Item response theory applied to the Beck Depression Inventory

Abstract

The Beck Depression Inventory (BDI), a scale that measures the latent trait intensity of depression symptoms, can be assessed by the Item Response Theory (IRT). This study used the Graded-Response model (GRM) to assess the intensity of depressive symptoms in 4,025 individuals who responded to the BDI, in order to efficiently use the information available on different aspects enabled by the use of this methodology. The fit of this model was done in PARSCALE software. We identified 13 items of the BDI in which at least one response category was not more likely than others to be chosen, so that these items had to be categorized again. The items with greater power of discrimination were sadness, pessimism, feeling of failure, dissatisfaction, self-hatred, indecision, and difficulty of work. The most serious items were weight loss, suicidal ideas, and social withdrawal. The group of 202 individuals with the highest levels of depressive symptoms was comprised by 74% of women and almost 84% had a diagnosis of a psychiatric disorder. The results show gains resulting from use of IRT in the analysis of latent traits.

Keywords: Item Response Theory. Latent trait. Intensity of depressive symptoms. Beck Depression Inventory.

Stela Maris de Jezus Castro
Clarissa Trentini
João Riboldi
Universidade Federal do Rio Grande do Sul

O estudo foi aprovado pelo Comitê de Ética em Pesquisa da UFRGS na reunião nº 37, ata nº 117, de 30 de outubro de 2008.

Correspondência: Stela Maris de Jezus Castro. Rua João Mendes Ouriques, 650, Ipanema, Porto Alegre, RS - CEP: 91760-450. E-mail: stela.castro@ufrgs.br.

Introduction

Latent variables, also termed latent traits, are non-observable quantities which must be inferred by observation of secondary variables associated with them. For this purpose, measurement procedures (scales) are generally used. They consist of a set of items, the responses to which are categories (ordered or otherwise) that are used to estimate the secondary variables, which in turn give estimates of a subject's latent traits.

The Beck Depression Inventory (BDI) is one example of such a measurement procedure. It consists of a group of items intended to assess the latent trait intensity of depressive symptoms; this is extremely important for establishing level of depression, and for predicting its likely outcome. Depression is a very common psychological, social and biological condition; epidemiological studies show that it affects between 3 and 11% of the general population[2,3] and that it is a lifetime condition for 16.2%[3]. In Brazil, studies show that depression occurs throughout the lives of between 2.8 and 19.2% of people.

Until recently, the most commonly used statistical model for latent trait prediction was the Classical Test Theory (CTT)[6], which uses a subject's total score as the estimate of a particular latent trait. This methodology has been reviewed by DeVellis[7]. Despite its great importance and the convenience of CTT use, a number of authors have drawn attention to its limitations, all of which are avoided by using the alternative measurement procedure known as Item Response Theory[9].

Largely used in psychiatry, the Item Response Theory (IRT) comprises a group of generalized linear models and associated statistical procedures which describe the association between item responses (such as an individual's behavior) and a latent trait. The aim of an IRT model is to link one individual to one item. The individual's pattern of response to a particular group of items provides the basis for estimating the latent trait. In IRT models, item parameters and the subject's latent trait levels are independent of each other; these parameters are expressed by the level of response observed for the item; the contribution of each item to the final scale is determined by the IRT information; there are powerful methods available for detecting differential item functioning (DIF) or item biases between populations or subgroups; and scores given by different subjects can be compared even when they have answered to different items[13].

IRT is particularly significant for the analysis of latent traits since it allows more efficient use of information, not only because the method groups individuals according to their latent traits, but also because it provides information about the measuring procedures themselves and, in particular, about each item used. IRT is therefore a more sophisticated method making more thorough use of the information available in each item and so giving improved measures of latent trait, since different items can be given specific degrees of importance according to their relevance to the trait being studied.

IRT models can be further categorized as cumulative or unfolding. Cumulative IRT models can be classified according to their dimensionality (unidimensional or multidimensional). Unidimensional IRT models describe the connection between observed item responses and a single underlying latent trait, usually represented by θ They are suitable for data in which a single common factor is assessed by the items. Unidimensional models include models for dichotomous data (such as where symptoms of depression are either present or absent; or success or failure) and models for polytomous data (items with more than two response categories for each BDI item). Such models can also be distinguished according to the number of parameters used. They may have one, two or three parameters: one related to item complexity (gravity of depressive symptoms in the case of BDI), another related to item discrimination, and a third representing the probability that

subjects exhibit the depressive symptom described by the item, even when the level of the latent trait is low.

All BDI items have four response categories, so that a unidimensional model for polytomous responses is appropriate. It is also unlikely that each of the 21 BDI items will discriminate equally well, for the latent trait, in every individual of a population; a model allowing each item to have a different discriminatory value, having a discrimination parameter for each item, is therefore an advantage. With this in mind, the objective of the present study is to explore the IRT Graded Response Model of Samejima[14] by using it to assess the intensity of depressive symptoms in individuals that have responded to the BDI, to make full use of the information given by this method.

Methods

Data sources

Subjects were taken from a cross-sectional study designed to adapt, normalize and validate a Portuguese version on the Beck Scale, made by Dra. Jurema Alcides Cunha and published in 2001[15]. The 4025 individuals taking part in the study are divided in three groups: group 1, consisting of psychiatric patients (n = 1138); group 2, patients undergoing medical treatment (n = 490); group 3, subjects not under medical treatment, drawn from the population at large (n = 2397). All of them responded to the BDI, consisting of a self-evaluated scale of 21 items, each with four assertions corresponding to increasing levels of depressive symptoms[15] and with scores ranging from 0 to 3. BDI items were chosen based on the most commonly observed behavior and attitudes reported in psychiatric patients with symptoms of depression[15]: sadness; pessimism; feelings of failure; dissatisfaction; guilt; punishment; self-hatred; self-accusation; suicidal thoughts; weeping; irritability; social withdrawal; indecision; changes in self-awareness; work difficulties; insomnia; fatigue; weight loss; somatic worries; loss of libido. A total score varying between 0 and 63 was used in the study[15] to estimate the intensity of subjects' depressive symptoms, so that the following intervals were defined to distinguish between the degree of symptom severity in different subjects: minimal (0-11), light (12-19), moderate (20-35) and heavy (36-63). As the IRT scores (estimating the intensity of depressive symptoms, in the graded response model) have zero mean and unit Standard deviation (i.e., are on scale (0,1)), a linear transformation is applicable, giving a change in scale on which IRT score can be related to total score, based on which individuals have been classified according to the intensity of their depressive symptoms.

Theory of Item Response[16,17] .

An IRT polytomous model suitable for BDI data is the Graded Response Model (GRM) proposed in 1969 by Samejima[14], which assumes that an item's response categories can be ordered. When fitting the GRM, the items in the procedure need not have the same number of response categories.

The response categories of an item are arranged with their scores in increasing order, denoted by $k=0,1,2\ldots m_i$, where (m_i+1) is the number of categories for the i-th item. Then the probability that the j-th subject chooses a given category i, or one larger than it, is given by

$$P^{+}_{i,k}(\theta_j) = \frac{1}{1+e^{-Da_i(\theta_j - b_{i,k})}}$$

where $i=1,2,\ldots,I$, $j=1,2,\ldots,n$, $k=1,2,\ldots,m_i$ and $b_{i,1} \leq b_{i,2} \leq \ldots \leq b_{i,mi}$; θ_i is the intensity of depressive symptoms (the latent trait) of the j-th subject; a_i is common slope parameter for all categories of item i; $b_{i,k}$ is the location parameter (a point on the latent trait continuum) for the k-th category of item i, i.e., each $b_{i,k}$ is the point of intersection amongst the categories of ordered responses, representing the degree of intensity of depressive symptoms needed for response category greater than or equal to the k-th

to be chosen with equal probability *0.5*; *D* is a scale factor equal to 1 or 1.7, the latter value pertaining when the logistic function is required to give results similar to those of the Normal distribution.

The probability that subject *j* responds to category *k* of item *i* is given by the difference

$$P_{i,k}(\theta_j) = P^{+}_{i,k}(\theta_j) - P^{+}_{i,k+1}(\theta_j)$$

where by definition $P^{+}_{i,0}(\theta_j) = 1$ and $P^{+}_{i,m_i+1}(\theta_j) = 0$.
Therefore

$$P_{i,k}(\theta_j) = \frac{1}{1+e-Da_i(\theta_j - b_{i,k})} - \frac{1}{1+e-Da_i(\theta_j - b_{i,k+1})} \quad (1)$$

The curves generated by (1) are termed Response Category Curves. They show the relationship between the response probabilities for the categories of each item and the level of latent trait, and from them it is possible to determine which response category has the highest probability of being chosen for each latent trait level.

The GRM was fitted using the software PARSCALE[18] version 4.1. This estimates the parameters of the GRM such that the $b_{i,k}$ between categories are partitioned into two terms, one being the location parameter (b_i) for each item, and a group of parameters $c_{i,k}$ for each item, so that $b_{i,k} = b_i - c_k$. In the case of the latent trait for Intensity of Depressive Symptoms, the location parameters (b_i) can be interpreted as a measure of the symptom severity measured by a given item[19], and the parameters $c_{i,k}$ represent the distances between the points of intersection between the response category curves for each item.

The scale on which estimates of depressive symptom intensity (termed the IRT scores) are measured is arbitrary, the important feature being the order relationship between points on the scale, and not necessarily their magnitude. They can therefore take any real value between -∞ and +∞, it being necessary to define an origin for measurement and a unit on the measurement scale. Here, the IRT scores were defined to have zero mean and unit Standard deviation ((0,1) scale). The parameter for symptom severity, b_i, is measured in the same units as the IRT scores, and so can be compared with them.

Figure 1(a) shows a graph of this model for item 1 (sadness) in the BDI, with four response categories measuring the intensity of depressive symptoms, and the following parameter estimates: $\hat{a}_1 = 1.478$, $\hat{b}_{1,1} = 0.153$, $\hat{b}_{1,2} = 1.280$ and $\hat{b}_{1,3} = 1.897$, where the response categories are:

1 I do not feel sad.
2 I feel sad.
3 I am always sad and can never shake it off.
4 I am so sad or unhappy that I cannot bear it.

It can be seen from this figure that subjects with intensity of depressive symptoms up to 0.153 are more likely to respond in category 1 (curve 1); subjects with intensity of depressive symptoms between 0.153 and 1.280 are more likely to respond in category 2 (curve 2), and subjects with intensity of depressive symptoms between 1.280 and 1.897 are more likely to choose category 3 (curve 3). Finally, subjects most likely to respond in category 4 (curve 4) are those with intensity of depressive symptoms greater than 1.897.

IRT models also yield Item Information Curves, which are widely used in conjunction with Category of Response Curves and the Test Information Curve. Using Item Information Curves, which are constructed from the information functions for each item, it is possible to analyze how much information on the measure of latent trait is contained in a given item; i.e., the curves show how much psychometric information a given depressive symptom contributes to the measure of intensity of depressive symptoms and, moreover, in which interval of this measure the symptom is most informative. It is with this characteristic that one can evaluate which depressive symptoms best discriminate within the population

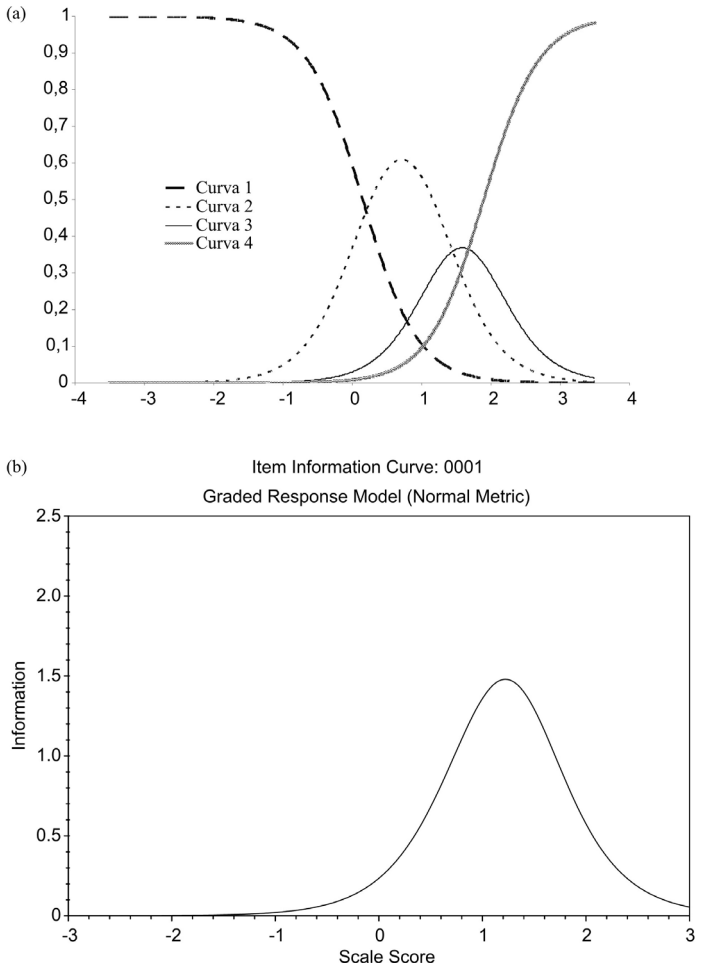

Figure 1 - *Graphic representation of item 1 (sadness) of the BDI on the graded response model: (a) Category of Response Curve; (b) Item Information Curve.*

with regard to the intensity of depressive symptoms[17]. In polytomous IRT models, the quantity of information yielded by an item depends on both the magnitude of the slope parameter a_i and the distribution of points of intersection between the response categories $b_{i,k}$ along the latent trait continuum. For example, the information curve for item 1 of the BDI, in Figure 1(b), shows that greatest concentration of information lies between 0.7 and 1.7 on the intensity scale of depressive symptoms. This shows that the depressive symptom 'Sadness' best discriminates between subjects whose intensity of depressive symptoms lie in this interval.

The Test Information Curve is a graphic representation of the information function. This is an additive function defined over the group of items which constitute the test (the BDI for example) which summarizes the contribution of each item to the information total. The total quantity of information yielded by a group of items at each latent trait level is inversely related to the standard error of its estimate. The test information function is a viable alternative to the concepts of confidence intervals and Standard errors in the Classical Theory of Tests, and by using the Test Information Curve one can determine in which latent trait interval the test works best.

Model assumptions.

For the GRM to be adequate, two assumptions must be satisfied: local inde-

pendence (meaning that when the levels of latent trait are held constant, the response yielded by any item is unrelated to the response yielded by the preceding item) and unidimensionality (all items in the test procedure exhibit the same latent trait). The two assumptions are related, implying that when a procedure is unidimensional, it also exhibits local independence; i.e., if the assumption of unidimensionality is satisfied, only a single latent trait is influencing responses to the items, and local independence is assured[16,20]. There is evidence that the assumption of unidimensionality can be relaxed, and that it need only be sufficient[21-24]; i.e., it is enough for one factor to be preponderant (such that the proportion of variance explained by the first factor in a principal component analysis is not less than 20%[21]) for IRT models to be usable. In view of this, the unidimensionality of the BDI was evaluated using the procedure termed Parallel Analysis[20,25-28], available in a macro[29] of SAS version 9.1.3 (SAS Institute, Cary, NC, USA). This procedure involves a comparison of the eigenvalues of a principal component analysis of real data with a statistic summarizing the eigenvalues of samples of simulated data having the same number of observations and variables as the real data-set (in this case, 4 025 observations on 21 variables). The simulated samples are uncorrelated and are generated by Monte Carlo methods (5000 samples were generated and the summary statistic was the median).

Results

Demographic characteristics of the sample are shown in Table 1. Just under half the subjects are men, with slightly more women. Almost all individuals defined themselves as white, and slightly more than half are single. In terms of education, subjects are distributed almost evenly between categories up to the completion of secondary school, subjects with higher education being less frequent. The mean age of subjects is about 32 (Standard deviation 15.1 years) but the

Table 1 - Description of the sample according to group of origin.

Socio-demographic characteristics**	Psychiatric (%)	Clinical (%)	Non-clinical (%)	Total
Sex (n=4025)				
Male	43,3	36,7	48,2	45,4
Femal	56,7	63,3	51,8	54,6
Skin color (n=3767)				
White	88,6	93,3	92,3	91,4
Non-white	11,4	6,7	7,7	8,6
Education (n=3816)				
Less than 5 years	28,3	31,8	27,5	28,2
Primary school completed	25,6	19,8	28,3	26,5
Secondary school completed	33,9	27,2	25,2	28,0
Higher education completed	12,2	21,2	19,0	17,3
Civil status (n=3898)				
Single	36,7	33,3	77,7	60,3
Married	42,4	50,1	15,3	27,5
Separated, divorced, widowed	20,9	16,6	7,0	12,2
Age (n=4014)				
Mean (DP*)	38,4 (12,3)	44,1 (14,4)	26,4 (13,6)	32,0 (15,1)

* *Standard deviation*
** *The number of individuals varies according to socio-demographic characteristics due to the occurrence of missing values.*

group undergoing medical treatment group is older, on average.

Results of the parallel analysis showed that the assumption of unidimensionality is adequate[19,30] because a preponderant factor was found which explained 38.7% of the total variation.

Based on curves derived from expression (1), fitting the Gradual Response model showed that for 13 of the 21 BDI items, at least one of the response categories has no probability greater than the others of being selected for any intensity of depressive symptoms, as demonstrated in Figure 2(a) for the item associated with pessimism (item 2). One possible explanation for this is that items related to pessimism, dissatisfaction, guilt, punishment, suicidal thoughts, weeping, irritability, changes in self-awareness, insomnia, appetite loss, weight loss, somatic worries and the loss of libido could be presenting problems of understanding the scale. These items were therefore re-categorized by combining together the categories adjacent to the problematic category, thus giving a scale with items having a different number of response categories. The Gradual Response model was then re-fitted, giving category of response curves which showed that all response categories have a chance of being selected for some interval in the latent trait continuum (Figure

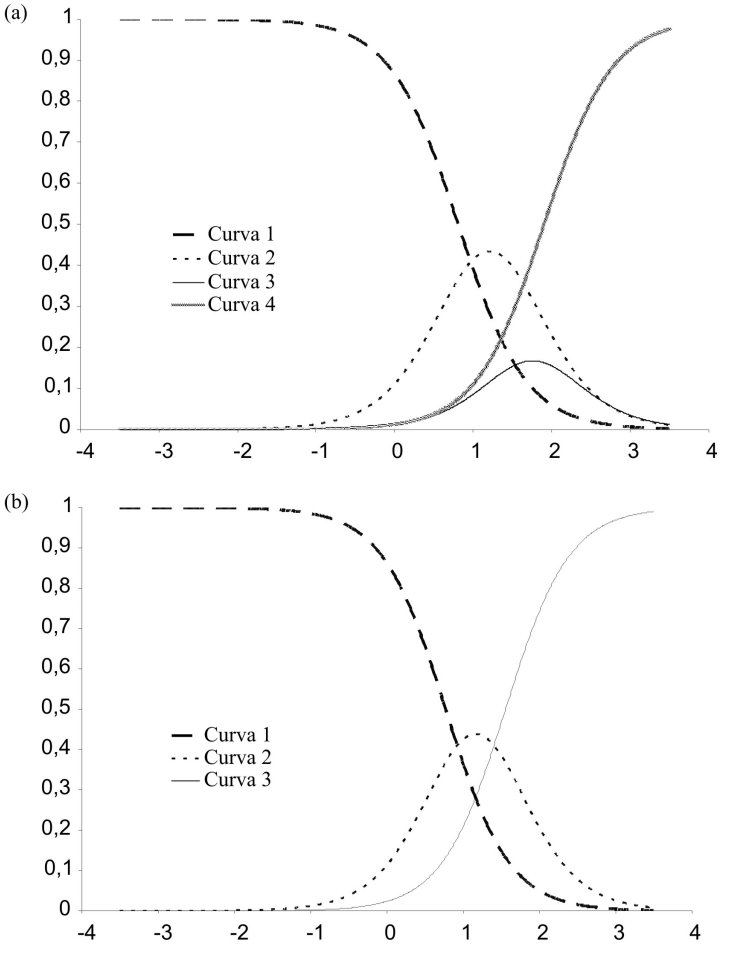

Figure 2 - *Category of Response Curve of item 2 of the BDI, mean scores: (a) Curve 1: I am not particularly discouraged about the future. Curve 2: I feel discouraged about the future. Curve 3: I think I have nothing to expect. Curve 4: I think there is no hope in the future and I have the impression that things can not improve. (b) Curve 1: I am not particularly discouraged about the future. Curve 2: I feel discouraged about the future. Curve 3: I think I have nothing to expect OR I think there is no hope for the future and I have the impression that things can not improve.*

2(b)). When the Gradual Response model was fitted with BDI items re-categorized, the estimates of model parameters were as shown in Table 2.

From the Item Information Curves, which are strongly influenced by the slope parameter (a_i in Table 2), it can be seen that when the cutoff point is defined as unity[16,19] for estimates of this parameter, so as to identify which items give good discrimination ($a_i > 1$), those items relating to sadness, pessimism, feelings of failure, dissatisfaction, self-hatred, indecision and work difficulties (items 1, 2, 3, 4, 7, 13 and 15 respectively) contribute most to the measure of depressive symptom intensity, and therefore give best discrimination in the population for this latent trait. It should be noted that items related to guilt (item 5) and suicidal thoughts (item 9) also have slope parameters greater than one ($\hat{a}_5$ = 1.172 and $\hat{a}_9$ = 1.078 in Table 2); however, the item information curves show that these items give poor discrimination. The item related to irritability (item 11) contributes least to the measure of depressive symptom intensity. Drawing a horizontal line from the point at which the item information function has value one, it can be seen that the item related to feelings of failure gives better discrimination for intensity of depressive symptoms in the population, when the IRT score is in the range [0.7;2], which corresponds approximately to the interval

Table 2 – Estimates of parameters in the graded response model.

Item		a_i (SP)	b_i (SP)	$b_{i,1}$ (SP)	$b_{i,2}$ (SP)	$b_{i,3}$ (SP)
1	Sadness	1,478 (0,036)	1,110 (0,023)	0,153 (0,028)	1,280 (0,031)	1,897 (0,039)
2	Pessimism	1,408 (0,041)	1,163 (0,028)	0,770 (0,033)	1,556 (0,038)	-
3	Feelings of failure	1,684 (0,052)	1,359 (0,025)	0,720 (0,030)	1,339 (0,033)	2,020 (0,041)
4	Dissatisfaction	1,574 (0,042)	0,667 (0,023)	0,102 (0,028)	1,232 (0,031)	-
5	Guilt	1,172 (0,032)	1,271 (0,028)	0,568 (0,035)	1,974 (0,045)	-
6	Punishment	0,850 (0,035)	0,671 (0,036)	0,671 (0,036)	-	-
7	Self-hate	1,393 (0,038)	1,596 (0,026)	0,529 (0,032)	1,826 (0,040)	2,433 (0,053)
8	Self-accusation	0,702 (0,014)	0,965 (0,031)	-1,048 (0,045)	1,150 (0,045)	2,793 (0,065)
9	Suicidal thoughts	1,078 (0,036)	1,726 (0,036)	1,152 (0,044)	2,300 (0,056)	-
10	Weeping	0,792 0,021)	1,067 (0,035)	0,447 (0,044)	1,687 (0,050)	-
11	Irritability	0,326 (0,007)	0,734 (0,060)	-1,011 (0,084)	2,479 (0,092)	-
12	Social withdrawal	0,988 (0,027)	1,767 (0,030)	0,650 (0,038)	1,790 (0,045)	2,861 (0,068)
13	Indecision	1,185 (0,029)	1,022 (0,025)	0,120 (0,032)	0,814 (0,033)	2,132 (0,046)
14	Change in self-awareness	0,950 (0,025)	1,336 (0,031)	0,549 (0,039)	2,123 (0,051)	-
15	Work difficulties	1,179 (0,027)	1,300 (0,026)	0,287 (0,032)	1,288 (0,036)	2,324 (0,051)
16	Insomnia	0,832 (0,021)	0,827 (0,030)	0,065 (0,039)	1,589 (0,045)	-
17	Tiredness	0,955 (0,022)	1,226 (0,028)	-0,254 (0,036)	1,535 (0,041)	2,397 (0,054)
18	Loss of appetite	0,711 (0,032)	0,961 (0,048)	0,961 (0,048)	-	-
19	Loss of weight	0,547 (0,033)	1,995 (0,104)	1,995 (0,104)	-	-
20	Somatic worries	0,745 (0,020)	1,087 (0,033)	0,318 (0,043)	1,856 (0,051)	-
21	Loss of libido	0,856 (0,024)	1,400 (0,035)	0,733 (0,044)	2,067 (0,044)	-

SE: *standard error of estimate*
a_i parameter of slope common to all categories of the item i
b_i measure of severity of symptoms assessed by the item i
$b_{i,1}$ point of intersection between the categories of response 1 and 2
$b_{i,2}$ point of intersection between the categories of response 2 and 3
$b_{i,3}$ point of intersection between the categories of response 3 and 4

[21;35] on the scale for total score.

The item related to Sadness discriminates best when the IRT score lies between 0.7 and 1.7, corresponding to the total score interval 21 to 32. To convert IRT scores to the scale for total score, the former were multiplied by the Standard deviation of total score and then added to its mean value.

Estimates of the parameter for gravity of depressive symptoms (b_i in Table 2) show that items related to loss of weight, social withdrawal and suicidal thoughts are considered more severe when estimating the intensity of depressive symptoms.

Levels of intensity of depressive symptoms estimated from the Gradual Response model (IRT scores) have the same severity scale as the BDI items, so that both groups are comparable. The 95% percentile for the level of intensity of depressive symptoms is 1.6, corresponding to a total score of 31. The group with higher IRT scores is formed by 202 subjects, corresponding to 5% of the analyzed population. Table 3 illustrates this group's profile: about 84% belong to the psychiatric group; the mean age is 39 years; 74% are women; roughly half studied for less than five years, and around 40% are married.

The Test Information Curve shows that the BDI is most efficient for individuals having intensity of depressive symptoms between 0.8 and 2.4, corresponding to an interval of 22 to 40 on the scale of total score.

Discussion

The 21 items of the BDI are representive of the symptoms most frequently observed in depressed people[15]. Evaluation of these items in terms of the amount psychometric information that they provide, and in terms of the severity of the latent trait being measured, is a significant advantage that IRT models have over CTT analyses for this kind of data, since relative weights can be determined for each depressive symptom; more importantly, they are taken into account when calculating the latent trait for each individual. By contrast, when total score is calculated using CTT, all items must be given equal levels of contribution.

Another substantial advantage of IRT models is that they generate Response Category Curves which reveal the association between the levels of intensity of depressive symptoms and the probability that a given

Table 3 - Description of subjects with estimated intensity of depressive symptoms above the 95th percentile ($\hat{\theta} = 1,6$**).

Socio-demographic characteristics **	Psiquiatric n=169	Clinical n=14	Non-clinical n=19	Total n=202
Sex (n=202)				
Male	34	9	9	25,7%
Female	135	5	10	74,3%
Education (n=196)				
Less than 5 years	71	10	8	45,4%
Primary school completed	47	2	4	27,1%
Secondary school completed	34	1	6	20,9%
Higher education completed	12	0	1	6,6%
Civil status (n=201)				
Single	50	8	13	35,4%
Married	75	4	1	39,8%
Separated, divorced, widowed	43	2	5	24,8%
Age (n=201)				
Mean (DP*)	40,1 (12,7)	39,8 (18,5)	33,7 (19,7)	39,5 (13,9)

* *Standard deviation*

** *On the total score scale this 95th percentile is equal to 31.*

category is selected; from such curves, it is possible to determine whether any item has categories that are poorly dimensioned, as occurred in the present study. This finding suggests that individuals responding to BDI cannot distinguish between the assertions defining the response categories for some of the items, demonstrating a need to rethink the measurement scale. Here, this problem occurred in thirteen items, of which two referred to symptoms with very high psychometric information content about the intensity of depressive symptoms namely, pessimism and dissatisfaction. One possible solution is that used in the present work, which was to combine response categories adjacent to the one giving the problem, since it appeared probable that subjects had not been able to distinguish between the information content of assertions in this category.

To establish which items hold more psychometric information about the intensity of depressive symptoms, i.e. those which best discriminate for this latent trait in the population, the point where the item's information function has unit value[16,17] can be taken as a cutoff point, since it is influenced by the magnitude of the slope parameter in the Gradual Response model. In this context, it is interesting to note that, of the seven items that discriminate most efficiently for the latent trait in the population, six performed best in the region where the intensity of depressive symptoms was moderate[15]. Only the item associated with self-hatred remained efficient in the region of very intense depressive symptoms region, since its interval on the IRT score was between 1.3 and 2.2, corresponding to an interval 27 to 37 in total score.

Uher et al.[11] used other cutoff points for the item slope parameter (a_i) in their study, this being used directly as the item discrimination parameter in the Gradual Response model. These cutoff points divide items into three groups: items giving poor discrimination ($a_i < 65$); items giving moderate discrimination ($0.65 \leq a_i \leq 1.34$) and items giving good discrimination ($a_i >$ 1.34). Using this criterion, the items giving good discrimination in the present study were sadness, pessimism, feelings of failure, dissatisfaction, and self-hatred (Table 2); except for the last, these items had the same classification as in Uher et al.[11] Indecision and work difficulties gave moderate discrimination, as did the majority of items; irritability and weight loss were the only items giving poor discrimination (Table 2).

The 95% percentile of the IRT score, corresponding to a total score of 31, was found to lie in the region of moderately intense depressive symptoms. The group with moderate to heavy depressive symptoms comprises 202 subjects (their IRT scores being not less than the 95% percentile) of which 74% were women, agreeing with previous evidence indicating that depression is two to three times more prevalent in women than in men[31].

Estimates of the points of intersection $b_{i,k}$ are divided in two parts; the first, b_i, is the location parameter, and corresponds to a measure of the gravity of the depressive symptom associated with a given BDI item. This allows results to be compared with those given by IRT dichotomous response models, such as models with two or three parameters models (both contain position and slope parameters) when these are used in the same evaluation procedure. Cúri[19] fitted a three-parameter logistic model to BDI data, and found results in moderate agreement with those concerning the severity of depression and discriminatory power of items, reported in the present work, particularly regarding the weight loss item, which is one of the most severe depressive symptoms but which shows poor discrimination (Table 2). Despite this, weight loss information may be extremely relevant for several reasons, for example: it could be a consequence of some clinical condition leading to depression (since there is evidence that depression is strongly associated with chronic pathological conditions such as hypertension, coronary diseases, diabetes, brain hemorrhage, terminal kidney diseases, chronic obstructive pulmonary

disease, congenital heart diseases, angina, asthma and arthritis[32-39]), or it might be a physical expression of other symptoms such as sadness and tiredness, which might lead to lack of attention to diet, with subsequent weight loss.

There are several advantages of using IRT. In the present study, these are illustrated by the possibility of differentiating and comparing depressive symptoms in terms of their discriminatory power and gravity; by the possibility of associating levels of depressive symptoms with the probability of response to each category, thus showing how well patients understand the scale; by the possibility of comparing the depressive symptoms inferred from each subject with the gravity of each symptom; and others. Even more important is the fact that different weights can be assigned to each item used for estimating the intensity of depressive symptoms (IRT score); this is not possible in CTT models, where subjects with the same total score values are considered identical, even when they responded differently to the questionnaire. This shows

Chart 1 - Comparative table of the IRT score* for individuals with total score 10com different patterns of response.

Item	Pattern of response of subjects								
	0005	0007	0028	0047	0073	0111	0150	0162	0163
1	1	0	0	0	0	0	1	0	0
2	0	0	0	0	0	0	0	0	0
3	0	1	0	1	0	0	0	0	0
4	1	1	0	0	1	0	1	1	0
5	1	0	0	0	1	3	0	0	3
6	0	0	0	3	0	0	1	0	0
7	0	0	0	0	1	0	0	0	1
8	1	1	0	0	1	2	2	1	1
9	0	0	0	0	0	0	0	0	0
10	0	3	0	3	0	2	0	0	1
11	0	0	3	0	3	0	0	3	2
12	1	0	3	0	0	0	0	1	0
13	0	2	0	1	2	0	0	1	0
14	0	0	0	2	0	0	0	0	0
15	1	0	1	0	1	0	0	0	0
16	1	0	1	0	0	3	3	1	0
17	1	0	1	0	0	0	1	1	1
18	0	0	0	0	0	0	0	0	0
19	0	0	1	0	0	0	0	0	0
20	1	2	0	0	0	0	1	1	1
21	1	0	0	0	0	0	0	0	0
TRI score	0,27	0,01	-0,28	-0,21	0,07	-0,41	-0,09	-0,03	-0,14
SE**	0,23	0,27	0,30	0,29	0,26	0,33	0,26	0,26	0,28
Change of scale***	16	13	10	10	14	8	12	13	11
Total score	10	10	10	10	10	10	10	10	10

** Intensity of depressive symptoms estimated by the fit of the Graded Response model.*
*** Standard error of the estimate of the intensity of depressive symptoms.*
**** Equivalent values of the IRT scores on the total score scale.*

up better in Chart 1: nine individuals had a final total score of ten (even though they had different profiles of response to the BDI items) but had different IRT scores, which take account the differences in response profiles to items with different weights when they are calculated (except for Rasch models, in which all items have the same weight in score calculations).

The relevance of IRT models to medical research has already been demonstrated. However, it is important to stress that before such models are used in medical practice, they must be shown to be appropriate for specific populations. Just as the BDI has been used to evaluate the intensity of depressive symptoms, their cutoff points for differentiating between individuals must be determined. Computational development and implementation of IRT models (especially for polytomous response items) which allow for the presence of DFI is also needed, since different groups of subjects often react differently when responding to a given item.

In the BDI, some items demonstrate a sex-differentiated function, in that men and women react differently to them. The fact that the IRT model used in the present work does not allow for the presence of DFI is one potential limitation; another is that the mean age of subjects in the sample is lower than that of the population from which it is drawn, especially for the group not receiving medical treatment.

References

1. Fleck MP, Lafer B, Sougey EB, Del Porto JA, Brasil MA, Juruena MF. [Guidelines of the Brazilian Medical Association for the treatment of depression (complete version)]. *Rev Bras Psiquiatr* 2003; 25: 114-22.

2. Kessler RC, Berglund P, Demler O et al. The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *JAMA* 2003; 289: 3095-105.

3. Almeida-Filho N, MariJ.J., Coutinho E, et al. Brazilian multicentric study of psychiatric morbidity. Methodological features and prevalence estimates. *Br J Psychiatry* 1997; 171: 524-9.

4. Theme-Filha MM, Szwarcwald CL, Souza-Junior PR. Socio-demographic characteristics, treatment coverage, and self-rated health of individuals who reported six chronic diseases in Brazil, 2003. *Cad Saúde Pública* 2005; 21: 43-53.

5. Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJ. Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *Lancet* 2006; 367: 1747-57.

6. Beck, AT, Steer, RA. *Beck Depression Inventory. Manual*. San Antonio, TX: Psychological Corporation; 1993.

7. Lord, FM, Novick MR. *Statistical theories of mental test scores*. Reading, MA; 1968.

8. DeVellis RF. Classical test theory. *Med Care* 2006; 44: S50-9.

9. Lord, FM. *Applications of Item Response Theory to ractical Testing Problems*. Hillsdale, NJ: Erlbaum; 1980.

10. Chachamovich, E. *Teoria de Resposta ao Item: Aplicação do modelo Rasch em desenvolvimento e validação de instrumentos em saúde mental* [tese de doutorado]. Rio Grande do Sul: Faculdade de Medicina da UFRGS; 2008.

11. Uher R, Farmer A, Maier W et al. Measuring depression: comparison and integration of three scales in the GENDEP study. *Psychol Med* 2008; 38: 289-300.

12. Nuevo R, Dunn G, Dowrick C, Vazquez-Barquero JL, Casey P, Dalgard OS et al. Cross-cultural equivalence of the Beck Depression Inventory: A five-country analysis from the ODIN study. *J Affect Disord* 2008.

13. Uttaro T, Lehman A. Graded response modeling of the Quality of Life Interview. *Eval Program Plann* 1999; 22: 41-52.

14. Samejima, F. *Estimation of latent ability using a response pattern of graded scores*. Psychometrika Monograph 17. 1969.

15. Cunha JA. *Manual da versão em português das ESCALAS BECK*. São Paulo: 2001.

16. Andrade, D. F., Tavares, H. R., & Valle, R. C. *Teoria da Resposta ao Item: conceitos e aplicações*. IN: 14º Simpósio Nacional de Probabilidade e Estatística; 2000 jul 28; Caxambu (BR). ABE - Associação Brasileira de Estatística.

17. Embretson SE, Reise SP. *Item Response Theory for Psychologists*. New Jersey: Lawrence Erlbaum Associates, Inc.; 2000.

18. *PARSCALE*. [computer program]. Versão 4.1. Chicago (Illinois): Scientific Software International, Inc.; 2003.

19. Cúri M. *Análise de questionários com itens constrangedores* [tese de doutorado]. São Paulo: Instituto de Matemática e Estatística da USP; 2006.

20. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Med Care* 2000; 38: II28-42.

21. McHorney CA, Cohen AS. Equating health status measures with item response theory: illustrations with functional status items. *Med Care* 2000; 38: II43-59.

22. Chan KS, Orlando M, Ghosh-Dastidar B, Duan N, Sherbourne CD. The interview mode effect on the Center for Epidemiological Studies Depression (CES-D) scale: an item response theory analysis. *Med Care* 2004; 42: 281-9.

23. Kim Y, Pilkonis PA, Frank E, Thase ME, Reynolds CF. Differential functioning of the Beck depression inventory in late-life patients: use of item response theory. *Psychol Aging* 2002; 17: 379-91.

24. Bernstein IH, Rush AJ, Carmody TJ, Woo A, Trivedi MH. Clinical vs. self-report versions of the quick inventory of depressive symptomatology in a public sector sample. *J Psychiatr Res* 2007; 41: 239-246.

25. Glorfeld LW. An improvement on Horn´s Parallel Analysis Methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement* 1995; 55: 377-93.

26. Hayton JC, Allen DG, Scarpello V. Factor Retention Decisions in Exploratory Factor Analysis: a Tutorial on Parallel Analysis. *Organizational Research Methods* 2004; 7: 191-205.

27. Ledesma RD, Valero-Mora P. Determining the Number of Factors to Retain in EFA: an easy-to-use computer program for carrying out Parallel Analysis. *Practical Assessment, Research & Evaluation* 2007; 12.

28. Franklin SB, Gibson DJ, Robertson PA, Pohlmann JT, Fralish JS. Parallel Analysis: a method for determining significant principal components. *J Vegetation Sci* 1995; 6: 99-106.

29. Determining the Dimensionality of Data: A SAS Macro for Parallel Analysis. [Portland.: SUGI 28, Paper 90-28; 2007].

30. Kirisci L, Moss HB, Tarter RE. Psychometric evaluation of the Situational Confidence Questionnaire in adolescents: fitting a graded item response model. *Addict Behav* 1996; 21: 303-317.

31. Beyer JL, Nash J, Shelton R, Loosen PT. Transtorno Depressivo maior. In: Artmed Editora, ed. *Manual Diagnóstico e Estatístico de Transtornos Mentais*. 4 ed. Porto Alegre; 2000. p. 288-324.

32. Dickens C, McGowan L, Percival C et al. Depression is a risk factor for mortality after myocardial infarction: fact or artifact*? J Am Coll Cardiol* 2007; 49: 1834-40.

33. Bogner HR, Morales KH, Post EP, Bruce ML. Diabetes, depression, and death: a randomized controlled trial of a depression treatment program for older adults based in primary care (PROSPECT). Diabetes Care 2007; 30: 3005-3010.

34. Collins-McNeil J, Holston EC, Edwards CL, Carbage-Martin J, Benbow DL, Dixon TD. Depressive symptoms, cardiovascular risk, and diabetes self-care strategies in African American women with type 2 diabetes. *Arch Psychiatr Nurs* 2007; 21: 201-209.

35. Golden SH, Lee HB, Schreiner PJ et al. Depression and type 2 diabetes mellitus: the multiethnic study of atherosclerosis. *Psychosom Med* 2007; 69: 529-36.

36. Kamphuis MH, Geerlings MI, Tijhuis MA et al. Physical inactivity, depression, and risk of cardiovascular mortality. *Med Sci Sports Exerc* 2007; 39: 1693-9.

37. Knol MJ, Heerdink ER, Egberts AC et al. Depressive symptoms in subjects with diagnosed and undiagnosed type 2 diabetes. *Psychosom Med* 2007; 69: 300-5.

38. Li C, Ford ES, Strine TW, Mokdad AH. Prevalence of depression among U.S. adults with diabetes: findings from the 2006 behavioral risk factor surveillance system. *Diabetes Care* 2008; 31: 105-7.

39. Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B. Depression, chronic diseases, and decrements in health: results from the World Health Surveys. *Lancet* 2007; 370: 851-8.

Recebido em: 20/07/09
Versão final reapresentada em: 26/06/10
Aprovado em: 12/07/10