IRT Application to Verify Psychometric Properties of the Beck Depression Inventory (BDI)*

Aplicación de la TRI para verificar las propriedades psicométricas del Inventario de Depresión de Beck

Received: August 29th, 2013 | Revised: January 15th 2014 | Accepted: June 15th 2014

Lucas de Francisco Carvalho**
Ricardo Primi
Makilim Nunes Baptista
Universidade São Francisco, Brazil

Abstract

The aim of this study was to evaluate the performance of the Brazilian version of the Beck Depression Inventory (BDI) using Rasch-based person and item analysis. For this purpose, 271 participants were recruited, between 18 and 51 years of age (M=23.61; *SD*=6.12), 187 (69%) female and 84 men, all Brazilian college students. Participants responded to the BDI on the assessment of depressive symptoms. Results suggest the adequacy of the psychometric properties of the instrument and demonstrate the Rasch model's applicability for clinical practices. Among the important tools offered by the Rasch model, we explore the use of the person-item map, which visually presents the intuitively understandable psychological construct along the dimensional scale of the instrument.

Keywords
Item response theory; psychometric properties; affective disorders; validity

Resumen

El objetivo de este estudio fue evaluar el desempeño de la versión brasileña del Inventario de Depresión de Beck (BDI) usando el modelo de Rasch. Para ello 271 participantes fueron reclutados, entre 18 y 51 años de edad (M = 23,61; SD = 6.12), 187 (69%) mujeres y 84 hombres, todos son estudiantes universitarios brasileños. Los participantes respondieron a la BDI sobre la evaluación de los síntomas de la depresión Los resultados sugieren la adecuación de las propiedades psicométricas del instrumento y demuestran la aplicabilidad del modelo de Rasch en prácticas clínicas. Entre las herramientas más importantes que ofrece el modelo de Rasch se explora el uso del mapa persona-artículos, que presenta visualmente la construcción psicológica intuitivamente comprensible a lo largo de la escala dimensional del instrumento.

Palabras clave
teoría de respuesta al ítem; propiedades psicométricas; trastornos afectivos; validez

doi:10.11144/Javeriana.upsy14-1.irta

To cite this article: Carvalho, L. de F., Primi, R., & Nunes Baptista, M. (2015). IRT application to verify psychometric properties of the Beck Depression Inventory (BDI). *Universitas Psychologica, 14*(1), 91-102. http://dx.doi.org/10.11144/Javeriana.upsy14-1.irta

[*] Original research article

[**] E-mails: Lucas@labape.com.br, rprimi@mac.com, makilim01@gmail.com

Introduction

Depression is a construct typically studied and evaluated by health professionals, especially psychologists and psychiatrists. Epidemiological studies show the prevalence of depression in the general population, ranging between 3% and 11% (Kessler et al., 2003) and, specifically in Brazil, the studies indicate the prevalence of depression between 2.8% and 19.2% (Theme-Filha, Szwarcwald, & Souza-Junior, 2005).

Generally, the term depression refers to specific symptoms present in various disorders or it is configured as a disorder itself. In the latter case, it is a set of diagnostic criteria or symptoms. In adults, according to the Diagnostic and Statistical Manual of Mental Disorders [DSM-IV-TR] (APA, 2013), major depressive disorder is characterized by five or more symptoms including necessarily the depressed mood and/or loss of interest or pleasure, at least two weeks and sometimes is related with other constructs like hopelessness, suicide ideation, social support and others (Baptista, Carneiro, & Cardoso, in press).

The literature is quite poor about the relatively position between the depression symptoms in terms of severity. One study (Castro, Trentini, & Riboldi, 2010), using the 2 parameter model based on Item Response Theory (IRT), evaluated 4025 subjects (psychiatric patients, *N*=1138; medical patients, *N*=490, and non-clinical subjects, *N*=2397) verified the hierarchical BDI items order related to the severity of depression symptom in item content. Through the Parscale software using the Graded-Response model, they established a specific hierarchical items order (see Table 2).

Also, there are are some studies that have attempted to identify the main symptoms of depression. Some findings suggested that the depressed mood and lack of interest in activities are core features of a major depressive episode (Kennedy, 2008; Nelson, Portera, & Leon, 2006), although psychic anxiety and guilt (Nelson, Portera, & Leon, 2006), as well as sleep disturbance, anhedonia, low self-steam and change in appetite are also pointed out (Brody et al., 1998).

Especially in the assessment of depression, it is noteworthy that although there are various instruments measuring depressive symptoms, one can expect a huge variation in several measures (Bauer & Hussong, 2009). Santor, Gregus and Welch (2006) conducted a survey between 1918 and 2008, selecting 280 evaluation measures of depression. They found differences between the measures in regard to the response format, content and objectives, identification theory, number of items to assess symptoms or specific components of depression, and other features.

In this field, the Beck Depression Inventory (BDI) is one among the most commonly used instruments to measure depression intensity. Originally created by Beck, Ward, Mendelson, Mock and Erlbaum (1961), the test was adapted and validated to different countries and cultures. In this study the BDI Brazilian version (Cunha, 2001) was applied. The BDI is a scale for measuring depression intensity and is not a diagnostic instrument. Each of the twenty-one items corresponds to a particular, putative symptom of depression, and is paired with a 4-point Likert response scale. In its structure, the BDI has a subgroup of cognitive-affective items (Cognitive-Affective subscale) and another that includes somatic and performance complaints (Somatic and Performance subscale).

A number of studies have attempted to show the joint validity of the BDI on patient and non-patient samples in terms of validity (e.g., Contreras, Fernandez, Malcarne, Ingram, & Vaccarino, 2004; Lykke, Hesse, Austin, & Oestrich, 2008). Specifically in relation to the Brazilian version (Cunha, 2001), there are favorable validity evidences (internal structure and based on external variables) and suitable reliability indexes (ranging between 0.79 and 0.91 with patients and 0.7 and 0.86 with non-patients).

Worth pointing out that the BDI was originally developed using Classical Test Theory (CTT). However, over the past few years, it has become increasingly common to find studies (Castro, Trentini, & Riboldi, 2010; Hammond, 1995) using an alternative mathematical model to evaluate and guide scale development, namely, Item Response

Theory (IRT). The IRT has emerged from criticisms to the classical model, mostly, concerning to the assumptions of CTT, that create problems known in the social sciences as arbitrary metrics (Embretson, 2006).

Typically, psychological tests are interpreted with reference standards, which give meaning to test scores by comparing them to normative groups. Although the importance of such information is recognized, normative referencing neither establishes not addresses the meaning of what is being measured *per se*, and therefore cannot reasonably explain changes in measures across the scale. In attempt to address this issue, recent investigations have successfully made use of Item Response Theory (IRT) for developing and testing psychometric properties of tests for psychiatric disorders assessment (Feske, Kirisci, Tarter, & Pilkonis, 2007; Olatunji et al., 2009; Samuel, Simms, Clark, Livesley, & Widiger, 2010; Stelmack et al., 2004).

The use of IRT models permits (a) an investigation of the structure and function of the categories used as test responses (especially for Likert and/or rating scales), (b) a comparison of the intensity level of the construct represented in the items of a test with the intensity level of the construct in persons (*theta*), (c) an investigation of the hierarchical organization of items according to the intensity represented by each of them, and (d) verification of the reliability of a test at the different levels at which the construct is measured (Embretson & Reise, 2000). While there are certainly other advantages and application possibilities of IRT, an extensive survey is beyond the scope of this work.

IRT proposes a mathematical model to represent the testing situation, in which one person answers a set of items. The more intense a given characteristic in the person, the greater the likelihood of agreement with a statement that measures this characteristic. Conversely, the less intense the feature, the smaller the probability that the person will agree. So, the likelihood of choosing a particular answer varies with the degree to which a given characteristic (θ, called theta) is present or not in the respondent. There are several models based on IRT, but the Rasch model stands out because of its simplicity and measurement properties. This model parameterizes items according to their intensity while measuring a latent trait; therefore, it has been named one-parameter Rasch model (Embretson & Reise, 2000).

For tests using rating scales, two alternative models are available, derived as extensions and further developments of the Rasch model: the rating scale and the partial credit model (Wright & Masters, 1982). In the present study, the rating scale model was used because the literature considers this model as more generalizable for working with rating scales.

Considering the possibility of using IRT in the field of assessment of depression, the aim of this study was to verify the parameters of the items and person for the Brazilian version of BDI obtained by the Rating Scale Model. The explanation of the procedures employed will be conveniently displayed throughout the work.

Method

Participants

A total of 271 people participated in the study. Age ranged between 18 and 51 years (*M*=23.61, *SD*=6.12); 69% (*N*=187) were female and 31% (*N*=84) were male. All participants were college students at a town in the Brazilian state of São Paulo.

Materials

In accordance with the objectives of this study, the Brazilian version of BDI was administered to all study participants. The BDI is a self-report inventory, which was adapted and validated in Brazil by Cunha (2001). It is an instrument for the assessment of symptoms of depression, consisting of 21 items, which should be completed in a Likert scale of 4 points referring to the intensity of symptoms considered typical of a depression. It is estimated that the approximate time of application of the instrument is 10 minutes.

Procedure and Data Analysis

Prior to initiation, the proposed study was submitted to the Ethics Committee and was approved (Protocol number CAAE: 0350.0.142.000-08). The instrument and the Informed Consent Form were administered to all participants. Only after agreeing to sign the form the participants were able to participate in the study.

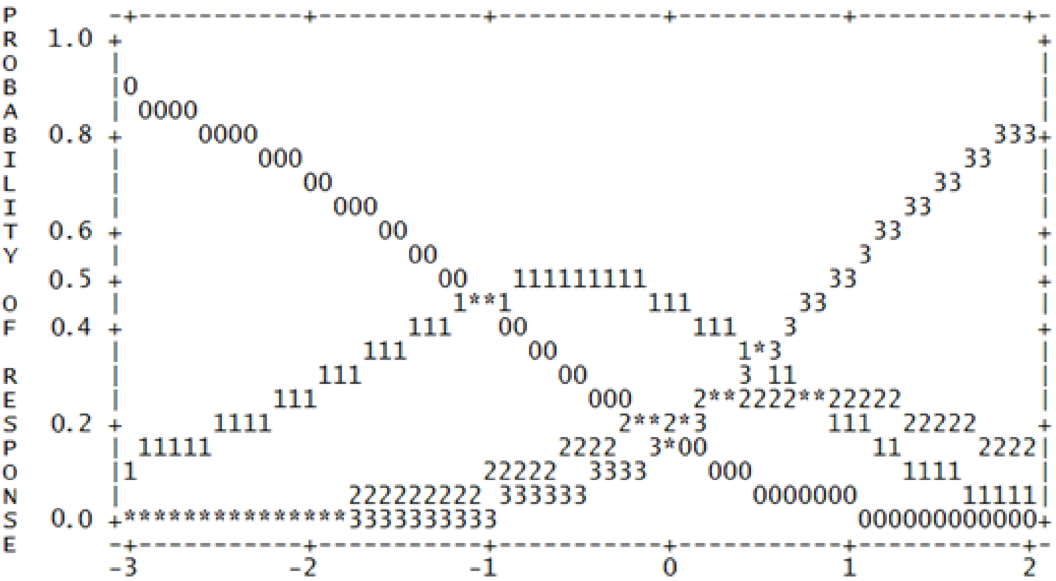
Participants in the study may have completed the entire instrument. The instrument was administered in classrooms of private universities from São Paulo. After collecting the data, statistical analyzes were performed to address the main questions raised in the study. The collected data were analyzed using the Rasch model, specifically the Rating Scale Model, using the statistical software Winsteps (Linacre, 2009) verifying the parameters of the items and respondents.

One of the basic postulates of modeling via IRT is unidimensionality, that is, the model assumes that items measure a primary dimension and secondary dimensions have a negligible influence (Swaminatham & Hambleton, 1985). Thus, the unidimensionality verification of the BDI was a necessary first step in the analysis.

Winsteps was used to calibrate the parameters of items, implementing a method of maximum likelihood estimation (Joint Maximum Likelihood Estimation). To analyze the model fit, we considered the model fit indexes, infit and outfit. These indexes consist of average values of the residues (observed score – modeled score) standardized and squared, (i.e., chi-square divided by degrees of freedom). Using the literature recommendations, we considered values above 1.3 and item-total correlations close to zero as indicative of misfit to the model (Linacre & Wright, 1994; Smith, 1996; Wright & Linacre, 1994). We also considered reliability indexes and local error, response categories of the scales, and, quantitative and visual analyses of the person-items map. It is worth noting that for purposes of analysis, the average difficulty of items (*b*) was set at zero.

Results and Discussion

This work aimed to evaluate the performance of the BDI using the Rasch Rating Scale Model. At first, we verified the response categories of BDI (Figure 1). *Theta* (x-axis) is paired with the response prob-



```
P      -+-----------+-----------+-----------+-----------+-----------+-
R  1.0 +                                                             +
O      |                                                             |
B      |0                                                            |
A      | 0000                                                        |
B  0.8 +     0000                                                 333+
I      |         000                                            33   |
L      |            00                                        33     |
I      |              000                                   33       |
T  0.6 +                 00                               33         +
Y      |                   00                            3           |
   0.5 +                     00   111111111            33            +
O      |                      1**1         111       33              |
F  0.4 +                   111   00           111   3                +
       |                111        00            1*3                 |
R      |             111             00          3 11                |
E      |          111                  000    2**2222**22222         |
S  0.2 +      1111                       2**2*3        111  22222    +
P      | 11111                       2222  3*00           11     2222|
O      |1                       22222  3333    000          1111     |
N      |               222222222 333333           0000000       11111|
S  0.0 +***************3333333333                        000000000000+
E      -+-----------+-----------+-----------+-----------+-----------+-
       -3          -2          -1           0           1           2
```

Figure 1. *BDI response categories pre-collapsing*

Source: own work

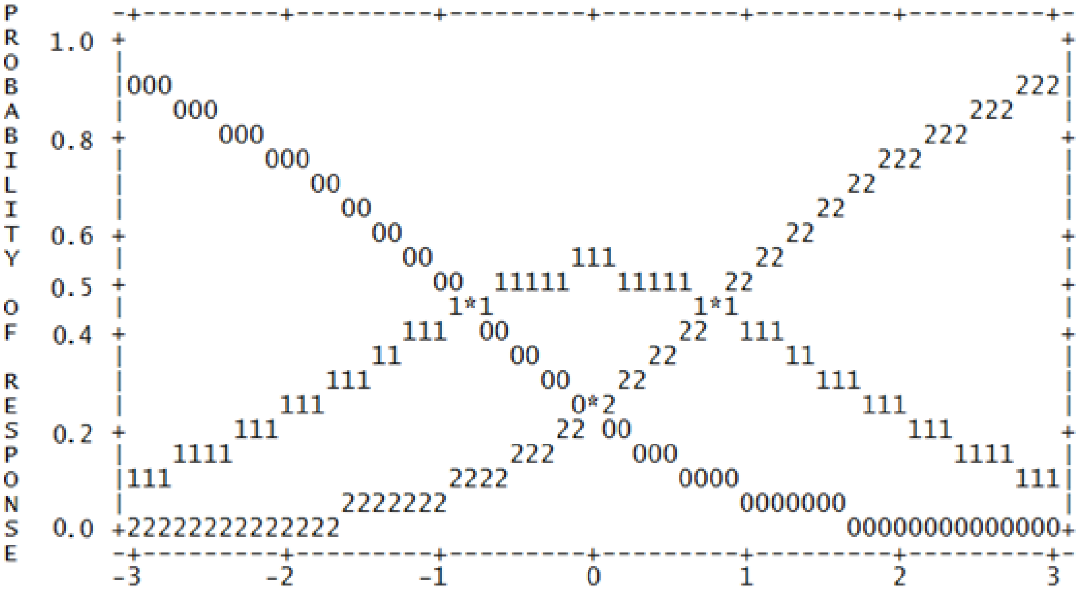
ability of participants at different levels of *theta* (y-axis) to describe each of the rating scale options. In Figure 1, the average *b* is centered on zero. Thus, it is possible to verify the likelihood of endorsement of participants in each category of response and their distributions in different levels of *theta* for an item $b_i = 0$ (i.e., the average level of difficulty equal to zero). The four response categories ranged from 0 to 3, meaning the increase of depression symptoms. The intersection between two categories can be interpreted as the threshold value of transition between these categories.

A clear representation of all categories was not observed, i.e., curve 2 is overlapped in all *theta* range. Because of that, we proceed to the collapsing procedure in which one or more categories can be aggregated allowing a clear representation in the theta range of the remaining categories. One possible explanation for the dysfunction observed in categories is that participants could not discriminate between all the categories labels, since this sample is not composed of people with known diagnoses (specifically, depression diagnoses). Figure 2 presents the BDI response categories pos-collapsing of category 2.

After the collapsing procedure, a clear representation of all categories was observed. The threshold between the first and second categories is equal to -0,81 and between 1 and 2 equal to 0,81. Separation of the curves in different regions of the theta scale is a desirable metric feature because it indicates that respondents demonstrate clear differentiation between each rating scale category, and the present empirical data shows that the response to stimuli (items) has been quantitatively modeled by means of an increasing monotonic relationship between theta and categories.

The specification of unidimensionality was, then, verified through a Rasch principal contrasts analysis implemented through Winsteps. Using the performance indicators associated with the item and person parameters it is possible to calculate an expected response for each subject for each item. The discrepancy between the modeled response (expected) and the observed is the residual.

The principal contrasts analysis is performed on this new residual data matrix, based on the portion of responses not predicted by the model. Thus, if a contrast composed by a set of items with a magnitude greater than 2 (according to guidelines of



```
P      -+---------+---------+---------+---------+---------+---------+-
R  1.0 +                                                             +
O      |                                                             |
B      |000                                                       222|
A      |   000                                                 222   |
B  0.8 +      000                                           222      +
I      |         000                                     222         |
L      |            00                                 22            |
I      |              00                             22              |
T  0.6 +                00                         22                +
Y      |                  00         111         22                  |
   0.5 +                    00  11111   11111  22                    +
O      |                     1*1             1*1                     |
F  0.4 +                  111  00           22  111                  +
       |                11       00       22       11                |
R      |             111           00   22           111             |
E      |          111                0*2                111          |
S  0.2 +       111                  22 00                  111       +
P      |   1111                  222     000                  1111   |
O      |111                  2222           0000                  111|
N      |              2222222                   0000000              |
S  0.0 +22222222222222                                 00000000000000+
E      -+---------+---------+---------+---------+---------+---------+-
       -3        -2        -1         0         1         2         3
```

Figure 2. BDI response categories pos-collapsing

Source: own work

Table 1.
Person and items summarized descriptive statistics

	Person			Items				
	Theta	Infit	Outfit	b	Infit	Outfit	r	Reliability
M (SD)	-1.97 (1.14)	1.02 (0.34)	0.99 (0.55)	0 (0.75)	1.02 (0.19)	1.0 (0.22)	0.6 -0.27	0.7 (0.72)
Max.	0.81	2.36	5.18	1.42	1.45	1.46		
Min.	-4.08	0.28	0.32	-1.41	0.69	0.59		

Source: own work

Linacre (2009) appears, it suggests a second dimension that may potentially affect the data in order to confound the meaning of the first dimension. This analysis seeks to determine values of components with eigenvalues greater than or equal to 2.0. However, in the present study, none of the contrasts reached eigenvalues of 2.0 or greater. Once assured of the BDI unidimensionality, the analysis could be continued.

Table 1 presents descriptive statistics summarizing the latent trait (*theta*) of the respondents, their fit indexes (infit and outfit) and the number of items answered. In addition, this table summarizes the descriptive data for the items (i.e., the difficulty level, the fit indexes, the correlation item-*theta*, and reliability indices - real and modeled).

In general, the average level (-1.97) of the latent trait suggests that the id not tend to be endorsed by the sample, although the standard deviation appoint to at least some respondents that tended to endorse some items. Also, the observed range of scores suggests that the sample is predominantly composed of people on a healthier continuum of depression characteristics. The Rasch model allows to intuitively infer that the scores of the subject, mild or more extreme, is indicative of the level of depression severity.

Also in relation to participants, through the fit indexes, infit and outfit, there were detected discrepancies between the observed and expected values with respect to the estimation of thetas. These values tended to be acceptable (Linacre & Wright, 1994), because the mean value was below 1.3. However, the fit indexes maximum values were higher than 1.3, suggesting discrepancies for some subjects according to what is expected by the model. Moreover, the reliability index of theta estimates calculated by the Rasch model was equal to 0.7 (real) and 0.72 (modeled). These indexes may be considered satisfactory, particularly considering the discrepancy between the items range and people range (Embretson & Reise, 2000). Through the item-person map (Figure 4) this discrepancy can be visually observed.

With respect to the items descriptive data, the difficulty index varied between -1.41 and 1.42. The items fit indexes were adequate (less than 1.3), although the maximum scores reached more than 1.3. Also, the item-theta correlations indicated high positive correlations between the items and the latent construct, which also suggests cohesion between the components (items) for *theta*. Complementing the information about the reliability of dimensions, we also calculated the local error (Figure 3).

One of the advantages of using IRT is to understand the conditioned reliability of each scale (i.e., to know in which level of the scale the instrument has a higher reliability rate). This is done by evaluating the local error curve that presents available information across the levels of *theta*. A way of expressing a standardized curve ranging from 0 to 1 is through the local error (Daniel, 1999).

This index allows assessing the levels of *theta* (latent trait) of items that are more error-free (i.e., more reliable). For example, a scale with a moderate reliability may be highly reliable in a certain range of latent trait, but less so at other levels. Figure 3 shows the reliability indexes for the BDI in accordance with the level of *theta* (local error).

In Figure 3, the x-axis (horizontal) refers to the *theta* (ranging between -5 and +4) and the y-axis

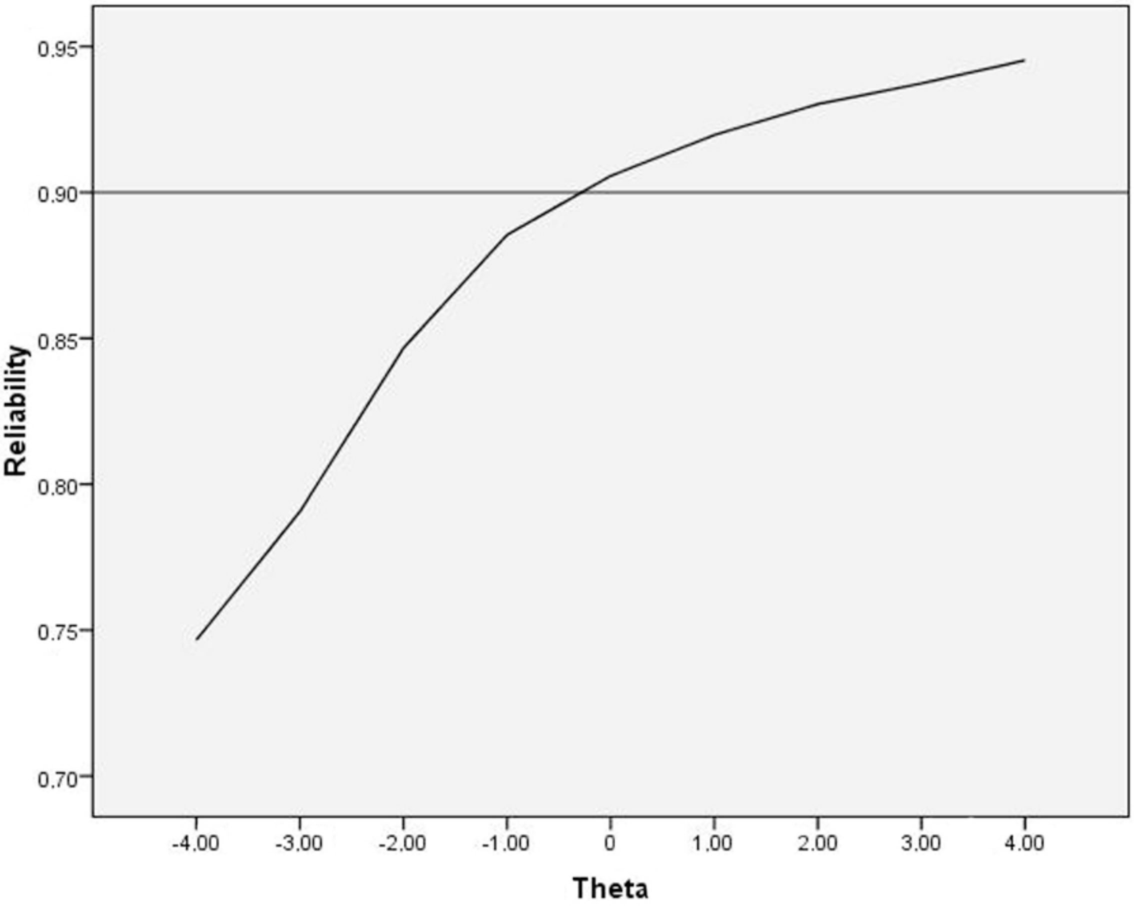


Figure 3. Local error

Source: own work

to the reliability indices. The horizontal line that cuts the graph is dividing the curve in reliability indices equal to or greater than 0.9, and indexes below this cutoff. From there, one can check in which range of *theta* the scale is more reliable. This range includes values of *theta* between -2.06 and 0.81, and the average reliability in this range is 0.93 (between 0.9 and 0.95). This finding contrasts with the "general" reliability of this dimension (0.7 real and 0.72 modeled), since the weighting for different latent trait levels can increase or decrease. As expected, the reliability index of the BDI is higher within higher levels in the latent trait because the instrument focuses on symptoms rather than health characteristics.

Figure 4 presents one important application of IRT to psychiatric disorders assessment, the person-item map. Through IRT it is possible to employ item referenced standard setting (Embretson & Reise, 2000), allowing one to assign meaning to the scores of respondents at different levels of scale. Items are presented, from the bottom up, starting with the most endorsed to the least endorsed ones. The number and content of each item can also be observed. The response categories (0-2) can be verified in the figure for each item of the dimension.

At the bottom of the figure the distribution of respondents is shown (number of responders in each theta level must be read vertically) and *theta* range (ranging from -5 to +4). Letters *T*,

```
-5   -4    -3    -2    -1     0     1     2     3     4
|-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM   ITEM
0                         0     :      1     :     2  2    9  BDI09
|                                                     |
|                                                     |
0                       0     :      1      :    2    2    3  BDI03
|                                                     |
0                      0    :      1      :    2      2   21  BDI21
0                     0     :      1      :    2      2   19  BDI19
0                     0     :     1      :     2      2    2  BDI02
0                     0    :      1      :    2       2    5  BDI05
|                                                     |
|                                                     |
0                   0    :      1      :    2         2   18  BDI18
0                  0     :     1      :     2         2    6  BDI06
0                  0    :      1      :    2          2   20  BDI20
0                 0     :      1      :    2          2   12  BDI12
|                                                     |
0                0     :      1      :    2           2    7  BDI07
0                0     :      1     :     2           2   15  BDI15
0                0    :      1      :     2           2   14  BDI14
|                                                     |
0               0    :      1      :     2            2   10  BDI10
0               0    :      1      :    2             2    1  BDI01
|                                                     |
0              0    :      1      :    2              2    4  BDI04
|                                                     |
0             0    :      1      :    2               2   13  BDI13
0             0    :      1      :    2               2   16  BDI16
|                                                     |
|                                                     |
0          0     :      1      :    2                 2   17  BDI17
|                                                     |
0          0    :      1      :    2                  2    8  BDI08
|                                                     |
0        0     :      1     :     2                   2   11  BDI11
|-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM   ITEM
-5   -4    -3    -2    -1     0     1     2     3     4

2     2   2  2 12 1111111
7     0   2  4 50259471437452443 332                     PERSON
       S        M        S        T
```

Figure 4. BDI Person-items map

Source: own work

S, *M* can be found below the distribution of participants, which refer to, respectively, two standard deviations (*T*=above or below the average), one standard deviation (*S*=above or below the average), and mean (*M*). For this study, a visual analysis was used for the items of BDI considering the theoretical perspective (Beck et al., 1961) underlying the construct in an attempt to bring clinical contributions to the items composing the instrument.

A higher concentration of respondents can be found between the theta range varying from approximately -3.0 to -1.0, which was expected according to the average theta observed (see Table 1). Moreover, there were a greater proportion of respondents in the lower theta categories of the sample, because most of the respondents had no diagnosis of major depression disorder known. Table 2 helps to compare the present findings with the data presented by Castro, Trentini and Riboldi (2010).

In one hand, there are evident discrepancies between the hierarchical arrangement of the items of this study and Castro, Trentini and Riboldi (2010). However, these discrepancies were expected considering the clear differences between the samples. Mainly, our sample is not composed by patients with known depression disorder, which probably entails important distinctions with the other research. Still, in other hand, some significant considerations can be done about the founded data (Figure 4).

First, considering that most of the sample is between -1 and -3 theta range, it can be stated that people tended to endorse categories 0 and 1 more than 2, i.e., categories related to a health functioning or to just light symptomatic depression characteristics. Further visual analysis of the map (Figure 4) were based on a subdivision of the items in four groups, according to the distances between items and the possibility to discriminate people with different level of *theta*.

The first items group is formed by 3 items (11.8 and 17), related to irritability, self-accusation and fatigability; group 2 is composed by items related to a negative perspective of self (7 and 14) or some kind of incapacity/difficulty (16, 13, 15), cry behavior and depressed mood (10 and 1) and dissatisfaction (4); the nest group is composed by 50% of items concerning, more or less directly, to somatic factors (20, 18, 19, 21), items related to a negative vision of the future of world (2 and 6),

Table 2.
BDI items according to difficult level

BDI item Nº	Content	Core Symptom	Our study item order	Castro, Trentini and Riboldi (2010) items order
1	Depressed mood	Depressed mood	7º	10º
2	Pessimism		17º	11º
3	Failure feeling		20º	16º
4	Dissatisfaction	Anhedonia	6º	1º
5	Guilt	Guilt	16º	13º
6	Punishment		14º	2º
7	Self-aversion		11º	18º
8	Self-accusation	Low self-steam	2º	6º
9	Suicidal ideation		21º	19º
10	Cry		8º	8º
11	Irritability		1º	3º
12	Social Retreat	Lack of interest in activities	12º	20º
13	Indecision		5º	7º
14	Low self-steem	Low self-steam	9º	15º
15	Difficulties to work		10º	14º
16	Sleep difficulty	Sleep disturbance	4º	4º
17	Fatigability		3º	12º
18	Loss of apetite	Change in apetite	15º	5º
19	Loss of weight	Change in appetive	18º	21º
20	Somatic concerns		13º	9º
21	Libidinal loss	Lack of interest in activities	19º	17º

Source: own work

and social retreat 12) and guilt (5); and, group 4 is formed by 2 items representing people who seems the self as a failure and report the presence of suicidal thoughts. So, it is possible to describe some kind of pattern of depression symptoms, although it appears to have no clear pattern, since the data is based on a health sample, as pointed before.

A significant part of the subjects tended to endorse the items forming group 1, and subjects in high range of *theta* (approximately 1) tended to endorse items of almost all groups, group 4 exceed. It seems to be coherent in a student without diagnostic sample. Based on the description of groups of items and the endorsement pattern of the sample, it can be noted that people with a similar profile of respondents in this study are likely to exhibit the same type of pattern. One possible explanation to the more endorsed items is related to sample specificity, i.e., college students that work in a period of the day and study in another period constitute most of the sample. This specificity could increase the sample probability to choose some of the items (e.g., irritability and fatigability). However, this hypothesis was not verified.

Furthermore, the symptoms viewed as depression core symptoms (Brody et al., 1998; Kennedy, 2008; Nelson, Portera, & Leon, 2006) were crowded especially in groups 3 (50%) and 2 (40%) of items. As one can see in Table 2, items in group 2 were related to sleep disturbance, anhedonia, depressed mood and low self-steam; and in group 3, to lack of activities interest, change in appetite and guilt.

It is interesting to note how the classical and item referenced standard setting procedures are complementary, allowing a better understanding of the reference points of the scale. In this sense, the presented analysis demonstrates that persons with certain levels of the latent trait (i.e., characteristics related to depression symptoms) tend to agree with some of the statements, in a less likely progressive fashion. Thus, the standardized scalar index (*theta*) is not an arbitrary number on the scale. Instead it is possible to infer which features are present or not in a person with a certain level in the latent trait (Embretson, 2006).

Conclusions

This study aimed to evaluate the item and person parameters and instrument (BDI) functioning obtained by the Rasch model, specifically, the Rating Scale Model. Overall, results suggest the adequacy of the psychometric properties of the instrument. Nay, data showed that through person-items map it could help clinicians in the use of BDI, since it focuses clinical understanding of the scores obtained by individuals who respond to a particular group of items on a continuum of latent trait development. But, it is important to consider the present results as limited, principally considering the size and characteristics (i.e., non-clinical) of the sample. The use of local error should be highlighted, because it is an addition to the reliability analyses conventionally used, offering the ability to check for different reliability indices that may vary across levels of the latent trait measured by the items.

Studies should focus on the use of other IRT models, searching for parameters not considered in this study (e.g., discrimination). Moreover, research involving people with major depressive disorder should be performed. Thus, we hoped that this research contributes to the field of assessment of depression, especially in light of modern psychometric procedures, which are scarce in the field of depression studies.

References

APA (2013) - American Psychiatric Association (2013). Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) American Psychiatric Association

Baptista, M. N., Carneiro, A. M., & Cardoso, H. F. (2014). Depression, Family Support and Hopelessness: a correlated study. *Universitas Psychologica. 13(2) 693-702. doi: 10.11144/Javeriana.UPSY13-2.dfsh*

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: traditional and new models. *Psychological Methods, 14*, 101-125. http://dx.doi.org/10.1037/a0015583

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erlbaum, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*, 561–569. http://dx.doi.org/10.1001/archpsyc.1961.01710120031004

Brody, D. S., Hahn, S. R., Spitzer, R. L., Kroenke, K., Linzer, M., deGruy III, F. V., & Williams, J. B. W. (1998). Identifying patients with depression in the primary care setting: a more efficient method. *Archives of Intern Medical, 158*, 2469-2475. http://dx.doi.org/10.1001/archinte.158.22.2469

Castro, S. M. de J., Trentini, C., & Riboldi, J. (2010). Teoria da Resposta ao Item aplicada ao Inventário de Depressão Beck [Item response theory applied to the Beck Depression Inventory]. *Revista Brasileira de Epidemiologia, 13*(3), 487-501. http://dx.doi.org/10.1590/S1415-790X2010000300012

Contreras, S., Fernandez, S., Malcarne, V. L., Ingram, R. E., & Vaccarino, V. R. (2004). Reliability and Validity of the Beck Depression and Anxiety Inventories in Caucasian Americans and Latinos. *Hispanic Journal of Behavioral Sciences, 26*(4), 446-462. http://dx.doi.org/10.1177/0739986304269164

Cunha, J. (2001). *Manual em português das Escalas Beck* [Beck Scales Brazilian Portuguese Manuals]. São Paulo: Casa do Psicólogo.

Daniel, M. H. (1999). Behind the scenes: using new measurement methods on the DAS and KAIT. In: Embretson, S. E., & Hershberger, S. L. *The new rules of measurement: What every psychologist and educator should know*. Mahwah, NJ: Lawrence Erlbaum.

Embretson, S. E. (2006). The continued search for nonarbitrary metrics in psychology. *American Psychologist, 61*(1), 50-55.

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

Feske, U., Kirisci, L., Tarter, R. E., & Pilkonis, P. A. (2007). An application of item response theory to the DSM-III-R criteria for borderline personality disorder. *Journal of Personality Disorders, 21*, 418-33. http://dx.doi.org/10.1521/pedi.2007.21.4.418

Hammond, S. M. (1995). An IRT Investigation of the validity of non-patient analogue research using the Beck Depression Inventory. *European Journal of Psychological Assessment, 11*(1), 14-20. http://dx.doi.org/10.1027/1015-5759.11.1.14

Kennedy, S. (2008). Core symptoms of major depressive disorder: relevance to diagnosis and treatment. *Dialogues in Clinical Neuroscience, 10*(3), 271-277.

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

Linacre, J. M. (2009). *WINSTEPS: Multiple-choice, rating scale, and partial credit Rasch analysis (Computer Software)*. Chicago, Illinois: MESA Press.

Linacre, J. M., & Wright, B. D. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(2), 370.

Lykke, J., Hesse, M., Austin, S. F., & Oestrich, I. (2008). Validity of the BPRS, the BDI and the BAI in dual diagnosis patients. *Addictive Behaviors, 33*(2), 292-300. http://dx.doi.org/10.1016/j.addbeh.2007.09.020

Nelson, J. C., Portera, L., & Leon, A. C. (2006). Assessment of outcome in depression. *Journal of Psychopharmacology, 20*(4), 47-53. http://dx.doi.org/10.1177/1359786806066046

Olatunji, B. O., Woods, C., Jong, P. J., Teachman, B., Sawchuk, C. N., & David, B. (2009). Development and initial validation of an abbreviated Spider Phobia Questionnaire using item response theory. *Behavior Therapy, 40*, 114-30. http://dx.doi.org/10.1016/j.beth.2008.04.002

Samuel, D., Simms, L. J., Clark, L. A., Livesley, J., & Widiger, T. A. (2010). An item response theory integration of normal and abnormal personality scales. *Personality Disorders: Theory, Research and Treatment, 1*, 5-21. http://dx.doi.org/10.1037/a0018136

Santor, D. A., Gregus, M., & Welch, A. (2006). Eight Decades of Measurement in Depression. *Measurement, 4*(3), 135-155. http://dx.doi.org/10.1207/s15366359mea0403_1

Smith R. M. (1996) Polytomous Mean-Square Fit Statistics. *Rasch Measurement Transactions, 10*(3), 516-517.

Stelmack, J., Szlyk, J. P., Stelmack, T., Babcock-Parziale, J., Demers-Turco, P., Williams, T. R., & Massof, R. W. (2004). Use of Rasch person-item map in exploratory data analysis: A clinical perspective. *Journal of Rehabilitation Research and Development, 41*(2), 233-41. http://dx.doi.org/10.1682/JRRD.2004.02.0233

Swaminatham, H., & Hambleton, H. K. (1985). *Item response theory: principles and applications*. Boston: Kluwer.

Theme-Filha M. M., Szwarcwald C. L., & Souza-Junior P. R. (2005). Socio-demographic characteristics, treatment coverage, and self-rated health of individuals who reported six chronic diseases in Brazil, 2003. *Caderno de Saúde Pública, 21*, 43-53. http://dx.doi.org/10.1590/S0102-311X2005000700006

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.