2022

Is the PHQ-9 a Unidimensional Measure of Depression? A 58,272-Participant Study

Renzo Bianchi
Norwegian University of Science and Technology

Jay Verkuilen
CUNY Graduate Center

Sharon Toker
Tel Aviv University

Irvin Sam Schonfeld
CUNY Graduate Center

Markus Gerber
University of Basel

See next page for additional authors

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Assessment

ISSN: 1040-3590

2022, Vol. 34, No. 6, 595–603
https://doi.org/10.1037/pas0001124

BRIEF REPORT

Is the PHQ-9 a Unidimensional Measure of Depression? A 58,272-Participant Study

Renzo Bianchi[1], Jay Verkuilen[2], Sharon Toker[3], Irvin Sam Schonfeld[4], Markus Gerber[5], Elmar Brähler[6, 7], and Kurt Kroenke[8]

[1] Institute of Work and Organizational Psychology, University of Neuchâtel
[2] Department of Educational Psychology, The Graduate Center of the City University of New York
[3] Coller School of Management, Tel Aviv University
[4] Department of Psychology, The City College of the City University of New York
[5] Department of Sport, Exercise and Health, University of Basel
[6] Department of Medical Psychology and Medical Sociology, Leipzig University Hospital, University of Leipzig
[7] Clinic and Polyclinic for Psychosomatic Medicine and Psychotherapy, University Medical Center of the Johannes Gutenberg-University Mainz
[8] Regenstrief Institute, Indiana University School of Medicine

The PHQ-9 has become a measure of reference in depression research and clinical practice. However, the issue of the PHQ-9's unidimensionality has not been fully elucidated, and the usability of the PHQ-9's total score requires clarification. In this study, we examined the dimensionality, scalability, and monotonicity properties of the PHQ-9 as well as the scale's total-score reliability. We did so based on exploratory structural equation modeling (ESEM) bifactor analysis and Mokken scale analysis (MSA). We relied on a total of 58,272 participants (63% female; $M_{age} = 43$, $SD_{age} = 13$) from 29 samples involving seven different countries (e.g., Germany, the U.S.) and five different languages (e.g., German, English). We found no concerning deviations from measurement invariance for our ESEM bifactor model, neither across samples nor across sexes, age groups, and languages. The PHQ-9 met the requirements for essential unidimensionality in the pooled sample and across sex-, age-, and language-based subsamples. In each case, the general factor was strong (e.g., factor loadings ranged from 0.725 to 0.893 in the pooled sample) and Omega Hierarchical values exceeded 0.900. The correlations between the general factor and the observed total scores were large (≥0.952). Our MSA, including multilevel MSA, revealed that the PHQ-9's scalability is satisfactory. No monotonicity violation was detected, suggesting that the scale's total score accurately orders respondents on the latent Depression variable. Total-score reliability was good. This study provides robust evidence that the PHQ-9 can be used as a unidimensional measure of depressive symptoms by researchers and practitioners.

Public Significance Statement

Because depression is a widespread condition with potentially lethal consequences, assessing depressive symptoms reliably and validly is of critical importance. This study indicates, with an unprecedented level of confidence, that researchers and practitioners are justified in employing the PHQ-9 as a unidimensional measure of depressive symptoms based on the scale's total score.

Keywords: depression, bifactor analysis, exploratory structural equation modeling, mokken scaling, unidimensionality

This article was published Online First March 31, 2022.

Renzo Bianchi https://orcid.org/0000-0003-2336-0407

Data and analysis codes are available upon reasonable request by emailing the corresponding author. The study was not preregistered.

The authors have no conflict of interest to disclose.

Renzo Bianchi played lead role in conceptualization, data curation, formal analysis, investigation, methodology, resources, supervision, writing of original draft, and writing of review and editing. Jay Verkuilen played supporting role in investigation and writing of original draft and equal role in formal analysis, methodology, and writing of review and editing. Sharon Toker played lead role in resources, supporting role in data curation, formal analysis and investigation, and equal role in writing of review and editing. Irvin Sam Schonfeld played supporting role in conceptualization, formal analysis, methodology, resources and writing of original draft, and equal role in investigation and writing of review and editing. Markus Gerber played supporting role in investigation and writing of review and editing, and equal role in resources. Elmar Brähler played supporting role in investigation and writing of review and editing, and equal role in resources. Kurt Kroenke played supporting role in investigation and methodology, and equal role in resources and writing of review and editing.

Correspondence concerning this article should be addressed to Renzo Bianchi, Institute of Work and Organizational Psychology, University of Neuchâtel, Émile-Argand 11, 2000 Neuchâtel, NE, Switzerland. Email: renzo.bianchi@unine.ch

Supplemental materials: https://doi.org/10.1037/pas0001124.supp

With over 300 million individuals affected worldwide, depression constitutes a major contributor to the global burden of disease (Gotlib & Hammen, 2014; James et al., 2018). The prevalence of major depressive disorder exceeds 15% in countries such as the U.S. and appears to be on the rise for several decades (American Psychiatric Association, 2013; Hasin et al., 2018; Kessler et al., 2003; Weinberger et al., 2018). In individuals with no noticeable vulnerability to the condition, depression is thought to reflect an adaptive breakdown in the face of insurmountable adversity (Pryce et al., 2011; Willner et al., 2013). Depression has been associated with systemic alterations (e.g., at endocrine, neurological, and immune levels) and is a prime risk factor for attempted and completed suicide (Chesney et al., 2014; Pryce et al., 2011; Willner et al., 2013).

Depression has long been approached categorically and constitutes an established nosological entity (American Psychiatric Association, 2013). However, evidence has emerged to indicate that depression can be considered a dimensional phenomenon, varying in severity along a continuum (Haslam et al., 2012; Kotov et al., 2017; Liu, 2016; Wichers, 2014). From this perspective, clinical forms of depression (i.e., diagnosable depressive disorders) represent the high end of the continuum. The interest in dimensional approaches to depression has led investigators to develop measures of depression severity. Among these measures, the PHQ-9 has become an instrument of reference in research as well as an international legacy measure in detecting and monitoring depression in clinical settings (Gliklich et al., 2020; Kroenke, 2021; Martin-Subero et al., 2017).

The PHQ-9 reflects the nine diagnostic criteria for major depression of the *Diagnostic and statistical manual of mental disorders*, fifth edition (*DSM-5*; American Psychiatric Association, 2013; Kroenke et al., 2001; Spitzer et al., 1999).[1] The scale thus covers both cognitive-affective and somatic manifestations of major depression. More specifically, the items of the PHQ-9 assess anhedonia, dysphoria, sleep alterations, fatigue/loss of energy, appetite alterations, feelings of worthlessness, cognitive impairment, psychomotor alterations, and thoughts of self-harm. Consistent with the *DSM-5* diagnostic criteria, the symptoms of interest are evaluated over a 2-week period. The PHQ-9 has gained considerable popularity in public health science, as illustrated by its incorporation into the *National Health and Nutrition Examination Survey* (NHANES) and the *Behavioral Risk Factor Surveillance System* supervised by the U.S. Centers for Disease Control and Prevention. The PHQ-9 has been used in thousands of studies and translated into more than 100 languages to date (Kroenke, 2021).

The PHQ-9 has demonstrated satisfactory psychometric properties and significant clinical utility across a wide array of settings, countries, and populations (Beard et al., 2016; Kroenke et al., 2010; Levis et al., 2019). However, findings on the factorial structure and dimensionality of the PHQ-9 remain equivocal (Lamela et al., 2020). Some confirmatory factor analytic studies concluded that a two-factor solution, distinguishing between Cognitive-Affective and Somatic factors, best characterizes the PHQ-9 (e.g., Keum et al., 2018; Patel et al., 2019). From one study to another, however, the PHQ-9 items assigned to Cognitive-Affective and Somatic factors varied greatly (e.g., Beard et al., 2016; Chilcot et al., 2013; Petersen et al., 2015). Moreover, between-factor correlations were often high enough (≥.80) to suggest a unidimensional structure (e.g., Beard et al., 2016; Bianchi et al., 2020; Bianchi & Mirkovic, 2020, Boothroyd et al., 2019; Keum et al., 2018; Patel et al., 2019; Schuler et al., 2018). Consistent with this observation, one-factor solutions occasionally showed an acceptable fit (e.g., Bianchi et al., 2020; Bianchi & Mirkovic, 2020; Schuler et al., 2018). Adding to the heterogeneity of findings, some investigators encountered difficulties converging on a factorial solution involving all PHQ-9 items (Turgeman-Lupo et al., 2020).

A few studies approached the factorial structure of the PHQ-9 relying on bifactor models, in which specific factors (bifactors) related to cognitive-affective and somatic items were considered in addition to a general Depression factor (Doi et al., 2018; Lamela et al., 2020; Stochl et al., 2020). Such models generally demonstrated an acceptable fit. Unidimensionality issues, however, were only examined superficially, based on the relative loadings of the items on the general and specific factors. Key indices of scale dimensionality, such as the Explained Common Variance (ECV) index (Rodriguez et al., 2016a, 2016b), were seldom considered. Moreover, the reliability of the PHQ-9 in relation to the general factor was rarely reported (for a notable exception, see Stochl et al., 2020). Additional limitations of these studies include a narrow linguistic scope (in terms of the PHQ-9 versions under consideration), an absence of sex- and age-specific analyses, and/or the use of relatively small samples.

This study reexamined the dimensionality and structural properties of the PHQ-9. Clarifying the dimensionality and structural properties of the PHQ-9 is crucial for evaluating the usability of the scale's total score. The study capitalized on 29 samples, yielding a total of 58,272 respondents. Seven countries and five language groups were represented. Because limited attention has been paid to the dimensionality and structural properties of the PHQ-9 across sexes, age groups, and languages in past research (Patel et al., 2019), we conducted sex-, age-, and language-specific analyses in addition to global analyses. We primarily relied on an exploratory structural equation modeling (ESEM) bifactor analytic framework (Marsh et al., 2014; Rodriguez et al., 2016a, 2016b). A bifactor model partitions the covariance among a set of items into a general factor and specific factors that can be defined in a theory-driven manner. The assumptions underlying ESEM bifactor analysis are less rigid, more complexity-compatible, and ultimately more realistic than those underlying "classical" confirmatory factor analysis (Marsh et al., 2014; Morin et al., 2016). Furthermore, ESEM bifactor analysis allows investigators to determine whether a measure is sufficiently unidimensional—a property referred to as *essential unidimensionality*—to be used based on its total score (Rodriguez et al., 2016a, 2016b).

We further investigated the dimensionality of the PHQ-9 using Mokken scale analysis (MSA; Mokken, 1971; Molenaar, 1982; van der Ark, 2012), a nonparametric method anchored in item response theory (IRT). Mokken scaling is founded on Loevinger's

[1] The PHQ-9 was developed in the era of the DSM-IV and DSM-IV-TR (American Psychiatric Association, 1994, 2000), but the diagnostic criteria for major depression have remained essentially unchanged since then.

homogeneity (H) coefficient, which itself is rooted in Guttman scaling. We focused on the scalability, monotonicity, and total-score reliability of the PHQ-9 (Stochl et al., 2012). Scalability in this context refers to the extent to which endorsing more severe symptoms (e.g., thoughts of self-harm) is related to a higher probability of endorsing less severe symptoms (e.g., fatigue/loss of energy). The importance of monotonicity lies in the fact that it justifies clinicians and researchers ordering respondents on the latent continuum according to the respondents' item sum score. Total-score reliability reflects the proportion of total-score variance that is due to true-score, that is, latent variable, variance. As is the case with ESEM bifactor analysis, methods attached to IRT such as MSA are thus particularly helpful in estimating the usability of scales' total scores. Given the widespread use of the PHQ-9 in research on depression for about 2 decades, it is important that the dimensionality and structural properties of the scale be elucidated. By involving 29 samples and 58,272 respondents, our study has a potential for unprecedented external validity.

Method

Datasets and Study Samples

The present study relied on 29 datasets (ns range = 257–17,176), reflecting a total of 58,272 PHQ-9 respondents (63% female; M_{age} = 43, SD_{age} = 13). Of these 29 datasets, 28 were directly pooled by our consortium of investigators, and one emanated from the NHANES 2017–2018.[2] The samples included participants from seven countries—France, Germany, Israel, New Zealand, Spain, Switzerland, and the U.S.—and the PHQ-9 was employed in five languages—French, German, Hebrew, English, and Spanish. We relied on nonclinical samples to maximize the probability of covering the entire continuum of latent depression. Most samples were used for different purposes in previously published studies; only two samples were not used in published studies to date (see Supplemental Material 1). Each study was conducted in accordance with the ethical standards of the main investigators' home institution. The study samples are presented individually in Supplemental Material 1. The study was not preregistered.

Measure of Interest

Our measure of interest was the PHQ-9 (Kroenke et al., 2001; Spitzer et al., 1999). As noted earlier, the PHQ-9 comprises nine core symptom items (e.g., "Feeling down, depressed, or hopeless") referencing the nine diagnostic criteria for major depression of the *DSM-5* (American Psychiatric Association, 2013). Each symptom is assessed within a 2-week time window. The PHQ-9 employs a 4-point rating scale, from 0 for "*not at all*" to 3 for "*nearly every day*." The instrument can be used at no cost and is available in a wide variety of languages (https://www.phqscreeners.com/). Of our 58,272 participants, 77% (n = 44,823) exhibited PHQ-9 total scores <1.000, 19% (n = 11,230) had PHQ-9 total scores between 1.000 and 1.999, and 4% displayed PHQ-9 total scores ≥2.000 (total scores refer to domain scores here, i.e., to average scores on the scale). Because a vast majority of our samples were nonclinical in nature, a positively skewed distribution was expected (Table 1). The distributions of PHQ-9 total scores across sexes, age groups, and languages are available in the form of boxplots in Supplemental Material 2.

Data Analyses

We primarily examined the dimensionality of the PHQ-9 based on ESEM bifactor analysis (Marsh et al., 2014). We conducted the analysis in Mplus 8.6 (Muthén & Muthén, 1998–2021). We considered two specific factors in addition to the general factor, on account of the scale's cognitive-affective and somatic symptom items. We treated all items as ordinal (i.e., all responses as ordered categories), employed the weighted least squares—mean and variance adjusted—estimator, and used a bigeomin rotation. We relied on a bigeomin, rather than a target, rotation because it is currently unclear what PHQ-9 items should be, respectively, assigned to Cognitive-Affective and Somatic factors (e.g., Beard et al., 2016; Chilcot et al., 2013; Petersen et al., 2015). Put differently, there was no clear information available to specify target loadings. Through our bifactor model, we considered the possibility that the PHQ-9 may exhibit a degree of multidimensionality on account of its cognitive-affective symptom items and somatic symptom items (a degree of multidimensionality that needs to be modeled), but that this degree of multidimensionality may not be high enough to prevent the use of the test's total score or call into question the unity of the depression construct—essential unidimensionality. In addition to relationally examining the factor loadings on the general and specific factors, we computed the ECV index at both an item and a scale level to have a more straightforward view of the structure and importance of the general factor. An item-level ECV is an estimate of the proportion of an item's communality that is accounted for by the general factor. A scale-level ECV is the average of all the item-level ECVs. Scale-level ECV indices ≥0.800 are suggestive of essential unidimensionality (Rodriguez et al., 2016a, 2016b). Furthermore, we computed coefficient Omega Hierarchical (ω_H). ω_H estimates the proportion of variance in total scores that can be attributed to a single general factor, thus treating variability in scores due to specific factors as measurement error (McDonald, 1999; Rodriguez et al., 2016b). Model fit was evaluated based on the Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Standardized Root Mean Squared Residual (SRMR) statistics. According to commonly applied rules of thumb (e.g., Kline, 2016), RMSEA values should be ≤0.080, and preferably ≤0.050; CFI and TLI values should be ≥0.950; SRMR values should be ≤0.080.

To make the most of our 29 sample, 58,272-respondent dataset, we were interested in analyzing the pooled sample as well as sex-, age-, and language-based subsamples. To ascertain whether such an analytical strategy was advisable, we first inquired into the measurement invariance of the above specified ESEM bifactor model across all individual samples as well as across sexes, age groups (based on a tercile split), and languages. Sex-based analyses involved the 21,658 male participants and the 36,614 female participants. Age-based groups distinguished between younger participants (≤37 years; n = 20,293), medium-age participants (38–49 years; n = 19,611), and older participants (≥50 years;

[2] https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017.

Table 1
PHQ-9: Descriptive Statistics (Pooled Sample)

PHQ-9 items and total score	% scoring 0	% scoring 1	% scoring 2	% scoring 3	*M*	*SD*	Skewness	Kurtosis
Item 1	55.945	30.828	8.452	4.776	0.621	0.831	1.303	1.034
Item 2	60.738	28.345	7.221	3.696	0.539	0.784	1.472	1.631
Item 3	47.043	28.214	12.927	11.815	0.895	1.030	0.859	−0.502
Item 4	30.148	40.079	16.483	13.289	1.129	0.991	0.560	−0.701
Item 5	59.052	21.410	11.364	8.174	0.687	0.966	1.204	0.232
Item 6	63.902	22.556	8.335	5.207	0.548	0.853	1.513	1.389
Item 7	66.179	21.983	7.923	3.914	0.496	0.802	1.613	1.832
Item 8	80.965	12.828	4.415	1.792	0.270	0.627	2.543	6.260
Item 9	92.243	5.845	1.177	0.734	0.104	0.401	4.646	24.319
Total score (*M*)	—	—	—	—	0.588	0.593	1.255	1.192

Note. $N = 58{,}272$. The standard error for skewness is 0.010; the standard error for kurtosis is 0.020. *M* = mean. *SD* = standard deviation. Scores ranged from 0 to 3 for each item of the PHQ-9 as well as for the scale's total score (mean). Item 1 = anhedonia; Item 2 = dysphoria; Item 3 = sleep alterations; Item 4 = fatigue/loss of energy; Item 5 = appetite alterations; Item 6 = feelings of worthlessness; Item 7 = cognitive impairment; Item 8 = psychomotor alterations; Item 9 = thoughts of self-harm.

n = 18,324).[3] Language-based analyses concerned French (n = 24,144), Hebrew (n = 17,176), English (n = 12,139), German (n = 3,024), and Spanish (n = 1,789) speakers. We focused on configural invariance (factor structure is similar across groups) and scalar invariance (item thresholds are similar across groups), noting that metric invariance (factor loadings are similar across groups) cannot be investigated using ESEM and categorical outcomes (Muthén & Muthén, 1998–2021). We relied on a threshold of 0.015 for assessing changes in fit indices, which, in the context of an analysis restricted to configural invariance and scalar invariance, can be regarded as quite conservative (Putnick & Bornstein, 2016). In our framework, RMSEA and SRMR increases exceeding 0.015 on the one hand, and CFI and TLI decreases exceeding 0.015 on the other hand, were considered indicative of problematic deviations from measurement invariance.

We further investigated the dimensionality of the PHQ-9 based on MSA. We conducted our MSA using the Mokken package Version 3.0.6 (van der Ark, 2012) in R Version 4.0.3 (R Core Team, 2020). We focused on the scalability, homogeneity, and total-score reliability properties (Stochl et al., 2012). We examined the scalability property relying on the H coefficient, considered at the level of items, item pairs, and the scale (van der Ark, 2012). H coefficients range from −1 to 1, with 0 indicating no relationship. We followed commonly applied rules of thumb for interpreting H coefficients. According to these rules of thumb, item-level Hs should be >0.30. Pairwise Hs should be >0. Scale-level Hs are considered weak if $0.30 \leq H < 0.40$; moderate, if $0.40 \leq H < 0.50$; and strong, if $H \geq 0.50$. The predicates "weak," "moderate," and "strong" indicate the extent to which the ordering of individuals by test score reflects the ordering on the latent variable. We also inquired into the PHQ-9's scalability using the Automated Item Selection Procedure (AISP). The AISP employs user-defined thresholds of homogeneity based on the scale H. We explored thresholds in increments of 0.05, starting with a reference threshold of 0.30 (Stochl et al., 2012). To consider the cross-sample scalability, we made use of multilevel Mokken scaling and computed the multilevel coefficient H (Crisan et al., 2016). This method decomposes the entire dataset's responses into between-sample and within-sample components. Given our use of samples speaking different languages and living in different countries, we anticipated a degree of heterogeneity among the samples and thus considered a between component likely. However, if the PHQ-9 mostly discriminated between samples and not participants within samples, the validity of the PHQ-9 would be called into question. We computed multilevel coefficient H for each item as well as the scale. We relied on the within-sample scalability coefficient (H_W), the between-sample scalability coefficient (H_B), and the ratio of between-sample scalability coefficients to within-sample scalability coefficients (BW). We interpreted these indices based on Crisan et al. (2016) heuristic guidelines. We examined the monotonicity property focusing on monotonicity violations considered in terms of their presence, statistical significance, and seriousness (by means of the *crit* statistic; van Schuur, 2003). Total-score reliability was indexed by Guttman's lambda-2 and the Molenaar–Sijtsma statistic (Sijtsma & Emons, 2011; van der Ark, 2012), which can be interpreted along the same lines as Cronbach's alpha.

Finally, we examined the correlations between PHQ-9 scores and both age and sex. We computed Pearson's r and Spearman's ρ correlation coefficients.[4]

Results and Discussion

Depression correlated weakly with age, $r = -.064$, $\rho = -.071$, $ps < .001$, with younger individuals being more likely to report depressive symptoms than older individuals. The correlation between depression and sex was moderate, $r = -.251$, $\rho = -.295$, $ps < .001$, with women being more likely to report depressive symptoms than men.[5] These results are in keeping with the observations ordinarily made regarding the association of depression with age and sex (American Psychiatric Association, 2013). Consistent with past research (e.g., Huang et al., 2006), the most frequently endorsed PHQ-9 item was item 4 (fatigue/loss of energy), and the least frequently endorsed PHQ-9 item was item 9 (thoughts of self-harm). Detailed descriptive statistics are available in Table 1.

No concerning deviations from measurement invariance were identified, neither across samples nor across sexes, age groups, and languages (Table 2).[6] RMSEA did not increase by more than 0.004; CFI did not decrease by more than 0.010; TLI did not decrease by

[3] There were 44 missing values for the age variable.
[4] Data and analysis codes are available upon reasonable request by emailing the corresponding author.
[5] Sex was coded 0 for women and 1 for men.
[6] Focusing on *countries* instead of languages led to similar results.

Table 2
Summary of Measurement Invariance Analysis

Measurement invariance	χ^2	*df*	RMSEA	ΔRMSEA	CFI	ΔCFI	TLI	ΔTLI	SRMR	ΔSRMR
Across samples										
Configural model	1244.230	360	0.036	—	0.997	—	0.992	—	0.013	—
Scalar model	5443.525	1,317	0.040	0.004	0.987	−0.010	0.990	−0.002	0.027	0.014
Across sexes										
Configural model	860.238	24	0.035	—	0.998	—	0.995	—	0.010	—
Scalar model	1450.354	57	0.029	−0.006	0.997	−0.001	0.996	0.001	0.013	0.003
Across age groups										
Configural model	905.690	36	0.035	—	0.998	—	0.995	—	0.010	—
Scalar model	1612.460	102	0.028	−0.007	0.997	−0.001	0.997	0.002	0.012	0.002
Across languages										
Configural model	952.490	60	0.036	—	0.997	—	0.992	—	0.011	—
Scalar model	3004.044	192	0.035	−0.001	0.992	−0.005	0.992	0.000	0.020	0.009

Note. RMSEA = root mean square error of approximation; ΔRMSEA = delta (change in) RMSEA; CFI = comparative fit index; ΔCFI = delta (change in) CFI; TLI = Tucker–Lewis index; ΔTLI = delta (change in) TLI; SRMR = standardized root mean squared residual; ΔSRMR = delta (change in) SRMR; *df* = degrees of freedom.

more than 0.002; SRMR, finally, did not increase by more than 0.014. In many cases, adding constraints (from configural to scalar) resulted in superior, rather than degraded, values, especially for RMSEA and TLI. A similar observation was made by Stochl et al. (2020) in an analysis of the temporal measurement invariance of the PHQ-9. Our results bearing on measurement invariance across sexes and age groups are in keeping with previous findings (e.g., Leung et al., 2020; Patel et al., 2019).

ESEM bifactor analysis indicated that the examined model had a satisfactory fit in the pooled sample as well as across sex-, age-, and language-based subsamples (Table 3). In each case, the general factor was strong (e.g., factor loadings ranged from 0.725 to 0.893 in the pooled sample). The PHQ-9 demonstrated that it was highly reliable; over 90% of the variance of total scores was attributable to the individual differences on the general factor. The correlations between the general factor and the observed total scores were large across all configurations (≥0.952). Scale-level ECV indices were consistently reflective of essential unidimensionality. The general factor accounted for about 90% of the common variance extracted in the pooled sample. In sex-, age-, and language-based subsamples, the general factor explained between 85% and 92% of the common variance extracted. Overall, item-level ECV indices were high (e.g., 0.815−0.993 in the pooled sample) and did not lead us to consider removing any particular item (Supplemental Material 3). The specific factors were relatively weak (e.g., the maximum factor loading was 0.381 in the pooled sample). Their structure showed a degree of consistency across subsamples; the items assessing anhedonia, dysphoria, cognitive impairment, and psychomotor alterations, for instance, loaded preferentially on the same bifactor in each (sub)sample (see Supplemental Material 3 for a detailed view of factor loadings). Considering the characteristics of the general factor and bifactors together, splitting the PHQ-9 into subscales may generally be of little benefit. In any case, it should be borne in mind that subscales' ω reliability is likely to be primarily attributable to individual differences on the general factor—as detectable when comparing subscale-level ω coefficients to subscale-level ω_H coefficients. Our results are in keeping with those of confirmatory factor analytic studies that (a) characterized the PHQ-9 structure through a single factor (e.g., Bianchi et al., 2020; Bianchi & Mirkovic, 2020; Schuler et al., 2018) or (b) found two *highly correlated* factors (e.g.,

Table 3
Exploratory Structural Equation Modeling Bifactor Analysis of the PHQ-9: Fit Statistics, Reliability, and Explained Common Variance

Sample	Sample size	χ^2 (*df*)	RMSEA	RMSEA 90% CI	Probability RMSEA ≤ .050	CFI	TLI	SRMR	ML-GF	ECV	ω_H	COR
Pooled	58,272	859.743 (12)	0.035	0.033, 0.037	1.000	0.998	0.995	0.009	0.784	0.898	0.941	0.970
Male	21,658	232.864 (12)	0.029	0.026, 0.032	1.000	0.999	0.996	0.009	0.805	0.915	0.948	0.973
Female	36,614	612.359 (12)	0.037	0.035, 0.039	1.000	0.998	0.994	0.010	0.761	0.884	0.932	0.965
Younger age	20,293	285.514 (12)	0.034	0.030, 0.037	1.000	0.998	0.995	0.010	0.767	0.885	0.937	0.968
Medium age	19,611	384.241 (12)	0.040	0.036, 0.043	1.000	0.998	0.994	0.011	0.787	0.897	0.941	0.970
Older age	18,324	232.046 (12)	0.032	0.028, 0.035	1.000	0.999	0.996	0.008	0.799	0.910	0.946	0.973
French	24,144	561.030 (12)	0.044	0.040, 0.047	1.000	0.996	0.989	0.012	0.710	0.851	0.910	0.954
Hebrew	17,176	104.622 (12)	0.021	0.018, 0.025	1.000	0.998	0.995	0.011	0.733	0.849	0.907	0.952
English	12,139	178.420 (12)	0.034	0.030, 0.038	1.000	0.999	0.996	0.010	0.798	0.906	0.947	0.973
German	3,024	37.840 (12)	0.027	0.017, 0.036	1.000	0.999	0.997	0.008	0.802	0.879	0.947	0.973
Spanish	1,789	49.328 (12)	0.042	0.030, 0.054	0.858	0.997	0.992	0.014	0.758	0.862	0.936	0.967

Note. RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker–Lewis index; SRMR = standardized root mean squared residual; *df* = degrees of freedom; ML-GF = mean loading on the General factor; ECV = explained common variance; ω_H = Omega Hierarchical; COR = correlation between the General factor and the observed total scores. According to commonly applied rules of thumb (e.g., Kline, 2016), RMSEA values should be ≤0.080, and preferably ≤0.050; CFI and TLI values should be ≥0.950; SRMR values should be ≤0.080. There were 44 missing values for the age variable. Factor loadings and item-level ECV indices are available in Supplemental Material 3.

Table 4
Mokken Scale Analysis of the PHQ-9 in the Pooled Sample and Per Sex

	Pooled sample (N = 58,272)			Male sample (n = 21,658)			Female sample (n = 36,614)		
Mokken scale analysis	Item H	*SE*	95% CI	Item H	*SE*	95% CI	Item H	*SE*	95% CI
Item 1	0.521	—	—	0.540	0.006	[0.529, 0.551]	0.504	0.004	[0.497, 0.511]
Item 2	0.594	—	—	0.606	0.005	[0.596, 0.616]	0.571	0.003	[0.565, 0.577]
Item 3	0.550	—	—	0.535	0.006	[0.524, 0.546]	0.527	0.003	[0.520, 0.534]
Item 4	0.616	—	—	0.618	0.005	[0.609, 0.627]	0.586	0.003	[0.580, 0.592]
Item 5	0.537	—	—	0.530	0.006	[0.519, 0.541]	0.513	0.004	[0.506, 0.520]
Item 6	0.564	—	—	0.571	0.005	[0.561, 0.582]	0.540	0.003	[0.533, 0.546]
Item 7	0.539	—	—	0.534	0.006	[0.522, 0.546]	0.518	0.003	[0.511, 0.525]
Item 8	0.507	—	—	0.513	0.007	[0.499, 0.527]	0.486	0.004	[0.478, 0.495]
Item 9	0.533	—	—	0.548	0.009	[0.530, 0.565]	0.522	0.006	[0.510, 0.533]
Scale H	0.554	—	—	0.557	0.005	[0.548, 0.566]	0.532	0.003	[0.526, 0.537]
α	0.880			0.880			0.872		
λ_2	0.889			0.888			0.881		
MS	0.893			0.893			0.887		
AISP	0.500			0.500			0.450		

Note. α = Cronbach's alpha; λ_2 = Guttman's lambda-2; MS = Molenaar–Sijtsma statistic; AISP = automated item selection procedure (the reported threshold is the threshold at which unidimensionality starts to crack); *SE* = standard error; 95% CI = 95% confidence interval. Item 1 = anhedonia; Item 2 = dysphoria; Item 3 = sleep alterations; Item 4 = fatigue/loss of energy; Item 5 = appetite alterations; Item 6 = feelings of worthlessness; Item 7 = cognitive impairment; Item 8 = psychomotor alterations; Item 9 = thoughts of self-harm. No violation of monotonicity was detected. *SE*s and 95% CIs for item H and scale H coefficients could not be computed with a sample as large as N = 58,272 (L. A. van der Ark, personal communication, October 18, 2021).

Beard et al., 2016; Bianchi et al., 2020; Bianchi & Mirkovic, 2020; Boothroyd et al., 2019; Keum et al., 2018; Patel et al., 2019; Schuler et al., 2018).

Our MSA indicated that the scalability of the PHQ-9 was globally strong (Tables 4–6). Local differences were observed, however. The scalability was (a) slightly stronger among male participants than among female participants, (b) slightly weaker among younger respondents than among medium-age and older participants, and (c) stronger for English- and German-speaking participants than for other participants. None of the pairwise H coefficients were low, and all item-level H coefficients exceeded 0.30. We did not detect any violation of monotonicity, suggesting that the scale's total score accurately ordered individuals on the latent variable. Total-score reliability, as indexed by Guttman's lambda-2 and the Molenaar–Sijtsma statistic, was good. Our results are consistent with those of the few studies that examined the PHQ-9 using MSA in the past (e.g., Adler et al., 2012; Boothroyd et al., 2019).

Results of our multilevel MSA are summarized in Supplemental Material 4. Scale-related H_W was 0.554 and scale-related H_B was 0.445. H_B suggested that, as anticipated, the samples themselves were heterogeneous—the baseline level of depression varied over the samples, but the high H_W indicated that within-sample scalability remained strong. If this were not true, the PHQ-9 would be mostly differentiating between samples, not differentiating

Table 5
Mokken Scale Analysis of the PHQ-9 Per Age Group

	Younger age (N = 20,293)			Medium age (n = 19,611)			Older age (n = 18,324)		
Mokken scale analysis	Item H	*SE*	95% CI	Item H	*SE*	95% CI	Item H	*SE*	95% CI
Item 1	0.489	0.005	[0.479, 0.499]	0.538	0.005	[0.528, 0.548]	0.538	0.006	[0.528, 0.549]
Item 2	0.574	0.004	[0.566, 0.583]	0.604	0.004	[0.595, 0.612]	0.602	0.005	[0.593, 0.611]
Item 3	0.530	0.005	[0.521, 0.540]	0.572	0.005	[0.563, 0.581]	0.548	0.005	[0.537, 0.558]
Item 4	0.591	0.004	[0.583, 0.599]	0.632	0.004	[0.624, 0.640]	0.623	0.005	[0.615, 0.632]
Item 5	0.518	0.005	[0.509, 0.527]	0.547	0.005	[0.538, 0.557]	0.542	0.005	[0.532, 0.553]
Item 6	0.549	0.005	[0.540, 0.558]	0.574	0.005	[0.565, 0.583]	0.573	0.005	[0.563, 0.583]
Item 7	0.509	0.005	[0.499, 0.519]	0.554	0.005	[0.545, 0.564]	0.555	0.005	[0.544, 0.565]
Item 8	0.473	0.006	[0.461, 0.486]	0.514	0.006	[0.502, 0.526]	0.536	0.007	[0.523, 0.549]
Item 9	0.510	0.009	[0.493, 0.528]	0.543	0.008	[0.527, 0.559]	0.549	0.008	[0.533, 0.566]
Scale H	0.531	0.004	[0.524, 0.539]	0.568	0.004	[0.560, 0.575]	0.565	0.004	[0.556, 0.573]
α	0.870			0.885			0.885		
λ_2	0.880			0.894			0.893		
MS	0.883			0.897			0.896		
AISP	0.450			0.500			0.500		

Note. α = Cronbach's alpha; λ_2 = Guttman's lambda-2; MS = Molenaar–Sijtsma statistic; AISP = automated item selection procedure (the reported threshold is the threshold at which unidimensionality starts to crack); *SE* = standard error; 95% CI = 95% confidence interval. Item 1 = anhedonia; Item 2 = dysphoria; Item 3 = sleep alterations; Item 4 = fatigue/loss of energy; Item 5 = appetite alterations; Item 6 = feelings of worthlessness; Item 7 = cognitive impairment; Item 8 = psychomotor alterations; Item 9 = thoughts of self-harm. Younger participants = ≤37 years. Medium-age participants = 38–49 years. Older participants = ≥50 years. There were 44 missing values for the age variable. No violation of monotonicity was detected.

Table 6
Mokken Scale Analysis of the PHQ-9 Per Language

	French sample (n = 24,144)			Hebrew sample (n = 17,176)			English sample (n = 12,139)			German sample (n = 3,024)			Spanish sample (n = 1,789)		
Mokken scale analysis	Item H	SE	95% CI	Item H	SE	95% CI	Item H	SE	95% CI	Item H	SE	95% CI	Item H	SE	95% CI
Item 1	0.448	0.004	[0.439, 0.456]	0.411	0.008	[0.396, 0.425]	0.562	0.007	[0.549, 0.574]	0.560	0.015	[0.530, 0.589]	0.489	0.018	[0.453, 0.522]
Item 2	0.529	0.004	[0.522, 0.536]	0.483	0.007	[0.469, 0.497]	0.605	0.006	[0.594, 0.616]	0.610	0.013	[0.582, 0.635]	0.582	0.015	[0.553, 0.610]
Item 3	0.468	0.004	[0.459, 0.476]	0.394	0.008	[0.379, 0.409]	0.545	0.006	[0.532, 0.557]	0.531	0.015	[0.500, 0.560]	0.482	0.018	[0.446, 0.516]
Item 4	0.525	0.004	[0.518, 0.533]	0.500	0.007	[0.486, 0.513]	0.602	0.006	[0.591, 0.613]	0.615	0.013	[0.589, 0.640]	0.582	0.014	[0.553, 0.608]
Item 5	0.448	0.004	[0.440, 0.456]	0.384	0.008	[0.368, 0.400]	0.546	0.007	[0.533, 0.559]	0.537	0.016	[0.506, 0.567]	0.488	0.017	[0.454, 0.521]
Item 6	0.480	0.004	[0.472, 0.488]	0.445	0.008	[0.428, 0.460]	0.578	0.006	[0.566, 0.590]	0.566	0.016	[0.533, 0.596]	0.502	0.018	[0.465, 0.537]
Item 7	0.452	0.004	[0.444, 0.460]	0.408	0.009	[0.391, 0.425]	0.554	0.007	[0.541, 0.567]	0.555	0.015	[0.524, 0.583]	0.510	0.018	[0.473, 0.543]
Item 8	0.430	0.005	[0.420, 0.440]	0.405	0.013	[0.379, 0.430]	0.514	0.008	[0.498, 0.530]	0.516	0.019	[0.476, 0.552]	0.465	0.019	[0.426, 0.501]
Item 9	0.484	0.007	[0.471, 0.496]	0.447	0.023	[0.400, 0.490]	0.534	0.011	[0.512, 0.556]	0.538	0.022	[0.493, 0.578]	0.478	0.028	[0.421, 0.530]
H	0.473	0.003	[0.467, 0.480]	0.430	0.007	[0.417, 0.443]	0.563	0.005	[0.553, 0.573]	0.561	0.012	[0.537, 0.585]	0.511	0.014	[0.484, 0.537]
α	0.852			0.791			0.883			0.876			0.869		
λ_2	0.858			0.807			0.892			0.883			0.876		
MS	0.861			0.818			0.897			0.886			0.885		
AISP	0.400			0.350			0.500			0.500			0.450		

Note. α = Cronbach's alpha; λ_2 = Guttman's lambda-2; MS = Molenaar–Sijtsma statistic; AISP = automated item selection procedure (the reported threshold is the threshold at which unidimensionality starts to crack); SE = standard error; 95% CI = 95% confidence interval. Item 1 = anhedonia; Item 2 = dysphoria; Item 3 = sleep alterations; Item 4 = fatigue/loss of energy; Item 5 = appetite alterations; Item 6 = feelings of worthlessness; Item 7 = cognitive impairment; Item 8 = psychomotor alterations; Item 9 = thoughts of self-harm. No violation of monotonicity was detected.

participants within each sample, a state of affairs that would be undesirable. With a value of 0.802, the ratio of between-sample scalability coefficients to within-sample scalability coefficients (BW) was deemed "excellent" according to Crisan et al. (2016) heuristic guidelines. The individual item and item pair coefficients were similar, with no obvious problems and no items or item pairs outside of the single-level Mokken scale guidelines, which are themselves higher than those for multilevel Mokken coefficients. To our knowledge, this study is the first to examine the psychometric properties of the PHQ-9 using multilevel Mokken scaling.

Our study has at least three limitations. First, we relied on cross-sectional data. Longitudinal data could have allowed us to examine, for instance, measurement invariance across time. We note that the measurement invariance across time of the PHQ-9 has proved highly satisfactory in past research (Stochl et al., 2020). In an effort to provide additional information on the instrument's temporal measurement invariance, however, we conducted an a posteriori, ancillary analysis based on five waves of PHQ-9 data collected in Israel. Detailed information about the analysis is available in Supplemental Material 5. As we added constraints, RMSEA never increased, CFI and TLI never decreased, and SRMR did not increase by more than 0.004. These results support the temporal measurement invariance of the PHQ-9, consistent with Stochl et al. (2020) findings. Second, although seven countries and five languages were represented in this study, an examination of additional countries and languages would have been an added advantage. Third, most of the study samples were samples of convenience and may not be representative of their populations of reference. However, Sample 4 was representative of the general U.S. population, and Sample 6 was representative of the general German population (see Supplemental Material 1). For the reader's information, analyses of each of the two representative samples are available in Supplemental Material 6. The results are highly consistent with those of our main analyses.

Our focus on nonclinical samples may be viewed as an additional limitation. Although an examination of how the PHQ-9 behaves among depressed patients (among individuals having received a formal diagnosis of clinical depression) could have been informative, we note that depressed patients are likely to populate the high end of the latent depression continuum and to exhibit a limited score range, that is, to cover only some of the response categories of a depression scale such as the PHQ-9. Such restrictions would have been problematic for our psychometric inquiry, leading us to concentrate on nonclinical samples. Our focus on nonclinical samples allowed us to examine respondents with various levels of depressive symptoms (from no symptoms to severe symptoms). We note that our samples, though best characterized as nonclinical, likely included respondents with clinically relevant levels of depressive symptoms.

Capitalizing on 29 samples and a total of 58,272 respondents, this study provides robust evidence that the PHQ-9 is "unidimensional enough" to allow researchers and practitioners to use the instrument based on its total score. Put differently, our findings suggest that, even when a degree of multidimensionality is identified in the PHQ-9, reliance on the scale's total score is likely to be warranted. All in all, our results consolidate the PHQ-9's status as a measure of reference in depression research and clinical practice. We recommend that investigators rely on ESEM bifactor analysis and its

related indices (e.g., ECV) to clarify the factorial structure of depression scales in the presence of (apparent) multidimensionality.

References

Adler, M., Hetta, J., Isacsson, G., & Brodin, U. (2012). An item response theory evaluation of three depression assessment instruments in a clinical sample. *BMC Medical Research Methodology*, *12*, Article 84. https://doi.org/10.1186/1471-2288-12-84

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.).

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.).

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.).

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

Bianchi, R., & Mirkovic, D. (2020). Is Machiavellianism associated with depression? A cluster-analytic study. *Personality and Individual Differences*, *152*, Article 109594. https://doi.org/10.1016/j.paid.2019.109594

Bianchi, R., Patthey, N., Mirkovic, D., Lemaitre, B., & Schlegel, K. (2020). Machiavellian males with high emotional intelligence exhibit fewer depressive symptoms. *Personality and Individual Differences*, *158*, Article 109867. https://doi.org/10.1016/j.paid.2020.109867

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

Chesney, E., Goodwin, G. M., & Fazel, S. (2014). Risks of all-cause and suicide mortality in mental disorders: A meta-review. *World Psychiatry*, *13*(2), 153–160. https://doi.org/10.1002/wps.20128

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. *Journal of Psychosomatic Research*, *75*(1), 60–64. https://doi.org/10.1016/j.jpsychores.2012.12.012

Crisan, D. R., van de Pol, J. E., & van der Ark, L. A. (2016). Scalability coefficients for two-level polytomous item scores: An introduction and an application. In L. van der Ark, D. Bolt, W. C. Wang, J. Douglas, & M. Wiberg (Eds.), *Quantitative psychology research. Springer proceedings in mathematics & statistics* (Vol. 167). Springer. https://doi.org/10.1007/978-3-319-38759-8_11

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), Article e0199235. https://doi.org/10.1371/journal.pone.0199235

Gliklich, R. E., Leavy, M. B., Cosgrove, L., Simon, G. E., Gaynes, B. N., Peterson, L. E., Olin, B., Cole, C., DePaulo, J. R., Jr., Wang, P., Crowe, C. M., Cusin, C., Nix, M., Berliner, E., & Trivedi, M. H. (2020). Harmonized outcome measures for use in depression patient registries and clinical practice. *Annals of Internal Medicine*, *172*(12), 803–809. https://doi.org/10.7326/M19-3818

Gotlib, I. H., & Hammen, C. L. (2014). *Handbook of depression* (3rd ed.). The Guilford Press.

Hasin, D. S., Sarvet, A. L., Meyers, J. L., Saha, T. D., Ruan, W. J., Stohl, M., & Grant, B. F. (2018). Epidemiology of adult *DSM-5* major depressive disorder and its specifiers in the United States. *JAMA Psychiatry*, *75*(4), 336–346. https://doi.org/10.1001/jamapsychiatry.2017.4602

Haslam, N., Holland, E., & Kuppens, P. (2012). Categories versus dimensions in personality and psychopathology: A quantitative review of taxometric research. *Psychological Medicine*, *42*(5), 903–920. https://doi.org/10.1017/S0033291711001966

Huang, F. Y., Chung, H., Kroenke, K., Delucchi, K. L., & Spitzer, R. L. (2006). Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. *Journal of General Internal Medicine*, *21*(6), 547–552. https://doi.org/10.1111/j.1525-1497.2006.00409.x

James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A., Abdollahpour, I., Abdulkader, R. S., Abebe, Z., Abera, S. F., Abil, O. Z., Abraha, H. N., Abu-Raddad, L. J., Abu-Rmeileh, N. M. E., Accrombessi, M. M. K., … the GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet*, *392*(10159), 1789–1858. https://doi.org/10.1016/S0140-6736(18)32279-7

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., Rush, A. J., Walters, E. E., Wang, P. S., & the National Comorbidity Survey Replication. (2003). The epidemiology of major depressive disorder: Results from the National Comorbidity Survey Replication (NCS-R). *Journal of the American Medical Association*, *289*(23), 3095–3105. https://doi.org/10.1001/jama.289.23.3095

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

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). The Guilford Press.

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., Brown, T. A., Carpenter, W. T., Caspi, A., Clark, L. A., Eaton, N. R., Forbes, M. K., Forbush, K. T., Goldberg, D., Hasin, D., Hyman, S. E., Ivanova, M. Y., Lynam, D. R., Markon, K., … Zimmerman, M. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology*, *126*(4), 454–477. https://doi.org/10.1037/abn0000258

Kroenke, K. (2021). PHQ-9: Global uptake of a depression scale. *World Psychiatry*, *20*(1), 135–136. https://doi.org/10.1002/wps.20821

Kroenke, K., Spitzer, R. L., Williams, J. B., & Löwe, B. (2010). The patient health questionnaire somatic, anxiety, and depressive symptom scales: A systematic review. *General Hospital Psychiatry*, *32*(4), 345–359. https://doi.org/10.1016/j.genhosppsych.2010.03.006

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Lamela, D., Soreira, C., Matos, P., & Morais, A. (2020). 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. *Journal of Affective Disorders*, *276*, 220–233. https://doi.org/10.1016/j.jad.2020.06.066

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

Levis, B., Benedetti, A., Thombs, B. D., & the DEPRESsion Screening Data (DEPRESSD) Collaboration. (2019). Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: Individual participant data meta-analysis. *BMJ (Clinical Research Ed.)*, *365*, Article l1476. https://doi.org/10.1136/bmj.l1476

Liu, R. T. (2016). Taxometric evidence of a dimensional latent structure for depression in an epidemiological sample of children and adolescents. *Psychological Medicine*, *46*(6), 1265–1275. https://doi.org/10.1017/S0033291715002792

Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*, 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

Martin-Subero, M., Kroenke, K., Diez-Quevedo, C., Rangil, T., de Antonio, M., Morillas, R. M., Loran, M. E., Mateu, C., Lupon, J., Planas, R., & Navarro, R. (2017). Depression as measured by PHQ-9 versus clinical diagnosis as an independent predictor of long-term mortality in a prospective cohort of medical inpatients. *Psychosomatic Medicine*, *79*(3), 273–282. https://doi.org/10.1097/PSY.0000000000000390

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. De Gruyter Mouton. https://doi.org/10.1515/9783110813203

Molenaar, I. W. (1982). Mokken scaling revisited. *Kwantitatieve Methoden*, *3*(8), 145–164.

Morin, A. J. S., Arens, A. K., & Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling*, *23*(1), 116–139. https://doi.org/10.1080/10705511.2014.961800

Muthén, L. K., & Muthén, B. O. (1998–2021). *Mplus user's guide* (8th ed.).

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*, *36*(9), 813–823. https://doi.org/10.1002/da.22940

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

Pryce, C. R., Azzinnari, D., Spinelli, S., Seifritz, E., Tegethoff, M., & Meinlschmidt, G. (2011). Helplessness: A systematic translational review of theory and evidence for its relevance to understanding and treating depression. *Pharmacology & Therapeutics*, *132*(3), 242–267. https://doi.org/10.1016/j.pharmthera.2011.06.006

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*(3), 223–237. https://doi.org/10.1080/00223891.2015.1089249

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150. https://doi.org/10.1037/met0000045

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

Sijtsma, K., & Emons, W. H. M. (2011). Advice on total-score reliability issues in psychosomatic measurement. *Journal of Psychosomatic Research*, *70*(6), 565–572. https://doi.org/10.1016/j.jpsychores.2010.11.002

Spitzer, R. L., Kroenke, K., & Williams, J. B. (1999). Validation and utility of a self-report version of PRIME-MD: The PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire *Journal of the American Medical Association*, *282*(18), 1737–1744. https://doi.org/10.1001/jama.282.18.1737

Stochl, J., Fried, E. I., Fritz, J., Croudace, T. J., Russo, D. A., Knight, C., Jones, P. B., & Perez, J. (2020). On dimensionality, measurement invariance, and suitability of sum scores for the PHQ-9 and the GAD-7. *Assessment*, Article 1073191120976863. Advance online publication. https://doi.org/10.1177/1073191120976863

Stochl, J., Jones, P. B., & Croudace, T. J. (2012). Mokken scale analysis of mental health and well-being questionnaire item responses: A non-parametric IRT method in empirical research for applied health researchers. *BMC Medical Research Methodology*, *12*(1), Article 74. https://doi.org/10.1186/1471-2288-12-74

Turgeman-Lupo, K., Toker, S., Ben-Avi, N., & Shenhar-Tsarfaty, S. (2020). The depressive price of being a sandwich-generation caregiver: Can organizations and managers help? *European Journal of Work and Organizational Psychology*, *29*(6), 862–879. https://doi.org/10.1080/1359432X.2020.1762574

van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, *48*(5), Article 27. https://doi.org/10.18637/jss.v048.i05

van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis*, *11*(2), 139–163. https://doi.org/10.1093/pan/mpg002

Weinberger, A. H., Gbedemah, M., Martinez, A. M., Nash, D., Galea, S., & Goodwin, R. D. (2018). Trends in depression prevalence in the USA from 2005 to 2015: Widening disparities in vulnerable groups. *Psychological Medicine*, *48*(8), 1308–1315. https://doi.org/10.1017/S0033291717002781

Wichers, M. (2014). The dynamic nature of depression: A new micro-level perspective of mental disorder that meets current challenges. *Psychological Medicine*, *44*(7), 1349–1360. https://doi.org/10.1017/S0033291713001979

Willner, P., Scheel-Krüger, J., & Belzung, C. (2013). The neurobiology of depression and antidepressant action. *Neuroscience and Biobehavioral Reviews*, *37*(10 Pt. 1), 2331–2371. https://doi.org/10.1016/j.neubiorev.2012.12.007

Received November 10, 2021
Revision received December 28, 2021
Accepted February 4, 2022 ■