DOI: https://doi.org/10.11144/Javeriana.upsy16-2.ccsb

Constructing a Common Scale Between Tests of Depression: The use of Item Response Theory for Transferring of Norms from the BDI to EBADEP-A [*]

Construcción de una escala común entre pruebas de la depresión: el uso de la teoría de respuesta al ítem para la transferencia de normas del BDI al EBADEP-A

Received: 24 February 2015 | Accepted: 25 January 2017

Makilim Nunes Baptista[a]
University of São Francisco, Brasil
ORCID: http://orcid.org/0000-0001-6519-254X
Lucas de Francisco Carvalho
University of São Francisco, Brasil
Ricardo Primi
University of São Francisco, Brasil
Juliana Gomes Oliveira
University of São Francisco, Brasil
Jon D. Elhai
University of Toledo, Estados Unidos

ABSTRACT
This study of depression assessment aims to demonstrate the process of joint calibration and transfer standards between the internationally recognized BDI and EBADEP-A, a new instrument recently validated in Brazil. In addition to the illustration of methodological procedures, our study is intended to contribute to the elaboration of normative references for the latter instrument as an assessment of depression symptoms. We included 1666 participants divided into subgroups of patients and non-patients. The respondents answered the EBADEP-A and the Brazilian version of BDI. Data were analyzed using the Rasch-Andrich Rating Scale Model. We performed concurrent calibration of items of both instruments. Next, for each instrument we performed calibration with item parameters fixed based on prior analysis. Based on that, norms of the BDI were transfered to EBADEP-A. This procedure can be applied to any test that measures the same construct. This procedure produces a scale on the same metric.

Keywords
IRT; psychological assessment; humor disorders; psychometric properties; depression

RESUMEN
Este estudio de evaluación de la depresión tiene como objetivo demostrar el proceso de calibración y transferencia conjunta entre el BDI y el EBADEP-A, un nuevo instrumento validado en Brasil. Además de la ilustración de los procedimientos metodológicos, este estudio pretende contribuir a la elaboración de referencias normativas para este último instrumento como una evaluación de los síntomas de depresión. Se

[a] Correspondance author. E-mail: makilim01@gmail.com

How to cite: Nunes, M., F., De Francisco Carvalho, L., Primi, R., Gomes, J., & Elhai, J. (2017). Constructing a Common Scale Between Tests of Depression: The use of Item Response Theory for Transferring of Norms from the BDI to EBADEP-A. *Universitas Psychologica*, *16*(2), 1-11.
https://doi.org/10.11144/Javeriana.upsy16-2.ccsb

incluyeron 1666 participantes divididos en subgrupos de pacientes y no pacientes. Los encuestados respondieron al EBADEP-A y a la versión brasileña del BDI. Los datos se analizaron utilizando el Rasch-Andrich Rating Scale Model. Se realizó la calibración simultánea de los ítems de ambos instrumentos. A continuación, para cada instrumento realizamos la calibración con parámetros de ítem fijados basados en análisis previos. Basándose en eso, las normas del BDI fueron transferidas al EBADEP-A. Este procedimiento se puede aplicar a cualquier prueba que mida el mismo constructo, además produce una escala en la misma métrica.

Palabras clave
IRT; evaluación psicológica; trastornos del humor; propiedades psicométricas; depresión

Literature presents a considerable number of instruments for depressive symptoms assessment. Among the most commonly used tests, the Beck Depression Inventory ([BDI]; Beck & Steer, 1987) is one of the most used in the world (Santor, Gregus, & Welch, 2006). In Brazil, the number of instruments to measure depressive symptomatology available for professional use is very limited. Recently, the Baptist Depression Scale Adult Version (EBADEP-A) was developed; a self-report instrument, that together with BDI, is one of the only instruments to measure symptoms of depression in adults in the country. Considering the large number of publications with the BDI, an instrument already well established in the literature, this study aimed to develop cutoff points for EBADEP-A based on the standards developed for the BDI using Item Response Theory (IRT) procedures. In other words, we use mathematical procedures to transfer the BDI norms to the EBADEP-A.

Item Response Theory (IRT) can be considered as one of the representatives of this new trend in the field of psychological assessment and successor of the classic models (Embretson & Reise, 2000). According to Thomas (2011), IRT has some advantages over the classical models, such as reduction of measurement error; creation of computerized adaptive tests; detailed assessment of item bias; accuracy in the evaluation of changes after therapeutic interventions; fit index of persons and items according to the mathematical models; and also allows for calibration and equalization measures. These advantages render IRT as substantially advantageous over classic methods of measurement that are seriously limited without these new features.

In practice, for example, it is possible to use IRT procedures to generate item construct maps (as it is possible to see in Figure 1). This map allows the researcher to verify how the construct evolves in terms of intensity in the latent trait. Nay, IRT opens the range of analyzes for researchers in the area of measurement (e.g., Uebelacker, Strong, Weinstock, & Miller, 2009; Forkmann et al., 2009; Gibbons et al., 2011).

Inside this perspective, two or more tests measuring the same latent variable, such as depression symptoms, can be calibrated in a single measurement scale, because the calibration process enables separate item and person parameters. One of the advantages of creating a single scale for different instruments is related to standardization, since it allows to stablish cutoffs for a relatively new test based on a widely and well-know test. In other words, the researcher can use specific procedures, such as equating and item calibration, to transfer the norms of a test to other if both were measuring the same latent variable. This procedure allows for the building of a more cumulative science (Bauer & Hussong, 2009; Thomas, 2011).

Equating procedures (Smith et al., 2006; Wyse & Rechase, 2011) determine how two different instruments can be treated in the same measurement scale, allowing them to have the same statistical meaning for an examinee with the same ability level. Therefore, the scores resulting from equating procedure are considered interchangeable and equiproportional, even in two different tests that measure the same latent variable. For example, even with two instruments measuring aggression (i.e., the same latent variable), since the measurement scales are not equal (e.g., the first instrument ranges from 0 to 20, and the latter instrument ranges between 10 and 40), the instruments are not directly comparable because they are not at the same measurement level. After the equating and item

calibration procedures, giving the stablishment of a single measurement scale for both tests, they became directly comparable.

The establishment of cutoffs for new tests is even more important in countries where there are just few possibilities of tools for assessment. This is the case of Brazil in relation to the depression symptoms assessment, since the range of tests for the assessment of this group of symptoms is very limited. Only the Beck Depression Inventory (both versions, BDI and BDI-II), one of the most used worldwide instruments to evaluate depression symptoms, is adapted and standardized and can be used in clinical adult evaluation in Brazil.

Recently, a new country-developed test for symptom depression assessment was developed in Brazil, the Baptist Depression Scale Adult Version ([EBADEP-A]; Baptista, 2012). The EBADEP-A is a self-report scale, containing 45 items to be answered on a 4-point Likert scale. The EBADEP-A items are distributed as 33% assessing cognitive symptoms, 20% mood symptoms, 18% vegetative symptoms, 18% social symptoms, and 4.5% motor symptoms and irritability, which is quite different from BDI items that evaluates cognitive symptoms (52%), vegetative symptoms (29%) and mood symptoms (9.5%). Besides that, the EBADEP-A is more appropriate to evaluate depression symptoms from the mild to moderate range, almost reaching the severe level of depression symptoms; and, BDI is more adjusted to measure more severe symptoms of depression (Baptista, 2012; see Kendall, Hollon, Beck, Hammen, & Ingram (1987) for a BDI use discussion). A series of studies with the EBADEP-A demonstrated validity evidence and adequate reliability (Baptista & Carneiro, 2011; Baptista & Gomes, 2011; Baptista, Carneiro & Sisto, 2010). The instrument has been approved for clinical and research use by a committee of experts in psychometrics in Brazil (Conselho Federal de Psicologia [CFP], 2014).

This study aims to demonstrate the process of joint calibration and transfer standards between the internationally recognized BDI and EBADEP-A, a new instrument recently validated in Brazil. In addition to the illustration of methodological procedures, it is intended to contribute to the elaboration of normative references for the latest instrument, in an effort to improve the cross-cultural assessment of depression.

Method

Participants

The study included 1666 participants, selected by convenience, with a minimum of eight years of educational attainment, divided into 6 subgroups: 1311 college students (normative sample), 40 patients with a major depressive disorder diagnosis (depressed patients), 40 subjects without a major depressive disorder diagnosis (non-depressed control group), 100 inpatients from a general hospital suffering of Crohn's disease, 100 companions of the subjects with Crohn's disease, and 75 patients diagnosed (by a psychiatric clinician and Structured Clinical Interview for DSM-IV Axis I (SCID-CV) with psychiatric disorders including depressive disorder diagnosis as principal disorder (49%) or comorbidity (51%) (for details relative to this sample, see Baptista (2012)). The normative sample was composed of 1082 graduate students who denied ever being diagnosed with a depressive episode in their lives. Considering the equating procedure (better explained in Procedure and Data Analysis), among the 1082 graduate students, 308 (equal in terms of gender) answered both instruments, the EBADEP-A and BDI. From that, the total sample was equalized. Table 1 presents the demographic data about the subgroups.

TABLE 1
Demographic data of six subgroups.

Groups	Gender	N (%)
Normative Sample	Men	316 (29.2%)
	Women	766 (70.8%)
Depressed patients	Men	6 (15%)
	Women	34 (85%)
Non-depressed control group	Men	6 (15%)
	Women	34 (85%)
Inpatients	Men	48 (48%)
	Women	52 (52%)
Accompanying	Men	48 (48%)
	Women	52 (52%)

Source: own work.

Instruments

Depressions symptoms were assessed using the Baptist Depression Scale Adult Version ([EBADEP-A]; Baptista, 2012) and the Brazilian version of the Beck Depression Inventory (Cunha, 2001). We note that the BDI was used rather than the BDI-II because when data collection was performed, the Brazilian version of the BDI-II was under development.

The EBADEP-A is a self-report inventory used for tracking depression symptomatology in psychiatric and non-psychiatric samples. The scale was developed according to depression models, such as the Becks's Cognitive Model (Beck, Rush, Shaw, & Emery, 1979) and Behavioral Model (Ferster, Culbertson, & Boren, 1977), and manuals such as the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* ([DSM-IV-TR]; American Psychiatric Association [APA], 2002) and the tenth edition of the *International Statistical Classification of Diseases and Related Health Problems* ([ICD-10]; Organização Mundial de Saúde [OMS], 1993). This scale consists of 90 questions presented in pairs, deriving 45 items. Each item is a depression symptomatology marker represented by a positive and negative statement. Each item must be answered on a specific 4-point Likert scale, with a minimum score of zero and maximum of 135. Regarding to the interpretation, the lower the score, the lower the depression symptomatology. Several studies were conducted based on CTT and IRT, showing suitable validity evidences and favorable reliability for EBADEP-A, specifically, the items set showed evidences for unidimensionality, with internal consistency reliability (alpha) of 0.94, and correlation of r = 0.75 with the BDI total score (Baptista, 2012; Baptista, Souza, & Alves, 2008; Baptista, Souza, Gomes, Alves, & Carneiro, 2012; Baptista & Gomes, 2011: Baptista et al., 2010; Carvalho, Primi, & Nunes Baptista, 2015). We also administered the BDI, an instrument to measure the intensity of depression. The total BDI score is obtained from the sum of the scores of the answers marked by examinees across the 21 items. The official Portuguese version of the instrument was used. In the adaptation to Brazil, Cunha (2001) found an alpha coefficient of 0.82 for the BDI in a sample of 1746 college students. Besides that, internal structure validity evidences and external validity were found in the Brazilian version of BDI.

Procedure and Data Analysis

This study was approved by the Ethics in Research Committee for Data Collection, and the Free and Informed Consent (IC) was presented to all participants. The instruments were administered collectively in classrooms with up to 40 college students per classroom.

Data were analyzed using the Rasch-Andrich Rating Scale Model (Wright & Masters, 1982). In this model the probability of choosing a specific Likert category $P_{ijx}(\theta_j)$, meaning the probability of a person j present score x in i item, is given by (Embretson & Reise, 2000) $P_{ijx}(\theta_j) = \frac{\exp\left(\sum_{j=0}^{x}\left[\theta_j - (\lambda_k + b_i)\right]\right)}{\sum_{x=0}^{m}\exp\left(\sum_{j=0}^{x}\left[\theta_j - (\lambda_k + b_i)\right]\right)}$. A distinctive feature of the Rating Scale Model is that these scalar intervals between points are relatively similar for all items. The difficulty parameter b_i represents the location of item i, or the average intensity of the thresholds of an item. Items that represent extremes in the latent dimension are represented with high

average thresholds because their thresholds are all located on the most intense theta levels.

Item and subject model parameters were calibrated by the Joint Maximum Likelihood Estimation method implemented in the Winsteps software (Linacre, 2011). This calibration was performed considering the items in the BDI and EBADEP jointly forming a single depression scale. The model parameters were estimated for the items (thresholds) and for the respondents. For each item of the Brazilian version of the BDI b_i values and three thresholds (λ_1, λ_2, λ_3) were estimated. For the EBADEP-A also b_i values and three thresholds were estimated. The fit of this calibration was assessed by the fit indexes, *infit* and *outfit*, that were calculated for all the items and subjects. These values are directly proportional to the residuals that reflect differences between the observed and expected responses as hypothesized from knowledge of the model parameters, thus providing evidence of how well the model fits the data. The *outfit* value is obtained by dividing the chi-square value by the degrees of freedom. The value for the degrees of freedom either is the number of subjects when the index is calculated for items, or the number of items when the index is calculated for subjects. Values greater than 1.3 indicate a misfit (Wright & Linacre, 1994). Thus, calibration with this analysis enabled a common metric between the scales. To enable calibration, the model requires that the theta mean or the difficulty *(b)* mean is fixed. We used the Winsteps default, i.e., the *b* mean was fixed to zero (which stands for an arbitrary zero, but not for an absolute zero).

Item linking and person equating that permitted the transfer of norms from the BDI to EBADEP-A was carried out in three steps. First, items from both instruments were calibrated concurrently. This calibration is known in IRT as common group equating. This process places item parameters in a common metric linking the items of the BDI to the EBADEP-A. In the second step, each instrument was calibrated separately, but, this time, fixing item parameters with the values found in Step 1. At this time, because items parameters are in a common metric, the two estimated subject theta parameters from the BDI or EBADEP-A are equated and reported in the same metric. Therefore, in the third step, each score table that maps total scores to thetas was examined to transfer expected cut points available for BDI total scores, reported in its manual and indicating subclinical, mild, moderate, and severe depression, to the EBADEP-A. This transfer is conducted by finding the theta value associated with each BDI total score point of interest and then, in the EBADEP-A table, doing the reverse, finding the total score associated with those theta values. With this procedure we can transfer these cut points between total scale scores.

Results

The first step was to perform a concurrent calibration of the BDI and EBADEP-A items (all protocols were answered completely). The total score on the BDI was M=7.1, SD=6.8 (N=329) and for the EBADEP-A was M=49.8, SD=28.0 (N=1069). The correlation between them was .70, indicating convergent validity for both scales. Even so, these raw scores are hardly comparable. The calibration of Rasch-Andrich Rating Scale model parameters was performed in WINSTEPS (Linacre, 2011). The 66 items (EBADEP-A and BDI) were calibrated concomitantly. Each test was allowed to have its own rating scale structure. For EBADEP-A the parameters were λ_1 =-.13, λ_2 =-.13 λ_3 =.26. For BDI the parameters were λ_1 =-.46, λ_2 =.74 λ_3 =-.28. The second and third thresholds of the BDI were not ordered. This is related to the low frequency of 2 points and indicates that points 2 and 3 are informing the same level of theta. At the same time the thresholds of the BDI are more dispersed than for the EBADEP-A. The summary results of the concurrent calibration are presented in Table 2.

TABLE 2

Descriptive statistics of the items and thetas parameters, and fit indices.

Items			
	b	S.E.	*Infit*
Mean (SD)	0 (0.75)	0.06 (0.04)	1.02 (0.37)
Maximum	1.97	0.18	2.51
Minimum	-1.59	0.03	0.50
Participants			
	theta	S.E.	*Infit*
Mean (SD)	-0.96 (0.71)	0.18 (.04)	1.05 (0.39)
Maximum	1.17	0.56	2.60
Minimum	-3.43	0.13	0.22

Source: own work.

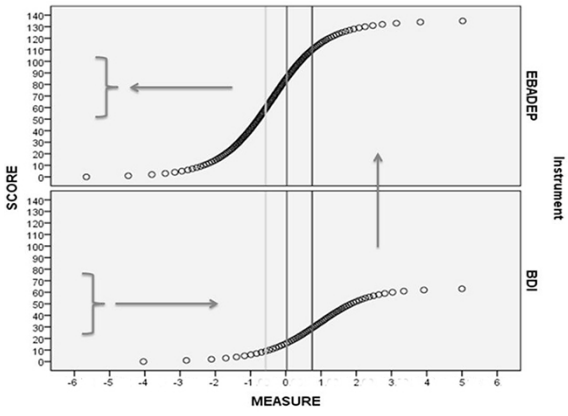
According to Table 2, the parameters of item difficulty (average of thresholds for each item) varied between -1.59 and 1.97 demonstrating that the items cover a wide range of the construct. The average fit indexes for the items and participants were shown to be adequate. However, twelve items showed *infit* and/or *outfit* indexes higher than expected, i.e., items 2, 3, 70, and 74 (EBADEP-A) and 2, 10, 11, and 19 (BDI) obtained both, *infit* and *outfit*, indexes above 1.30; item 50 (EBADEP-A) and 6 (BDI) obtained *infit* indexes above 1.30; and items 65 and 67 (EBADEP-A) obtained *outfit* indexes above 1.30 (Wright & Linacre, 1994). In addition, just twelve items showed item-theta correlations less than 0.40 and the average of correlations was .48. In general, results indicated an adequate fit for the majority of items. The average level of the latent trait was M = -0.96. Overall items tend to be difficult for people in the sample to endorse as is expected for this scale's presenting symptoms. The reliability of the theta estimates calculated by the Rasch Model were 0.92 (real value) and 0.94 (model value), which can be considered as very satisfactory.

The construct map including EBADEP-A and BDI items was generated, showing item expected scores related to the level on theta. This makes it possible to verify the construct representation by both instruments. In general, the map showed that BDI items tend to be more difficult to endorse by respondents than EBADEP-A items. BDI items seem to evaluate the latent construct (depressive symptoms) in more severe levels compare to EBADEP-A items. From this, we can see that, in general, BDI items tend to be more difficult for endorsement by participants in relation to items of the EBADEP-A. This suggests that the BDI assesses the latent construct (depressive symptoms) at levels more stringent than the EBADEP-A.

Next, for each instrument we performed calibration with item parameters fixed based on prior analysis. Thus the calibration of EBADEP-A items was fulfilled fixing items parameters according to the parameters found in previous analysis; the same procedure was done with the BDI items. With this procedure the estimated values of theta for participants were equated and obtained on the same metric, allowing the comparison between theta of both instruments. We obtained two conversion tables, one for each instrument that indicates for each raw score the corresponding equated theta scale. These conversion tables are based on Test Characteristic Curves (TCC) that show the relationship between theta and expected total raw scores on each instrument. Therefore, at the next step, we transferred the criterion-referenced normative expectation – cutoffs that were discovered in a Brazilian normative study (Baptista, 2012) – for the EBADEP-A scale.

There were three cutoffs separating the categories for minimal, mild, moderate, and severe depression. First, we converted these cutoffs from BDI raw scale scores to their corresponding theta values using the BDI conversion table. Then we used the EBADEP-A conversion table to obtain raw scores that corresponded to the theta values. Figure 1 shows the conversion process. It shows TCCs for the BDI and EBADEP-as well as cutoff values.

Figure 1.
BDI and EBADEP-A Test Characteristic Curves and cutoff transferring process



Source: own work.

The *theta* values corresponding to BDI cutoffs can be verified through Figure 1 (the three vertical lines). Minimal symptomatology of depression (ranging from a score of zero to nine), suggests the minimal-mild threshold equals 9 (i.e., a cutoff of 9 separates minimal from mild depression), where *theta* is equal -0.57; mild depression (scores ranging from 10 to 16) with a mild-moderate threshold (cutoff) of 16 and a *theta* value equal to 0.03; moderate depression (scores of 17 to 29) with a threshold of 29 separating moderate to severe depression and equivalent to a *theta* value of 0.75; and, severe depression (score 30 to 63) with theta values above 0.75. In the figure the arrows indicate the transferring process that starts from raw scores, identified in normative studies of the BDI. The raw scores were converted to theta scores through the BDI's Test Characteristic Curve; based on those theta scores, the EBADEP-A's raw scores were converted to theta scores. As a direct product of this procedure, Table 3 shows the corresponding equivalent raw scales resulting from this process.

TABLE 3
Conversion table of normative references from BDI to EBADEP-A

Instruments	Mininal depression	Mild depression	Moderate depression
BDI	0 to 9	10 to 16	17 to 29
EBADEP-A	0 to 59	60 to 85	86 to 110

Source: own work.

We next present Table 3 that is based on the EBADEP-A manual (Baptista, 2012). The table shows the distribution of the EBADEP-A normative reference group and selected groups from the validity studies (depressive patients, non-depressed control group matched to the depressed patients, inpatients, participants accompanying inpatient, and psychiatric patients). We tested the transference of norms by comparing the distributions of depressive patients as compared to the non-depressed control group and also as compared with the normative sample. If the norm transferring is successful it will be able to differentiate depressed patients from other groups. Therefore this tests the criterion validity of the EBADEP-A using the norms that were transferred from the BDI.

The distributions of participants across the four levels of depression were highlighted (bold) for the entire sample, for the clinical group of people diagnosed with depression and for the control group (non-depressed). Most of clinically depressed individuals would be categorized in the moderate depression range (47.5%), followed by equal percentages in the mild and severe categories (22.5%) and, lastly, 7.5% would be categorized as not depressed against 70.8% in the normative reference group and 97.5% in the non-depressed control group. The statistical test comparing the distributions across the four categories of depressive patients with the control group showed a very large effect: $\chi^2 = 65.3$, $df = 3$, Somer's $d = 0.74$, Spearman $r = 0.87$, $p < 0.001$; and for the depressive patients with normative reference group the tests showed a moderate effect: $\chi^2 = 201.3$, $df = 3$, Somer's d = 0.09, Spearman $r = 0.33$, $p < 0.001$. This relative lower effect is expected because in the normative group it will be expected that a proportion of

the sample will show signs of mild and moderate depression. This was not the case for the control group that was systematically selected in order to include only healthy individuals. Therefore these results show positive validity evidence for the EABADEP-A and their new normative criteria transferred from BDI.

TABLE 4

Normative groups distribution according to BDI categories in EBADEP-A by equating.

Groups	N and %	BDI categories in EBADEP-A		
		Minimal	Mild	Moderate
Normative Sample	N	928	282	89
	%	**70.8%**	**21.5%**	**6.8%**
Validity Sample				
Depressed patients	N	3	9	19
	%	**7.5%**	**22.5%**	**47.5%**
Non-depressed control group	N	39	1	0
	%	**97.5%**	**2.5%**	**0%**
Inpatients	N	56	29	14
	%	56.0%	29.0%	14.0%
Accompanying	N	76	18	6
	%	76.0%	18.0%	6.0%
Psychiatric patients	N	39	16	18
	%	52.0%	21.3%	24.0%

Source: own work.

Considering the availability of the criterion information in the present study we also performed a Receiver Operation Curve (ROC) analysis trying to identify the optimal cut score for the EBADEB-A, which can identify the clinically depressed individuals as compared to the control and normative groups. Because the BDI's cut scores were themselves based on criterion validity studies done by Cunha (2001), this study is a replication and enhancement of the earlier studies as well as an enhancement for the EBADEP-A criterion-referenced interpretations.

We performed two ROC analyses - one contrasting the clinically depressed group with the control group and other with the normative group. By analyzing the coordinates of ROC curves, the first analysis (depressed vs non-depressed control group) showed an overall area under the curve of 98% ($p < 0.001$). A cut score of 66 would result in a sensitivity index of 90% and specificity of 97.5%. The second analysis (depressed vs normative reference group) resulted in an overall area under the curve of 91.8% ($p < 0.001$). A cut score of 77 would result in a sensitivity index of 80% and specificity of 88%.

A score of 77 corresponds to mild depression according to the criterion-referenced interpretations transferred from the BDI. The actual criteria of 86 for moderate depression results in a reduction of sensitivity to 68% (specificity of 100% in the control group, and 93% in the normative reference group). This reduction in specificity is due to the fact that some patients have scores lower than 86. Therefore, the adjustment of the cut-score that separates mild from moderate depression will improve the capacity of the EBADEP-A to identify depressive patients in the category of moderate depression. These patients would, otherwise, be placed in the mild depressed category when using the actual cut-score. So, this cutoff was revised leading to the following ranges: 0 to 59 (minimal depression), 60 to 76 (mild depression), 77 to 110 (moderate depression) and 111 or higher (severe depression). Table 5 shows the new sample distributions across these four levels of depression with the new revised cut-scores.

TABLE 5

Normative groups distribution according to BDI categories in EBADEP-A by equating after category relocation based on ROC curves.

Groups	N and %	BDI categories in EBADEP-A		
		Minimal	Mild	Moderate
Normative Sample	N	928	217	154
	%	70.8%	16.6%	11.7%
Validity Samples				
Depressive patients	N	3	5	23
	%	**7.5%**	**12.5%**	**57.5%**
Non-depressed	N	39	1	0
	%	97.5%	2.5%	0%
Inpatients	N	56	16	27
	%	56.0%	16%	27.0%
Accompanying	N	76	13	11
	%	76%	13%	11.0%
Psychiatric patients	N	39	10	24
	%	52%	13.3%	32%

Source: own work.

Discussion

Overall, this study aimed to demonstrate the process of joint calibration and transfer standards between the BDI, an internationally recognized depression instrument, and EBADEP-A, a new instrument recently developed and validated in Brazil. In addition, this kind of objective contributes to the elaboration of normative references for the EBADEP-A. As indicated by

Thomas (2011) IRT is a good tool to develop new instruments in the mental health field. The methodology used in the present study is a propitious and relevant tool when one has a gold standard scale (as BDI) and wants to compare and transfer its standards to an instrument (e.g., EBADEP-A) in its initial development/validation states.

The BDI was developed initially to assess persons with depressive pathology (Beck & Steer, 1987) and probably this explains the finding that when both scales are placed into an item constructing map, the BDI items evaluated more severe symptomatology and the EBADEP-A evaluated the mild and moderate ones. This is clinically useful information because the EBADEP-A, while perhaps not serving as well as the BDI in assessing more severe depression symptoms, could be more useful as a screening tool in more general samples. For example, research demonstrates that the primary care health sector is more often accessed by persons presenting emotional problems than the mental health specialty sector is (Wang et al., 2006); also, patients report preferring to discuss mental health problems with their primary care physician rather than visiting a mental health professional (Del Piccolo, Saltini, & Zimmerman, 1998). Thus in primary care settings, depression will not likely be as prevalent as in at-risk mental health settings, but it is still worthwhile to screen for depression among the minority of individuals presenting with such symptoms, and the EBADEP-A may be useful in this regard. As pointed out by Uebelacker, Strong, Weinstock, & Miller (2009), IRT also could provide information about peculiarities of expression of symptoms.

The procedure presented here can be applied to any test that measures the same construct of depression as was the case of EBADEP-A. With these cases a marker instrument that is the BDI is used to anchor the cut points that are used as an aid to the diagnosis. It would only be necessary to apply the new instrument and the BDI together and to calibrate jointly both instruments fixing the parameters of the BDI. This procedure will produce a scale on the same metric.

In particular, this kind of research is extremely important because, until now, there has been no scale that measures depression symptoms that was developed, validated, and normed in Brazil. It is possible to see, by equating, how two different instruments can be treated in the same measurement scale and cover several levels of symptomatology (Smith et al., 2006). Considering the clinical point of view, through the use of measures assessing the same construct (depression symptoms) at a different level, more information can be aggregated by the clinician in terms of determining the relative localization of the patient in the latent construct.

Two main limitations of this study should be pointed out. The first relates to the equating procedure, which can add more biases than cases where all subjects respond to all items of the tests. Future research should check whether the data encountered replicates in samples answering all tests completely. The second limitation relates to the size of the clinical sample, compared to the healthy sample is much lower. Future research should continue for this type of study using larger clinical samples.

References

American Psychiatric Association [APA] (2002). *Diagnostic and statistical Manual of mental Disorders*. (4th Ed.) Text Revision. Washington, DC.

Baptista, M. N., Souza, M. S., & Alves, G. A. S. (2008). Evidências de Validade entre a escala de Depressão (EDEP), o BDI e o Inventário de Percepção de Suporte Familiar (IPSF). *PSICO-USF, 13*(2), 211-220.

Baptista, M. N., Carneiro, A. M., & Sisto, F. F. (2010). Estudo Psicométrico de Escalas de Depressão (EDEP e BDI) e o Inventário de Percepção de Suporte Familiar (IPSF). *Psicologia em pesquisa (UFJF), 4*, 65-73. https://doi.org/10.1590/S1413-82712012000300007

Baptista, M. N., & Carneiro, A. M. (2011). Validade da Escala de Depressão (EDEP):

relação com ansiedade (BAI) e estresse laboral (EVENT). *Estudos de Psicologia* (PUCCAMP. Impresso), *28*(3), 345-352.

Baptista, M. N., & Gomes, J. O. (2011). *Escala Baptista de Depressão* (Versão Adulto) - EBADEP-A: evidências de validade de construto e de critério. *Psico-USF* (Impresso), 16, 151-161.

Baptista, M. N. (2012). *Escala Baptista de Depressão* (*Versão Adulto* - EBADEP-A. Vetor Editora: São Paulo – Brazil.

Bauer, D. J., & Hussong, A. M. (2009). Psychometric Approaches for Developing Commensurate Measures Across Independent Studies: Traditional and New Models. *Psychological Methods, 14*(2), 101-125. https://doi.org/10.1037/a0015583

Beck, A. T., & Steer, R. A. (1987). *Beck Depression Inventory manual*. San Antonio, TX: Psychological Corporation.

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression*. New York: Guilford.

Carvalho, L. F., Primi, R., & Nunes Baptista, M. N. (2015). IRT Application to verify psychometric properties of the Beck Depression Inventory (BDI), *Universitas Psychologica, 14*(1), 91-102.

Conselho Federal de Psicologia [CFP] (2014). Sistema de Avaliação de Testes Psicológicos (SATEPSI). Lista de testes com parecer favorável. Disponível em: http://www2.pol.org.br/satepsi/sistema/admin.cfm?lista1=sim . Acesso em 22 jan. 2014.

Cunha, J. (2001). *Manual em português das Escalas Beck*. São Paulo: Casa do Psicólogo.

Del Piccolo, L., Saltini, A., & Zimmerman, C. (1998). Which patients talk about stressful events and social problems to the general practitioner? *Psychological Medicine, 28*, 1289-1299. https://doi.org/10.1111/j.1600-0447.1983.tb09716.x

Embretson S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah: Lawrence Erlbaum.

Ferster, C. B., Culbertson, S., & Boren, M. C. (1977). *Princípios do comportamento*. São Paulo: Hucitec.

Forkmann, T., Boecker, M., Wirtz, M., Eberle, N., Westhofen, M., Schauerte, P. ... Norra, C. (2009). Development and validation of the Rasch-based Depression Screening (DESC) using Rasch analysis and structural equation modelling. *Journal of Behavior Therapy and Experimental Psychiatry, 40*(3), 468-478. https://doi.org/10.1016/j.jbtep.2009.06.003

Gibbons, L. E., Feldman, B. J., Crane, H. M., Mugavero, M., Willig, J. H., Patrick, D. … Crane, P. K. (2011). Migrating from a legacy fixed-format measure to CAT administration: Calibrating the PHQ-9 to the PROMIS depression measures. *Quality of Life Research, 20*(9), 1349-1357. https://10.1007/s11136-011-9882-y

Kendall, P. C., Hollon, S. D., Beck, A. T., Hammen, C. L., & Ingram, R. E. (1987). Issues and recommendations regarding use of the Beck Depression Inventory. *Cognitive Therapy and Research, 11*(3), 289-299. https://doi.org/10.1023/A:1021233215457

Linacre, J. M. (2011). *Winsteps*® (Version 3.72.3) [*Computer Software*]. Beaverton, Oregon: Winsteps.com. Retrieved January 1, 2011 from http://www.winsteps.com/

Organização Mundial de Saúde [OMS] (1993). *Classificação de Transtornos Mentais e de Comportamento da CID-10: descrições clínicas e diretrizes diagnósticas*. Porto Alegre: Artes Médicas.

Santor, D. A., Gregus, M., & Welch, A. (2006). Eight Decades of Measurement in Depression. *Measurement, 4*(3), 135-155. https://doi.org/10.15689/ap.2015.1401.06

Smith, A. B., Wright, E. P., Rush, R., Stark, D. P., Velikova, G., & Selby, P. J. (2006). Rasch analysis of the dimensional structure of the Hospital Anxiety and Depression Scale. *Psycho-Oncology, 15*, 817-827. https://doi.org/10.1002/pon.1015

Thomas, M. L. (2011). The Value of Item Response Theory in Clinical Assessment: A

Review. *Assessment, 18*(3), 291-307. https://doi.org/10.1177/1073191110374797

Uebelacker, L. A., Strong, D., Weinstock, L. M., & Miller, W. (2009). Use of item response theory to understand differential functioning of DSM-IV major depression symptoms by race, ethnicity and gender. *Psychological Medicine, 39*, 591-601. https://doi.org/10.1017/S0033291708003875

Wang, P. S., Demler, O., Olfson, M., Pincus, H. A., Wells, K. B., & Kessler, R. C. (2006). Changing profiles of service sectors used for mental health care in the United States. *American Journal of Psychiatry, 163*, 1187-1198.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA.

Wyse, A. E., & Reckase, M. D. (2011). A Graphical Approach to Evaluating Equating Using Test Characteristic Curves. *Applied Psychological Measurement, 35*(3), 217-234. https://doi.org/10.1177/0146621610377082

Notes

* Research article