MEDICINA CLÍNICA Y SOCIAL

Med. clín. soc. 2025;9:e521. e-ISSN: 2521-2281

https://doi.org/10.52379/mcs.v9.521

ORIGINAL

Construct validity of the PHQ-9 in university students in Colombia: A Rasch analysis approach

Validez constructo del PHQ-9 en estudiantes universitarios de Colombia: Análisis tipo Rasch

Carlos Arturo Cassiani-Miranda[1], Orlando Scoppetta[2], María Alejandra Barrios-Villadiego[3]
Andrés Felipe Tirado-Otálvaro[4], Andrea Carolina Duran-Bedoya[4]

[1] Universidad de Santander, Faculty of Health Sciences, Bucaramanga, Colombia
[2] Fundación Nuevos Rumbos, Bogotá, Colombia
[3] Universidad de Cartagena, Faculty of Health Sciences, Cartagena, Colombia
[4] Universidad Pontificia Bolivariana, Medellín, Colombia

Received: 26/09/2024
Accepted: 17/01/2025
Published: 27/02/2025

Corresponding author
Andrés Felipe Tirado-Otálvaro PhD
Universidad Pontificia Bolivariana, Colombia
felipe.tirado@upb.edu.co

Responsable Editor
Iván Barrios, PhD
Universidad Nacional de Asunción
San Lorenzo, Paraguay

Conflicts of interests
The authors declare that there is no conflict of interest.

Funding
The authors received no financial support for the research, authorship, and/or publication of this article.

ABSTRACT

Introduction: The Patient Health Questionnaire 9 (PHQ-9) is one of the most widely used screening instruments for major depressive episodes. However, there are no published studies on Rasch-type analysis of the PHQ-9 among Spanish-speaking university students. **Objective:** To evaluate the psychometric properties of the PHQ-9 in university students using Rasch-type models and to assess possible biases of the items according to gender. **Methods**: This cross-sectional observational study evaluated the psychometric performance of the PHQ-9 in health sciences students at the University of Cartagena (Colombia). A random sampling stratified by academic program, semester, and sex was used, obtaining a sample of 550 participants (9 excluded for incomplete responses). Participants signed an informed consent, and the study was approved by an ethics committee. Rasch analysis was used to assess model fit, differential item functioning, dimensionality, local independence, and reliability. Adequate internal consistency (α=0.83, ω=0.89) and factorial validity were found. **Results:** A cross-sectional study was conducted with 550 health science students from Colombian university. The data were analyzed using a Rasch model, in which the following psychometric characteristics were verified: 1) differential item functioning, 2) dimensionality and local independence, and 3) overall fit. Only item 2 showed a tendency toward differential functioning. **Conclusions:** One-dimensionality and local independence of the items, moderate reliability, and good general fit were found, although there was a gap between the degree of depression measured by the PHQ-9 and the participants' responses. The Spanish version of the PHQ-9 for Colombian university students showed adequate item-level psychometric properties for screening for major depressive episodes.

Keywords: Depression; Psychometrics; Patient Health Questionnaire; Reproducibility of Results.

RESUMEN

Introducción: El Cuestionario de Salud del Paciente 9 (PHQ-9) es uno de los instrumentos de detección más utilizados para episodios depresivos mayores. Sin embargo, no existen estudios publicados sobre el análisis tipo Rasch del PHQ-9 entre estudiantes universitarios de habla hispana. **Objetivo:** Evaluar las propiedades psicométricas del PHQ-9 en estudiantes universitarios utilizando modelos tipo Rasch y analizar posibles sesgos en los ítems según el género. **Metodología**: Este estudio observacional transversal evaluó el rendimiento psicométrico del PHQ-9 en estudiantes de ciencias de la salud de la Universidad de Cartagena (Colombia). Se utilizó un muestreo aleatorio estratificado por programa académico, semestre y sexo, obteniendo una muestra de 550 participantes (9 excluidos por respuestas incompletas). Los participantes firmaron un consentimiento informado, y el estudio fue aprobado por un comité de ética. Se empleó el análisis de Rasch para evaluar el ajuste del modelo, funcionamiento diferencial de los ítems, dimensionalidad, independencia local y confiabilidad. Se encontró una adecuada consistencia interna (α=0.83, ω=0.89) y validez factorial. **Resultados:** Se realizó un estudio transversal con 550 estudiantes de ciencias de la salud de una universidad colombiana. Los datos fueron analizados mediante un modelo Rasch, en el que se verificaron las siguientes características psicométricas: 1) funcionamiento diferencial de los ítems, 2) dimensionalidad e independencia local, y 3) ajuste general. Solo el ítem 2 mostró una tendencia hacia el funcionamiento diferencial. **Conclusiones:** Se encontró unidimensionalidad e independencia local de los ítems, fiabilidad moderada y buen ajuste general, aunque hubo una discrepancia entre el grado de depresión medido por el PHQ-9 y las respuestas de los participantes. La versión en español del PHQ-9 para estudiantes universitarios colombianos mostró propiedades psicométricas adecuadas a nivel de ítems para la detección de episodios depresivos mayores.

Palabras clave: Depresión; Psicometría; Cuestionario de Salud del Paciente; Reproducibilidad de los resultados.

How to cite thi article: Cassiani-Miranda CA, Scoppetta O, Barrios-Villadiego MA, Tirado-Otálvaro AF, Duran-Bedoya AC. Construct validity of the PHQ-9 in university students in Colombia: A Rasch analysis approach. Med. clín. soc. 2025;9:e521.

INTRODUCTION

The Patient Health Questionnaire 9 (PHQ-9) is a brief, easy-to-administer, and interprets a major depressive episode (MDE) screening instrument (1). Because of its brevity and good reliability and validity (1,2), the PHQ-9 is one of the most widely used MDE screening instruments in clinical and non-clinical settings (3,4). Two recent systematic reviews have concluded that the PHQ-9 is the most reliable and accurate tool for MDE screening, so it is enjoying increasing acceptance in the scientific community every day (5-8). Overall, the PHQ-9 has demonstrated adequate psychometric properties in both general and medical populations (1,9,10), including university populations (11-16).

The psychometric performance of the PHQ-9 has generally been evaluated using the classical test theory (CTT) framework, which focuses on the concepts of true score and measurement error (17). From the CTT perspective, the comparability of test scores has the limitation that scores are "test dependent" (18), implying that a low score may be due to a person's low ability or a difficult test (19,20). Based on an analysis of the articles included in the systematic reviews on the PHQ-9 published to date, approximately 800 articles evaluated the psychometric properties of the PHQ-9 using the CTT model (4,5,8,21-29). CTT-based techniques could provide inaccurate diagnoses and do not consider the heterogeneity in each specific item (30, 31). Currently, these CCT-based methods are being complemented and, in some cases, replaced by item response theory (IRT) approaches, considered by many researchers to be the "measurement paradigm of the 21st century" (19, 32-35). IRT-based models offer several advantages over CTT and are considered the most appropriate and robust methods for assessing the psychometric properties of screening scales such as the PHQ-9 (36,37). The idea of IRT is that test-response behavior (i.e., solving an item or choosing a specific category) can be explained by underlying person parameters (latent ability or trait) and item parameters (difficulty) (34).

The relationship between the latent trait and probability of response to a given item can be expressed by different estimation models for both dichotomous and polytomous items (32,38). This approach estimates item and person parameters from the obtained data and is considered independent of the sample and test (39). In addition, IRT-based models provide a richer description of the performance of each item, greater detail on the accuracy of a measure, and when assumptions are met, scores are item-independent and invariant across different samples (33,34,40,41). Consequently, the use of IRT models may increase the validity and utility of depression screening when the PHQ-9 is used in culturally diverse settings (42,43). To date, only 30 published studies have used IRT models to assess the psychometric properties of the PHQ-9 and only one in college students (37,42-68).

Given the high prevalence of depression among college students (69,70) it is important to evaluate the validity and utility of the PHQ-9 in this population. The appropriate detection and management of depression in health science students could translate into significant potential benefits if the psychometric properties of screening instruments such as the PHQ-9 are optimized (71,72). Therefore, this study aimed to evaluate the psychometrics of the Colombian version of the PHQ-9 adapted for university students using IRT models and to assess possible item biases as a function of gender.

METHODOLOGY

Population and sample

This cross-sectional observational study evaluated the psychometric performance of the PHQ-9. Stratified random sampling was applied according to academic program, semester, and sex, where each stratum had several subjects proportional to its size. The participants were health science students from the University of Cartagena (Colombia). A non-probabilistic sample size of 550 participants was considered good for confirmatory factor analysis (73). Of these 550 subjects, 9 were excluded because of a significant number of failed responses.

Procedures and instruments

All the participants were informed of the research objectives and signed an informed consent form. This project was approved by the Ethics Committee of a university in Colombia. A sociodemographic characteristics instrument was applied to all participants, and all of them completed the PHQ-9 (1). The PHQ-9 is a screening scale that measures the presence and severity of depressive symptoms (1) and consists of nine symptoms of the DSM-IV MDE criterion A (74). These nine items are arranged in the form of an adjectival-type scale assessing the presence of the symptom in the last two weeks ("not at all," "several days," "more than half of the days" and "almost every day"), which are scored from 0 to 3, reaching a score between 0 and 27 (75). It can be self- or hetero-administered and is used either algorithmically to make a probable diagnosis of an MDE or as a continuous measure of scores ranging from 0 to 27 and cut-off points (CP) of 5, 10, 15, and 20, representing levels of depressive symptoms as mild, moderate, moderately severe, and severe (1). These scores can also be used dichotomously from a PC to classify subjects with or without clinically significant depressive symptoms (CSDS) (76).

The psychometric characteristics of the PHQ-9, according to Kroenke et al., present a sensitivity of 88% and a specificity of 88%, an adequate internal consistency (Cronbach's α of 0.86-0.89), a test-retest score of 0.84, a concordance between the self-administered test and the one performed by the evaluator of 84%, and an area under the curve (AUC) of 0.95 (1). In this study, a version adapted to this population was used (77). Preliminary analyses showed that the PHQ-9 in Colombian university students had adequate internal consistency with a Cronbach's alpha of 0.83 and McDonald's omega of 0.89 (76); a two-factor model with adequate fit indicators: CFI, 0.98; NFI, 0.96; NNFI (TLI), 0.97; RMSEA, 0.045; and a metric, structural, and residual invariance by sex with a ΔCFI of 0.002, 0.001, and 0.003 (p = 0.67, 0.27, and 0.78, respectively) (11).

In Colombia, the criterion validity of the PHQ-9 in primary care has also been evaluated compared with the mini international neuropsychiatric interview (MINI), which found that for a PC of 7 or more, the PHQ-9 showed the following indicators: AUC of 0.92 (95% CI, 0.880-0.963), sensitivity of 90.38 (95% CI: 81.41-99.36); specificity of 81.68 (95% CI: 75.93-87.42) (78).

Statistical Analysis: The Rasch analysis is based on a mathematical model in which the probability of passing an item is a logistic function of the difference between the person's level of depression and the level of depression expressed by the item (item difficulty) (53,79,80). For the Rasch analysis, the following psychometric characteristics were verified.

Differential item functioning: For the analysis of differential item functioning, the following criterion was considered: the value of the difference between males and females was greater than 0.5, and the probability of the Mantel statistic (for polytomous tests) was less than 0.05 (40); thus, items whose score was due to gender and not to the construct itself could be identified.

Dimensionality and local independence: Regarding the assumptions of dimensionality and local independence, the principal component analysis allowed us to establish how much of the variability in the responses was explained by the instrument (41). Here, we used the indicator of the number of eigenvalues of the non-exponential variance (41). Here, we used the indicator of the number of eigenvalues of the unexplained variance in the first contrast, which must be less than two (58,71,81). Local dependence occurs when the response to one item depends on the response to another item (67). Residual correlations above 0.3 are indicative of local dependence (82).

General fit: The general fit was studied from the joint analysis of several measures: in the first instance, the close fit (infit) and far fit (outfit). For this type of analysis, the items are expected to have a measure of fit close to one. A result between 0.5 and 1.5 indicates a measure with an appropriate balance between information and noise (83). Difficulty, measured in logits, refers to the probability that a person will respond to an item indicating a sign of depression (34). Concerning the discrimination index, each item is expected to have a positive, nonzero value of at least 0.3 and not much above one (33). Correlation refers to the association between an item and a scale. Positive correlations were expected with values of at least 0.2. Rasch’s reliability is analogous to the internal consistency indicators in the CTT. Similarly, the person separation index (PSI) was estimated, establishing acceptable values above 0.7 (52), although this is relative to the sample size and the number of items. These provisions jointly show the adequacy of the data to a model that allows a consistent measurement of the construct and are the recommendations for the use of the Rasch model to establish the psychometric characteristics of instruments (81,84,85). Winsteps® version 3.80.1 was used for this analysis.

RESULTS

Differential Item Functioning

First, a differential behavioral analysis of the items was performed. Item 2, "felt discouraged, sad, irritable or hopeless" showed a tendency toward differential behavior. This difference is close to the established limit and is because female participants tended to report experiencing the condition referred to in the item more (mean of 0.9 and for males 0.6) (Table 1).

TABLE 1. DIFFERENTIAL ITEM FUNCTIONING.

ITEM	SIZE OF THE DIFFERENCES	MANTEL (PROBABILITY)
Disinterest	0.03	0.89
Discouragement, sadness	-0.46	0.00
Trouble sleeping	0.15	0.26
Tiredness	0.13	0.33
Problems with appetite	-0.11	0.27
Feeling bad about oneself	0.00	0.62
Difficulty in concentrating	0.12	0.39
Alteration of movement	0.19	0.11
Ideation of death or harm	-0.18	0.29

Dimensionality and Local Independence

The total variance explained by these measures was 46.8%. The residual variance in the first contrast was 1.7, which is below the undesirable level of 2.0 (81). Similarly, the percentage of variance explained by the items was 22.8%, which was higher than the variance not explained by the first contrast. The residual standardized correlations had a negative sign, which is compatible with the evidence of the local independence of the items.

General Fit

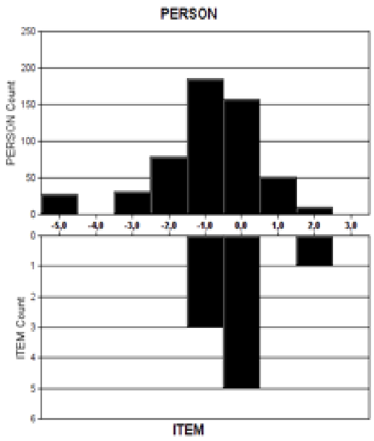
All items fell within the range in which they are considered productive (between 0.5 and 1.5 in the infit - outfit measures). The correlations between item scores and the test were compatible with good test performance. There were no negative or very low correlations (Table 2).

TABLE 2. MEASURES OF GENERAL FIT OF THE PHQ-9

ITEM	DIFFICULTY	MEASUREMENT ERROR	INFIT	OUTFIT	DISCRIMINATION	CORRELATION
Disinterest	-0.1	0.1	1.12	1.18	0.90	0.63
Discouragement, sadness	-0.1	0.1	0.66	0.69	1.33	0.63
Sleeping problems	-0.7	0.1	1.23	1.22	0.75	0.67
Tiredness	-0.8	0.1	0.86	0.85	1.21	0.68
Appetite problems	-0.8	0.1	1.29	1.29	0.65	0.67
Feeling bad about oneself	0.1	0.1	0.90	0.84	1.12	0.62
Difficulty concentrating	-0.1	0.1	0.95	0.98	1.02	0.63
Movement disturbance	0.5	0.1	0.90	0.86	1.08	0.59
Ideation of death or harm	2.1	0.1	1.16	0.82	0.98	0.43

Person separation and reliability were 1.75 and 0.75. Overall, these data indicate that the instrument allows for discrimination between the two levels. Person separation and reliability were 1.75 and 0.75. Taken together, these data indicate that the instrument allows for discrimination between the two levels. The discrimination index of the items is good, but not negative in all cases (above 0.30), although there is a tendency for most of the items to over-discriminate persons at high and low risk of depression, although in no case, the dataset indicates a possible unproductive item. The analysis of person-item fit revealed a mismatch between the degree of depression measured by the PHQ-9 in this group and the participants' responses (Figure 1).

FIGURE 1. HISTOGRAM AND WRIGHT'S MAP WITH THE PERSON-ITEM RELATIONSHIP



```
MEASURE    PERSON - MAP - ITEM
                  <more>|<rare>
    2                .#  +
                         |
                         |
                     .#  |  Death ideation
                        T|
                      .  |
                     .#  |T
                         |
    1                 .  +
                     ##  |
                    ###  |
                   .###  |S
                     .#  |
                .###### S|
                    ###  | Alteration of movement
                 .#####  |
    0             #####  +M Disinterest  - Feeling bad about oneself
               .#######  |  Discouragement, sadness  - Difficulty to  concentrate
                .######  |
                 ######  |
              .########  |  Trouble sleeping   -  Tiredness
                         |S Problems with appetite
              .######## M|
              #########  |
   -1                    +
              .########  |
                         |T
                         |
           .###########  |
                         |
                         |
           ############ S|
   -2                    +
                         |
                         |
               .#######  |
                         |
                         |
                         |
                         |
   -3                   T+
                         |
               .#######  |
                         |
                         |
                         |
                         |
                         |
   -4           #######  +
                  <less>|<frequent>
 EACH "#" IS 4: EACH "."  IS 1 TO 3
```

Note: On the right is the person-item histogram, and on the left is Wright's map. The map shows the locations of items above those with the highest response frequencies. Each "#" represents four persons, and each ".", one to three persons.

DISCUSSION

Evidence of one-dimensionality, local independence, adequate overall fit for the Rasch model, and moderate reliability was found for the Colombian Spanish version of the PHQ-9 among university students. These findings provide complementary information about its psychometric characteristics that reinforce the arguments in favor of its validity, but at the same time identify areas that require further research and, eventually, the development of better tests.

First, it is important to analyze the results of differential item functioning. The lack of differential item functioning (DIF) along with the Rasch model results supports the validity of an instrument across diverse cultural contexts with different characteristics when assessing depressive symptoms (40,63,68,85). DIF analysis showed that PHQ-9 responses were sex invariant in this sample of college students. These results are comparable to those reported for a sample of university students in Japan (65). Although our work item 2 of the PHQ-9 showed some tendency toward differential behavior, a previous study in Colombia from the perspective of CTT did not find invariance by gender (11). In addition, it should be considered that in this analysis, a more restrictive criterion was used (0.5) than in other studies on the PHQ-9, such as that of Jiraniramai (37) in which the decision was taken from 0.64 logists. Differential functioning findings in other research are mixed: in Thailand, in a population of healthcare workers, slight differential functioning was found for item 6 (feeling bad about oneself) according to the sex of the participants (37); in research in India with people with visual impairment (54), it was found that there was differential functioning of item 3, referring to problems with sleep when dividing the groups according to the duration of their disability; in Germany, in a sample of older adults, differential functioning of items 3 and 6 was found (53). In summary, the differential behavior attributable to the PHQ-9 items depends on the specificities of the populations studied.

When analyzing the dimensionality of the construct in our sample, it is important to consider that the variance explained by the measures was below 50% and that the residual variance was 1.7, not quite 2, but not so far from that number. This evidence could prompt further study on whether the set of items measured the same construct. This interpretation has a place given the flexibility of this type of analysis, which does not focus on one indicator but invites consideration of the data within a larger set of information. Additionally, it is necessary to consider that all psychometric analyses should be analyzed concerning contextual information: in this case, it is known that, although the test works well as a whole, there is also a somatic and a non-somatic dimension within it (11) and although one could delve from the above into the question of whether the instrument serves to assess the degree of depression or not, the reality is that the construct "depression" itself requires including behavioral, somatic and cognitive issues of a different nature, but which are expressions of that same entity (86,87). This statement is consistent with data from theory and practice, suggesting that depressive syndrome is a multidimensional construct (88).

The values of the standardized residual correlations in our study indicated that the PHQ-9 items showed local independence. This finding is compatible with those of studies on pregnant women in Peru (68) and the rural population in Brazil (45).

The reliability and separation indices were low in this study, although the reliability of individuals was within an acceptable limit above 0.7 for this type of analysis (41,89). According to Linacre (83), this indicates that the test can discriminate between two levels of people, which, in principle, would be sufficient. However, this indicates that the test could be improved to better discriminate between different levels of a major depressive episode. In this sense, the analysis of the location of the persons and the items along the measure (logist) shows that the test measures the depressive episode at a high level of occurrence for a non-clinical population, with item 9 "Have you thought you would be better off dead or have you thought about hurting yourself or hurting yourself in some way" corresponding to the highest level of the construct and the items referring to the feeling of tiredness and loss of appetite, those that measure a lower level. In this section, several readings are possible. On the one hand, the test does not appropriately measure the construct in people with low levels of depression. This is important because a mild form of depression precedes a moderate form and can develop into a severe form when an individual's coping strategies fail. Screening tests are useful to the extent that they detect a mild form of depression, but it is believed that the PHQ-9 did not capture this well enough in this population, as has been observed in other non-clinical populations (37).

However, the fact that the PHQ-9 is better able to identify more severe levels of the construct probably does not affect whether individuals with significant levels of depression are adequately detected and subsequently diagnosed by the professional in charge. Findings related to items measuring feelings of tiredness and changes in appetite could be explained by the nosological relationship between anxiety symptoms and depressive symptoms (90,91); for example, some somatic symptoms of depressive disorder, such as easy tiredness and fatigue, are part of the diagnostic criteria of other disorders, such as generalized anxiety disorder (92). A gap between death ideation and movement impairment, as measured by logist distance, was also evident. This observation raises the possibility that some intermediate psychopathological manifestations are missing between death ideation and movement disturbance, which could be interpreted as a failure of the test to measure some aspects of a construct that we accept exists and that is accepted by the scientific community. This raises questions about further refinement of the PHQ-9: is there an aspect of the depressive syndrome on this continuum that perhaps the test is not measuring, and can the PHQ-9 uniformly measure all persons who would meet the criteria for a depressive episode? In terms of reliability, inter-person separation was not excellent, although it was acceptable (0.74), while Cronbach's alpha was 0.83. This difference could be explained by the fact that Pearson's separation index is lower than Cronbach's alpha because the reliability in this model is based on a linear scale at the interval level when a good fit between the model and the data is observed, whereas alpha is based solely on the assumption of linear measures (93).

To our knowledge, this is the first study to report the validity of the PHQ-9 using Rasch analysis on a sample of Spanish-speaking university students. The large sample size results in high statistical power, which could technically increase the probability of model misfit or DIF detection. The results of this study should be analyzed considering several limitations. First, we did not analyze the concurrent validity of the PHQ-9 with other screening tests for depression or using a reference criterion. Second, we could not effectively control for response bias. Because the respondents were health science students, it is possible that some underestimated their actual symptoms for fear of stigmatization. Third, we did not assess the test-retest reliability, which can be useful when repeated measures of the instrument are required. Finally, it is worth noting that the Rasch model is a one-parameter IRT model that assumes that discrimination is the same for all items (36,57,94). Under this assumption, easy-to-endorse items (e.g., Item 4, low energy) discriminate as well as difficult-to-endorse items (e.g., Item 9, suicidal ideation) for subjects with a lower level of depressive symptoms (95), which is not always true.

Based on the limitations of this study, future research should verify the construct validity of the PHQ-9 by contrasting the results of the application of the test through a diagnosis of depressive disorder using structured psychiatric interviews. Further research is needed to verify whether invariance is maintained in the successive measurements of the instrument. This is important because there is work proposing the use of the PHQ-9 to assess the severity of a depressive episode and to monitor response to treatment (96). To further advance construct validity, future studies using two- or 3-parameter IRT models are warranted to validate the PHQ-9 in university populations.

Considering the above results, it can be concluded that the Rasch model analysis of instruments such as the PHQ-9 offers a detailed perspective on psychometric behavior, with evidence of its strengths and weaknesses. A way for the improvement of the instrument could be an analysis of the construct and the construction of items that represent a more continuous measure of depression in its different levels of severity.

AUTHORS CONTRIBUTIONS

Carlos Arturo Cassiani-Miranda: Conceptualization, Data curation, Investigation, Methodology, Project administration, Supervision, Validation, Writing - original draft, Writing - review & editing. Orlando Scoppetta: Conceptualization, Formal analysis, Methodology, Software, Validation, Writing - original draft, Writing -review & editing. María Alejandra Barrios-Villadiego: Conceptualization, Investigation, Writing - original draft, Writing - review & editing. Andrés Felipe Tirado-Otálvaro: Conceptualization, Investigation, Writing - original draft, Writing - review & editing. Andrea Carolina Duran-Bedoya: Conceptualization, Investigation, Writing - original draft, Writing - review & editing.

DATA AVAILABILITY

Data is available upon request from the corresponding author.

REVIEWER COMMENTS

The identities of the external reviewers and their respective evaluations are accessible via the following link: Opinion 521.pdf

REFERENCES

1. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16(9):606-13. https://doi.org/10.1046/j.1525-1497.2001.016009606.x
2. Kroenke K, Stump TE, Chen CX, Kean J, Damush TM, Bair MJ et al. Responsiveness of PROMIS and Patient Health Questionnaire (PHQ) depression scales in three clinical trials. Health Qual. Life Outcomes. 2021;19(1):41. https://doi.org/10.1186/s12955-021-01674-3
3. Levis B, Benedetti A, Thombs BD; DEPRESsion Screening Data (DEPRESSD) Collaboration. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. BMJ. 2019;365:l1476. https://doi.org/10.1136/bmj.l1476
4. Negeri ZF, Levis B, Sun Y, He C, Krishnan A, Wu Y et al. Accuracy of the Patient Health Questionnaire-9 for screening to detect major depression: Updated systematic review and individual participant data meta-analysis. BMJ. 2021;375:n2183. https://doi.org/10.1136/bmj.n2183
5. El-Den S, Chen TF, Gan YL, Wong E, O'Reilly CL. The psychometric properties of depression screening tools in primary healthcare settings: A systematic review. J Affect Disord. 2018;225:503-22. https://doi.org/10.1016/j.jad.2017.08.060
6. Levis B, Benedetti A, Levis AW, Ioannidis JPA, Shrier I, Cuijpers P, et al. Selective Cutoff Reporting in Studies of Diagnostic Test Accuracy: A Comparison of Conventional and Individual-Patient-Data Meta-Analyses of the Patient Health Questionnaire-9 Depression Screening Tool. Am J Epidemiol. 2017185(10):954-64. https://doi.org/10.1093/aje/kww191
7. Manea L, Boehnke JR, Gilbody S, Moriarty AS, McMillan D. Are there researcher allegiance effects in diagnostic validation studies of the PHQ-9? A systematic review and meta-analysis. BMJ Open. 2017;7(9):e015247. https://doi.org/10.1136/bmjopen-2016-015247
8. Wu Y, Levis B, Riehm KE, Saadat N, Levis AW, Azar M, 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):1368-80. https://doi.org/10.1017/s0033291719001314
9. 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(11):1596-602. https://doi.org/10.1007/s11606-007-0333-y
10. 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(1):71-7. https://doi.org/10.1016/j.genhosppsych.2005.07.003
11. Cassiani-Miranda CA, Scoppetta O. Factorial structure of the Patient Health Questionnaire-9 as a depression screening instrument for university students in Cartagena, Colombia. Psychiatry Res. 2018;269:425-49. https://doi.org/10.1016/j.psychres.2018.08.071
12. Ghazisaeedi M, Mahmoodi H, Arpaci I, Mehrdar S, Barzegari S. Validity, reliability, and optimal cut-off scores of the WHO-5, PHQ-9, and PHQ-2 to screen depression among university students in Iran. Int J Ment Health Addict. 2022;20(3):1824-33. https://doi.org/10.1007/s11469-021-00483-5
13. Keum BT, Miller MJ, Inkelas KK. Testing the factor structure and measurement invariance of the PHQ-9 across racially diverse U.S. college students. Psychol Assess. 2018;30(8):1096-106. https://doi.org/10.1037/pas0000550
14. Makhubela M, & Khumalo IP. Psychometric evaluation of the PHQ-9 in university students: Factorial validity and measurement equivalence across three African countries. Current psychology. 2022. https://doi.org/10.1007/s12144-022-02997-0
15. Rahman MA, Dhira TA, Sarker AR, Mehareen J. Validity and reliability of the Patient Health Questionnaire scale (PHQ-9) among university students of Bangladesh. PLoS One. 2022;17(6):e0269634. https://doi.org/10.1371/journal.pone.0269634
16. Zhang YL, Liang W, Chen ZM, Zhang HM, Zhang JH, Weng XQ et al. Validity and reliability of Patient Health Questionnaire-9 and Patient Health Questionnaire-2 to screen for depression among college students in China. Asia Pac Psychiatry. 2013;5(4):268-75. https://doi.org/10.1111/appy.12103
17. Marosszeky N, Shores EA, Jones MP, Sadeghi R. A Psychometric replication of fan (1998) item response theory and classical test theory: An empirical comparison of their Item/Person Statistics. J Appl Meas. 2020;21(4):456-80. URL
18. Zhao Y, Hambleton RK. Practical consequences of item response theory model misfit in the context of test equating with mixed-format test data. Front Psychol. 2017;8:484. https://doi.org/10.3389/fpsyg.2017.00484
19. Thomas ML. Advances in applications of item response theory to clinical assessment. Psychol Assess.

2019;31(12):1442-55. https://doi.org/10.1037/pas0000597

20. Tijmstra J, Bolsinova M, Jeon M. General mixture item response models with different item response structures: Exposition with an application to Likert scales. Behav Res Methods. 2018;50(6):2325-44. https://doi.org/10.3758/s13428-017-0997-0

21. Carroll HA, Hook K, Perez OFR, Denckla C, Vince CC, Ghebrehiwet S, et al. Establishing reliability and validity for mental health screening instruments in resource-constrained settings: Systematic review of the PHQ-9 and key recommendations. Psychiatry Res. 2020;291:113236. https://doi.org/10.1016/j.psychres.2020.113236

22. de Joode JW, van Dijk SEM, Walburg FS, Bosmans JE, van Marwijk HWJ, et al. Diagnostic accuracy of depression questionnaires in adult patients with diabetes: A systematic review and meta-analysis. PLoS One. 2019;14(6):e0218512. https://doi.org/10.1371/journal.pone.0218512

23. Fekadu A, Demissie M, Birhane R, Medhin G, Bitew T, Hailemariam M, et al. Under detection of depression in primary care settings in low and middle-income countries: a systematic review and meta-analysis. Syst Rev. 2022;11(1):21. https://doi.org/10.1186/s13643-022-01893-9

24. Kaggwa MM, Najjuka SM, Ashaba S, Mamun MA. Psychometrics of the Patient Health Questionnaire (PHQ-9) in Uganda: A systematic review. Front Psychiatry. 2022 Mar 7;13:781095. https://doi.org/10.3389/fpsyt.2022.781095

25. Kim S, Lee K. Screening for depression in mobile devices using Patient Health Questionnaire-9 (PHQ-9) Data: A diagnostic meta-analysis via machine learning methods. Neuropsychiatr Dis Treat. 2021;17:3415-30. https://doi.org/10.2147/ndt.s339412

26. Manea L, Gilbody S, McMillan D. A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. Gen Hosp Psychiatry. 2015;37(1):67-75. https://doi.org/10.1016/j.genhosppsych.2014.09.009\1.

27. 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(6):567-76. https://doi.org/10.1016/j.genhosppsych.2015.06.012

28. Pettersson A, Boström KB, Gustavsson P, Ekselius L. Which instruments to support diagnosis of depression have sufficient accuracy? A systematic review. Nord J Psychiatry. 2015;69(7):497-508. https://doi.org/10.3109/08039488.2015.1008568

29. Yin L, Teklu S, Pham H, Li R, Tahir P, Garcia ME. Validity of the Chinese language Patient Health Questionnaire 2 and 9: A systematic review. Health Equity. 2022 ;6(1):574-94. https://doi.org/10.1089/heq.2022.0030

30. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. Lancet Neurol. 2007;6(12):1094-105. https://doi.org/10.1016/s1474-4422(07)70290-9

31. Packham T, MacDermid JC. Measurement properties of the patient-rated wrist and hand evaluation: rasch analysis of responses from a traumatic hand injury population. J Hand Ther. 2013 Jul-Sep;26(3):216-23; quiz 224. https://doi.org/10.1016/j.jht.2012.12.006

32. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care. 2000;38(9 Suppl):II28-42. https://doi.org/10.1097/00005650-200009002-00007

33. Nguyen TH, Lee CS, Kim MT. Using item response theory to develop and refine patient-reported outcome measures. Eur J Cardiovasc Nurs. 2022;21(5):509-15. https://doi.org/10.1093/eurjcn/zvac020

34. Stover AM, McLeod LD, Langer MM, Chen WH, Reeve BB. State of the psychometric methods: patient-reported outcome measure development and refinement using item response theory. J Patient Rep Outcomes. 2019;3(1):50. https://doi.org/10.1186/s41687-019-0130-5

35. Ware JE Jr. Conceptualization and measurement of health-related quality of life: comments on an evolving field. Arch Phys Med Rehabil. 2003;84(4 Suppl 2):S43-51. https://doi.org/10.1053/apmr.2003.50246

36. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res. 2007;16 Suppl 1:5-18. https://doi.org/10.1007/s11136-007-9198-0

37. Jiraniramai S, Wongpakaran T, Angkurawaranon C, Jiraporncharoen W, Wongpakaran N. Construct validity and differential item functioning of the phq-9 among health care workers: rasch analysis approach. Neuropsychiatr Dis Treat. 2021;17:1035-45. https://doi.org/10.2147/ndt.s271987

38. Cappelleri JC, Jason Lundy J, Hays RD. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. Clin Ther. 2014;36(5):648-62. https://doi.org/10.1016/j.clinthera.2014.04.006

39. Hambleton RK, Russell WJ. Comparison of classical test theory and item response theory and their applications to test development. An NCME instructional module. 2005. http://dx.doi.org/10.1111/j.1745-3992.1993.tb00543.x

40. He Q, Wheadon C. Using the dichotomous Rasch model to analyze polytomous items. J Appl Meas. 2013;14(1):44-56. https://pubmed.ncbi.nlm.nih.gov/23442327/

41. 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(8):1358-62. https://doi.org/10.1002/art.23108

42. Christensen KS, Oernboel E, Zatzick D, Russo J. Screening for depression: Rasch analysis of the structural validity of the PHQ-9 in acutely injured trauma survivors. J Psychosom Res. 2017;97:18-22. https://doi.org/10.1016/j.jpsychores.2017.03.117

43. Reich H, Rief W, Brähler E, Mewes R. Cross-cultural validation of the German and Turkish versions of the PHQ-9: An IRT approach. BMC.2018; 6(1), 26. https://doi.org/10.1186/s40359-018-0238-z

44. Adler M, Hetta J, Isacsson G, Brodin U. An item response theory evaluation of three depression assessment instruments in a clinical sample. BMC Med Res Methodol. 2012;12:84. https://doi.org/10.1186/1471-2288-12-84

45. Martins Barroso S, Souto Melo AP, da Silva MA, Crosland Guimarães M. D. Evaluation of the Brazilian version of Patient Health Questionnaire (PHQ-9) in Quilombola population using the Item Response Theory. Salud Ment. 2019;42(1):43-50. https://doi.org/10.17711/SM.0185-3325.2019.006

46. Barthel D, Barkmann C, Ehrhardt S, Schoppen S, Bindt C; International CDS Study Group. Screening for depression in pregnant women from Côte d'Ivoire and Ghana: Psychometric properties of the Patient Health Questionnaire-9. J Affect Disord. 2015;187:232-40. https://doi.org/10.1016/j.jad.2015.06.042

47. Boulton AJ, Tyner CE, Choi SW, Sander AM, Heinemann AW, Bushnik T, et al. Linking the GAD-7 and PHQ-9 to the TBI-QOL anxiety and depression item banks. J Head Trauma Rehabil. 2019;34(5):353-63. https://doi.org/10.1097/htr.0000000000000529

48. Brodey BB, Goodman SH, Baldasaro RE, Brooks-DeWeese A, Wilson ME, Brodey IS, et al. Development of the Perinatal Depression Inventory (PDI)-14 using item response theory: a comparison of the BDI-II, EPDS, PDI, and PHQ-9. Arch Womens Ment Health. 2016;19(2):307-16. https://doi.org/10.1007/s00737-015-0553-9

49. Choi SW, Schalet B, Cook KF, Cella D. Establishing a common metric for depressive symptoms: linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. Psychol Assess. 2014;26(2):513-27. https://doi.org/10.1037/a0035768

50. Cumbe VFJ, Muanido A, Manaca MN, Fumo H, Chiruca P, Hicks L, et al. Validity and item response theory properties of the Patient Health Questionnaire-9 for primary care depression screening in Mozambique (PHQ-9-MZ). BMC Psychiatry. 2020;20(1):382. d https://doi.org/10.1186/s12888-020-02772-0

51. Donk LJ, Bickel EA, Krijnen WP, Tovote KA, Sanderman R, Schroevers MJ, et al. The value of distinct depressive symptoms (PHQ-9) to differentiate depression severity in cancer survivors: An item response approach. Psychooncology. 2019;28(11):2240-3. https://doi.org/10.1002/pon.5192

52. Fischer HF, Tritt K, Klapp BF, Fliege H. How to compare scores from different depression scales: equating the Patient Health Questionnaire (PHQ) and the ICD-10-Symptom Rating (ISR) using item response theory. Int J Methods Psychiatr Res. 2011;20(4):203-14. https://doi.org/10.1002/mpr.350

53. Forkmann T, Gauggel S, Spangenberg L, Brähler E, Glaesmer H. Dimensional assessment of depressive severity in the elderly general population: psychometric evaluation of the PHQ-9 using Rasch Analysis. J Affect Disord. 2013;148(2-3):323-30. https://doi.org/10.1016/j.jad.2012.12.019

54. Gothwal VK, Bagga DK, Sumalini R. Rasch validation of the PHQ-9 in people with visual impairment in South India. J Affect Disord. 2014;167:171-7. https://doi.org/10.1016/j.jad.2014.06.019

55. Horton M, Perry AE. Screening for depression in primary care: a Rasch analysis of the PHQ-9. BJ Psych Bull. 2016;40(5):237-43. https://doi.org/10.1192/pb.bp.114.050294

56. Huang XJ, Ma HY, Wang XM, Zhong J, Sheng DF, Xu MZ. Equating the PHQ-9 and GAD-7 to the HADS depression and anxiety subscales in patients with major depressive disorder. J Affect Disord. 2022;311:327-35. https://doi.org/10.1016/j.jad.2022.05.079

57. Kendel F, Wirtz M, Dunkel A, Lehmkuhl E, Hetzer R, Regitz-Zagrosek V. Screening for depression: Rasch analysis of the dimensional structure of the PHQ-9 and the HADS-D. J Affect Disord. 2010;122(3):241-6. https://doi.org/10.1016/j.jad.2009.07.004.

58. Lamoureux EL, Tee HW, Pesudovs K, Pallant JF, Keeffe JE, Rees G. Can clinicians use the PHQ-9 to assess depression in people with vision loss? Optom Vis Sci. 2009;86(2):139-45. https://doi.org/10.1097/opx.0b013e318194eb47

59. Liegl G, Wahl I, Berghöfer A, Nolte S, Pieh C, Rose M, et al. Using Patient Health Questionnaire-9 item parameters of a common metric resulted in similar depression scores compared to independent item response theory model reestimation. J Clin Epidemiol. 2016;71:25-34. https://doi.org/10.1016/j.jclinepi.2015.10.006

60. Ma S, Yang J, Yang B, Kang L, Wang P, Zhang N, et al. The Patient Health Questionnaire-9 vs. the Hamilton rating scale for depression in assessing major depressive disorder. Front Psychiatry. 2021;12:747139. https://doi.org/10.3389/fpsyt.2021.747139

61. Pedersen SS, Mathiasen K, Christensen KB, Makransky G. Psychometric analysis of the Patient Health Questionnaire in Danish patients with an implantable cardioverter defibrillator (The DEFIB-WOMEN study). J Psychosom Res. 2016;90:105-112. https://doi.org/10.1016/j.jpsychores.2016.09.010

62. Smith AB, Rush R, Wright P, Stark D, Velikova G, Sharpe M. Validation of an item bank for detecting and assessing psychological distress in cancer patients. Psychooncology. 2009;18(2):195-9. https://doi.org/10.1002/pon.1423

63. 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(3):355-66. https://doi.org/10.1177/1073191120976863

64. Teymoori A, Real R, Gorbunova A, Haghish EF, Andelic N, Wilson L, et al. Measurement invariance of assessments of depression (PHQ-9) and anxiety (GAD-7) across sex, strata and linguistic backgrounds in a European-wide sample of patients after traumatic brain injury. J Affect Disord. 2020;262:278-85. https://doi.org/10.1016/j.jad.2019.10.035

65. Umegaki Y, Todo N. Psychometric properties of the Japanese CES-D, SDS, and PHQ-9 depression scales in university students. Psychol Assess. 2017;29(3):354-9. https://doi.org/10.1037/pas0000351

66. Wahl I, Löwe B, Bjorner JB, Fischer F, Langs G, Voderholzer U, et al. Standardization of depression measurement: a common metric was developed for 11 self-report depression measures. J Clin Epidemiol. 2014;67(1):73-86. https://doi.org/10.1016/j.jclinepi.2013.04.019

67. Williams RT, Heinemann AW, Bode RK, Wilson CS, Fann JR, Tate DG. Improving measurement properties of the Patient Health Questionnaire-9 with rating scale analysis. Rehabil Psychol. 2009;54(2):198-203. https://doi.org/10.1037/a0015529

68. Zhong Q, Gelaye B, Fann JR, Sanchez SE, Williams MA. Cross-cultural validity of the Spanish version of PHQ-9 among pregnant Peruvian women: a Rasch item response theory analysis. J Affect Disord. 2014;158:148-53. https://doi.org/10.1016/j.jad.2014.02.012

69. Akhtar P, Ma L, Waqas A, Naveed S, Li Y, Rahman A, et al. Prevalence of depression among university students in low and middle income countries (LMICs): a systematic review and meta-analysis. J Affect Disord. 2020;274:911-9. https://doi.org/10.1016/j.jad.2020.03.183

70. Liu XQ, Guo YX, Zhang WJ, Gao WJ. Influencing factors, prediction and prevention of depression in college students: A literature review. World J Psychiatry. 2022;12(7):860-73. https://doi.org/10.5498/wjp.v12.i7.860

71. Arun P, Ramamurthy P, Thilakan P. Indian medical students with depression, anxiety, and suicidal behavior: why do they not seek treatment? Indian J Psychol Med. 2022;44(1):10-6. https://doi.org/10.1177/0253717620982326

72. Atienza-Carbonell B, Guillén V, Irigoyen-Otiñano M, Balanzá-Martínez V. Screening of substance use and mental health problems among Spanish medical students: A

multicenter study. J Affect Disord. 2022;311:391-8. https://doi.org/10.1016/j.jad.2022.05.090

73. Kyriazos TA. Applied psychometrics: Sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. Psychology. 2018;09(08):2207-30. https://doi.org/10.4236/psych.2018.98126

74. Kocalevent RD, Hinz A, Brähler E. Standardization of the depression screener patient health questionnaire (PHQ-9) in the general population. Gen Hosp Psychiatry. 2013;35(5):551-5. https://doi.org/10.1016/j.genhosppsych.2013.04.006

75. Spitzer RL, Williams JB, Kroenke K, Hornyak R, McMurray J. Validity and utility of the PRIME-MD patient health questionnaire in assessment of 3000 obstetric-gynecologic patients: the PRIME-MD Patient Health Questionnaire obstetrics-gynecology study. Am J Obstet Gynecol. 2000;183(3):759-69. https://doi.org/10.1067/mob.2000.106580

76. Cassiani-Miranda CA, Vargas-Hernández MC, Pérez-Anibal E, Herazo-Bustos MI, Hernández-Carrillo M. Confiabilidad y dimensionalidad del PHQ-9 en el cribado de síntomas de depresión en estudiantes de ciencias de la salud de Cartagena, 2014. Biomedica. 2017;37(0):112-20. https://doi.org/10.7705/biomedica.v37i0.3221

77. Cassiani-Miranda CA, Pérez-Aníbal E, Vargas-Hernández MC, Herazo-Bustos M. Validez de apariencia y adaptación del PHQ-9 para la detección de síntomas depresivos en estudiantes universitarios de ciencias de la salud Cartagena (Colombia). Salud uninorte. 2018; 34(1):75-87. http://dx.doi.org/10.14482/sun.34.1.9154

78. Cassiani-Miranda CA, Cuadros-Cruz AK, Torres-Pinzón H, Scoppetta O, Pinzón-Tarrazona JH, López-Fuentes WY, et al. Validity of the Patient Health Questionnaire-9 (PHQ-9) for depression screening in adult primary care users in Bucaramanga, Colombia. Rev Colomb Psiquiatr (Engl Ed). 2021;50(1):11-21. https://doi.org/10.1016/j.rcp.2019.09.001

79. Bond T. Applying the Rasch Model: Fundamental Measurement in the Human Sciences, (3rd ed.). Routledge. 2015. https://doi.org/10.4324/9781315814698

80. Rasch G. Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. Nielsen & Lydiche. 1960.

81. Linacre JM. WINSTEPS Rasch measurement computer program User's Guide. 2022.Version 5.3.2.

82. Lambert S, Pallant JF, Girgis A. Rasch analysis of the Hospital Anxiety and Depression Scale among caregivers of cancer survivors: implications for its use in psycho-oncology. Psychooncology. 2011;20(9):919-25. https://doi.org/10.1002/pon.1803

83. Linacre JM. Advancing the Metrological Agenda in the Social Sciences. In W. P. Fisher, & S. J. Cano (Eds.), Person-Centered Outcome Metrology (pp. 165-193). Springer International Publishing. 2023. https://doi.org/10.1007/978-3-031-07465-3_7

84. Scoppetta O, Pardo Adames C, Aguilar-Pardo D, & Barreto I. Evidencia Complementaria para la Validación de la Escala SRAS de Altruismo en Colombia. [Complementary evidence for the validation of the altruism SARS scale in Colombia]. Rev. Iberoam. de Diagnostico y Evaluacion Psicol. 2021;58(1):47-55. https://doi.org/10.21865/RIDEP58.1.04

85. Zwick R. A review of its differential item functioning assessment procedures: flagging rules, minimum sample size requirements, and criterion refinement. ETS research report. 2012;2012(1), i-30. https://doi.org/10.1002/j.2333-8504.2012.tb02290.x

86. Bi Y, Wang L, Cao C, Fang R, Li G, Liu P, et a. The factor structure of major depressive symptoms in a sample of Chinese earthquake survivors. BMC Psychiatry. 2021;21(1):59. https://doi.org/10.1186/s12888-020-02993-3

87. Cassiani-Miranda CA, Scoppetta O, Cabanzo-Arenas DF. Validity of the Hospital Anxiety and Depression Scale (HADS) in primary care patients in Colombia. Gen Hosp Psychiatry. 2022;74:102-9. https://doi.org/10.1016/j.genhosppsych.2021.01.014

88. Tsai J, Elhai JD, Pietrzak RH, Hoff RA, Harpaz-Rotem I. Comparing four competing models of depressive symptomatology: a confirmatory factor analytic study of 986,647 U.S. veterans. J Affect Disord. 2014;165:166-9. https://doi.org/10.1016/j.jad.2014.04.075

89. 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(2):256-63. https://doi.org/10.1080/02813432.2019.1608039

90. Belmans E, Bastin M, Raes F, Bijttebier P. Temporal associations between social anxiety and depressive symptoms and the role of interpersonal stress in adolescents. Depress Anxiety. 2019;36(10):960-7. https://doi.org/10.1002/da.22939

91. Zbozinek TD, Rose RD, Wolitzky-Taylor KB, Sherbourne C, Sullivan G, Stein MB, et al. Diagnostic overlap of generalized anxiety disorder and major depressive disorder in a primary care sample. Depress Anxiety. 2012;29(12):1065-71. https://doi.org/10.1002/da.22026

92. Stein DJ, Scott KM, de Jonge P, Kessler RC. Epidemiology of anxiety disorders: from surveys to nosology and back. Dialogues Clin Neurosci. 2017;19(2):127-36. https://doi.org/10.31887/dcns.2017.19.2/dstein

93. Wind SA, & Gale JD. Diagnostic opportunities using Rasch measurement in the context of a misconceptions-based physical science assessment: distractor analysis with Rasch measurement theory. Science education. 2015;99(4):721-41. https://doi.org/10.1002/sce.21172

94. Reise SP, Du H, Wong EF, Hubbard AS, Haviland MG. Matching IRT models to patient-reported outcomes constructs: the graded response and log-logistic models for scaling depression. Psychometrika. 2021;86(3):800-24. https://doi.org/10.1007/s11336-021-09802-0

95. Harris D. Comparison of 1-, 2-, and 3-parameter IRT models. Educational measurement: issues and practice. 1989; 8(1), 35-41. https://doi.org/10.1111/j.1745-3992.1989.tb00313.x

96. Kroenke K, Spitzer RL, Williams JB, Löwe B. The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review. Gen Hosp Psychiatry. 2010;32(4):345-59. https://doi.org/10.1016/j.genhosppsych.2010.03.006