Brazilian Journal of Psychiatry
bjp
Revista Brasileira de Psiquiatria

Braz J Psychiatry. 2024;46:e20233449
doi:10.47626/1516-4446-2023-3449
Brazilian Psychiatric Association

ORIGINAL ARTICLE

Patient Health Questionnaire-9: a clinimetric analysis

Fiammetta **Cosci**,[1,2,3] Kaj Sparle **Christensen**,[4] Sara **Ceccatelli**,[5] Chiara **Patierno**,[1] Danilo **Carrozzino**[6]

[1]*Department of Health Sciences, University of Florence, Florence, Italy.* [2]*International Lab of Clinical Measurements, University of Florence, Florence, Italy.* [3]*Department of Psychiatry & Neuropsychology, Maastricht University, Maastricht, The Netherlands.* [4]*Research Unit for General Practice and Section for General Medical Practice, Department of Public Health, Aarhus University, Aarhus, Denmark.* [5]*Department of Experimental and Clinical Medicine, University of Florence, Florence, Italy.* [6]*Department of Psychology Renzo Canestrari, University of Bologna, Bologna, Italy.*

Objective: The Patient Health Questionnaire-9 (PHQ-9) is a widely used self-report measure of depression whose psychometric properties were found to be merely acceptable. Insufficient attention has been devoted to its clinimetric validity, while its clinical utility is still debated, particularly for assessing depression severity. This is the first study to test the PHQ-9 construct validity and clinical utility based on clinimetric principles.
Methods: An online survey of 3,398 participants was conducted. Item response theory models (Rasch and Mokken analyses) were used to assess the PHQ-9 validity and determine its clinical utility.
Results: Fit to the Rasch model was achieved after adjusting the sample size. Items 2, 4, 6, and 9 over-discriminated, while items 1, 5, and 7 under-discriminated. Local dependency between items 2 and 6 was indicated. The PHQ-9 was not unidimensional. A Loevinger's coefficient of 0.49 was found, indicating an acceptable level of scalability.
Conclusion: The PHQ-9 is an instrument with potential clinical utility as an overall index of depression, mainly for screening purposes. Substantial revision, particularly in the wording of over- and under-discriminating items, is needed.

Keywords: PHQ-9; depression; patient-reported outcome measure; validity; Rasch analysis

Introduction

The scientific and clinical interest in patient-reported outcome measures (PROMs) has grown in recent decades.[1] PROMs are self-report tools assessing data directly from patients about how they function or feel in relation to a health condition or treatment.[2-4] Ideally, they were developed to assess symptom burden and disease severity, biopsychosocial functioning, quality of life, and well-being.[4] A number of studies have examined the measurement properties of these assessment instruments, although most have applied a classical psychometric approach rather than clinimetric principles.[5,6]

Clinimetrics was originally introduced by Alvan R. Feinstein in 1980s[7,8] as an innovative clinically-based evaluation method for a wide range of clinical issues not included in customary taxonomy.[9-14] Over the decades, this scientific discipline has provided instruments for the identification, classification, and temporal distinction of clinical phenomena. The CLIPROM criteria, which consider clinimetric criteria relevant for PROMs, are a recent contribution.[15] Unlike classical psychometrics, clinimetrics applies unidimensionality to the assessment of construct validity and uses it to: 1) evaluate whether each item of a rating scale covers unique clinical information, 2) test whether symptoms belong to an underlying syndrome, and 3) determine the extent to which the total score of a tool is a statistically sufficient measure of the severity of the investigated clinical condition.[10,15,16]

Several PROMs have been developed for depression, and their dimensionality has been largely documented from a psychometric point of view. Among them, the Patient Health Questionnaire-9 (PHQ-9)[17,18] is a widely used self-report measure of depression in general[19-21] and clinical populations (e.g., cardiology,[22] dermatology,[23] diabetology,[24] gastroenterology,[25] neurology,[26,27] oncology,[28] primary care,[29-31] rheumatology,[32] and psychiatry[33-35]). The psychometric properties of the PHQ-9 were considered good, particularly its internal consistency, test-retest reliability, and factorial validity.[17,36-39] However, evidence is lacking about its clinimetric validity. In recent years, there has also been a debate about the clinical utility of the PHQ-9. Some authors recommended it as a dimensional assessment of depression,[40,41] while others have reported measurement problems with the current version.[42,43] In particular, Christensen & Sparle-Christensen[42] found that the PHQ-9 was an overall misfit to the Rasch model, mainly because of its disordered item threshold (items 1, 3,

Correspondence: Fiammetta Cosci, Department of Health Sciences, University of Florence, Via di San Salvi 12, 50135, Florence, Italy.
E-mail: fiammetta.cosci@unifi.it
Submitted Oct 25 2023, accepted Dec 09 2023.

How to cite this article: Cosci F, Christensen KS, Ceccatelli S, Patierno C, Carrozzino D. Patient Health Questionnaire-9: a clinimetric analysis. Braz J Psychiatry. 2024;46:e20233449. http://doi.org/10.47626/1516-4446-2023-3449

4, 5, 6, and 9) and local dependency between items 2 and 6. Thus, such controversial issues should be clarified.

The present study proposes the first clinimetric analysis of the PHQ-9. Based on CLIPROM criteria,[15] its main aims were to examine its construct validity using both Rasch and Mokken analyses and to determine the PHQ-9 clinical utility.[10,12,14,15]

Methods

Participants

Students enrolled at the University of Florence during the 2021-2022 academic year with an active institutional e-mail address were eligible for participation. No exclusion criteria were applied. Participation was voluntary and uncompensated. The age distribution of the participants was as follows: 18-20 years, 344 (9.9%); 21-25 years, 1,060 (30.6%); 26-30 years, 310 (8.9%); $<$ 18 years, four (0.1%), and $>$ 30 years, 183 (5.3%). Age was not reported by 1,497 students, since this information was not made compulsory.

Procedure

An invitation with a link to the online survey was sent on May 13, 2022 to the institutional email addresses of University of Florence (Italy) students. The online survey was kept active until May 30, 2022. A total of 3,464 students participated, who were instructed about the research protocol (i.e., a brief description of the study, the involved investigators, the duration and content of the survey, the assessment method and time required as well as a guarantee of anonymity). The students provided informed consent to participate via an online form. Due to the need to collect data anonymously, no strategies were applied to limit duplicate responses, although the length of the survey and its limited period of availability likely discouraged this practice. PHQ-9 data were collected from 3,398 participants and analyzed in the present paper. This study, called Mental Health Literacy among students (MATTERS), was supported by the European University for Well-being consortium via the 2021 second call for Seed Funding.

Measures

The PHQ-9 is a self-report measure for diagnostic, monitoring, and screening purposes, as well as for assessing depression severity.[17,18] The instrument consists of two parts. The first section investigates nine depression symptoms listed in the DSM-IV: 1) lack of interest, 2) depressed mood, 3) sleeping difficulties, 4) tiredness, 5) appetite problems, 6) negative feelings about self, 7) concentration problems, 8) psychomotor agitation/retardation, and 9) suicidal ideation.[17,18] Respondents are asked to indicate how often they have been bothered by any of these symptoms in the last 2 weeks.[17,18] Each symptom-item is rated on a 4-point Likert scale ranging from 0 (Not at all) to 3 (Nearly every day). Because the second part of the scale assesses functional impairment caused by depression, it was not used in the present study and, thus, the total PHQ-9 score was based on the first section alone.[17,18] Total scores of 5, 10, 15, and 20 represent cutoff points for mild, moderate, moderately severe, and severe depression, respectively.[44] A cut-off of $\geqslant$ 10 points showed a sensitivity of 88% and a specificity of 88% for major depression.[17]

Statistical analyses

The Rasch analysis was conducted using Rasch Unidimensional Measurement Models (RUMM2030) software[45] to test the following clinimetric properties:

1. Overall fit to the model, which was evaluated using the chi-square item-trait interaction statistics.[46,47] Such statistics provided a summary measure of how the PHQ-9 conforms to Rasch model expectations.[48] A non-significant chi-square probability value indicated a good level of overall fit.[46,47]
2. Individual item and person fit: standardized fit residual values for items and participants were examined for any indication of misfit.
3. Dimensionality testing: principal component analysis of residuals was conducted to identify the two most different subsets of items (i.e., the most positively and negatively factor-loading items in the first component). Paired *t*-tests were then performed to compare scores on the item subsets. If more than 5% of *t*-tests were significant, the instrument would not be considered a unidimensional measure of depression.[48,49]
4. The Person separation reliability index was assessed to estimate the clinimetric sensitivity of the PHQ-9, i.e., its ability to discriminate among respondents with different levels of depression.[47,50]

Mokken analysis, a non-parametric version of item response theory models,[10,51] was performed to further assess the PHQ-9 dimensionality or scalability. Mokken analysis was conducted using Stata version 7. The Stata LoevH command was used to compute Loevinger's coefficients of homogeneity. According to Mokken,[51] Loevinger's coefficients of homogeneity[52] between 0.30 and 0.39 are considered acceptable, while a value $\geqslant$ 0.40 is a clear demonstration of a rating scale scalability.[10]

Ethics statement

This study was approved by the University of Florence ethics commission (no. 184, November 23, 2021).

Results

Overall and individual item fit to the Rasch model

Rasch model fit statistics are shown in Table 1. A significant item-trait interaction statistic (χ^2 = 356.00, degrees of freedom = 81, $p < 0.001$) was found, thus revealing an initial misfit to the Rasch model expectations (Table 1, Analysis 1). Misfit to the Rasch model remained after rescoring the disordered response categories (Table 1, Analysis 2). However, fit to the Rasch model

Table 1 Model fit statistics for Patient Health Questionnaire-9 (n=3,398)

Sample	Analysis	Model fit (overall)	Item fit residual, (mean [SD])	Person fit residual (mean [SD])	PSI	Dimensionality, significant *t* tests (95%CI)	Local dependency (residual correlation > 0.20 above average)	Differential item functioning:age group, degree
All items	1	$\chi^2(81) = 356.00$, $p < 0.001$	-0.41 (4.91)	-0.27 (0.95)	0.85	7.09 (6.36-7.83)	Items 2 and 6	Items 8 (age group) and 6 (degree)
Rescoring all items (0112)	2	$\chi^2(81) = 286.94$, $p < 0.001$	-1.56 (3.69)	-0.39 (1.04)	0.82	6.24 (5.51-6.97)	Items 2 and 6	None
Adjusted sample (n=500)	3	$\chi^2(81) = 42.22$, $p = 0.999$	-1.56 (3.69)	-0.39 (1.04)	0.82	6.24 (5.51-6.97)	Items 2 and 6	None

PSI = person separation index (with extremes).

was achieved after adjusting the sample size to 500 respondents (χ^2 = 42.22, degrees of freedom = 81, p = 0.999). The summary fit residuals for items and respondents were found to be within the acceptable limits of ± 2.5 (Table 1, Analyses 1-3). Table 2 shows the Rasch model fit statistics for individual PHQ-9 items. Items 2, 4, 6, and 9 over-discriminated, while items 1, 5, and 7 under-discriminated (Table 2).

Dimensionality and scalability

Significant *t*-tests outside the critical value of 5% were found for opposing residuals, indicating that the PHQ-9 was multidimensional (Table 1, Analyses 1-3). Mokken analysis showed that the total score had acceptable scalability, with Loevinger's coefficient of homogeneity being 0.49. As shown in Table 3, individual PHQ-9 items also showed acceptable scalability, with Loevinger's coefficients of homogeneity ranging from 0.44 to 0.56.

Local dependency

Local dependency was detected between items 2 (Feeling down, depressed or hopeless) and 6 (Feeling bad about yourself – or that you are a failure or have let yourself or your family down).

Differential item functioning

Item 8 (Moving or speaking so slowly that other people have noticed – or the opposite – being so fidgety or restless that you have been moving around a lot more than usual) showed significant differential item functioning for age. Significant differential item functioning for degree was observed in item 6.

Person separation reliability index

Person separation reliability indices ranged from 0.82 to 0.85 (Table 1, Analyses 1-3), indicating that the PHQ-9 could be reliably used to distinguish between different groups of individuals.

Discussion

The PHQ-9 was found to have potential clinical utility, despite requiring substantial improvement. Local dependency was observed, and several over- and under-discriminating items were found. Item 2 (Feeling down, depressed or hopeless) had the largest misfit to the Rasch model, probably because it is a compound question, which makes it difficult for respondents to provide precise answers. Because feeling depressed and feeling hopeless might be core symptoms of different clinical pictures,[53] they should be investigated in separate items.

Item 6 (Feeling bad about yourself – or that you are a failure or have let yourself or your family down) had the same over-discrimination problem. Once again, this is a compound question referring to two different clinical dimensions.[53-55] Such problematic over-discrimination

Table 2 Individual item fit statistics for Patient Health Questionnaire-9 items (n=3,398)

Item	Location	Fit residual	χ^2	Probability[†]
1. Little interest or pleasure in doing things	-0.687	4.859	11.929	0.217
2. Feeling down, depressed, or hopeless	-0.360	-7.723	135.448	0.000
3. Trouble falling or staying asleep, or sleeping too much	-0.481	2.458	7.984	0.536
4. Feeling tired or having little energy	-1.312	-4.287	51.839	0.000
5. Poor appetite or overeating	0.061	5.171	22.584	0.007
6. Feeling bad about yourself – or that you are a failure or have let yourself or your family down	-0.051	-5.697	54.322	0.000
7. Trouble concentrating on things, such as reading the newspaper or watching television	0.065	3.226	14.399	0.109
8. Moving 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	1.240	1.922	20.443	0.015
9. Thoughts that you would be better off dead or of hurting yourself in some way	1.526	-3.635	37.048	0.000

[†] Bonferroni-adjusted at 1%.

Table 3 Mokken analysis of the Patient Health Questionnaire-9

Item	Mean score	Loevinger's coefficient of scalability
1	1.45	0.44
2	1.36	0.56
3	1.47	0.48
4	1.76	0.54
5	1.17	0.45
6	1.22	0.53
7	1.16	0.46
8	0.55	0.44
9	0.39	0.52

might be the reason that the PHQ-9 overestimates depression severity, particularly when used as a dimensional measure.[56-58] Indeed, it has been observed that, compared to other scales of depression severity assessment, the PHQ-9 is overinclusive in identifying patients with severe depression and underinclusive in identifying those with mild depression.[58] The present findings consistently show that the PHQ-9 includes a mixture of over- and under-discriminating items. The challenge lies in establishing appropriate cut-off values, since setting them too low can result in numerous false positive diagnoses of depression, which might result in antidepressant overprescription.[59,60]

The present study also identified double- and triple-barreled questions,[61-63] i.e., questions incorporating different clinical variables in a single item. For instance, item 9 (Thoughts that you would be better off dead or of hurting yourself in some way) combines two separate clinical entities (i.e., suicidal ideation and self-harm thinking) into a single question. This could bias responses and become a source of misfit. A rewording of double- and triple-barreled items is needed[42,61,63,64] to improve the PHQ-9 clinical validity and clinimetric sensitivity in discriminating symptoms that might belong to different conditions (e.g., demoralization and depression) and distinguish between different levels of depression severity.[12,15,65] Thus, one solution for item 9, for example, would be to split it into two subitems (Thoughts that you would be better off dead and Thoughts of hurting yourself in some way). The local dependency between items 2 and 6 (in line with Christensen & Sparle-Christensen[42]) can be solved by reformulating them. Item 2 might be split into two subitems (Feeling down and depressed and Feeling hopeless), while item 6 needs simplification (e.g., Feeling guilty or blaming yourself).

Paired *t*-tests of opposing residuals indicated that the PHQ-9 is multidimensional, while Mokken analysis showed that the items and the total score have acceptable scalability, which is an acceptable level of unidimensionality. This confirms the instrument's conflicting construct validity results, which have been reported in the literature.[36,40-42,64,66-71] The PHQ-9 multidimensionality might be related to the fact that it was found to cover more than one dimension of depression severity. Each item measures a different depressive symptom (e.g., item 1 measures anhedonia, item 2 measures depressed mood, item 3 measures sleep problems, etc.), but all nine items measure the same underlying construct, depression, which explains the acceptable level of unidimensionality in the total score.

Therefore, future studies are needed, although it should be noted that the Person separation reliability indices were acceptable and that the PHQ-9 fit Rasch model expectations after adjustment for sample size. There is, thus, empirical justification for using the PHQ-9 as an overall index of depression, particularly when supplemented by other clinimetric indices, such as the Major Depression Inventory[50,72,73] and the 6-item version of the Hamilton Rating Scale for Depression,[74,75] which were found to be unidimensional. This is in line with the proposed use of the PHQ-9 as a screening, rather than a severity, measure.[66,68]

The present findings should be interpreted in light of some limitations. First, participants were recruited by convenience sampling, thus limiting generalizability of results. Future research using a sample of patients with

depression is needed. Second, due to the cross-sectional design, the PHQ-9 incremental and predictive validity could not be evaluated. Third, no other measures of depression were used. In future investigations, depression should be assessed with other PROMs to test the PHQ-9 clinical and concurrent validity.

The present clinimetric analysis suggests that the PHQ-9 can be used as an overall index of depression, mainly for screening purposes. Substantial revision, particularly in item wording, is needed to improve its construct validity and clinical utility.

Acknowledgements

This study was supported by the EUniWell (European University for Well-being) consortium via the 2021 second call of Seed Funding.

Disclosure

The authors report no conflicts of interest.

References

1 Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care. 2007;45: S22-31.
2 Basch E. Patient-reported outcomes – Harnessing patients' voices to improve clinical care. N Engl J Med. 2017;376:105-8.
3 Deshpande PR, Rajan S, Sudeepthi BL, Nazir CPA. Patient-reported outcomes: A new era in clinical research. Perspect Clin Res. 2011;2:137-44.
4 Kristensen S, Mainz J, Baandrup L, Bonde M, Videbech P, Holmskov J, et al. Conceptualizing patient-reported outcome measures for use within two Danish psychiatric clinical registries: Description of an iterative co-creation process between patients and healthcare professionals. Nord J Psychiatry. 2018;72:409-19.
5 Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD, et al. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. JAMA. 2013;309:814-22.
6 Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol. 2010;63: 1179-94.
7 Feinstein AR. T. Duckett Jones Memorial Lecture. The Jones criteria and the challenges of clinimetrics. Circulation. 1982;66:1-5.
8 Feinstein AR. Clinimetrics. New Haven: Yale University Press; 1987.
9 Bech P. Modern psychometrics in clinimetrics: impact on clinical trials of antidepressants. Psychother Psychosom. 2004;73:134-8.
10 Bech P. Clinical psychometrics. Oxford: Wiley Blackwell; 2012.
11 Carrozzino D, Christensen KS, Patierno C, Woźniewicz A, Møller SB, Arendt ITP, et al. Cross-cultural validity of the WHO-5 Well-Being Index and Euthymia Scale: A clinimetric analysis. J Affect Disord. 2022;311:276-83.
12 Cosci F. Clinimetric perspectives in clinical psychology and psychiatry. Psychother Psychosom. 2021;90:217-21.
13 Fava GA, Ruini C, Rafanelli C. Psychometric theory is an obstacle to the progress of clinical research. Psychother Psychosom. 2004;73: 145-8.
14 Fava GA, Tomba E, Sonino N. Clinimetrics: the science of clinical measurements. Int J Clin Pract. 2012;66:11-5.
15 Carrozzino D, Patierno C, Guidi J, Montiel CB, Cao J, Charlson ME, et al. Clinimetric criteria for patient-reported outcome measures. Psychother Psychosom. 2021;90:222-32.
16 Fava GA, Carrozzino D, Lindberg L, Tomba E. The clinimetric approach to psychological assessment: A tribute to Per Bech, MD (1942-2018). Psychother Psychosom. 2018;87:321-6.
17 Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16: 606-13.
18 Spitzer RL, Kroenke K, Williams JB. 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. JAMA. 1999;282:1737-44.
19 Martin A, Rief W, Klaiberg A, Braehler E. Validity of the brief patient health questionnaire mood scale (PHQ-9) in the general population. Gen. Hosp. Psychiatry. 2006;28:71-7.
20 Schomerus G, Matschinger H, Angermeyer MC. Attitudes that determine willingness to seek psychiatric help for depression: a representative population survey applying the Theory of Planned Behaviour. Psychol Med. 2009;39:1855-65.
21 Wang SY, Singh K, Lin SC. Prevalence and predictors of depression among participants with glaucoma in a nationally representative population sample. Am J Ophthalmol. 2012;154:436-44.
22 Stafford L, Berk M, Jackson HJ. Validity of the Hospital Anxiety and Depression Scale and Patient Health Questionnaire-9 to screen for depression in patients with coronary artery disease. Gen Hosp Psychiatry. 2007;29:417-24.
23 Picardi A, Amerio P, Baliva G, Barbieri C, Teofoli P, Bolli S, et al. Recognition of depressive and anxiety disorders in dermatological outpatients. Acta Derm Venereol. 2004;84:213-7.
24 Khamseh ME, Baradaran HR, Javanbakht A, Mirghorbani M, Yadollahi Z, Malek M. Comparison of the CES-D and PHQ-9 depression scales in people with type 2 diabetes in Tehran, Iran. BMC Psychiatry. 2011;16:11-61.
25 Persoons P, Vermeire S, Demyttenaere K, Fischler B, Vandenberghe J, van Oudenhove L, et al. The impact of major depressive disorder on the short- and long-term outcome of Crohn's disease treatment with infliximab. Aliment Pharmacol Ther. 2005;22:101-10.
26 Rathore JS, Jehi LE, Fan Y, Patel SI, Foldvary-Schaefer N, Ramirez MJ, et al. Validation of the Patient Health Questionnaire-9 (PHQ-9) for depression screening in adults with epilepsy. Epilepsy Behav. 2014;37:215-20.
27 Williams LS, Jones WJ, Shen J, Robinson RL, Kroenke K. Outcomes of newly referred neurology outpatients with depression and pain. Neurology. 2004;63:674-7.
28 Ell K, Xie B, Quon B, Quinn DI, Dwight-Johnson M, Lee PJ. Randomized controlled trial of collaborative care management of depression among low-income patients with cancer. J Clin Oncol. 2008;26:4488-96.
29 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;58:32-6.
30 Mitchell AJ, Yadegarfar M, Gill J, Stubbs B. Case finding and screening clinical utility of the Patient Health Questionnaire (PHQ-9 and PHQ-2) for depression in primary care: a diagnostic meta-analysis of 40 studies. BJPsych Open. 2016;2:127-38.
31 Sherina MS, Arroll B, Goodyear-Smith F. Criterion validity of the PHQ-9 (Malay version) in a primary care clinic in Malaysia. Med J Malaysia. 2012;67:309-15.
32 Löwe B, Willand L, Eich W, Zipfel S, Ho AD, Herzog W, et al. Psychiatric comorbidity and work disability in patients with inflammatory rheumatic diseases. Psychosom Med. 2004;66:395-402.
33 Beard C, Hsu KJ, Rifkin LS, Busch AB, Björgvinsson T. Validation of the PHQ-9 in a psychiatric sample. J Affect Disord. 2016;193:267-73.
34 Inoue T, Tanaka T, Nakagawa S, Nakato Y, Kameyama R, Boku S, et al. Utility and limitations of PHQ-9 in a clinic specializing in psychiatric care. BMC Psychiatry. 2012;12:73.
35 Löwe B, Gräfe K, Kroenke K, Zipfel S, Quenter A, Wild B, et al. Predictors of psychiatric comorbidity in medical outpatients. Psychosom Med. 2003;65:764-70.
36 Boothroyd L, Dagnan D, Muncer S. PHQ-9: One factor or two? Psychiatry Res. 2019;271:532-4.
37 Gelaye B, Williams MA, Lemma S, Deyessa N, Bahretibeb Y, Shibre T, et al. Validity of the Patient Health Questionnaire-9 for depression screening and diagnosis in East Africa. Psychiatry Res. 2013;210: 653-61.
38 He C, Levis B, Riehm KE, Saadat N, Levis AW, Azar M, et al. The accuracy of the Patient Health Questionnaire-9 algorithm for screening to detect major depression: An individual participant data meta-analysis. Psychother Psychosom. 2020;89:25-37.

39 Ryan TA, Bailey A, Fearon P, King J. Factorial invariance of the Patient Health Questionnaire and Generalized Anxiety Disorder Questionnaire. Br J Clin Psychol. 2013;52:438-49.
40 Bianchi R, Verkuilen J, Toker S, Schonfeld IS, Gerber M, Brähler E, et al. Is the PHQ-9 a unidimensional measure of depression? A 58,272-participant study. Psychol Assess. 2022;34:595-603.
41 Stochl J, Fried EI, Fritz J, Croudace TJ, Russo DA, Knight C, et al. On Dimensionality, Measurement Invariance, and Suitability of Sum Scores for the PHQ-9 and the GAD-7. Assessment. 2022;29:355-66.
42 Christensen KS, Sparle-Christensen M. Comparing the construct validity of the Patient Health Questionnaire (PHQ-9) and the Major Depression Inventory (MDI) using Rasch analysis. J Affect Disord. 2023;333:44-50.
43 Ford J, Thomas F, Byng R, McCabe R. Use of the Patient Health Questionnaire (PHQ-9) in Practice: Interactions between patients and physicians. Qual Health Res. 2020;30:2146-59.
44 Gilbody S, Richards D, Brealey S, Hewitt C. Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. J Gen Intern Med. 2007;22:1596-1602.
45 Andrich D, Lyne A, Sheridan B, Luo G. RUMM 2030. Perth: RUMM Laboratory; 2010.
46 Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). Br J Clin Psychol. 2007;46:1-18.
47 Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum. 2007;57:1358-62.
48 Nielsen MG, Ørnbøl E, Vestergaard M, Bech P, Christensen KS. The construct validity of the Major Depression Inventory: A Rasch analysis of a self-rating scale in primary care. J Psychosom Res. 2017;97:70-81.
49 Christensen KS, Oernboel E, Nielsen MG, Bech P. Diagnosing depression in primary care: a Rasch analysis of the Major Depression Inventory. Scand J Prim Health Care. 2019;37:256-63.
50 Carrozzino D, Christensen KS, Cosci F. Construct and criterion validity of patient-reported outcomes (PROs) for depression: A clinimetric comparison. J Affect Disord. 2021;283:30-5.
51 Mokken RJ. A theory and procedure of scale analysis. The Netherlands: Mouton; 1971.
52 Loevinger J. A systematic approach to the construction and evaluation of tests of ability. Psychol Monogr. 1947;61:i-49.
53 Woźniewicz A, Cosci F. Clinical utility of demoralization: A systematic review of the literature. Clin Psychol Rev. 2023;99:102227.
54 De Figueiredo JM, Frank JD. Subjective incompetence, the clinical hallmark of demoralization. Compr Psychiatry. 1982;23:353-63.
55 Fava GA, Guidi J. Clinical characterization of demoralization. Psychother Psychosom. 2023;92:139-47.
56 Bech P, Christensen EM, Vinberg M, Østergaard SD, Martiny K, Kessing LV. The performance of the revised Major Depression Inventory for self-reported severity of depression--implications for the DSM-5 and ICD-11. Psychother Psychosom. 2013;82:187-8.
57 Cameron IM, Cardy A, Crawford JR, du Toit SW, Hay S, Lawton K, et al. Measuring depression severity in general practice: discriminatory performance of the PHQ-9, HADS-D, and BDI-II. Br J Gen Pract. 2011;61:e419-26.
58 Zimmerman M. Symptom severity and guideline-based treatment recommendations for depressed patients: implications of DSM-5's potential recommendation of the PHQ-9 as the measure of choice for depression severity. Psychother Psychosom. 2012;81:329-32.
59 Cosci F, Guidi J, Tomba E, Fava GA. The emerging role of clinical pharmacopsychology. Clin Psychol Eur. 2019;1:32158.
60 Fava GA, Cosci F. Understanding and managing withdrawal syndromes after discontinuation of antidepressant drugs. J Clin Psychiatry. 2019;80:19com12794.
61 Gothwal VK, Bagga DK, Bharani S, Sumalini R, Reddy SP. The patient health questionnaire-9: validation among patients with glaucoma. PLoS One. 2014;9:e101295.
62 Smith AB, Rush R, Fallowfield LJ, Velikova G, Sharpe M. Rasch fit statistics and sample size considerations for polytomous data. BMC Med Res Methodol. 2008;8:33.
63 Williams RT, Heinemann AW, Bode RK, Wilson CS, Fann JR, Tate DG. Improving measurement properties of the Patient Health Questionnaire-9 with rating scale analysis. Rehabil Psychol. 2009;54:198-203.
64 Christensen KS, Oernboel E, Zatzick D, Russo J. Screening for depression: Rasch analysis of the structural validity of the PHQ-9 in acutely injured trauma survivors. J Psychosom Res. 2017;97:18-22.
65 Fava GA. Forty Years of clinimetrics. Psychother Psychosom. 2022;91:1-7.
66 Barthel D, Barkmann C, Ehrhardt S, Schoppen S, Bindt C, International CDS Study Group. Screening for depression in pregnant women from Côte d'Ivoire and Ghana: Psychometric properties of the Patient Health Questionnaire-9. J Affect Disord. 2015;187:232-40.
67 Elhai JD, Contractor AA, Tamburrino M, Fine TH, Prescott MR, Shirley E, et al. The factor structure of major depression symptoms: a test of four competing models using the Patient Health Questionnaire-9. Psychiatry Res. 2012;199:169-73.
68 Forkmann T, Gauggel S, Spangenberg L, Brähler E, Glaesmer H. Dimensional assessment of depressive severity in the elderly general population: psychometric evaluation of the PHQ-9 using Rasch Analysis. J Affect Disord. 2013;148:323-30.
69 González-Blanch C, Medrano LA, Muñoz-Navarro R, Ruíz-Rodríguez P, Moriana JA, Limonero JT, et al. Factor structure and measurement invariance across various demographic groups and over time for the PHQ-9 in primary care patients in Spain. PLoS One. 2018;13:e0193356.
70 Guo B, Kaylor-Hughes C, Garland A, Nixon N, Sweeney T, Simpson S, et al. Factor structure and longitudinal measurement invariance of PHQ-9 for specialist mental health care patients with persistent major depressive disorder: Exploratory Structural Equation Modelling. J Affect Disord. 2017;219:1-8.
71 Lamoureux EL, Tee HW, Pesudovs K, Pallant JF, Keeffe JE, Rees G. Can clinicians use the PHQ-9 to assess depression in people with vision loss? Optom Vis Sci. 2009;86:139-45.
72 Bech P, Rasmussen NA, Olsen LR, Noerholm V, Abildgaard W. The sensitivity and specificity of the Major Depression Inventory, using the Present State Examination as the index of diagnostic validity. J Affect Disord. 2001;66:159-64.
73 Olsen LR, Jensen DV, Noerholm V, Martiny K, Bech P. The internal and external validity of the Major Depression Inventory in measuring severity of depressive states. Psychol Med. 2003;33:351-6.
74 Carrozzino D, Patierno C, Fava GA, Guidi J. The Hamilton Rating Scales for Depression: A critical review of clinimetric properties of different versions. Psychother Psychosom. 2020;89:133-50.
75 Timmerby N, Andersen JH, Søndergaard S, Østergaard SD, Bech P. A systematic review of the clinimetric properties of the 6-item version of the Hamilton Depression Rating Scale (HAM-D6). Psychother Psychosom. 2017;86:141-9.