Construct and criterion validity of patient-reported outcomes (PROs) for depression: A clinimetric comparison

Construct and criterion validity of patient-reported outcomes (PROs) for depression: A clinimetric comparison

Danilo Carrozzino[a], Kaj Sparle Christensen[b], Fiammetta Cosci[c,d,*]

[a]Department of Psychology, Alma Mater University of Bologna, Bologna, Italy

[b]Research Unit for General Practice and Section for General Medical Practice, Department of Public Health, Aarhus University, Aarhus, Denmark

[c]Department of Health Sciences, University of Florence, Florence, Italy

[d]Department of Psychiatry & Neuropsychology, Maastricht University, Maastricht, The Netherlands

ABSTRACT

Background: A number of patient-reported outcomes (PROs) have been developed but insufficient attention has been devoted to the assessment of their clinimetric properties. Clinimetrics, the science of clinical measurements, has been considered an emerging approach for evaluating reliability and validity of PROs. This is the first study using clinimetric principles to compare the construct and criterion validity of the Major Depression Inventory (MDI), the Beck Depression Inventory-II (BDI-II), the World Health Organization Well-Being Index (WHO-5), three of the most widely used PROs for the assessment of depression. Methods: Construct validity was evaluated via Item Response Theory (IRT) models (i.e., combining Rasch and Mokken analyses). Using the ICD-10 diagnostic algorithm for any depression as the gold standard, Receiver Operating Characteristic (ROC) curves were performed to examine the criterion validity of PROs. Results: One hundred healthy subjects (73% females, 32.6 ±10.5 years) participated in the study, giving a response rate of 90.1%. When using IRT analyses, MDI and WHO-5 were found to be reliable and unidimensional, while BDI-II showed lack of unidimensionality. ROC analyses supported the diagnostic accuracy of MDI and the screening properties of WHO-5. Limitations: The main limitations of the present study are that healthy subjects were assessed only via only self-reported measures and a cross-sectional design was used. Conclusions: WHO-5 and MDI outperformed BDI-II in terms of construct and criterion validity. WHO-5 should be considered when screening for depression, while MDI should be used as a valid diagnostic instrument and as a unidimensional measure to assess depression severity.

1. Introduction

In clinical research and practice, there has been growing interest on patient reported outcomes (PROs), any report coming directly from patients about how they function or feel in relation to a health condition or its therapy (Basch, 2017; Deshpande et al., 2011; Fava et al., 2019). Emphasizing the importance of assessing what matters to patients, PROs were developed to ideally cover the following aspects: 1) symptom burden and severity; 2) biopsychosocial functioning; 3) quality of life and well-being (Kristensen et al., 2018). A number of studies have examined the reliability and validity of PROs (Calvert et al., 2013; Cella et al., 2010; Mokkink et al., 2010; Pilkonis et al., 2011). The Patient-Reported Outcomes Measurement Information System (PROMIS), a project commissioned by the US National Institutes of Health (NIH), was one of the most important initiatives which aimed to improve the precision and efficiency of PROs (Reeve et al., 2007). The major promise of PROMIS was to develop unidimensional outcome measures using Item Response Theory (IRT) models (Fries et al. 2005; Thomas, 2011).

The assessment of unidimensionality has been considered a central issue in clinimetrics (Bech, 2004, 2012; Carrozzino et al., 2020; Fava and Belaise, 2005; Fava et al., 2004; Tomba and Bech, 2012), an innovative evaluation method originally introduced by Alvan R. Feinstein (1982, 1987) and, more recently, refined as the science of clinical measurements (Fava et al., 2012). According to the clinimetric approach, the concept of unidimensionality applies to the assessment of construct validity and can be used (i) for evaluating whether each item of the rating scale covers unique clinical information, (ii) for testing if symptoms (i.e., items) belong (i.e., correspond) to an underlying syndrome, and (iii) for determining the extent to which the total score of the assessment instrument is a statistically sufficient outcome measure of the severity of the clinical condition under examination (Bech, 2012; Carrozzino et al., 2020; Fava et al., 2018). On the contrary, in psychometrics, the concept of unidimensionality relies to homogeneity of components. Thus, to be

included in a rating scale, items have to be highly correlated and display the same clinical weight (Fava and Belaise, 2005; Fava et al., 2004).

The unidimensionality of PROs has been largely documented from a psychometric point of view (Irwin et al., 2010; Kwan et al., 2019; Millier et al., 2014; Pilkonis et al., 2011; Rose and Devine, 2014). PROs assessing depression, such as the Beck Depression Inventory (BDI) (Beck et al., 1961), the Zung Self-Rating Depression Scale (SDS) (Zung, 1965) and the Patient Health Questionnaire-9 (PHQ-9) (Kroenke and Spitzer, 2002; Kroenke et al., 2001) were developed and tested according to psychometric principles, with only one recent clinimetric analysis for PHQ-9 (He et al., 2020). Not surprisingly, BDI (Konstantinidis et al., 2011) and SDS (Bech and Wermuth, 1998) showed poor construct validity and the items of PHQ-9 were able to detect the prevalence but not the severity of depression symptoms (Bech and Timmerby, 2018). Thus, a number of clinimetric dilemmas (Bech, 2016a) are still in need of being addressed when using PROs in the clinical process of assessment of mental disorders.

Bech (2016b) recommended the Major Depression Inventory (MDI) (Bech et al., 2001; Bech and Wermuth, 1998; Olsen et al., 2003) and the five-item version of the World Health Organization Well-Being Index (WHO-5) (Topp et al., 2015) as PROs to be used for a comprehensive assessment of depression. MDI is a self-report questionnaire that can be used both as a diagnostic tool covering ICD-10 (World Health Organization, 1993) and DSM-IV (American Psychiatric Association, 1994) diagnostic criteria and as a unidimensional measure of depression severity (Bech et al., 2001; Olsen et al., 2003). WHO-5 is a self-rating scale assessing a subjective state of well-being which can be also used as a screening measure of depression (Topp et al., 2015).

In the present study, MDI and WHO-5 were compared with the Beck Depression Inventory-II (BDI-II) (Beck et al., 1996), one of the most widely used PROs for the assessment of self-reported symptoms of depression. The aim was to identify PROs assessing depression which

performed better in terms of construct and criterion validity. Construct validity was tested using IRT models (i.e., combining Rasch and Mokken analyses), criterion validity was tested via the Receiver Operating Characteristic (ROC) curve (i.e., using the ICD-10 diagnostic algorithm for any depression as the criterion standard).

2. Materials and methods

2.1. Sample

Potential participants were informed about the research through advertisements and flyers posted on bulletin boards placed in public areas (e.g., post office, library, university campus). The study was run in a convenience sample of healthy subjects consecutively recruited from the general population of Florence (Italy). The optimal number of participants to recruit was determined using methodological recommendations, which suggest a sample size of at least 100 subjects for conducting validation studies on PROs (Anthoine et al., 2014). Inclusion criteria were: (a) self-reported declaration of being healthy; (b) age ranging from 18 to 64 years. The exclusion criteria were: (a) any current self-reported chronic medical disease; (b) any current psychiatric disorder as assessed via the Mini International Neuropsychiatric Interview - MINI (Sheehan et al., 1998); (c) self-reported declaration of cognitive deficits or other intelligence problems negatively affecting the ability of reading and understanding the Italian language; (d) mother tongue other than Italian.

2.2. Procedure

Participants who agreed to take part to the study and met the inclusion criteria provided a signed written informed consent of privacy protection disclaimer. Participants filled out PROs in a

paper-pencil format and data were analyzed anonymously according to the Italian law on the treatment of personal data (i.e., Law no. 196, June 30, 2003).

2.3. Measures

Participants completed the following PROs:

The Major Depression Inventory (Bech, 1997; Bech and Wermuth, 1998), a 10-item self- rating scale covering both the ICD-10 (World Health Organization, 1993) and DSM-IV (American Psychiatric Association, 1994) symptoms of depression. Items are rated on a 6-point Likert scale ranging from 0 ("at no time") to 5 ("all the time") and refer to the last 14 days. The total score of MDI ranges from 0 to 50 (Bech, 2012). The MDI version used in the present study was translated into Italian by the first author (D.C.) and back-translated into English by an independent English-speaking clinical psychologist. Each item of the English back-translation was checked for accuracy, compared with the original MDI (Bech, 1997; Bech and Wermuth, 1998), and approved by the developer of MDI (Prof. Per Bech). The Italian version of MDI is published in the appendix of the Italian edition of Clinical Psychometrics (Bech, 2018).

The *Beck Depression Inventory*-II (Beck et al., 1996), a 21-item self-administered rating scale assessing depressive symptoms. Subjects score how they felt in the last two weeks, including the day of the assessment. Each item is rated on a 4-point Likert scale ranging from 0 to 3 and the total score ranges from 0 to 63. We used the Italian version of BDI-II (Ghisi et al., 2007).

The World Health Organization Well-Being Index (Topp et al., 2015), a 5-item self-reported questionnaire for the assessment of psychological well-being. Each item is positively worded and rated on a 6-point Likert scale ranging from 0 ("at no time") to 5 ("all of the time"). WHO-5 raw score ranges from 0 to 25 (Topp et al., 2015). Multiplying the raw score by 4, the percentage score

of WHO-5 ranges from 0 (representing the worst imaginable well-being) to 100, reflecting the best imaginable well-being (Topp et al., 2015). The Italian version of WHO-5 (Montella et al., 2016), available on https://www.psykiatri-regionh.dk/who-5/who-5-questionnaires/Pages/default.aspx, was used in the present study.

2.4. Statistical analyses

IRT analyses were conducted to examine the construct validity of PROs. The Rasch analysis was performed using Rasch Unidimensional Measurement Models (RUMM2030) software (Andrich et al., 2010) and the following measurement properties were tested:

1. *Overall fit to the model*, which was evaluated using the chi-square item-trait interaction statistics (Pallant and Tennant, 2007; Tennant and Conaghan, 2007). Such statistics provided a summary measure of how the scale under examination conforms to the Rasch model expectations (Nielsen et al., 2017). A non-significant chi-square probability value indicated a good level of overall fit (Pallant and Tennant, 2007; Tennant and Conaghan, 2007).
2. *Individual item and person fit*: standardized fit residual values for items and subjects were examined for any indication of misfit.
3. *Unidimensionality*: to determine whether the rating scale is a valid measure of an underlying unidimensional construct, Principal Component Analysis (PCA) of residuals was conducted to identify the two most different subsets of items (i.e., the most positively and negatively factor-loading items on the first component). Paired *t*-tests were then performed to compare scores on the two subsets of items. If more than 5% of *t*-tests were significant, the scale under assessment was not considered unidimensional (Christensen et al., 2019; Nielsen et al., 2017).

4. *Person Separation Reliability Index (PSI)*, which was examined to evaluate the internal consistency of PROs and their ability to discriminate among subjects with different levels of the underlying trait (Tennant and Conaghan, 2007).

Mokken analysis, which is the non-parametric version of IRT models (Bech, 2012; Mokken, 1971), was also performed to further examine the unidimensionality or scalability of PROs under evaluation. The Mokken analysis was conducted using Stata statistical software, version 7 (Stata Corporation, College Station, TX). The Stata LoevH command was used to compute Loevinger's coefficients of homogeneity. According to Mokken (Mokken, 1971), Loevinger's coefficients of homogeneity (Loevinger, 1947) ranging from 0.30 to 0.39 are considered just acceptable, while a coefficient ≥ 0.40 is a clear demonstration of the scalability of the rating scale under assessment (Bech, 2012). ROC analyses were conducted to assess the criterion validity of PROs. Using the ICD-10 diagnostic algorithm for any depression (i.e., having at least 2 core [central] symptoms most of the time or all the time and 2 accompanying symptoms more than half of the time, most of the time, or all the time) as the gold standard (Bech, 2012), the Area Under the Curve (AUC) was calculated to assess the diagnostic accuracy of PROs (Deyo and Centor, 1986). An AUC of at least 0.70 indicated acceptable diagnostic accuracy (Terwee et al., 2007). Optimal cut-off scores, which maximized the sensitivity and specificity of PROs, were also identified (Deyo and Centor, 1986).

3. Results

3.1. Sample

One hundred and ten consecutive subjects were invited to take part in the study. Of them, 100 accepted, giving a response rate of 90.1%. They had a mean age ± SD of 32.64 ± 10.48 years,

73% were females. Further details on sample characteristics are provided elsewhere (Carrozzino et al., 2019).

3.2. Overall fit to the Rasch model

Model fit statistics for WHO-5, BDI-II, and MDI are reported in Table 1. Rasch analysis of WHO-5 revealed a non-significant item-trait interaction statistic ($\chi^2 = 12.63$, degrees of freedom [df] = 10, p = 0.245), indicating adequate fit to the model, with no misfitting items. Standardized fit residual values for items (SD = 0.80) and persons (SD = 1.09) were within acceptable limits. The initial analysis of BDI-II revealed a non-significant item-trait interaction statistic ($\chi^2 = 26.37$, df = 42, p = 0.97), indicating a good fit of the data to the Rasch model. Standardized fit residual values for items (SD = 0.62) and persons (SD = 1.29) were within acceptable limits. The initial analysis of MDI showed a significant item-trait interaction statistic ($\chi^2 = 33.85$, df = 20, p = 0.03), which indicated misfit to the Rasch model. Standardized fit residual values for items (SD = 1.11) and persons (SD = 1.32) were within acceptable limits.

3.3. Unidimensionality

The unidimensionality of WHO-5, BDI-II, and MDI was tested using both Mokken and Rasch analyses. The total score of WHO-5 had an acceptable scalability (Loevinger's coefficient of homogeneity of 0.41). Paired *t*-tests comparisons indicated that 5% of *t*-tests were significant, suggesting that WHO-5 had a just acceptable unidimensionality (Table 1). Concerning BDI-II, the total score was found to have a just acceptable scalability (Loevinger's coefficient of homogeneity of 0.32). Paired *t*-tests comparisons revealed that more than 5% of *t*-tests were significant (Table 1), indicating that BDI-II was not unidimensional. The Mokken analysis of MDI showed that its total score had an acceptable scalability (Loevinger's coefficient of homogeneity of 0.39). Paired *t*-tests

comparisons demonstrated that 4% of *t*-tests were significant, indicating that MDI was a unidimensional measure (Table 1).

3.4. Person separation reliability index (PSI)

PSI indices of WHO-5 (0.73) and MDI (0.79) were just acceptable, suggesting that these PROs could reliably distinguish between different groups but not between different subjects, at least when healthy individuals are considered. Concerning BDI-II, PSI was 0.65, suggesting that it could not be reliably used to discriminate between different groups of subjects with different levels of the underlying construct.

3.5. Criterion validity

ROC statistics are presented in Table 2. When using the ICD-10 diagnostic algorithm for depression (i.e., minor, moderate, severe depression), 3% of the sample (n = 3) satisfied the diagnostic criteria for depression. MDI had an AUC of 0.96, which indicated excellent diagnostic accuracy. The AUC for WHO-5 was 0.72, indicating good diagnostic accuracy. Concerning BDI-II, the AUC was 0.69, which indicated insufficient diagnostic accuracy.

3.6. Cut-off scores

The sensitivity and specificity for different cut-off scores of MDI are illustrated in Table 3. When using the ICD-10 diagnostic algorithm for depression, the optimal cut-off value for MDI was 23, giving a sensitivity of 100% and a specificity of 96%. The sensitivity and specificity for different cut-off scores of WHO-5 are presented in Table 4. The optimal cut-off point of WHO-5 was 10, which had a sensitivity of 100% and a specificity of 42%. The sensitivity and specificity for

different cut-off values of BDI-II are showed in Table 5. The optimal cut-off score for BDI-II was 4, giving a sensitivity of 100% and a specificity of 49%.

4. Discussion

This is the first study applying a clinimetric approach to compare the construct and criterion validity of three PROs assessing depression head a head. Based on the present results, WHO-5 and MDI outperformed BDI-II in terms of construct and criterion validity. WHO-5 might, thus, be considered a valid screening measure of depression and MDI a sensitive diagnostic instrument and a unidimensional measure to assess depression severity.

When using IRT analyses, WHO-5 had acceptable unidimensionality, while BDI-II showed lack of unidimensionality. Bech and his research group (Konstantinidis et al., 2011) reported similar findings, demonstrating that the first version of BDI (Beck et al., 1961) is a multidimensional measure of depression. Other authors (Konstantinidis et al., 2011) also noted that BDI items are particularly problematic (i.e., invasive) when used in general population as they were generated based on descriptions of symptoms reported by depressed patients (Naughton and Wiklund, 1993).

When applying IRT models to MDI, an initial misfit to the model was found but the Rasch and Mokken analyses test of unidimensionality were accepted, indicating that the total score and items were a statistically sufficient and clinically valid measure of the underlying construct of depression severity. An additional analysis (not presented) showed that eliminating the least fitting item (i.e., the question no. 9), MDI resulted in a fit to the Rasch model. Such findings are in line with previous studies, suggesting that MDI is a unidimensional measure of depression severity (Bech et al., 2001; Bech and Wermuth, 1998; Konstantinidis et al., 2011; Olsen et al., 2003). Contrasting results were, however, also reported (Amris et al., 2016; Christensen et al., 2019; Nielsen et al., 2017). Amris et al. (2016) revealed lack of unidimensionality in a sample of patients

with chronic widespread pain while Nielsen et al. (2017) and Christensen et al. (2019) showed similar problems especially with item number 9 when assessing the construct validity of MDI in a sample of primary care patients. The population under study may be one of the reasons of such inconsistency. Future studies are needed to further investigate the construct validity of MDI in different clinical populations.

When estimating the internal consistency of PROs, MDI had the highest PSI value, indicating that it could reliably discriminate between groups of subjects with different levels of depression severity. Kellner and Sheffield (1973) introduced the clinimetric concept of sensitivity to describe the ability of a rating scale to discriminate between different subgroups of patients suffering from the same illness (e.g., depressed inpatients and outpatients) and to differentiate the severity of symptoms (e.g., certain symptoms may be more troublesome or incapacitating than others). In the clinimetric approach, the assessment of sensitivity has been considered a central issue, particularly when treatment effects are small and in case of subclinical symptoms (Fava and Belaise, 2005; Fava et al., 2004). MDI was found to entail such a clinimetric property. However, studies are needed to further evaluate MDI sensitivity using different subgroups of patients suffering from the same illness (e.g., patients with major and minor depression).

When using the ICD-10 algorithm for any depression, 3% of the sample satisfied the diagnostic criteria for depression. This result may be surprising since all subjects were assessed via the MINI and did not satisfy the diagnostic criteria for any psychiatric disorder. However, the ICD-10 algorithm refers to minor-moderate-severe depression while the MINI assesses major depression and dysthymia. Indeed, the 3% of the sample who satisfied the diagnostic criteria for depression according to the ICD-10 algorithm corresponded to those subjects who had clinical depressive manifestations but were under the diagnostic threshold according to the MINI.

When applying ROC statistics, MDI demonstrated excellent diagnostic accuracy. The optimal cut-off score of 23 revealed a sensitivity of 100% and a specificity of 96% in detecting depression. These findings are consistent with previous studies (Bech et al., 2001, 2015) and should be commented under the light of the poor diagnostic accuracy we found for BDI-II. BDI-II is a common measure of depression in clinical studies although its sensitivity is questioned (Demyttenaere and De Fruyt, 2003). The present investigation cast some further doubts on the validity of choosing BDI-II, instead for instance of MDI, in clinical studies where both sensitivity and specificity are relevant. However, a proper knowledge of the different tools and their backgrounds is the basis for researchers and clinicians to choose the correct scale for their purposes (Demyttenaere and De Fruyt, 2003). Under this light, and based on the current findings, while the BDI-II seems a poor choice in a drug trial where differences are expected to be minimal, it may be a good choice in a psychotherapy trial that uses the constructs that were used by Beck for building the scale.

WHO-5 showed good criterion validity. The optimal cut-off score of 10 had a sensitivity of 100% and a specificity of 42%, as found also by Christensen et al. (2015), thus supporting the clinical utility of WHO-5 as a screening measure of depression. Indeed, already Bech and his research group (Topp et al., 2015) noted: “For a screening instrument such as the WHO-5, having sufficiently high sensitivity (i.e., a very high proportion of depressed individuals screen positive) is a key factor, whereas high specificity is less important. This is due to the fact that the second step of the diagnostic process, after an initial positive screening with the WHO-5, consists of a diagnostic interview performed by a trained clinician, during which false positives (patients screening positive on the WHO-5 but not meeting criteria for depression) will be detected”.

4.1. Limitations

The present study has some limitations. First, we enrolled healthy subjects; future studies including depressive patients are needed to further evaluate the construct and criterion validity of PROs for depression. Second, a cross-sectional design was used, thus precluding the evaluation of test-retest reliability, predictive, and incremental validity of PROs. Third, only self-report measures were proposed. Future studies, making use of the clinical judgment of experienced clinicians or including clinician-rated scales as the main indices of validity are highly encouraged to further examine the clinimetric properties of PROs for depression.

5. Conclusions

The findings of this study indicate that WHO-5 and MDI are valid and sensitive PROs. It is, however, important to note that they were found to entail different clinimetric properties. Further, there are no perfect or ideal PROs. In choosing a scale the investigators should have in mind a number of clinical factors: the pros and cons of each tool, study aims, and clinical characteristics of the population under examination. The clinimetric perspective may therefore offer important insights for performing an informed choice. Investigators, on the contrary, tend to choose an instrument as a result of the popularity and not of specific indications.

A few indications emerge from the present study. The WHO-5 appears to be particularly useful when screening for depressive symptoms but, if the aim of the investigation is to diagnose depression and assess its severity, the MDI should be used.

Conflicts of interest:

Authors have no conflicts of interest to declare.

Acknowledgements:

None.

Funding:

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Statement Contributors:

- Conception and design: Fiammetta Cosci.
- Data analysis: Kaj Sparle Christensen.
- Drafting the manuscript: Danilo Carrozzino.
- Revising the manuscript and final approval of the submitted version: All authors.

References

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders, Fourth Ed. (DSM-IV). American Psychiatric Association, Washington, DC.

Amris, K., Omerovic, E., Danneskiold-Samsøe, B., Bliddal, H., Wæhrens, E.E., 2016. The validity of self-rating depression scales in patients with chronic widespread pain: a Rasch analysis of the Major Depression Inventory. Scand. J. Rheumatol. 45, 236-246. https://doi.org/10.3109/03009742.2015.1067712.

Andrich, D., Lyne, A., Sheridan, B., Luo, G., 2010. RUMM 2030. RUMM Laboratory, Perth, Australia.

Anthoine, E., Moret, L., Regnault, A., Sébille, V., Hardouin, J.B., 2014. Sample size used to validate a scale: a review of publications on newly-developed patient reported outcomes measures. Health. Qual. Life. Outcomes. 12, 1-10. https://doi.org/10.1186/s12955-014-0176-2.

Basch, E., 2017. Patient–Reported Outcomes - Harnessing Patients' voices to improve clinical care. N. Eng. J. Med. 376, 105-108. https://doi.org/10.1056/NEJMp1611252.

Bech, P., 1997. Quality of life instruments in depression. Eur. Psychiatry. 12, 194-198. https://doi.org/10.1016/S0924-9338(97)89104-3.

Bech, P., 2004. Modern psychometrics in clinimetrics: impact on clinical trials of antidepressants. Psychother. Psychosom. 73, 134-138. https://doi.org/10.1159/000076448.

Bech, P., 2012. Clinical Psychometrics. Wiley Blackwell, Oxford.

Bech, P., 2016a. Clinimetric dilemmas in outcome scales for mental disorders. Psychother. Psychosom. 85, 323-326. https://doi.org/10.1159/000448810.

Bech, P., 2016b. Measurement-based care in mental disorders. Springer, New York.

Bech, P., 2018. L'assessment in clinica psicologica: la Psicodiagnostica Clinimetrica. Giovanni Fioriti Editore, Roma.

Bech, P., Rasmussen, N.A., Olsen, L.R., Noerholm, V., Abildgaard, W., 2001. The sensitivity and specificity of the Major Depression Inventory, using the Present State Examination as the index of diagnostic validity. J. Affect. Disord. 66, 159-164. https://doi.org/10.1016/S0165-0327(00)00309-8.

Bech, P., Timmerby, N., 2018. An overview of which health domains to consider and when to apply them in measurement-based care for depression and anxiety disorders. Nord. J. Psychiatry, 72, 367-373. https://doi.org/10.1080/08039488.2018.1465592.

Bech, P., Timmerby, N., Martiny, K., Lunde, M., Søndergaard, S., 2015. Psychometric evaluation of the Major Depression Inventory (MDI) as depression severity scale using the LEAD (Longitudinal Expert Assessment of All Data) as index of validity. BMC. Psychiatry. 15, 190. https://doi.org/10.1186/s12888-015-0529-3.

Bech, P., Wermuth, L., 1998. Applicability and validity of the Major Depression Inventory in patients with Parkinson's disease. Nord. J. Psychiatry. 52, 305-310. https://doi.org/10.1080/08039489850149741.

Beck, A.T., Steer, R. A., Brown, G.K., 1996. Beck Depression Inventory (BDI-II). Pearson, Canada.

Beck, A.T., Ward, C., Mendelson, M., Mock, J., Erbaugh, J., 1961. Beck Depression Inventory (BDI). Arch. Gen. Psychiatry. 4, 561-571. https://doi.org/10.1001/archpsyc.1961.01710120031004.

Calvert, M., Blazeby, J., Altman, D.G., Revicki, D.A., Moher, D., Brundage, M.D., CONSORT PRO Group, 2013. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. JAMA. 309, 814-822. https://doi.org/10.1001/jama.2013.879.

Carrozzino, D., Patierno, C., Fava, G.A., Guidi, J., 2020. The Hamilton Rating Scales for Depression: A critical review of clinimetric properties of different versions. Psychother. Psychosom. 89, 133-150. https://doi.org/10.1159/000506879.

Carrozzino, D., Svicher, A., Patierno, C., Berrocal, C., Cosci, F., 2019. The euthymia scale: a clinimetric analysis. Psychother. Psychosom. 88, 119-121. https://doi.org/10.1159/000496230.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Dagmar, A., Bode, R., Buysse, D., Choi, S., Cook, K., Devellis, R., DeWalt, D., Fries, J.F., Gershon, R., Hahn, E.A., Lai, J.S., Pilkonis, P., Revicki, D., Rose, M., Weinfurt, K., Hays, R., PROMIS Cooperative Group, 2010. 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. 63, 1179-1194. https://doi.org/10.1016/j.jclinepi.2010.04.011.

Christensen, K.S., Haugen, W., Sirpal, M.K., Haavet, O.R., 2015. Diagnosis of depressed young people - criterion validity of WHO-5 and HSCL-6 in Denmark and Norway. Fam. Pract. 32, 359-363. https://doi.org/10.1093/fampra/cmv011.

Christensen, K.S., Oernboel, E., Nielsen, M.G., Bech, P., 2019. Diagnosing depression in primary care: a Rasch analysis of the Major Depression Inventory. Scand. J. Prim. Health. Care. 37, 256-263. https://doi.org/10.1080/02813432.2019.1568703.

Demyttenaere, K., De Fruyt, J., 2003. Getting what you ask for: on the selectivity of depression rating scales. Psychother. Psychosom. 72, 61-70. https://doi.org/10.1159/000068690.

Deshpande, P. R., Rajan, S., Sudeepthi, B.L., Abdul Nazir, C.P., 2011. Patient-reported outcomes: A new era in clinical research. Perspect. Clin. Res. 2, 137–144. https://doi.org/10.4103/2229-3485.86879.

Deyo, R.A., Centor, R.M., 1986. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J. Chronic Dis. 39, 897-906. https://doi.org/10.1016/0021-9681(86)90038-X.

Fava, G.A., Belaise, C., 2005. A discussion on the role of clinimetrics and the misleading effects of psychometric theory. J. Clin. Epidemiol. 58, 753-756. https://doi.org/10.1016/j.jclinepi.2004.12.006.

Fava, G.A., Carrozzino, D., Lindberg, L., Tomba, E., 2018. The Clinimetric Approach to Psychological Assessment: A Tribute to Per Bech, MD (1942–2018). Psychother. Psychosom. 87, 321-326. https://doi.org/10.1159/000493746.

Fava, G.A., Ruini, C., Rafanelli, C., 2004. Psychometric theory is an obstacle to the progress of clinical research. Psychother. Psychosom. 73, 145-148. https://doi.org/10.1159/000076451.

Fava, G.A., Tomba, E., Brakemeier, E.L., Carrozzino, D., Cosci, F., Eöry, A., Leonardi, T., Schamong, I., Guidi, J., 2019. Mental pain as a transdiagnostic patient-reported outcome measure. Psychother. Psychosom. 88, 341-349. https://doi.org/10.1159/000504024.

Fava, G.A., Tomba, E., Sonino, N., 2012. Clinimetrics: the science of clinical measurements. Int. J. Clin. Pract. 66, 11-15. https://doi.org/10.1111/j.1742-1241.2011.02825.x.

Feinstein, A.R., 1982. The Jones criteria and the challenge of clinimetrics. Circulation. 66, 1-5. https://doi.org/10.1161/01.CIR.66.1.1.

Feinstein, A.R., 1987. Clinimetrics.Yale University Press, New Haven.

Fries, J.F., Bruce, B., Cella, D., 2005. The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. Clin. Exp. Rheumatol. 23, S53-57.

Ghisi, M., Flebus, G.B., Montano, A., Sanavio, E., Sica, C., 2007. BDI-II Manual. Giunti OS Organizzazioni Speciali, Firenze.

He, C., Levis, B., Riehm, K.E., Saadat, N., Levis, A.W., Azar, M., Rice, D.B., Krishnan, A., Wu, Y., Sun, Y., Imran, M., Boruff, J., Cuijpers, P., Gilbody, S., Ioannidis, J.P.A., Kloda, L.A., McMillan, D., Patten, S.B., Shrier, I., Ziegelstein, R.C., Akena, D.H., Arroll, B., Ayalon, L., Baradaran, H.R., Baron, M., Beraldi, A., Bombardier, C.H., Butterworth, P., Carter, G., Chagas, M.H.N., Chan, J.C.N., Cholera, R., Clover, K., Conwell, Y., de Man-van Ginkel, J.M., Fann, J.R., Fischer, F.H., Fung, D., Gelaye, B., Goodyear-Smith, F., Greeno, C.G., Hall, B.J., Harrison, P.A., Härter, M., Hegerl, U., Hides, L., Hobfoll, S.E., Hudson, M., Hyphantis, T.N., Inagaki, M., Ismail, K., Jetté, N.,

Khamseh, M.E., Kiely, K.M., Kwan, Y., Lamers, F., Liu, S.I., Lotrakul, M., Loureiro, S.R., Löwe, B., Marsh, L., McGuire, A., Mohd-Sidik, S., Munhoz, T.N., Muramatsu, K., Osório, F.L., Patel, V., Pence, B.W., Persoons, P., Picardi, A., Reuter, K., Rooney, A.G., da Silva Dos Santos, I.S., Shaaban, J., Sidebottom, A., Simning, A., Stafford, L., Sung, S., Tan, P.L.L., Turner, A., van Weert, H.C.P.M., White, J., Whooley, M.A., Winkley, K., Yamada, M., Thombs, B.D., Benedetti, A., 2020. The accuracy of the Patient Health Questionnaire-9 algorithm for screening to detect major depression: an individual participant data meta-analysis. Psychother Psychosom. 89, 25-37. https://doi.org/10.1159/000502294.

Irwin, D.E., Stucky, B., Langer, M.M., Thissen, D., DeWitt, E.M., Lai, J.S., Varni, J.W., Yeatts, K., DeWalt, D.A., 2010. An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. Qual. Life Res. 19, 595-607. https://doi.org/10.1007/s11136-010-9619-3.

Kellner, R., Sheffield, B.F., 1973. A self-rating scale of distress. Psychol. Med. 3, 88-100. https://doi.org/10.1017/S0033291700046377.

Konstantinidis, A., Martiny, K., Bech, P., Kasper, S., 2011. A comparison of the Major Depression Inventory (MDI) and the Beck Depression Inventory (BDI) in severely depressed patients. Int. J. Psychiatry. Clin. Pract. 15, 56-61. https://doi.org/10.3109/13651501.2010.507870.

Kristensen, S., Mainz, J., Baandrup, L., Bonde, M., Videbech, P., Holmskov, J., Bech, P., 2018. 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. 72, 409-419. https://doi.org/10.1080/08039488.2018.1492017.

Kroenke, K., Spitzer, R.L., 2002. The PHQ-9: a new depression diagnostic and severity measure. Psychiatr. Ann. 32, 509-515. https://doi.org/10.3928/0048-5713-20020901-06.

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

Kwan, A., Marzouk, S., Ghanean, H., Kishwar, A., Anderson, N., Bonilla, D., Vitti, M., Su, J., Touma, Z., 2019. Assessment of the psychometric properties of patient-reported outcomes of depression and anxiety in systemic lupus erythematosus. Semin. Arthritis Rheum 49, 260-266. https://doi.org/10.1016/j.semarthrit.2019.03.004.

Loevinger, J., 1947. A systematic approach to the construction and evaluation of tests of ability. Psychol. Monogr. 61, i-49. https://doi.org/10.1037/h0093565.

Millier, A., Clay, E., Charaf, I., Chauhan, D., Murthy, V., Toumi, M., Cadi-Soussi, N., 2014. Patient Reported Outcomes Instruments in Schizophrenia: A Review of Psychometric Properties. Open J. Med. Psychol. 3, 141-156. https://doi.org/10.4236/ojmp.2014.32017.

Mokken, R.J., 1971. A theory and procedure of scale analysis. Mouton, The Netherlands.

Mokkink, L.B., Terwee, C.B., Patrick, D.L., Alonso, J., Stratford, P.W., Knol, D.L., Bouter, L.M., de Vet, H.C.W., 2010. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J. Clin. Epidemiol. 63, 737-745. https://doi.org/10.1016/j.jclinepi.2010.02.006.

Montella, S., Baraldi, E., Cazzato, S., Aralla, R., Berardi, M., Brunetti, L.M., Cardinale, F., Cutrera, R., de Benedictis, F.M., di Palmo, E., Di Pillo, S., Fenu, G., La Grutta, S., Lombardi, E., Piacentini, G., Santamaria, F., Ullmann, N., Rusconi, F., Italian Pediatric Severe Asthma Network (IPSAN) on behalf of the Italian Society of Pediatric Respiratory Diseases (SIMRI), 2016. Severe asthma features in children: a case–control online survey. Ital. J. Pediatr. 42, 9. https://doi.org/10.1186/s13052-016-0217-z.

Naughton, M.J., Wiklund, I., 1993. A critical review of dimension-specific measures of health-related quality of life in cross-cultural research. Qual. Life. Res. 2, 397-432. https://doi.org/10.1007/BF00422216.

Nielsen, M.G., Ørnbøl, E., Vestergaard, M., Bech, P., Christensen, K.S., 2017. The construct validity of the Major Depression Inventory: A Rasch analysis of a self-rating scale in primary care. J. Psychosom. Res. 97, 70-81. https://doi.org/10.1016/j.jpsychores.2017.04.001.

Olsen, L.R., Jensen, D.V., Noerholm, V., Martiny, K., Bech, P., 2003. The internal and external validity of the Major Depression Inventory in measuring severity of depressive states. Psychol. Med. 33, 351-356. https://doi.org/10.1017/S0033291702006724.

Pallant, J.F., Tennant, A., 2007. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). Br. J. Clin. Psychol. 46, 1-18. https://doi.org/10.1348/014466506X96931.

Pilkonis, P.A., Choi, S.W., Reise, S.P., Stover, A.M., Riley, W.T., Cella, D., PROMIS Cooperative Group, 2011. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. Assessment. 18, 263-283. https://doi.org/10.1177/1073191111411667.

Reeve, B.B., Hays, R.D., Bjorner, J.B., Cook, K.F., Crane, P.K., Teresi, J.A., Thissen, D., Revicki, D.A., Weiss, D.J., Hambleton, R.K., Liu, H., Gershon, R., Reise, S.P., Lai, J.S., Cella, D., PROMIS Cooperative Group, 2007. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med. Care 45, S22-S31. https://doi.org/10.1097/01.mlr.0000250483.85507.04.

Rose, M., Devine, J., 2014. Assessment of patient-reported symptoms of anxiety. Dialogues. Clin. Neurosci. 16, 197-211.

Sheehan, D.V., Lecrubier, Y., Sheehan, K.H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., Dunbar, G.C., 1998. The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J. Clin. Psychiatry. 59, 22-33.

Tennant, A., Conaghan, P.G., 2007. 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. Care. Res. 57, 1358-1362. https://doi.org/10.1002/art.23108.

Terwee, C.B., Bot, S.D.M., de Boer, M.R., van der Windt, D.A.W.M., Knol, D.L., Dekker, J., Bouter, L.M., de Vet, H.C.W., 2007. Quality criteria were proposed for measurement properties of health status questionnaires. J. Clin. Epidemiol. 60, 34-42. https://doi.org/10.1016/j.jclinepi.2006.03.012.

Thomas, M.L., 2011. The value of item response theory in clinical assessment: a review. Assessment. 18, 291-307. https://doi.org/10.1177/1073191110374797.

Tomba E., Bech, P., 2012. Clinimetrics and Clinical Psychometrics: Macro- and Micro-Analysis. Psychother. Psychosom. 81, 333-343. https://doi.org/10.1159/000341757.

Topp, C.W., Østergaard, S.D., Søndergaard, S., Bech, P., 2015. The WHO-5 Well-Being Index: a systematic review of the literature. Psychother. Psychosom. 84, 167-176. https://doi.org/10.1159/000376585.

World Health Organization, 1993. The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic criteria for research (Vol. 2). World Health Organization, Geneva.

Zung, W.W., 1965. A self-rating depression scale. Arch. Gen. Psychiatry. 12, 63-70. https://doi.org/10.1001/archpsyc.1965.01720310065008.

Table 1. Model fit statistics for WHO-5, BDI-II, MDI (n = 100)

Scale	K	Model fit (overall)	Item fit residual, mean (SD)	Person fit residual, mean (SD)	PSI	Unidimensionality, significant *t*-tests (%)
WHO-5	5	χ2(10)=12.63, p=0.245	0.24 (0.80)	0.39 (1.09)	0.73	5.00
BDI-II	21	χ2(42)=26.37, p=0.97	-0.35 (0.62)	-4.23 (1.29)	0.65	24.00
MDI	10	χ2(20)=33,85, p=0.03	-0.37 (1.11)	-1.95 (1.32)	0.79	4.00

WHO-5: World Health Organization Well-Being Index; BDI-II: Beck Depression Inventory-II; MDI: Major Depression Inventory

K: number of items; χ2: chi-square; p: probability; SD: standard deviation; PSI: person separation index (with extremes)

Table 2. ROC statistics for the WHO-5, BDI-II, MDI

Scale	Observations	AUC	Standard error	95% CI
WHO-5	100	0.72	0.13	0.45-0.99
BDI-II	100	0.69	0.10	0.49-0.89
MDI	100	0.96	0.01	0.92-1.00

WHO-5: World Health Organization Well-Being Index; BDI-II: Beck Depression Inventory-II;

MDI: Major Depression Inventory

AUC: Area Under the Curve; CI: Confidence Interval

Table 3. Sensitivity and specificity of Major Depression Inventory cut-off values using the ICD-10 diagnostic algorithm

Cut-points	Sensitivity	Specificity	Correctly classified	LR+	LR-
≥ 0	100.00%	0.00%	3.00%	1.0000	-
≥ 2	100.00%	13.40%	16.00%	1.1548	0.0000
≥ 3	100.00%	20.62%	23.00%	1.2597	0.0000
≥ 4	100.00%	29.90%	32.00%	1.4265	0.0000
≥ 5	100.00%	35.05%	37.00%	1.5397	0.0000
≥ 6	100.00%	43.30%	45.00%	1.7636	0.0000
≥ 7	100.00%	45.36%	47.00%	1.8302	0.0000
≥ 8	100.00%	49.48%	51.00%	1.9796	0.0000
≥ 9	100.00%	56.70%	58.00%	2.3095	0.0000
≥ 10	100.00%	61.86%	63.00%	2.6216	0.0000
≥ 11	100.00%	71.13%	72.00%	3.4643	0.0000
≥ 12	100.00%	76.29%	77.00%	4.2174	0.0000
≥ 13	100.00%	79.38%	80.00%	4.8500	0.0000
≥ 14	100.00%	80.41%	81.00%	5.1053	0.0000
≥ 15	100.00%	84.54%	85.00%	6.4667	0.0000
≥ 17	100.00%	87.63%	88.00%	8.0833	0.0000
≥ 18	100.00%	90.72%	91.00%	10.7778	0.0000
≥ 19	100.00%	91.75%	92.00%	12.1250	0.0000

≥ 21	100.00%	92.78%	93.00%	13.8571	0.0000
≥ 22	100.00%	93.81%	94.00%	16.1667	0.0000
≥ 23	100.00%	95.88%	96.00%	24.2500	0.0000
≥ 24	66.67%	95.88%	95.00%	16.1667	0.3477
≥ 25	33.33%	95.88%	94.00%	8.0833	0.6953
≥ 28	33.33%	96.91%	95.00%	10.7778	0.6879
≥ 29	33.33%	97.94%	96.00%	16.1666	0.6807
≥ 30	0.00%	97.94%	95.00%	0.0000	1.0211
≥ 32	0.00%	98.97%	96.00%	0.0000	1.0104

LR: Likelihood ratios

Table 4. Sensitivity and specificity of World Health Organization Well-Being Index cut-off values using the ICD-10 diagnostic algorithm

Cut-points	Sensitivity	Specificity	Correctly classified	LR+	LR-
≥ 4	100.00%	0.00%	3.00%	1.0000	-
≥ 5	100.00%	3.09%	6.00%	1.0319	0.0000
≥ 6	100.00%	7.22%	10.00%	1.0778	0.0000
≥ 7	100.00%	11.34%	14.00%	1.1279	0.0000
≥ 8	100.00%	19.59%	22.00%	1.2436	0.0000
≥ 9	100.00%	30.93%	33.00%	1.4478	0.0000
≥ 10	100.00%	42.27%	44.00%	1.7321	0.0000
≥ 11	66.67%	57.73%	58.00%	1.5772	0.5774
≥ 12	66.67%	65.98%	66.00%	1.9596	0.5052
≥ 13	33.33%	77.32%	76.00%	1.4697	0.8622
≥ 14	33.33%	83.51%	82.00%	2.0208	0.7984
≥ 15	33.33%	88.66%	87.00%	2.9394	0.7519
≥ 16	33.33%	91.75%	90.00%	4.0417	0.7266
≥ 17	33.33%	92.78%	91.00%	4.6190	0.7185
≥ 18	33.33%	94.85%	93.00%	6.4667	0.7029
≥ 19	0.00%	97.94%	95.00%	0.0000	1.0211
≥ 22	0.00%	98.97%	96.00%	0.0000	1.0104

LR: Likelihood ratios

Table 5. Sensitivity and specificity of Beck Depression Inventory-II cut-off values using the ICD-10 diagnostic algorithm

Cut-points	Sensitivity	Specificity	Correctly classified	LR+	LR-
≥ 0	100.00%	0.00%	3.00%	1.0000	-
≥ 1	100.00%	23.71%	26.00%	1.3108	0.0000
≥ 2	100.00%	34.02%	36.00%	1.5156	0.0000
≥ 3	100.00%	42.27%	44.00%	1.7321	0.0000
≥ 4	100.00%	49.48%	51.00%	1.9796	0.0000
≥ 5	66.67%	57.73%	58.00%	1.5772	0.5774
≥ 6	66.67%	63.92%	64.00%	1.8476	0.5215
≥ 7	33.33%	72.16%	71.00%	1.1975	0.9238
≥ 8	33.33%	78.35%	77.00%	1.5397	0.8509
≥ 9	33.33%	81.44%	80.00%	1.7963	0.8186
≥ 10	33.33%	86.60%	85.00%	2.4872	0.7698
≥ 11	0.00%	87.63%	85.00%	0.0000	1.1412
≥ 12	0.00%	88.66%	86.00%	0.0000	1.1279
≥ 13	0.00%	90.72%	88.00%	0.0000	1.1023
≥ 14	0.00%	93.81%	91.00%	0.0000	1.0659
≥ 17	0.00%	94.85%	92.00%	0.0000	1.0543
≥ 20	0.00%	95.88%	93.00%	0.0000	1.0430
≥ 24	0.00%	97.94%	95.00%	0.0000	1.0211
≥ 29	0.00%	98.97%	96.00%	0.0000	1.0104

LR: Likelihood ratios