. (2019)	4348	17-93	Primary care patients	England	Yes	One-factor[a]	Acceptable
Miranda and Scoppetta (2018)	541	20.18 (2.59)	College students	Colombia	Yes	Two-factor	Close
Keum et al. (2018)	857	*na*	College students	USA	Yes	One-factor[a]	Acceptable
Doi et al. (2018)	2205	19-79	General population	Japan	Yes	Bifactor	Acceptable
Marcos-Nájera et al. (2018)	445	19-45	Pregnant women	Spain	Yes	Three-factor	Acceptable
Schuler et al. (2018)	561	56.7 (7.2)	Chronic obstructive pulmonary disease patients	Germany	Yes	One-factor	Acceptable
González-Blanch et al. (2018)	836	19-60+	Primary care patients	Spain	Yes	One factor[a]	Acceptable
Galenkamp et al. (2017)	23182	18-70	Multi-ethnic population	Netherlands	Yes	One-factor	Acceptable
Arrieta et al. (2017)	215	38 (16)	Rural population	Mexico	Yes	One-factor	Mediocre
Janssen et al. (2016)	2997	40-75	Type 2 Diabetes	Netherlands	Yes	Two-factor	Close*
Beard et al. (2016)	1023	34.30 (13.36)	Psychiatric patients	USA	No	Two-factor	Acceptable
Hinz et al. (2016)	2058	18-94	Cancer patients	Germany	Yes	Two-factor	Acceptable
Nguyen et al. (2016)	2498	18-25	Sexual minority women	Vietnam	No	One-factor	Acceptable

Barthel et al. (2015)	639	18-45	Pregnant women	Côte d'Ivoire	No	One-factor	Acceptable
	389	18-42	Pregnant women	Ghana	No	One-factor	Acceptable
Familiar et al. (2015)	55555	25-55+	Women general population	Mexico	Yes	One-factor	Mediocre
Petersen et al. (2015)	626	18-80	Primary care patients with MDD	Germany	Yes	Two-factor	Close
(Zhong et al., 2014)	1517	18-49	Pregnant women	Peru	No	Two-factor	Acceptable
Amtmann et al. (2014)	455	52.9 (10.8)	Multiple sclerosis patients	USA	No	One-factor	Mediocre
Chilcot et al. (2013)	300	68.5 (13.6)	Palliative care population	UK	Yes	Two-factor	Acceptable
Forkmann et al. (2013)	1631	50-85	Elderly general population	Germany	No	One-factor	Acceptable
Granillo (2012)	8377	18-31+	Female college students	USA	No	Two-factor	Acceptable
Arthurs et al. (2012)	960	56.6 (11.5)	Systemic sclerosis patients	Canada	Yes	One-factor[b]	Acceptable
Elhai et al. (2012)	2615	17-61	National Guard soldiers	USA	Yes	Two-factor	Close
Baas et al. (2011)	1772	18-70	Primary care patients	Netherlands	No	One-factor	Close
Merz et al. (2011)	479	18-80	Hispanic Americans females	USA	No	One-factor	Close
Krause et al. (2011)	7296	31.8 (13.9)	Spinal Cord Injury patients	USA	Yes	Two-factor	Acceptable[b]
Crane et al. (2010)	1467	18-50+	Patients with HIV infection	USA	No	One-factor	Acceptable
Williams et al. (2009)	202	18-80	Spinal Cord Injury patients	USA	No	One-factor	Acceptable

Note. MDD = Major depressive disorder.
[a]Despite a two-factor model showed best fit to data, the one-factor structure was preferred due to the high intercorrelation between the two factors [b]Only a fit measure was reported (Root-mean-square error of approximation).

References included in the systematic review of the factor structure of the PHQ-9

Amtmann, D., Kim, J., Chung, H., Bamer, A. M., Askew, R. L., Wu, S., … Johnson, K. L. (2014). Comparing CESD-10, PHQ-9, and PROMIS depression instruments in individuals with multiple sclerosis. *Rehabilitation Psychology*, *59*(2), 220–229. https://doi.org/10.1037/a0035919

Arnold, S., Uljarević, M., Hwang, Y. I., Richdale, A. L., Trollor, J. N., & Lawson, L. P. (2019). Psychometric properties of the Patient Health Questionaire-9 (PHQ-9) in autistic adults. *Journal of Autism and Developmental Disorders*, 1–9. https://doi.org/10.1007/s10803-019-03947-9

Arrieta, J., Aguerrebere, M., Raviola, G., Flores, H., Elliott, P., Espinosa, A., … Franke, M. F. (2017). 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. *Journal of Clinical Psychology*, *73*(9), 1076–1090. https://doi.org/10.1002/jclp.22390

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

Baas, K. D., Cramer, A. O. J., Koeter, M. W. J., van de Lisdonk, E. H., van Weert, H. C., & Schene, A. H. (2011). Measurement invariance with respect to ethnicity of the Patient Health Questionnaire-9 (PHQ-9). *Journal of Affective Disorders*, *129*(1–3), 229–235. https://doi.org/10.1016/J.JAD.2010.08.026

Barthel, D., Barkmann, C., Ehrhardt, S., Schoppen, S., & Bindt, C. (2015). Screening for depression in pregnant women from Côte d'Ivoire and Ghana: Psychometric properties of the Patient Health Questionnaire-9. *Journal of Affective Disorders*, *187*, 232–240. https://doi.org/10.1016/J.JAD.2015.06.042

Beard, C., Hsu, K. J., Rifkin, L. S., Busch, A. B., & Björgvinsson, T. (2016). Validation of the PHQ-9 in a psychiatric sample. *Journal of Affective Disorders*, *193*, 267–273. https://doi.org/10.1016/J.JAD.2015.12.075

Bélanger, E., Thomas, K. S., Jones, R. N., Epstein-Lubow, G., & Mor, V. (2019). Measurement validity of the Patient-Health Questionnaire-9 in US nursing home residents. *International Journal of Geriatric Psychiatry*, *34*(5), 700–708. https://doi.org/10.1002/gps.5074

Boothroyd, L., Dagnan, D., & Muncer, S. (2019). PHQ-9: One factor or two? *Psychiatry Research*, *271*, 532–534. https://doi.org/10.1016/J.PSYCHRES.2018.12.048

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

Crane, P. K., Gibbons, L. E., Willig, J. H., Mugavero, M. J., Lawrence, S. T., Schumacher, J. E., … Crane, H. M. (2010). Measuring depression levels in HIV-infected patients as part of routine clinical care using the nine-item Patient Health Questionnaire (PHQ-9). *AIDS Care*, *22*(7), 874–885. https://doi.org/10.1080/09540120903483034

Doi, S., Ito, M., Takebayashi, Y., Muramatsu, K., & Horikoshi, M. (2018). Factorial validity and invariance of the Patient Health Questionnaire (PHQ)-9 among clinical and non-clinical populations. *PLOS ONE*, *13*(7), e0199235. https://doi.org/10.1371/journal.pone.0199235

Elhai, J. D., Contractor, A. A., Tamburrino, M., Fine, T. H., Prescott, M. R., Shirley, E., … Calabrese, J. R. (2012). The factor structure of major depression symptoms: A test of four competing models using the Patient Health Questionnaire-9. *Psychiatry Research*, *199*(3), 169–173. https://doi.org/10.1016/J.PSYCHRES.2012.05.018

Familiar, I., Ortiz-Panozo, E., Hall, B., Vieitez, I., Romieu, I., Lopez-Ridaura, R., & Lajous, M. (2015). Factor structure of the Spanish version of the Patient Health Questionnaire-9 in Mexican women. *International Journal of Methods in Psychiatric Research*, *24*(1), 74–82. https://doi.org/10.1002/mpr.1461

Forkmann, T., Gauggel, S., Spangenberg, L., Brähler, E., & Glaesmer, H. (2013). Dimensional assessment of depressive severity in the elderly general population: Psychometric evaluation of the PHQ-9 using Rasch Analysis. *Journal of Affective Disorders*, *148*(2–3), 323–330. https://doi.org/10.1016/J.JAD.2012.12.019

Galenkamp, H., Stronks, K., Snijder, M. B., & Derks, E. M. (2017). Measurement invariance testing of the PHQ-9 in a multi-ethnic population in Europe: the HELIUS study. *BMC Psychiatry*, *17*(1), 349. https://doi.org/10.1186/s12888-017-1506-9

González-Blanch, C., Medrano, L. A., Muñoz-Navarro, R., Ruíz-Rodríguez, P., Moriana, J. A., Limonero, J. T., … Group, on behalf of the P. R. (2018). Factor structure and measurement invariance across various demographic groups and over time for the PHQ-9 in primary care patients in Spain. *PLOS ONE*, *13*(2), e0193356. https://doi.org/10.1371/journal.pone.0193356

Granillo, M. T. (2012). Structure and function of the Patient Health Questionnaire-9 among latina and non-Latina

white female college students. *Journal of the Society for Social Work and Research*, *3*(2), 80–93. https://doi.org/10.5243/jsswr.2012.6

Harry, M. L., & Waring, S. C. (2019). The measurement invariance of the Patient Health Questionnaire-9 for American Indian adults. *Journal of Affective Disorders*, *254*, 59–68. https://doi.org/10.1016/J.JAD.2019.05.017

Hinz, A., Mehnert, A., Kocalevent, R.-D., Brähler, E., Forkmann, T., Singer, S., & Schulte, T. (2016). Assessment of depression severity with the PHQ-9 in cancer patients and in the general population. *BMC Psychiatry*, *16*(1), 22. https://doi.org/10.1186/s12888-016-0728-6

Janssen, E. P. C. J., Köhler, S., Stehouwer, C. D. A., Schaper, N. C., Dagnelie, P. C., Sep, S. J. S., … Schram, M. T. (2016). The Patient Health Questionnaire-9 as a screening tool for depression in individuals with type 2 diabetes mellitus: The Maastricht study. *Journal of the American Geriatrics Society*, *64*(11), e201–e206. https://doi.org/10.1111/jgs.14388

Keum, B. T., Miller, M. J., & Inkelas, K. K. (2018). Testing the factor structure and measurement invariance of the PHQ-9 across racially diverse U.S. college students. *Psychological Assessment*, *30*(8), 1096–1106. https://doi.org/10.1037/pas0000550

Krause, J. S., Saunders, L. L., Bombardier, C., & Kalpakjian, C. (2011). Confirmatory factor analysis of the Patient Health Questionnaire-9: A study of the participants from the Spinal Cord Injury Model Systems. *PM&R*, *3*(6), 533–540. https://doi.org/10.1016/J.PMRJ.2011.03.003

Marcos-Nájera, R., Le, H.-N., Rodríguez-Muñoz, M. F., Olivares Crespo, M. E., & Izquierdo Mendez, N. (2018). The structure of the Patient Health Questionnaire-9 in pregnant women in Spain. *Midwifery*, *62*, 36–41. https://doi.org/10.1016/J.MIDW.2018.03.011

Merz, E. L., Malcarne, V. L., Roesch, S. C., Riley, N., & Sadler, G. R. (2011). A multigroup confirmatory factor analysis of the Patient Health Questionnaire-9 among English- and Spanish-speaking Latinas. *Cultural Diversity and Ethnic Minority Psychology*, *17*(3), 309–316. https://doi.org/10.1037/a0023883

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

Nguyen, T. Q., Bandeen-Roche, K., Bass, J. K., German, D., Nguyen, N. T. T., & Knowlton, A. R. (2016). 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. *Journal of Gay & Lesbian Mental Health*, *20*(2), 173–191. https://doi.org/10.1080/19359705.2015.1080204

Patel, J. S., Oh, Y., Rand, K. L., Wu, W., Cyders, M. A., Kroenke, K., & Stewart, J. C. (2019). Measurement invariance of the patient health questionnaire-9 (PHQ-9) depression screener in U.S. adults across sex, race/ethnicity, and education level: NHANES 2005–2016. *Depression and Anxiety*, da.22940. https://doi.org/10.1002/da.22940

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

Saldivia, S., Aslan, J., Cova, F., Vicente, B., Inostroza, C., & Rincón, P. (2019). Psychometric characteristics of the Patient Health Questionnaire (PHQ-9). *Revista Médica de Chile*, *147*(1), 53–60. https://doi.org/10.4067/S0034-98872019000100053

Schuler, M., Strohmayer, M., Mühlig, S., Schwaighofer, B., Wittmann, M., Faller, H., & Schultz, K. (2018). 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). *Journal of Affective Disorders*, *232*, 268–275. https://doi.org/10.1016/J.JAD.2018.02.037

Williams, R. T., Heinemann, A. W., Bode, R. K., Wilson, C. S., Fann, J. R., & Tate, D. G. (2009). Improving measurement properties of the Patient Health Questionnaire–9 with rating scale analysis. *Rehabilitation Psychology*, *54*(2), 198–203. https://doi.org/10.1037/a0015529

Zhong, Q., Gelaye, B., Rondon, M., E. Sánchez, S., J. García, P., Sánchez, E., … A. Williams, M. (2014). Comparative performance of Patient Health Questionnaire-9 and Edinburgh Postnatal Depression Scale for screening antepartum depression. *Journal of Affective Disorders*, *162*, 1–7. https://doi.org/10.1016/J.JAD.2014.03.028

General references cited in Appendix A

Hu, L., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Little, T. (2013). *Longitudinal structural equation modeling.* New York: Guilford.

Patel, J. S., Oh, Y., Rand, K. L., Wu, W., Cyders, M. A., Kroenke, K., & Stewart, J. C. (2019). Measurement invariance of the patient health questionnaire-9 (PHQ-9) depression screener in U.S. adults across sex, race/ethnicity, and education level: NHANES 2005–2016. *Depression and Anxiety*, da.22940. https://doi.org/10.1002/da.22940

Appendix B

Study 2: Flow Diagram and Summary of Findings in the Systematic Review of Measurement Invariance of the PHQ-9

Table B1

Details of Studies Included in the Systematic Review of the Measurement Invariance of the PHQ-9

Study	*N*	Age range or *M* (*SD*)	Participant characteristics	Country	Tested model
Patel et al. (2019)	31366	18+	Representative general population	USA	Two-factor
Harry & Waring (2019)	8886	18-98	American Indian & Caucasian American	USA	One-factor
Miranda & Scoppetta (2018)	541	20.18 (2.59)	College students	Colombia	Two-factor
Keum et al. (2018)	857	*na*	College students	USA	One-factor
Doi et al. (2018)	2205	19-79	General population	Japan	Bifactor
Schuler et al. (2018)	561	56.7 (7.2)	Chronic obstructive pulmonary disease (COPS) patients	Germany	One-factor
González-Blanch et al. (2018)	836	19-60+	Primary care patients	Spain	One factor
(Galenkamp et al. (2017)	23182	18-70	Multi-ethnic population	Netherlands	One-factor
Chung et al. (2015)	8297	40.39 (15.96) to 51.81 (11.49)	Neurologic patients and general population	USA	One-factor
Merz et al. (2011)	479	18-80	Hispanic Americans females	USA	One-factor

Table B2

Summary of Findings Reported in the Tests of Measurement Invariance of the PHQ-9

Study	Groups	Steps of measurement invariance					
		Dimensional	Configural	Metric	Scalar	Strict	Factor variances and covariances
Patel et al. (2019)	Sex	+	+	NT	+	+	NT
	Race/ethnicity	+	+	NT	+	+	NT
	Education level	+	+	NT	+	+	NT
Harry et al. (2019)	Race/ethnicity	NT	+	+	+	NT	NT
Miranda et al. (2018)	Sex	NT	+	+	+	+	NT
Keum et al. (2018)	Sex	NT	+	+	+	NT	NT
	Race/ethnicity	NT	+	+	+	NT	NT
Doi et al. (2018)	Non-clinical vs. MDD	NT	+	+	+	+	+
	MDD vs. MDD + AD	NT	+	+	+	+	+
Schuler et al. (2018)	Sex	NT	+	+	P	+	+
	COPS stages	NT	+	+	+	+	+
	Over time	NT	+	+	P	P	+
González-Blanch et al. (2018)	Sex	NT	+	+	+	+	NT
	Age	NT	+	+	+	+	NT
	Marital status	NT	+	+	+	+	NT
	Education level	NT	+	+	+	+	NT
	Employment status	NT	+	+	+	+	NT
	Over time	NT	+	+	+	+	NT
Galenkamp et al. (2017)	Race/ethnicity	NT	+	+	+	+	NT
Chung et al. (2015)	Neurologic patients vs. general population	NT	+	+	+	NT	NT
Merz et al. (2011)	Race/ethnicity	NT	+	+	NT	+	NT

Note. + = the step of measurement invariance was supported; P = the step of measurement invariance was partially supported; NT = the step of measurement invariance was not tested.

References included in the systematic review of the measurement invariance of the PHQ-9

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

Chung, H., Kim, J., Askew, R. L., Jones, S. M. W., Cook, K. F., & Amtmann, D. (2015). Assessing measurement invariance of three depression scales between neurologic samples and community samples. *Quality of Life Research*, *24*(8), 1829–1834. https://doi.org/10.1007/s11136-015-0927-5

Doi, S., Ito, M., Takebayashi, Y., Muramatsu, K., & Horikoshi, M. (2018). Factorial validity and invariance of the Patient Health Questionnaire (PHQ)-9 among clinical and non-clinical populations. *PLOS ONE*, *13*(7), e0199235. https://doi.org/10.1371/journal.pone.0199235

Galenkamp, H., Stronks, K., Snijder, M. B., & Derks, E. M. (2017). Measurement invariance testing of the PHQ-9 in a multi-ethnic population in Europe: the HELIUS study. *BMC Psychiatry*, *17*(1), 349. https://doi.org/10.1186/s12888-017-1506-9

González-Blanch, C., Medrano, L. A., Muñoz-Navarro, R., Ruíz-Rodríguez, P., Moriana, J. A., Limonero, J. T., … Group, on behalf of the P. R. (2018). Factor structure and measurement invariance across various demographic groups and over time for the PHQ-9 in primary care patients in Spain. *PLOS ONE*, *13*(2), e0193356. https://doi.org/10.1371/journal.pone.0193356

Harry, M. L., & Waring, S. C. (2019). The measurement invariance of the Patient Health Questionnaire-9 for American Indian adults. *Journal of Affective Disorders*, *254*, 59–68. https://doi.org/10.1016/J.JAD.2019.05.017

Keum, B. T., Miller, M. J., & Inkelas, K. K. (2018). Testing the factor structure and measurement invariance of the PHQ-9 across racially diverse U.S. college students. *Psychological Assessment*, *30*(8), 1096–1106. https://doi.org/10.1037/pas0000550

Merz, E. L., Malcarne, V. L., Roesch, S. C., Riley, N., & Sadler, G. R. (2011). A multigroup confirmatory factor analysis of the Patient Health Questionnaire-9 among English- and Spanish-speaking Latinas. *Cultural Diversity and Ethnic Minority Psychology*, *17*(3), 309–316. https://doi.org/10.1037/a0023883

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

Patel, J. S., Oh, Y., Rand, K. L., Wu, W., Cyders, M. A., Kroenke, K., & Stewart, J. C. (2019). Measurement invariance of the patient health questionnaire-9 (PHQ-9) depression screener in U.S. adults across sex, race/ethnicity, and education level: NHANES 2005–2016. *Depression and Anxiety*, da.22940. https://doi.org/10.1002/da.22940

Schuler, M., Strohmayer, M., Mühlig, S., Schwaighofer, B., Wittmann, M., Faller, H., & Schultz, K. (2018). 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). *Journal of Affective Disorders*, *232*, 268–275. https://doi.org/10.1016/J.JAD.2018.02.037

Appendix C
Study 3: Validation of the European Portuguese Version of the PHQ-9

Table C1

Sociodemographic Description of each subsample, n (%) for Categorical Variables and M (SD) for Continuous Variables

	Total sample (*N* = 1479)	**Subsample 1 (*n* = 514)**	**Subsample 2 (*n* = 738)**	**Subsample 3 (*n* = 227)**
Sex	**(*n* = 1479)**	**(*n* = 514)**	**(*n* = 738)**	**(*n* = 227)**
Women	1032 (69.8)	272 (52.9)	598 (81.0)	162 (71.4)
Men	447 (30.2)	241 (47.1)	140 (19.0)	65 (28.6)
Age	**(*n* = 1479)**	**(*n* = 514)**	**(*n* = 738)**	**(*n* = 227)**
18-34 years	513 (34.7)	384 (74.7)	129 (17.5)	0
35-60 years	730 (49.4)	119 (23.2)	606 (82.1)	5 (2.2)
> 61 years	236 (16.0)	11 (2.1)	3 (0.4)	222 (97.8)
Mean (*SD*)	42.2 (19.5)	28.5 (12.3)	29.9 (6.3)	80.6 (8.9)
Marital status	**(*n* = 1468)**	**(*n* =503)**	**(*n* = 738)**	**(*n* = 227)**
Single	372 (25.3)	243 (48.3)	104 (14.1)	25 (11.0)
Married/Cohabiting	829 (56.5)	251 (49.9)	534 (72.4)	44 (19.4)
Divorced/Widowed	267 (18.2)	9 (1.8)	100 (13.5)	158 (69.6)
Education level	**(*n* =1448)**	**(*n* = 497)**	**(*n* = 733)**	**(*n* = 218)**
≤ 9th grade	345 (23.8)	104 (20.9)	54 (7.4)	187 (85.8)
High school graduate or equivalent	470 (32.5)	281 (54.7)	179 (24.4)	10 (4.6)
College degree	371 (25.6)	103 (20.7)	261 (35.6)	7 (3.2)
Master or doctorate degree	262 (18.1)	9 (1.8)	239 (32.6)	14 (6.4)

Table C2

Group Differences Tests in PHQ-9 Total Score, PHQ-9 Cognitive/affective Factor, and PHQ-9 Somatic Factor

	PHQ-9 total score		PHQ-9 cognitive/affective factor		PHQ-9 somatic factor	
	statistics (effect size)	Group contrasts[a]	statistics (effect size)	Group contrasts[a]	statistics (effect size)	Group contrasts[a]
Sex						
1. Women (*n* = 1032) 2. Men (*n* = 447)	*t* = 6.84*** (*d* = 0.42)	1 > 2	*t* = 5.09*** (*d* = 0.30)	1 > 2	*t* = 7.15*** (*d* = 0.41)	1 > 2
Age						
3. 18-34 years (*n* = 513) 4. 35-60 years (*n* = 730) 5. > 61 years (*n* = 236)	*F* = 2.81 (η^2 = 0.004)	–	*F* = 0.22 (η^2 = 0.000)	–	*F* = 9.52*** (η^2 = 0.013)	3, 4 > 5
Marital status						
6. Single (*n* = 372) 7. Married/Cohabiting (*n* = 829) 8. Divorced/Widowed (*n* = 267)	*F* = 8.24*** (η^2 = 0.011)	7 > 6, 8	*F* = 6.24** (η^2 = 0.008)	7 > 8	*F* = 7.13*** (η^2 = 0.010)	7 > 8
Education level						
9. ≤ 9th grade (*n* = 345) 10. High school graduate or equivalent (*n* = 470) 11. College degree (*n* = 371) 12. Master or doctorate degree (*n* = 262)	*F* = 1.61 (η^2 = 0.003)	–	*F* = 0.74 (η^2 = 0.002)	–	*F* = 3.52* (η^2 = 0.007)	12 > 9
Administration format						
13. Pencil-and-paper (*n* = 741) 14. Internet (*n* = 738)	*t* = -0.15 (*d* = 0.007)	–	*t* = -1.36 (*d* = 0.07)	–	*t* = 0.87 (*d* = 0.005)	–

Note. Means and standard deviations are presented in Table 4. Rule of thumb for Cohen's *d* effect size: 0.20 = small effect size; 0.50 = medium effect size; 0.80 = large effect size. Rule of thumb for η^2 effect size: 0.01 = small effect size; 0.06 = medium effect size; 0.14 = large effect size.

[a] Significant group differences at least $p < .05$ using Tukey–Kramer test after ANOVAs.

*** $p < .001$. ** $p < .01$. * $p < .05$.