Reliability and cross-country equivalence of the 8-item version of the Patient Health Questionnaire (PHQ-8) for the assessment of depression: results from 27 countries in Europe

Jorge Arias de la Torre,[a,b,c,*] Gemma Vilagut,[b,d] Amy Ronaldson,[e] Jose M. Valderas,[f,g,h] Ioannis Bakolis,[e] Alex Dregan,[e] Antonio J. Molina,[b,c] Fernando Navarro-Mateu,[b,i,j] Katherine Pérez,[b,k,l] Xavier Bartoll-Roca,[k,m] Matilde Elices,[n] Víctor Pérez-Sola,[n] Antoni Serrano-Blanco,[b,o] Vicente Martín,[b,c] and Jordi Alonso[b,d,p]

[a]Care in Long Term Conditions Research Division, King's College London, London, UK
[b]CIBER Epidemiology and Public Health (CIBERESP), Madrid, Spain
[c]Institute of Biomedicine (IBIOMED), University of Leon, Leon, Spain
[d]Health Services Research Group, Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain
[e]Institute of Psychiatry, Psychology and Neuroscience (IoPPN), King's College London, London, UK
[f]Department of Medicine, National University of Singapore, Singapore
[g]Department of Family Medicine, National University Health System, Singapore
[h]Centre for Research in Health Systems Performance (CRiHSP), National University Health System, Singapore
[i]Unidad de Docencia, Investigación y Formación en Salud Mental (UDIF-SM), Servicio Murciano de Salud, Murcia, Spain
[j]IMIB-Arrixaca, Murcia, Spain
[k]Agència de Salut Pública de Barcelona (ASPB), Barcelona, Spain
[l]Institut d'Investigació Biomèdica Sant Pau (IIB SANT PAU), Barcelona, Spain
[m]Institut de Recerca de l'Hospital de laSanta Creu i Sant Pau, Barcelona, Spain
[n]CIBER Mental Health (CIBERSAM), Madrid, Spain
[o]Institut de Recerca Sant Joan de Déu, Parc Sanitari Sant Joan de Déu, Barcelona, Spain
[p]Department of Medicine and Life Sciences, Universitat Pompeu Fabra (UPF), Barcelona, Spain

Summary

Background The 8-item version of the Patient Health Questionnaire (PHQ-8) is one of the self-reported questionnaires most frequently used worldwide for the screening and severity assessment of depression. However, in some European countries its reliability is unknown, and it is unclear whether its psychometric properties vary between European countries. Therefore, the aim of this study was to assess the internal structure, reliability and cross-country equivalence of the PHQ-8 in Europe.

Methods All participants from the 27 countries included in the second wave of the European Health Interview Survey (EHIS-2) between 2014 and 2015 with complete information on the PHQ-8 were included (n = 258,888). The internal structure of the PHQ-8 was assessed using confirmatory factor analyses (CFA) for categorical items. Additionally, the reliability of the questionnaire was assessed based on the internal consistency, Item Response Theory information functions, and item-discrimination (using Graded Response Models), and the cross-country equivalence based on multi-group CFA.

Findings The PHQ-8 shows high internal consistency for all countries. The countries in which the PHQ-8 was more reliable were Romania, Bulgaria and Cyprus and less reliable were Iceland, Norway and Austria. The PHQ-8 item with highest discrimination was item 2 (feeling down, depressed, or hopeless) in 24 of the 27 countries. Measurement invariance between countries in Europe was observed from multigroup CFA at the configural, metric and scalar levels.

Interpretation The results from our study, likely the largest study to the date assessing the internal structure, reliability and cross-country comparability of a self-reported mental health assessment measure, shows that the PHQ-8 has an adequate reliability and cross-country equivalence across the 27 European countries included. These results highlight the suitability of the comparisons of the PHQ-8 scores in Europe. They could be helpful to improve the screening and severity assessment of depressive symptoms at the European level.

The Lancet Regional Health - Europe 2023;31: 100659

Published Online xxx https://doi.org/10.1016/j.lanepe.2023.100659

DOI of original article: https://doi.org/10.1016/j.lanepe.2023.100668

*Corresponding author. JCMB, Second Floor, Office 2.16. 57 Waterloo Road, London SE1 8WA, UK.
E-mail address: Jorge.arias_de_la_torre@kcl.ac.uk (J. Arias de la Torre).

Funding This work was partially funded by CIBER Epidemiology and Public Health (CIBERESP) as part of the Intramural call of 2021 (ESP21PI05).

Keywords: Depression; Patient Health Questionnaire; Europe; Reliability; Validity; Health surveys

Research in context

Evidence before this study

PubMed and Google Scholar were searched from inception using the key term "PHQ-8", alone and combined with "reliability" and "validity". Different articles about the metric properties of the PHQ-8 were identified. However, most of them were focused on specific population groups, e.g., patients with a specific health condition, or in specific countries. Additionally, the search show that the PHQ-8 has increasingly gained relevance worldwide during the last years, standing out as one of the self-reported questionnaires for the screening and severity assessment of depression most widely used in Europe and worldwide. Despite being one of the most used questionnaires to determine the burden of depression at the population level, the psychometric properties of the PHQ-8 remain unclear for some European countries and across them. Furthermore, the evidence about its cross-country equivalence between countries is very limited.

Added value of this study

This study, the largest evaluation to date, assessed the internal structure, reliability and cross-country validity of the PHQ-8 in Europe. It demonstrates the consistent unidimensional structure of the PHQ-8 across all the countries assessed, and it shows that the PHQ-8 is a reliable measure to be used for the assessment of depressive symptoms in each of the 27 countries included. Additionally, it shows that the scores from the PHQ-8 are comparable across Europe.

Implications of all the available evidence

These results provide further evidence supporting the use of the PHQ-8 for the screening and severity assessment of depressive symptoms at the European level, and at the country level in the 27 countries included in the study. This evidence could be helpful to improve the accuracy of the comparisons derived from its use.

Introduction

Depression is one of the leading causes of disability, dependency, and health expenditure worldwide,[1–4] with an overall prevalence in Europe of 6.4%.[5] Given its relevance and burden, in order to plan and develop preventive measures to reduce its impact, it is necessary to assess the magnitude of depressive symptoms. This assessment should be made using valid and reliable questionnaires, and both at the individual level to detect at-risk cases eligible for prevention strategies, as well as at the population level to determine the magnitude of the problem. Moreover, at the population level, the measurement equivalence between different populations of the questionnaires used for assessing depressive symptoms must be ensured to permit adequate and relevant comparisons. Ensuring this equivalence would be also helpful at the clinical level to compare the possible need for healthcare and plan healthcare services, i.e., improving the identification of individuals at risk for whom it could be beneficial to contact a clinician, and improving country comparisons allowing a more effective and efficient allocation of mental healthcare resources.

Because of the subjective nature of the symptoms of depression, to date, there is no measure available for their independent objective assessment.[6] Therefore, it is particularly relevant that the self-reported questionnaires used for the assessment of depressive symptoms are adapted for their use in the population under study, and that their reliability and validity are ensured. Furthermore, the measurement invariance across populations should be also ensured, i.e., that items are measuring the same construct and in the same way in different contexts.[7–10] Otherwise, the conclusions derived from them could be biased, especially those related to differences in the construct to be measured, the interpretation of their items, as well as with response trends or atypical or inappropriate values of their scores depending on the population evaluated.[8,11]

One of the questionnaires most commonly used for the screening and severity assessment of depression in clinical settings and epidemiological studies is the PHQ.[12] The original version of the PHQ, the 9-item version or PHQ-9, is a self-reported measure of depressive symptoms composed of 9 Likert-type items each of them corresponding to one of the 9 criteria for major depressive disorder of the 4th edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV).[13] These criteria (and the corresponding symptoms) are the same than for the 5th version of the manual and its revision (DSM-5 and DSM-5 TR, respectively).[14] Additionally, an 8-item version (the PHQ-8) has been proposed to avoid possible ethical problems and implications surrounding

positive response to the 9th item of the PHQ-9 about suicide and suicidal ideation.[15] Remarkably, despite being a slightly shorter version, the PHQ-8 has shown very similar psychometric properties when has been compared to the PHQ-9.[15,16] Besides, despite being a screening questionnaire, i.e., not a diagnostic measure at the individual level, the PHQ-8 has proven prognostic ability to identify at-risk groups at the population level. Hence, the PHQ-8 could be a suitable option to be used in large epidemiological studies in which the use of clinical interviews could not be feasible.[17]

It should be noted that the PHQ-8 has gained relevance during the last years, emerging as one of the most widely used self-reported questionnaires for the screening and severity assessment of depressive symptoms worldwide. Some examples of relevant population-based studies in which the PHQ-8 has been included are the Behavioral Risk Factor Surveillance Survey (BRFSS) in the US,[15] or the European Health Interview Survey (EHIS) in Europe.[5] However, despite being widely used, the psychometric properties of the PHQ-8 remain unclear for some of the European countries in which it is used, i.e., the measurement model and reliability, of the PHQ-8 have not been assessed in some European counties. Additionally, the evidence about its internal structure, cross-country equivalence and about whether its psychometric properties could vary across countries is limited, with studies usually including a restricted number of countries.[18,19] This information could be key to determining the suitability of its use in the different countries, and of the comparability between them.

The aim of this study was to assess the internal structure, reliability and cross-country equivalence of the PHQ-8 for the assessment of depressive symptoms at the country level for 27 European countries, and at the population level in Europe.

Methods

Data and study population

Data from the second wave of the European Health Interview Survey (EHIS-2) were used for this study.[20] EHIS-2 is a population-based representative survey of individuals aged 15 years or older (16 years and older for the UK and Sweden) carried out between 2014 and 2015 in 31 European countries, including the 27 EU member states, the UK, Norway, Iceland, and Turkey. EHIS-2 captures relevant data about the health status of the population of the different countries participating in it, and about different relevant sociodemographic characteristics of participants (e.g., age and sex), habits and lifestyle factors (e.g., diet and physical activity), and about the use of health services (e.g., the medications prescribed or the number of hospital admissions during the last year). The sample was selected using single- and multi-stage probability sampling methods depending on the specific country and the distribution of the general characteristics of the population and accounting for primary strata and primary sampling units (PSU).[20] The non-response rates before substitutions, i.e., the rate of rebuttal to participate in the survey of participants selected before considering substituting participants, were higher than 50% in five of the countries included in the study (Austria, Denmark, Finland, Germany and Luxembourg), while in 5 countries these rates were lower than 20% (Cyprus, Greece, Italy, Portugal and Romania).[21] These non-response rates were corrected using substituting participants i.e., participants considered in the sampling design of EHIS-2 to replace the possible lack of answer of primary participants.[21] In some of the countries, substitution of sampled individuals that could not be contacted was allowed, although not recommended.[20] To ensure the quality of the data collection and maximise the response rate, in all countries all participants were interviewed by trained interviewers. Interviewers were just involved in data collection and administered the questionnaires to the participants ensuring that self-reported measures were correctly filled, but without any influence or rating on them.[20] Further information and details about EHIS-2 data, its sampling strategy, representativeness, response rate, and data collection can be found in the EHIS wave 2 methodological manual.[20] EHIS-2 microdata were provided by Eurostat through a signed agreement considering different aspects to safeguard their security (confidentiality, accessibility, and use of data) after their anonymisation and harmonisation. Despite the anonymous nature of the data used for this study, ethical approvals were obtained from the Ethics Committees of the Hospital del Mar (2021/9896) and the Universidad de León (ETICA-ULE-032-2021).

Out of the 31 European countries participating in the EHIS-2, due to lack of the PHQ-8 item-level data in the harmonised microdata files provided by Eurostat, data from 3 of them could not be included in this study (Belgium, the Netherlands, and Spain). Furthermore, due to the lack of information about their quality, data from Turkey were not included.[21] Hence, from the overall sample of 316,333 participants in EHIS-2, 39,608 belonged to countries not included in the study, and 17,837 were additionally excluded from the analysis due to missing responses in PHQ-8, leading to a final sample included in our analyses of 258,888 participants from 27 out of the 31 European countries participating in EHIS-2.

Measure

The PHQ-8 is a self-reported measure of depressive symptoms composed of 8 Likert type items with a response scale ranging from 0 (Not at all) to 3 (Nearly every day), that refer to the presence of that symptom during the previous 2 weeks.[15] Each item corresponds to the first 8 symptoms of the 4th edition of the DSM-IV diagnostic criteria for major depressive disorder (Appendix 1). The PHQ-8 final score is obtained by adding the score for each of the items, ranging from 0 to 24 (higher scores corresponding to higher levels of depression).

Data analyses

Descriptive analyses

Descriptive analyses of the distribution of the PHQ-8 scores for the whole of Europe and for each country were carried out. The latent mean PHQ-8 score, standard deviation, 95% confidence interval (95% CI), median, and the 25th and 75th percentiles (P25–P75) were calculated. Non-parametric tests (Kruskal–Wallis) were performed to compare the scores between countries.

Internal structure

Confirmatory Factor Analyses for categorical items (iCFA) were performed to assess the unidimensional structure of the PHQ-8 both for all the countries combined and for each of the countries included in the study. The model was fitted on the polychoric correlation matrix, using weighted least squares estimator, with adjustment for mean and variance for robust model testing and standard errors. To assess the goodness of fit of the one-factor model, the following statistics were used: Chi-Square, Comparative Fit Index (CFI), Tucker–Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA) and Standardized Root Mean Square Residual (SRMR). Given the sensitivity of the Chi-Square statistic to the sample size, we additionally examined the following goodness of fit indices: the CFI (values higher than 0.95 indicating a suitable goodness of fit), TLI values higher than 0.95 indicating a suitable goodness of fit), the RMSEA values lower than 0.06 indicating a suitable goodness of fit, and than 0.06 adequate) and the SRMR (values lower than 0.08 indicating a suitable goodness of fit).[22] Additionally, the standardized item loadings (λ) for each item were obtained from the iCFA models.

Reliability

The internal consistency of the scale was calculated using Cronbach's Alpha (α) and McDonald's Omega (Ω) coefficients, considering acceptable values higher than 0.7 when the scale is used for group level comparisons, and 0.90 when it is used at the individual level.[23,24] To determine the discrimination capacity of the items for the whole Europe and for each country, three polytomic item response theory (IRT) models were adjusted: a partial credit model (PCM), a generalized partial credit model (GPCM), and a graded response model (GRM). From these models, the one with higher goodness of fit was selected. To select the one with the best fit to the data and obtain from it the discrimination parameter for each item and for each country, the Akaike Information Criterion (AIC) was used. For each model and item, higher values of the discrimination parameter (α) indicate higher ability of the item to discriminate the symptom to which the item refers. In addition, to evaluate the test information and each of the items for all the countries combined, and for each country, the Item Information Function (IIF) and the Test Information Function (TIF) were obtained. Based on the test information function, score reliability for each latent trait level can be estimated as 1–(1/Information [θi]).[25]

Cross-country equivalence and measurement invariance

Measurement invariance of the PHQ-8 was assessed for all the countries combined using a multigroup Confirmatory Factor Analysis (mCFA) considering the country as the group variable. According to the conventions and reporting of measurement invariance,[26] three steps of invariance were assessed: Configural invariance (i.e., the consistency of the latent structure of the scale across all the countries included in the study), metric invariance (i.e., whether the items were related to the latent trait of the scale in an equivalent way in all countries), and scalar invariance (i.e., whether the items presented the same expected response across countries). Additionally, residual invariance was assessed restricting the residual variances from scalar models equal to 1. To evaluate the measurement invariance at the different levels, goodness of fit for the configural invariance level was assessed. To evaluate metric, scalar and residual invariance, the absolute change (difference) of the goodness of fit statistics of the configural and metric model, of the metric and scalar models, and of the scalar and residual models were compared, respectively: CFI and TLI (absolute change <0.010 indicating a suitable goodness of fit), RMSEA (absolute change <0.015), and SRMR (absolute change values < 0.030).[27]

Sensitivity analyses

To assess the measurement invariance by sex and by age were performed using a mCFA considering the sex (men and women) and age (categorised in 15–29, 30–44, 45–59, 60–74, and ≥75 years) as the group variable. Additionally, sensitivity analyses to assess the robustness of the main findings (i.e., measurement invariance by country) using imputed data of those participants with up to 2 missing responses to the PHQ-8 items were carried out. For these analyses, multiple imputation models using chained equations were carried out. After imputation, cross-country equivalence and measurement invariance analyses were replicated, i.e., the mCFA considering the country as the group variable.

All analysis were weighted, using sampling weights derived from the complex sampling strategy; and additional post-stratification weights to adjust the sample to external data relating to distribution of persons in the target population. Descriptive and reliability analyses were carried out using Stata v.17 M.P. Additionally, cross-country equivalence and measurement invariance of the PHQ-8 analyses were performed using M-plus v.8.

Role of the funding source

The funders of this work had no role related to the design of the study, the data management and analysis

and the writing of the results. The first and last authors had full access to all the data in the study and affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as originally planned (and, if relevant, registered) have been explained.

Results

The countries that had a greater number of participants in the total sample analysed were (Appendix 2): Germany (weighted %: 20.48, 95% CI: 20.21–20.75), France (weighted %: 13.82, 95% CI: 13.58–14.06), Italy (weighted %: 13.51, 95% CI: 13.36–13.77) and the UK (weighted %: 12.89, 95% CI: 12.66–13.12). Additionally, Appendix 3 shows the distribution of the sample by sex and by age for the whole Europe and by country. Besides, Fig. 1 shows the distribution of the latent mean PHQ-8 scores overall for the whole Europe and by country, showing a weighted latent mean score for all Europe of 2.77 (95% CI: 2.75–2.78). The lowest latent mean scores were observed in Cyprus (1.54, 95% CI: 1.44–1.63) and Greece (1.58, 95% CI: 1.51–1.65), and the highest in Luxemburg (3.97, 95% CI: 3.83–4.11) and Iceland (4.12, 95% CI: 3.99–4.26) (Appendix 4).

According to all the goodness-of-fit statistics except the Chi-Square test (Table 1), the one-factor model was acceptable both for all 27 European countries taken together and for each of the individual countries. In addition, Appendix 5 shows the standardized item loadings (λ) from the iCFAs and that Item 2 ("*Feeling down, depressed or hopeless*") has the highest ability to discriminate whether a person has symptoms of depression in both all the countries combined, and in all the specific countries, except for Ireland (which was item 1 "*Little interest or pleasure in doing things*" λ = 0.901), Malta (item 6 "*Feeling bad about yourself, or that you are a failure, or have let yourself or your family down*": λ = 0.902), and Romania (item 8 "*Moved or speaking so slowly that other people could have noticed. Or the opposite – being so fidgety or restless that you have been moving around a lot more than usual*": λ = 0.923). As for internal consistency of the PHQ-8 (Table 2), both collectively and for each of the countries included, the α coefficient presented a value higher than 0.70), and the Ω coefficient was higher than 0.90 in all cases, except for Austria (α = 0.772 and Ω = 0.88).

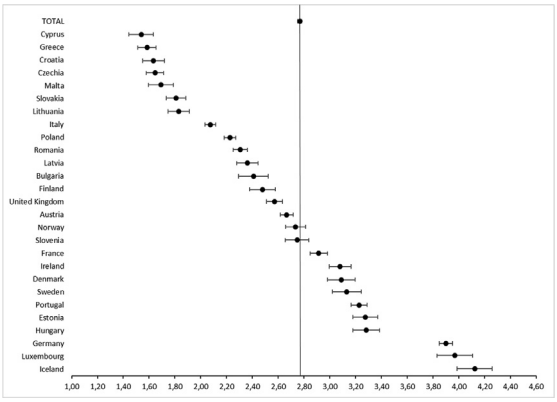

Fig. 1: **Latent mean score (95% CI) of the PHQ-8 scale for the whole Europe and by country**. 95% CI: 95% Confidence Interval; PHQ-8: Patient Health Questionnaire-8. The vertical line indicates the latent mean score of the PHQ-8 scale in all countries (latent mean: 2.77, 95% CI: 2.75–2.78).

According to the AIC (Appendix 6), the IRT models with a better fit to the data for all the countries combined and for each of all the countries considered were the GRM. From the results of the GRM (Table 3), it can be observed that item 2 has the greater ability to discriminate whether a person presents symptoms of depression for all the countries combined, and in all individual countries, except for Bulgaria and Ireland (which is item 1), Malta (item 6), and Romania (item 8) (Fig. 2). In addition, considering the test information functions both globally and for each country, Fig. 3 shows that the reliability of the test was higher than 0.90 in the positive part of the latent trait continuum, and that the countries with the highest reliability of the test were Bulgaria, Cyprus and Romania.

Table 4 shows the goodness of fit of the invariance mCFA models for the PHQ-8 in all the countries combined, using country as group variable. An adequate goodness of fit was observed at the configural level, meaning that the latent structure is consistent among the countries analysed. Invariance was also observed at the metric level, with absolute differences in the goodness of fit statistics between the configural and metric models below the stablished cut-off points, (ΔCFI = −0.008, ΔTLI = −0.004, ΔRMSEA = 0.006, ΔSRMR = 0.005), indicating that the factor loadings were comparable across countries; Additionally, the absolute differences in the goodness of fit statistics between the metric and scalar models indicated comparable item thresholds across countries (ΔCFI = 0.002, ΔTLI = 0.009, ΔRMSEA = −0.015, ΔSRMR = 0.002). Finally, residual invariance was also observed at the residual level comparing the differences in the goodness of fit statistics between the scalar and residual models (ΔCFI = 0.011, ΔTLI = 0.004, ΔRMSEA = −0.007, ΔSRMR = 0.007).

Sensitivity analyses to assess measurement invariance by sex (Appendix 7) and by age (Appendix 8), shows the measurement equivalence of the PHQ-8 at the configural, metric and scalar levels. The differences found in the goodness of fit statistics at the configural and metric levels and at the metric and scalar levels were suitable both by sex (configural vs metric: ΔCFI = 0.000, ΔTLI = 0.004, ΔRMSEA = −0.005, ΔSRMR = 0.001, and

	χ^2	CFI	TLI	RMSEA (90% CI)	SRMR
Austria	376.522	0.982	0.975	0.034 (0.031, 0.037)	0.031
Bulgaria	737.577	0.989	0.984	0.083 (0.078, 0.088)	0.024
Croatia	321.504	0.988	0.983	0.055 (0.050, 0.060)	0.025
Cyprus	248.911	0.995	0.993	0.049 (0.044, 0.055)	0.021
Czechia	294.473	0.987	0.982	0.046 (0.041, 0.050)	0.026
Denmark	464.992	0.989	0.985	0.064 (0.059, 0.069)	0.026
Estonia	744.585	0.971	0.960	0.082 (0.077, 0.087)	0.038
Finland	283.698	0.992	0.988	0.051 (0.045, 0.056)	0.023
France	1189.875	0.985	0.979	0.064 (0.061, 0.067)	0.028
Germany	1971.429	0.983	0.976	0.063 (0.061, 0.066)	0.029
Greece	348.102	0.989	0.985	0.046 (0.042, 0.050)	0.023
Hungary	649.047	0.977	0.967	0.074 (0.069, 0.079)	0.036
Iceland	433.145	0.975	0.966	0.074 (0.068, 0.080)	0.034
Ireland	887.966	0.981	0.974	0.069 (0.065, 0.073)	0.031
Italy	1065.959	0.989	0.984	0.049 (0.046, 0.051)	0.024
Latvia	480.412	0.983	0.977	0.059 (0.055, 0.064)	0.028
Lithuania	462.687	0.985	0.979	0.067 (0.061, 0.072)	0.036
Luxembourg	375.161	0.985	0.979	0.070 (0.064, 0.076)	0.027
Malta	170.753	0.992	0.989	0.044 (0.038, 0.050)	0.025
Norway	327.054	0.984	0.978	0.044 (0.040, 0.048)	0.026
Poland	987.347	0.990	0.986	0.047 (0.044, 0.049)	0.023
Portugal	753.756	0.988	0.983	0.045 (0.042, 0.048)	0.025
Romania	1864.581	0.990	0.986	0.075 (0.072, 0.078)	0.027
Slovakia	602.560	0.975	0.965	0.073 (0.068, 0.078)	0.033
Slovenia	437.414	0.985	0.978	0.059 (0.055, 0.064)	0.028
Sweden	582.177	0.985	0.978	0.070 (0.065, 0.075)	0.026
United Kingdom	1051.058	0.985	0.979	0.054 (0.051, 0.057)	0.027
Total	**10,372.488**	**0.986**	**0.980**	**0.045 (0.044, 0.045)**	**0.025**

χ^2: chi-square test; CFI: comparative fit index; TLI: Tucker–Lewis index; RMSEA (90% CI): Root Mean Square Error of approximation, and 90% confidence interval; SRMR: Standardized Root Mean Square Residual; CI: confidence interval. χ^2 with 20 degrees of freedom and $p < 0.001$ for iCFA model for all countries.

Table 1: **Fit statistics for Confirmatory Factor Analysis for categorical items (iCFA) total and by country.**

metric vs scalar: ΔCFI = 0.003, ΔTLI = 0.006, ΔRMSEA = –0.009, ΔSRMR <0.001) and by age (configural vs metric: ΔCFI = –0.003, ΔTLI = 0.001, ΔRMSEA = –0.001, ΔSRMR = 0.003, and metric vs scalar: ΔCFI = 0.004, ΔTLI = 0.008, ΔRMSEA = –0.012, ΔSRMR <0.001). Finally, sensitivity analyses using imputed data show the number of participants with missing data in the PHQ-8 (Appendix 9), and consistency with the main results, i.e., a similar goodness of fit of the invariance mCFA models for the PHQ-8 using both imputed and non-imputed data.

Discussion

To our knowledge, this is the largest study carried out to date to determine the internal structure, reliability and cross-country validity of a mental health questionnaire and, specifically, for the PHQ-8, one of the self-reported questionnaires most widely used nowadays worldwide for the screening and severity assessment of depression. The results show that the PHQ-8 is a reliable measure to assess symptoms of depression at the population level among 27 European countries. Importantly, our findings show measurement invariance of PHQ-8 scores across all the countries studied (i.e., equivalence at configural, metric, and scalar levels) and, hence, the suitability of the comparisons of the PHQ-8 scores between countries. These findings suggest that the PHQ-8 could be a particularly relevant questionnaire to assess and compare the presence of depressive symptoms at the population level in Europe and, consequently, to inform and guide public mental health policies to reduce their burden.

It should be noted the magnitude of our study, including a large sample from 27 different countries and more than a quarter of million participants (likely to be representative of more than 400 million people) is a very significant strength.[21] As far as we know, this is the largest study carried out to date assessing psychometric properties of a mental health evaluation measure in general, and of a measure for the assessment of depression in particular, even larger than all previous meta-analyses.[10,12,19,28] Additionally, further to the evidence from previous studies carried out in specific countries or in a reduced number of them,[29–31] given the robustness of the methods used (verified with the

	N	Cronbach's alpha coefficient[a]	Omega coefficient[b]
Austria	15,701	0.772	0.883
Bulgaria	5258	0.918	0.967
Croatia	5016	0.863	0.951
Cyprus	4695	0.905	0.969
Czechia	6607	0.832	0.941
Denmark	5449	0.876	0.941
Estonia	5439	0.836	0.917
Finland	5146	0.878	0.944
France	14,191	0.876	0.945
Germany	24,404	0.863	0.918
Greece	7834	0.878	0.952
Hungary	5777	0.832	0.922
Iceland	3812	0.827	0.894
Ireland	9046	0.890	0.950
Italy	21,934	0.847	0.936
Latvia	6607	0.838	0.928
Lithuania	4982	0.849	0.946
Luxembourg	3629	0.875	0.929
Malta	3974	0.837	0.946
Norway	8069	0.814	0.904
Poland	22,076	0.866	0.942
Portugal	17,974	0.866	0.936
Romania	16,422	0.919	0.970
Slovakia	5489	0.837	0.937
Slovenia	5914	0.851	0.929
Sweden	5737	0.878	0.943
United Kingdom	17,706	0.878	0.945
Total	**258,888**	**0.870**	**0.939**

N: number of participants; PHQ-8: Patient Health Questionnaire-8. [a]It must be higher than 0.85 for the scale to be considered reliable (calculated using standardized factor loadings from a weighted factor analysis for categorical items). [b]It must be higher than 0.70 for the scale to be considered reliable (Calculated using the weighted variance of each item and of the total items).

Table 2: **Cronbach's alpha and omega coefficients for the PHQ-8 scale for the whole Europe and by country.**

sensitivity analyses), the large sample size, and the likely representativeness of the data of EHIS-2 at the country level,[21] our results could be considered as justification for the use of the PHQ-8 at the population level in all the countries included within this study.

The evidence of the comparability of depression measures across European countries shows that from a broad variety of questionnaires for the assessment of depression, such as the Beck Depression Inventory II (BDI-II), or the Centre for Epidemiological Studies Depression Scale (CES-D), the PHQ-9 has been identified as the most extensively evaluated measure.[32,33] It should be also noted that previous studies to assess the psychometric properties of the PHQ-9 including samples from different countries,[28,34] and to compare the PHQ-8 and the PHQ-9,[15,16] have shown that they could be considered as equivalent measures. The availability of a measure for the assessment of depressive symptoms with a suitable cross-country comparability, opens a window of opportunity for the development of a common framework for their assessment worldwide. In this sense, the reliability and cross-country comparability of the PHQ-8 (showing invariance between countries in Europe at all levels), together with its low burden and extended use,[17] makes it a relevant candidate to consider as reference (at least when the use of clinical interviews is not feasible) for the assessment of depressive symptoms in large population-based studies.

Our results are consistent with those from previous studies carried out using samples from a single country or some of those included in our study, such as Portugal and the UK.[30,31] Additionally, the adequate reliability of the PHQ-8 for each of the 27 countries included, and the measurement invariance between countries found must be highlighted. These results, together with its current use in large epidemiological studies worldwide,[5,15] and the lower number of items of the PHQ-8 compared to some other questionnaires with adequate psychometric properties widely used for the screening and severity assessment of depression, such as the Center for Epidemiologic Studies

	N	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
Austria[a]	15,701	2.04	2.78	1.26	1.94	1.47	1.92	1.57	1.39
Bulgaria	5258	4.82	4.76	3.01	3.25	3.01	3.87	4.22	4.15
Croatia	5016	3.68	4.09	2.47	3.01	2.65	3.05	3.08	3.26
Cyprus	4695	5.28	5.46	3.37	3.93	3.48	4.07	3.77	3.91
Czechia	6607	2.96	3.40	1.98	2.35	2.04	3.28	3.33	2.72
Denmark	5449	3.31	4.55	1.55	2.83	2.07	3.24	2.61	2.26
Estonia	5439	2.87	3.26	1.64	1.92	1.51	2.48	1.99	1.88
Finland	5146	3.21	4.33	2.04	2.85	2.41	3.24	2.50	2.05
France	14,191	3.57	3.90	2.01	2.89	2.40	2.92	2.68	2.60
Germany	24,404	2.65	3.25	1.42	2.67	1.93	2.26	2.00	1.88
Greece	7834	3.98	3.99	2.51	2.80	2.59	3.68	3.03	3.11
Hungary	5777	2.48	3.17	1.51	1.82	1.65	2.80	2.20	2.07
Iceland	3812	2.91	3.54	1.30	1.57	1.29	2.14	1.50	1.38
Ireland	9046	3.86	2.82	2.33	2.81	2.51	3.58	2.93	2.68
Italy	21,934	3.11	3.28	1.88	2.63	1.96	2.76	2.56	2.92
Latvia	6607	3.19	3.83	1.66	2.23	1.80	2.45	2.50	2.11
Lithuania	4982	3.90	3.99	2.33	2.51	2.08	3.44	2.82	2.67
Luxembourg	3629	2.65	3.38	1.55	2.45	1.97	2.38	2.12	2.12
Malta	3974	1.94	3.37	2.47	2.99	2.81	4.20	3.86	3.57
Norway	8069	1.53	2.63	1.65	2.19	1.59	2.23	2.10	2.05
Poland	22,076	2.99	3.48	2.12	2.62	2.36	3.47	2.92	2.47
Portugal	17,974	2.92	3.67	1.66	2.51	1.82	3.66	2.43	2.48
Romania	16,422	4.14	3.95	3.56	3.76	3.37	3.77	4.28	4.37
Slovakia	5489	2.79	3.50	2.26	2.29	2.36	2.80	3.39	1.95
Slovenia	5914	3.15	3.38	1.44	2.25	1.96	3.06	2.20	2.23
Sweden	5737	2.65	4.32	1.98	2.88	2.37	2.96	2.57	2.44
United Kingdom	17,706	3.34	3.88	2.07	2.40	2.29	3.09	2.50	2.46
Total	**258,888**	**3.06**	**3.42**	**1.91**	**2.67**	**2.22**	**2.86**	**2.55**	**2.43**

N: number of participants. α1: discrimination capacity for item 1; α2: discrimination capacity for item 2; α3: discrimination capacity for item 3; α4: discrimination capacity for item 4; α5: discrimination capacity for item 5; α6: discrimination capacity for item 6; α7: discrimination capacity for item 7; α8: discrimination capacity for item 8. [a]The higher the discrimination parameter, the greater the ability of the item to discriminate whether a person presents symptoms of depression.

Table 3: **Discrimination capacity of the items from Graded Response Model (GRM) for the whole Europe and by country.**

Depression Scale (CES-D, composed by 20 items),[10] makes it a relevant and valuable option to consider for the assessment of depressive symptoms and their impact at the population level. Further research assessing the reliability, validity and, particularly, the cross-county comparability of the PHQ-8 including other countries from both Europe and worldwide as well as other measures for the screening and severity assessment of depressive symptoms, will enhance the evidence base to determine the suitability of their use within the specific countries and to compare the estimations derived from its use.

Our study has several limitations. First, it should be noted that the PHQ-8 instead of the PHQ-9 (the original version) was used. However, the PHQ-8 have shown a very high equivalence with the PHQ-9,[16] and our results support the use of the PHQ-8 in Europe, a suitable option to be used in population-based studies in all the countries included within EHIS-2 following their different regulations. It should be also noted that our study focuses on the population level and not on the individual level, i.e., it is focused on determining the internal structure, reliability, and measurement invariance across countries of the PHQ-8, and not on determining optimal cut-off scores for the screening of depressive disorders at the individual or patient levels. While these scores are necessary to maximise the number of cases detected in a specific population, they should be always determined after considering at least its internal structure, i.e., the number of dimensions that the questionnaire is assessing. Furthermore, to determine optimal cut-off scores, a clinical interview, i.e., an approach to a gold standard measure in the case of mental disorders, needs to be administered to the participants in the survey in addition to the PHQ-8, and this could not be feasible in population-based studies with large samples, as it is the case of EHIS-2.[17,25,28] Another limitation is related to the lack of data from Belgium, the Netherlands, Spain, and Turkey and the uncertainty on whether the PHQ-8 is cross-country comparable in these countries. Further studies including data from

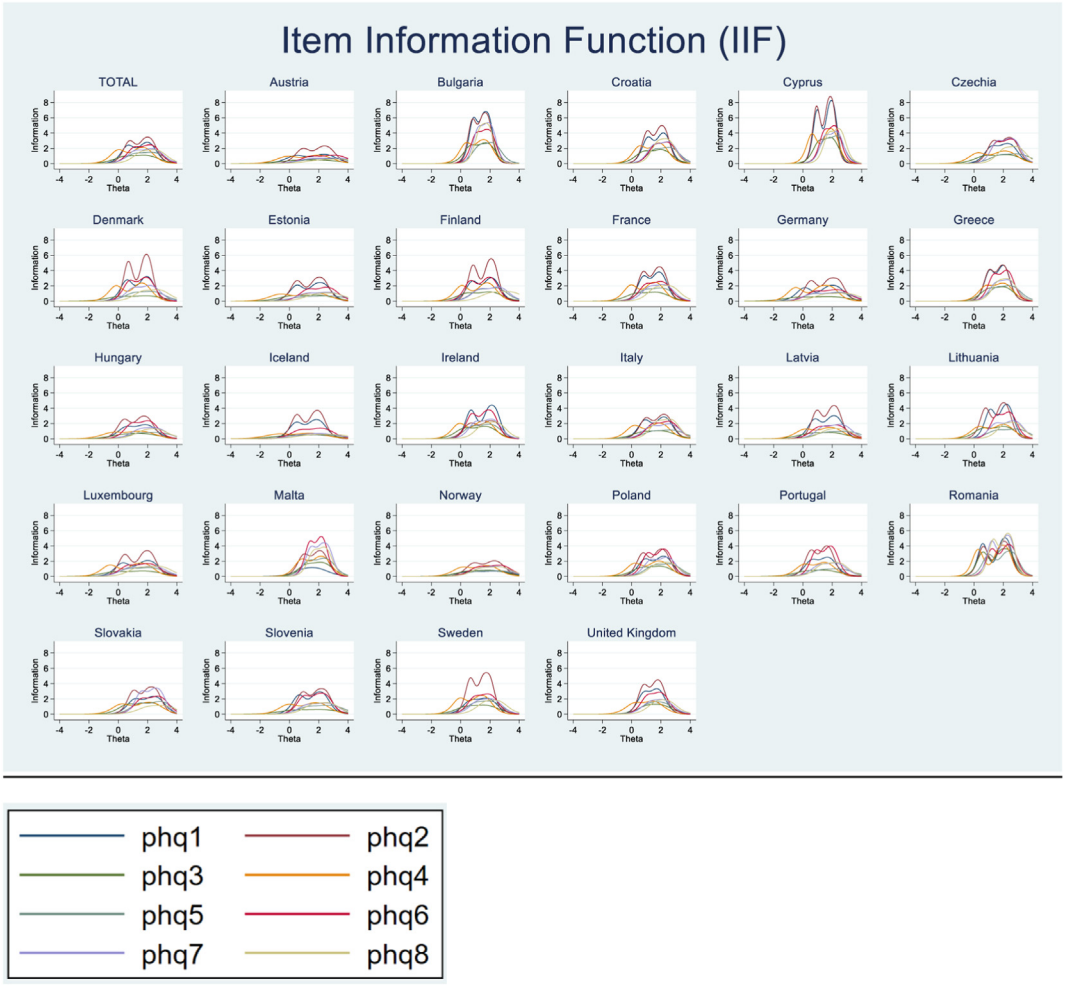

Fig. 2: **Item Information Function (IIF) for each item for the whole Europe and by country.**

these and other countries worldwide, should provide a wider perspective of the cross-country equivalence of the PHQ-8. Finally, the differences found in the coefficients from the graded response models must be commented. Despite the measurement invariance found across countries, variability in the item with a higher discrimination capacity in Bulgaria, Ireland, Malta and Romania was found, i.e., item 2 ("*Feeling down, depressed or hopeless*") might not have the largest discriminating ability for these countries. While sensitivity analyses show that these differences could not be explained by the sex or age distribution of the population of the different countries included. However, they may be explained by an error in the context of a survey using pooled data collected in different countries, and to a potential cross-cultural differences (including linguistic-related aspects) in the concept of depression and their symptoms (corresponding to each of the items of the PHQ).[18,35] Therefore, these results could serve as starting point for further research to explore this potential source of error, the possible cultural differences in the concept of depression and their possible clinical relevance,[17] and to highlight the necessity of further cross-country validation studies to better understand depression and how to assess it.

In conclusion, the PHQ-8 could be considered a reliable and valid self-reported measure for the screening and severity assessment of depressive symptoms in Europe, with a suitable comparability between countries at all levels. New research considering other countries and these results could be helpful to develop a common framework for the assessment of depression both in Europe and worldwide. This will be helpful to improve the screening and severity assessment of depression, the knowledge about its determinants, to inform and focus preventive measures and, hence, reduce its burden.

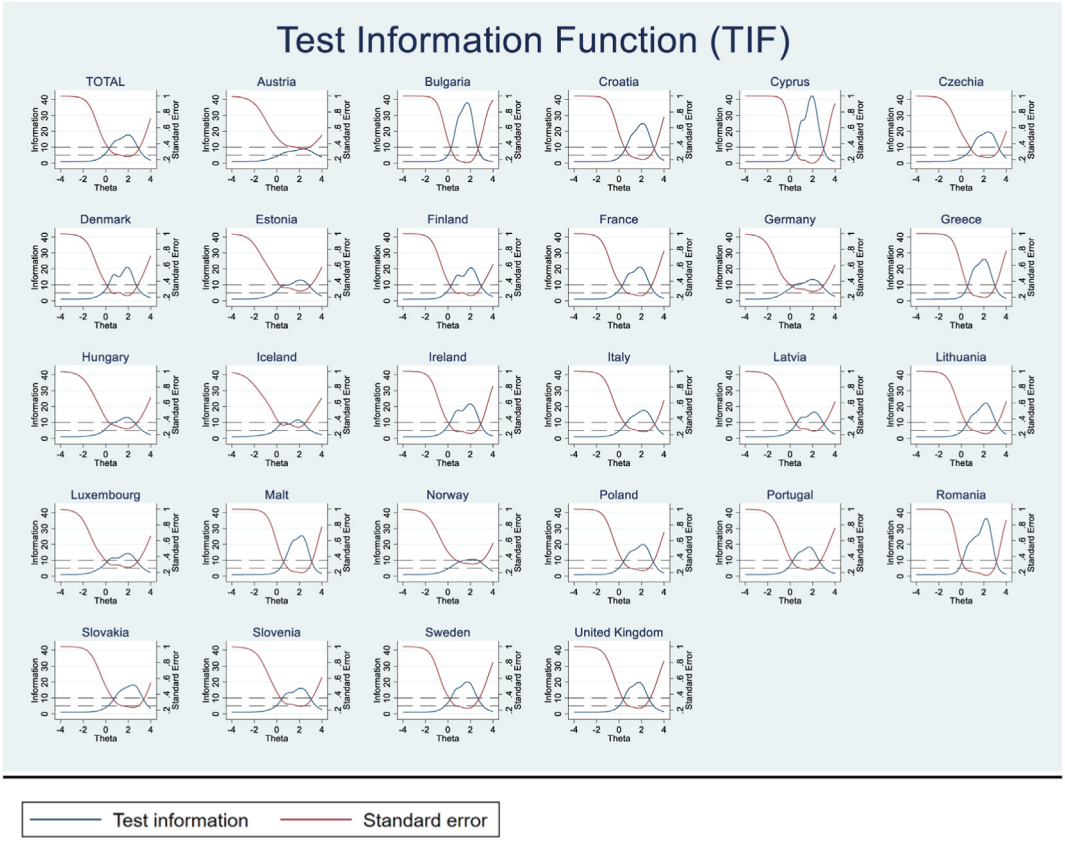

***Fig. 3*: Test Information Function (TIF) for the whole Europe and by country**. Dash line in *Information* = 10 indicates a reliability of 0.90; Dash line in *Information* = 5 indicates a reliability of 0.80.

Invariance	χ^2	CFI	TLI	RMSEA	SRMR
Configural	18,112.640	0.987	0.982	0.058 (0.058, 0.059)	0.027
Metric	28,697.329	0.979	0.978	0.064 (0.063, 0.064)	0.032
Scalar	26,220.679	0.981	0.987	0.049 (0.048, 0.049)	0.034
Residual	40,873.880	0.970	0.983	0.056 (0.055, 0.056)	0.041

Fit statistics: χ^2: chi-square test; CFI: comparative fit index; TLI: Tucker–Lewis index; RMSEA: Root Mean Square Error of approximation; SRMR: Standardized Root Mean Square Residual; χ^2 with 540 *degrees of freedom* and $p < 0.001$ for configural invariance model; χ^2 with 722 *degrees of freedom* and $p < 0.001$ for metric invariance model; χ^2 with 11,112 *degrees of freedom* and $p < 0.001$ for scalar invariance model.

***Table 4*: Configural, metric and scalar invariance for the items of the PHQ-8 scale for the whole Europe from multigroup confirmatory factor analyses.**

Contributors

JAT and JA had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. All authors conceived and designed the study. JAT analysed and JAT, JA and GV interpreted data. JAT, AR, VM, ASB and JMV drafted the manuscript. All authors critically revised the manuscript for intellectual content. JA and VM supervised the study.

Data sharing statement

Upon request to Eurostat, data from the EHIS-2 are publicly available for research purposes.

Declaration of interests

All authors declare that they have no conflict of interest. Data of EHIS-2 is publicly available for different purposes under request to Eurostat.

Acknowledgements

The present work is partially funded by CIBER Epidemiology and Public Health (CIBERESP) as part of the Intramural call of 2021 (ESP21PI05), the SGR Agencia de Gestió Ajuts Universitaris de Recerca (AGAUR 2021 SGR 00624), and the Medical Research Council (MRC) and Guy's Charity. This article represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.lanepe.2023.100659.

References

1 Kessler RC, Birnbaum HG, Shahly V, et al. Age differences in the prevalence and co-morbidity of DSM-IV major depressive episodes: results from the WHO World Mental Health Survey Initiative. *Depress Anxiety*. 2010;27:351–364.
2 Herrman H, Kieling C, McGorry P, Horton R, Sargent J, Patel V. Reducing the global burden of depression: a Lancet–World Psychiatric Association Commission. *Lancet*. 2019;393(10189):e42–e43. https://doi.org/10.1016/S0140-6736(18)32408-5.
3 Ferrari AJ, Charlson FJ, Norman RE, et al. Burden of depressive disorders by country, sex, age, and year: findings from the Global Burden of Disease Study 2010. *PLoS Med*. 2013;10(11):e1001547. https://doi.org/10.1371/journal.pmed.1001547.
4 GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Psychiatry*. 2022;9:137–150. https://doi.org/10.1016/S2215-0366(21)00395-3.
5 Arias-de la Torre J, Vilagut G, Ronaldson A, et al. Prevalence and variability of current depressive disorder in 27 European countries: a population-based study. *Lancet Public Health*. 2021;6(10):e729–e738. https://doi.org/10.1016/S2468-2667(21)00047-5.
6 Regier DA, Kaelber CT, Rae DS, et al. Limitations of diagnostic criteria and assessment instruments for mental disorders: implications for research and policy. *Arch Gen Psychiatry*. 1998;55:109–115. https://doi.org/10.1001/archpsyc.55.2.109.
7 Fayers P. Item response theory for psychologists. *Qual Life Res*. 2004;13(3):715–716. https://doi.org/10.1023/b:qure.0000021503.45367.f2.
8 Reise SP, Waller NG. Item response theory and clinical measurement. *Annu Rev Clin Psychol*. 2009;5:27–48. https://doi.org/10.1146/annurev.clinpsy.032408.153553.
9 Sousa VD, Rojjanasrirat W. Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. *J Eval Clin Pract*. 2011;17(2):268–274. https://doi.org/10.1111/j.1365-2753.2010.01434.x.
10 Vilagut G, Forero CG, Barbaglia G, Alonso J. Screening for depression in the general population with the center for epidemiologic studies depression (CES-D): a systematic review with meta-analysis. *PLoS One*. 2016;11:e0155431.
11 Poortinga YH. Equivalence of cross-cultural data: an overview of basic issues. *Int J Psychol*. 1989;24:737–756. https://doi.org/10.1080/00207598908247842.
12 Moriarty AS, Gilbody S, McMillan D, Manea L. Screening and case finding for major depressive disorder using the Patient Health Questionnaire (PHQ-9): a meta-analysis. *Gen Hosp Psychiatry*. 2015;37:567–576.
13 American Psychiatric Association. *DSM-IV*. 2000.
14 Uher R, Payne JL, Pavlova B, Perlis RH. Major depressive disorder in DSM-5: implications for clinical practice and research of changes from DSM-IV. *Depress Anxiety*. 2014;31:459–471. https://doi.org/10.1002/da.22217.
15 Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. *J Affect Disord*. 2009;114:163–173.
16 Wu Y, Levis B, Riehm KE, et al. Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis. *Psychol Med*. 2020;50(8):1368–1380. https://doi.org/10.1017/S0033291719001314.
17 Arias-de la Torre J, Vilagut G, Ronaldson A, Serrano-Blanco A, Alonso J. PHQ-8 scores and estimation of depression prevalence – author's reply. *Lancet Public Health*. 2021;6(11):e794. https://doi.org/10.1016/S2468-2667(21)00226-7.
18 Dreher A, Hahn E, Diefenbacher A, et al. Cultural differences in symptom representation for depression and somatization measured by the PHQ between Vietnamese and German psychiatric outpatients. *J Psychosom Res*. 2017;102:71–77. https://doi.org/10.1016/j.jpsychores.2017.09.010.
19 Shevlin M, Butter S, McBride O, et al. Measurement invariance of the Patient Health Questionnaire (PHQ-9) and generalized anxiety disorder scale (GAD-7) across four European countries during the COVID-19 pandemic. *BMC Psychiatry*. 2022;22:154. https://doi.org/10.1186/s12888-022-03787-5.
20 Eurostat. European health interview survey (EHIS wave 2). Methodological manual. https://ec.europa.eu/eurostat/documents/3859598/5926729/KS-RA-13-018-EN.PDF/26c7ea80-01d8-420e-bdc6-e9d5f6578e7c; 2013
21 Eurostat. Quality report of the second wave of the European health interview survey. https://ec.europa.eu/eurostat/documents/7870049/8920155/KS-FT-18-003-EN-N.pdf/eb85522d-bd6d-460d-b830-4b2b49ac9b03; 2018.
22 Hu L-T, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct Equ Modeling*. 2009;6(1):1–55. https://doi.org/10.1080/10705519909540118.
23 Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. *Chest*. 2008;96(5):1161–1164. https://doi.org/10.1378/chest.96.5.1161.
24 Revelle W, Condon D. Reliability from alpha to omega: a tutorial. *Psychol Assess*. 2019;31(12):1395–1411.
25 Stenbeck M, Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory. *Contemp Sociol*. 1992;21:289–290. https://doi.org/10.2307/2075521.
26 Putnick DL, Bornstein MH. Measurement invariance conventions and reporting: the state of the art and future directions for psychological research. *Dev Rev*. 2016;41:71–90.
27 Chen FF. Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct Equ Model*. 2007;14:464–504. https://doi.org/10.1080/10705510701301834.
28 Negeri ZF, Levis B, Sun Y, et al. Accuracy of the Patient Health Questionnaire-9 for screening to detect major depression: updated systematic review and individual participant data meta-analysis. *BMJ*. 2021;375:n2183. https://doi.org/10.1136/BMJ.N2183.
29 Galenkamp H, Stronks K, Snijder MB, Derks EM. Measurement invariance testing of the PHQ-9 in a multi-ethnic population in Europe: the HELIUS study. *BMC Psychiatry*. 2017;17:349. https://doi.org/10.1186/s12888-017-1506-9.
30 Lamela D, Soreira C, Matos P, Morais A. Systematic review of the factor structure and measurement invariance of the Patient Health Questionnaire-9 (PHQ-9) and validation of the Portuguese version in community settings. *J Affect Disord*. 2020;276:220–233. https://doi.org/10.1016/j.jad.2020.06.066.
31 Cameron IM, Crawford JR, Lawton K, Reid IC. Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. *Br J Gen Pract*. 2008;58:32–36. https://doi.org/10.3399/bjgp08X263794.
32 Missinne S, Vandeviver C, Van de Velde S, Bracke P. Measurement equivalence of the CES-D 8 depression-scale among the ageing population in eleven European countries. *Soc Sci Res*. 2014;46:38–47. https://doi.org/10.1016/j.ssresearch.2014.02.006.
33 El-Den S, Chen TF, Gan YL, Wong E, O'Reilly CL. The psychometric properties of depression screening tools in primary healthcare settings: a systematic review. *J Affect Disord*. 2018;225:503–522.
34 Bianchi R, Verkuilen J, Toker S, et al. Is the PHQ-9 a unidimensional measure of depression? A 58,272-participant study. *Psychol Assess*. 2022;34:595–603.
35 Alarcón RD. Culture, cultural factors and psychiatric diagnosis: review and projections. *World Psychiatry*. 2009;8:131–139. https://doi.org/10.1002/j.2051-5545.2009.tb00233.x.