DOI: 10.1590/1980-5497201500010005

ORIGINAL ARTICLE / *ARTIGO ORIGINAL*

Differential Item Functioning in the Beck Depression Inventory

Funcionamento Diferencial do Item no Inventário de Depressão Beck

Stela Maris de Jezus Castro[I], Mariana Cúri[II], Vanessa Bielefeldt Leotti Torman[I], João Riboldi[I]

ABSTRACT: ***Introduction:*** There are several studies showing the presence of Differential Item Functioning (DIF) in some items of the Beck Depression Inventory (BDI), when comparing men and women. The presence of a large number of items with DIF in BDI is a severe threat to the validity of measurement of the intensity of depressive symptoms obtained by Item Response Theory (IRT) and to the conclusions based on the scores derived from the items with or without DIF. ***Objective:*** The objectives of this study were to identify these items from the BDI, adjust the IRT model for embarrassing items (model 2), which accommodates items with the presence of DIF, and compare these results with the fit of the traditional two-parameter logistic IRT model (model 1). ***Methods:*** The results obtained with the both models were compared. ***Results:*** Items with DIF were: sadness, feeling of failure, dissatisfaction, guilty, punishment, crying, fatigability and loss of libido. The results of the adjustment of the two models are similar in discrimination, gravity (except for items with DIF), and in the calculation of scores for individuals. Nevertheless, model 2 is beneficial because it shows the differences in gravity of depressive symptoms for groups evaluated, thus providing more information to the researcher on the study population. ***Conclusion:*** This model, which has a broader scope in terms of target population, may be a good alternative to the identification and follow-up of individuals with potential depression.

Keywords: Item Response Theory. Differential Item Functioning. Intensity of Depressive Symptoms. Beck Depression Inventory. Latent trait. IRT Model for embarrassing items.

[I] *Universidade Federal do Rio Grande do Sul* – Porto Alegre (RS), Brazil.
[II] *Universidade de São Paulo* – São Carlos (SP), Brazil.
Corresponding author: Stela Maris de Jezus Castro. Rua João Mendes Ouriques, 650, Ipanema, CEP: 91760-450, Porto Alegre, RS, Brasil. E-mail: stela.castro@ufrgs.br
Conflict of interests: nothing to declare – **Financing source:** none.

RESUMO: ***Introdução:*** Diversos estudos mostram o Funcionamento Diferencial do Item (DIF) em itens do Inventário de Depressão Beck (BDI), ao compararem homens e mulheres. A presença de um grande número de itens com DIF no BDI é uma severa ameaça à validade da medida da intensidade de sintomas depressivos obtida pela Teoria da Resposta ao Item (TRI) e às conclusões baseadas nos escores derivados dos itens com e sem DIF. ***Objetivo*****:** Os objetivos deste estudo foram identificar esses itens do BDI, ajustar o modelo de TRI para itens constrangedores (modelo 2), o qual acomoda itens com a presença de DIF, e comparar esses resultados com os do ajuste do modelo logístico de dois parâmetros tradicional da TRI (modelo 1). ***Métodos***: Os resultados obtidos com ambos os modelos foram comparados. ***Resultados***: Os itens que apresentaram DIF foram: tristeza, sentimento de fracasso, insatisfações, culpa, punição, choro, fatigabilidade e perda da libido. Os resultados do ajuste dos dois modelos são similares quanto à discriminação, gravidade (à exceção dos itens com DIF) e no cálculo de escores para os indivíduos. Apesar disso, o modelo 2 é vantajoso, pois mostra as diferenças em gravidade do sintoma depressivo para os grupos avaliados, trazendo, dessa forma, mais informação ao pesquisador sobre a população estudada. ***Conclusão:*** Esse modelo, que tem um alcance mais amplo em termos de população-alvo, pode ser uma ótima alternativa na identificação e acompanhamento de indivíduos com potencial depressivo.

Palavras-chave: Teoria da Resposta ao Item. Funcionamento Diferencial do Item. Intensidade de Sintomas Depressivos. Inventário de Depressão Beck. Traço latente. Modelo TRI para itens constrangedores.

INTRODUCTION

A latent trait is a variable that may be observed directly. In an attempt to measure it, it is necessary the use an instrument consisted of items which, presumably, reflect it. Establishing measure equivalence between groups which differ as to their characteristics such as school education, gender and race, for example, is important in the evaluation of mental health, so that these groups may be compared in terms of their measures of interesting traits, such as intensity of depressive symptoms, physical functioning or satisfaction with care, for instance[1]. Therefore, before comparing groups of respondents (according to their age or gender, for example) in terms of latent trait being measured, one must be confident that the items comprising the measure operate equivalently between the different groups[1]. In other words, there is a possibility that some items, specially psychological and/or psychiatric measures, work differently or have biases according to the different respondent groups[2]. If an item has a different response function for both groups, this item then is said to be biased[3].

In the literature on Item Response Theory (IRT), the term bias has been essentially replaced by the expression Differential Item Functioning (DIF). The DIF occurs when the probability of a determined response to an item of the instrument does not relate

to the latent trait in two or more respondent groups, i.e., when the probability of choosing as a response a category of an item does not depend only on the latent trait of the individual, but also on the fact that they belong to a given group (for example, the probability of choosing a response category is different between men and women with the same latent trait). More specifically, the DIF occurs when an item represents a different Item Characteristic Curve (ICC) for each group or, equivalently, when any parameter of the item differs between the groups. If there is a bias-free item, the answers to this item will be related only to the level of the latent trait that the item is trying to measure. If the item has a bias, then the answers to it will be related to some other factor besides the latent trait.

Many measuring instruments, especially in psychiatry, have items which may work in different ways within the different groups. Among those, the Beck Depression Inventory may be mentioned. It is an instrument which estimates the latent trait of the Intensity of Depressive Symptoms. Some studies report the presence of items with DIF in the BDI concerning gender[4-6]. The difference between the responses' distribution of men and women were observed in the items regarding crying, punishment, loss of libido, dissatisfaction, guilt and fatigability.

The presence of a great number of items with DIF in the BDI is a severe threat to the validity of the measure for intensity of depressive symptoms obtained by the IRT and to the conclusions based on the scores resulted from the items with and without DIF. A possible solution for this problem would be the elimination of those from the measuring instruments. However, this could compromise the measure of the latent trait, because for the items have information considered relevant, since the BDI was built in order to encompass all observable depressive symptoms[7]. The use of a model which allows the maintenance of all items in the instrument and, at the same time, contemplates the differences between the groups is actually a great alternative for the analysis of BDI data.

The IRT model for embarrassing items, proposed by Cúri et al.[8], is within this perspective, since it preserves such characteristics. Thus, this study aimed at identifying BDI items which have a DIF for gender, i.e., which have biases comparing men and women through the differential analysis of the item, adjusting the model for embarrassing items for the sample considered and comparing these results with the ones from the adjustment of the traditional two parameter logistic IRT model.

METHODS

SAMPLE

The individuals come from a cross-sectional study conducted in order to perform the adaption, normatization and validation of the Beck Scales into Portuguese, in a study conducted by Dr. Jurema Alcides Cunha and published in 2001[9].

The BDI scale, originally with 4 points, for the objectives of this work, was dichotomized in a way the response takes over the value 1 ($X_{ij} = 1$) when the individual *j* reports having the symptom described in item *i* (i.e., chooses one of the categories with scores 1, 2 or 3 of the determined item) and 0 ($X_{ij} = 0$) in case it does not represent that symptom.

CONSIDERED MODELS

Two IRT models were adopted for dichotomous variables (in this case being the absence or presence of the depressive symptom).

Unidimensional logistic model of 2 parameters (Model 1)

This is a IRT model for the dichotomic response, appropriated for the measures in which the item does not equally discriminate the levels of the latent trait[2,10]. The two parameters model predicts that the probability that the individual *j* presents the symptoms measured in item *i*, conditioned to its intensity on depressive symptoms, i.e., $P(X_{ii} = 1 \mid \theta_j, \zeta_i)$, as follows:

$$P(X_{ij} = 1 \mid \theta_j, \zeta_j) = \frac{1}{1+e^{-a_i(\theta_j, \zeta_j)}} \quad (1)$$

where: $i = 1, ..., 21$ items, $j = 1, ..., n$ individuals, $\zeta_j = (a_i, b_i)^t$, θ_j is the intensity of the depressive symptoms (latent trait) of the individual *j* (parameter of the individual); b_i is the parameter of gravity (position) of the item *i* and it represents the gravity of the depressive symptom described by the item *i* (when $\theta_j = b_i$, the probability of presence of the symptom *i* is 0.5); a_i is the discrimination (or inclination) parameter of the item *i*.

IRT model for embarrassing items (Model 2)

This model for dichotomous items, proposed by Cúri et al.[8], allow to differentiate the severity of the presence of depressive symptoms among individuals who are embarrassed and not embarrassed by a specific item so they have different behaviors face their respective ICC. The probability that individual *j* has or not the symptom measured in item *i* ($X_{ij} = 1$ or 0, respectively) and feel embarrassed or not by the item *i* ($C_{ij} = 1$ or 0, respectively) is:

$$\begin{aligned} P(X_{ij} = x_{ij}, C_{ij} = c_{ij} \mid \theta_j, \zeta_i) &= P(X_{ij} = x_{ij} \mid C_{ij} = c_{ij}, \theta_j, \zeta_i) \times P(C_{ij} = c_{ij} \mid \theta_j, \zeta_i) \\ &= [(P^{\star}_{ij})^{x_{ij}} (1-P^{\star}_{ij})^{(1-x_{ij})} \zeta_i]^{c_{ij}} \times [P_{ij}^{x_{ij}} (1-P_{ij})^{(1-x_{ij})} (1-\zeta_i)]^{(1-c_{ij})} \end{aligned} \quad (2)$$

where:
$\zeta_i=(a_{ij},b_{1j},b_{2j},\gamma_i,\delta_i)^t$; $P^{\star}_{ij}=\frac{\gamma_i}{1+e^{-a_i(\theta_j,\ b_{2i})}}=P(X_{ij}=1 \mid C_{ij}=1,\ \theta j,\ \zeta j)$; $P_{ij}=\frac{1}{1+e^{-a_i(\theta_j,\ b_{1i})}}=P(X_{ij}=1 \mid C_{ij}=0,\ \theta j,\ \zeta j)$; θ_j is the intensity of the depressive symptoms (latent trait) of the individual j (individual's parameter); b_{1i} is the severity parameter of item i for individuals who are not embarrassed, named, from now on, as the group with standard behaviors (women); b_{2i} is the severity parameter of item i for embarrassed individuals, named, from now on, as the group with different behavior (men); a_i is the discrimination (or inclination) parameter of item i; γ_i is the probability that the individual in the different behavior group states having the depressive symptom, i.e., the probability of an embarrassed individual saying they actually have the given symptom (notice that, in the not embarrassed group, it is assumed that this possibility is 1); δ_i is the probability of an individual presenting different behavior in relation to the symptom i. In this study, it will be assumed that the classification of embarrassed and not embarrassed individuals will be given according to gender, meaning, $C_{ij}=1$, for men, or 0, for women.

This model, in addition to the discrimination parameter of the item, common to the other IRT models, estimates other parameters which regard the different functioning of those items presenting DIF. For those items, the groups are comparable among each other, but you cannot do this when looking at the severity parameters. The parameters b_{1i} and b_{2i} express different probabilities of an individual presenting the symptom. The proper comparison between severities should be done between b_{1i} and $\theta^{\star}_{0,5}$.

Notice that b_{1i}, as in the 2-parameter logistic model, may be interpreted as the intensity of the depressive symptoms of an individual with standard behavior, such that the possibility of having the symptom i is 0.5 (when $\theta_j = b_i$, $P_{ij}=0.5$). On the other hand, for individuals with different behavior, when $\theta_j = b_{2i}$, $P^{\star}_{ij}=\gamma_i/2$. For this reason, comparing b_{1i} and b_{2i} does not make sense. In this work, the interpretation of the severity of the assessed symptoms by an item with DIF will be made through the comparison of intensities of the depressive symptoms of the individuals in each of the 2 groups, to whom the probability of having the symptoms is 0.5. In the group with standard behavior, it is $\theta_j = b_i$ and, in the group with different behavior, $\theta_j = -(1/a_i)\ln[(\gamma_i - 0.5)\ /\ 0.5] + b_{2i}$ (whose estimative is $\theta^{\star}_{0,5}$).

ANALYSIS STRATEGY

The analysis of the differential functioning of the item was performed with the use of the technique known as Item Response Theory Log-Likelihood Ratio (IRTLR), version 2.0b[11] using the IRTLRDIF software, developed by Dave Thissen and available in his homepage[12]. This procedure comes from the definition of Frederic Lord on DIF (then called the item's bias) and uses the log-likelihood ratio test as a significance test for the null hypothesis that the parameters of a response function of an item does not differ between groups — a significant result indicates the detection of DIF. As for IRT parameter models, the parameter group of the item is isomorphic (it has the same shape) to the response function of the item. The IRTLRDIF software has implemented two of the most used IRT models: the 3-parameter

logistic model and the graded response polytomous model of Samejima[13]. The 2-parameter logistic model (used in order to identify DIF items) is a special case of both previous models and, in this software, it is implemented as a gradual response model with two response categories. Because of the sample size, the significance level used for the identification of the DIF items was 1%.

The adjust of the 2-parameter IRT logistic model (model 1) and some embarrassing items (model 2) was performed through elaborate routines in WinBUGS, version 1.4.3[14]. The routines regarding both models used a Bayesian method of parameter estimation through Markov Chain Monte Carlo simulation (MCMC).

This study was submitted and approved by the Research Ethics Committee of the Universidade Federal do Rio Grande do Sul (UFRGS), meeting No. 37, minute No. 117, October 30, 2008.

RESULTS

The demographic characteristics of the sample may be found in the article *Teoria da resposta ao item aplicada ao Inventário de Depressão Beck*[15], wherethe Samejima's graded response model was adjusted to those data. It is noteworthy that the individuals in the samples are divided almost equally between men and women, with a slight advantage for the later ones.

The items presenting DIF, according to the log-likelihood ratio technique, were: sadness, dissatisfaction, guilt, punishment, crying, fatigability and loss of libido. The results of the adjusted model 1 are in Table 1 and the results of the adjusted model 2, considering the 8 items presenting DIF and the male group as the individuals embarrassed by those items, in Table 2.

The estimative of the discrimination parameters in models 1 and 2 (Tables 1 and 2, respectively) indicate that basically all items may be considered appropriate regarding this characteristic ($a_i > 1$[8,10]), except weight loss and self reproaching. The items with higher discrimination power are related to feeling of failure and dissatisfaction.

From the severity estimatives (b_i) of the depressive symptoms (Table 1), it is observed that symptoms of self reproaching and irritability are less severe and symptoms such as weight loss and suicidal ideas, the most severe ones. It is noteworthy that weight loss is the most severe depressive symptom and, at the same time, the one that less discriminate the population ($\hat{a}_{19}$=1.20). However, the suicidal ideas symptom is the second most severe one ($\hat{b}_9$=0.93) and it discriminates well the population for the severity level of the depressive symptoms ($\hat{a}_9$=1.71).

As for the severity of the symptoms, the results are the same of model 1 for all BDI items which do not present DIF. The difference occur in the eight remaining items. It is noticeable that guilt is more likely to be observed in higher levels of the depressive symptoms intensity ($\hat{b}_{1,5}$=0.58) among women and lower among men ($\theta^{\star}_{0,5}$ = 0.53). This is the opposite for the

Table 1. Mean and standard deviation of the posterior distribution of parameters in the 2-parameter logistic model (model 1).

Item	a_i (SD)	b_i (SD)
1 Sadness	2.38 (0.09)	0.16 (0.02)
2 Pessimism	2.41 (0.10)	0.76 (0.03)
3 Feeling of failure	2.90 (0.12)	0.71 (0.03)
4 Dissatisfaction	2.79 (0.11)	0.10 (0.02)
5 Guilt	1.98 (0.08)	0.57 (0.03)
6 Punishment	1.48 (0.06)	0.66 (0.04)
7 Self-loathing	2.50 (0.10)	0.51 (0.02)
8 Self-reproaching	0.97 (0.05)	-1.19 (0.06)
9 Suicidal thoughts	1.71 (0.08)	1.20 (0.04)
10 Crying	1.69 (0.06)	0.36 (0.03)
11 Irritability	1.07 (0.05)	-0.65 (0.04)
12 Social withdrawal	1.51 (0.06)	0.71 (0.03)
13 Indecision	1.99 (0.08)	0.13 (0.03)
14 Change in self-image	1.72 (0.07)	0.52 (0.03)
15 Difficulty to work	1.94 (0.07)	0.30 (0.03)
16 Insomnia	1.44 (0.06)	0.06 (0.03)
17 Fatigability	1.41 (0.06)	-0.25 (0.03)
18 Appetite loss	1.22 (0.06)	0.96 (0.04)
19 Weight loss	0.93 (0.06)	2.00 (0.11)
20 Somatic worries	1.19 (0.05)	0.35 (0.04)
21 Libido loss	1.43 (0.06)	0.74 (0.04)

a_i: discrimination parameter of the item *i*; b_i: severity parameter of the item *i*; SD: Standard deviation.

feelings of sadness, feeling of failure, dissatisfaction, punishment, loss of libido, crying and fatigability. For instance, the loss of libido has a higher chance of being observed in the lowest intensity levels of depressive symptoms ($\hat{b}_{1,21}$=0.48) among women and higher ones among men ($\theta^{\star}_{0,5} = 1.50$). Still as a result of model 2, it is estimated that the probability of a man with high intensity of depressive symptoms to express symptoms related to sadness, feelings of failure, dissatisfaction, guilt, punishment, crying, fatigability and loss of libido is higher or equal to 88% ($\hat{\gamma}_i \geq 0.88$). Figure 1 show the ICCs produced by models 1 and 2 for

Table 2. Mean and standard deviation of the posterior distribution of parameters of the model for embarrassing items (model 2).

Item	a_i (SD)	b_{1i} (SD)	b_{2i} (SD)	$\theta^*_{0,5}$	γ_i (SD)	δ_i (SD)
1 Sadness	2.37 (0.12)	0.07 (0.03)	0.16 (0.07)	0.28	0.88 (0.04)	0.45 (0.008)
2 Pessimism	2.34 (0.10)	0.76 (0.03)	–	–	–	–
3 Feeling of failure	2.94 (0.15)	0.68 (0.03)	0.70 (0.06)	0.76	0.92 (0.06)	0.45 (0.007)
4 Dissatisfaction	2.87 (0.14)	0.04 (0.03)	0.06 (0.05)	0.14	0.90 (0.04)	0.45 (0.007)
5 Guilt	2.06 (0.09)	0.58 (0.03)	0.50 (0.05)	0.53	0.97 (0.03)	0.45 (0.007)
6 Punishment	1.53 (0.07)	0.65 (0.04)	0.54 (0.10)	0.67	0.91 (0.06)	0.45 (0.007)
7 Self-loathing	2.53 (0.10)	0.51 (0.02)	–	–	–	–
8 Self-reproaching	0.99 (0.05)	-1.17 (0.06)	–	–	–	–
9 Suicidal thoughts	1.71 (0.08)	1.20 (0.04)	–	–	–	–
10 Crying	1.62 (0.07)	0.29 (0.04)	0.42 (0.07)	0.49	0.95 (0.04)	0.45 (0.008)
11 Irritability	1.08 (0.05)	-0.65 (0.04)	–	–	–	–
12 Social withdrawal	1.51 (0.06)	0.71 (0.03)	–	–	–	–
13 Indecision	1.99 (0.08)	0.13 (0.03)	–	–	–	–
14 Change in self-image	1.71 (0.07)	0.52 (0.03)	–	–	–	–
15 Difficulty to work	1.91 (0.07)	0.30 (0.03)	–	–	–	–
16 Insomnia	1.43 (0.06)	0.06 (0.03)	–	–	–	–
17 Fatigability	1.38 (0.06)	-0.33 (0.05)	-0.25 (0.07)	-0.19	0.96 (0.03)	0.45 (0.007)
18 Appetite loss	1.21 (0.06)	0.96 (0.05)	–	–	–	–
19 Weight loss	0.93 (0.06)	2.00 (0.10)	–	–	–	–
20 Somatic worries	1.18 (0.05)	0.35 (0.03)	–	–	–	–
21 Libido loss	1.19 (0.06)	0.48 (0.04)	1.39 (0.11)	1.50	0.94 (0.05)	0.45 (0.007)

a_i: discrimination parameter of the item *i*; b_{1i}: severity parameter of the item *i* for individuals in the female group; SD: standard deviation; b_{2i}: severity parameter of the item *i* for individuals in the male group; $\theta^*_{0,5}$: intensity level of depressive symptoms of an individual in the groups with differentiated behavior, where the probability of having the symptom is 0.5; γ_i: probability of a individual in the male group saying he has a depressive symptom, i.e., probability of having symptoms among men with a high intensity level of the depressive symptoms; δ_i: probability of an individual having differentiated behavior in relations to the symptom *i*.

item 21, regarding loss of libido. Here, it is evident the advantage of the use of model 2 in relation to model 1, since differences in behavior in a DIF item, in relation to their severity, is clearly shown for both compared groups.

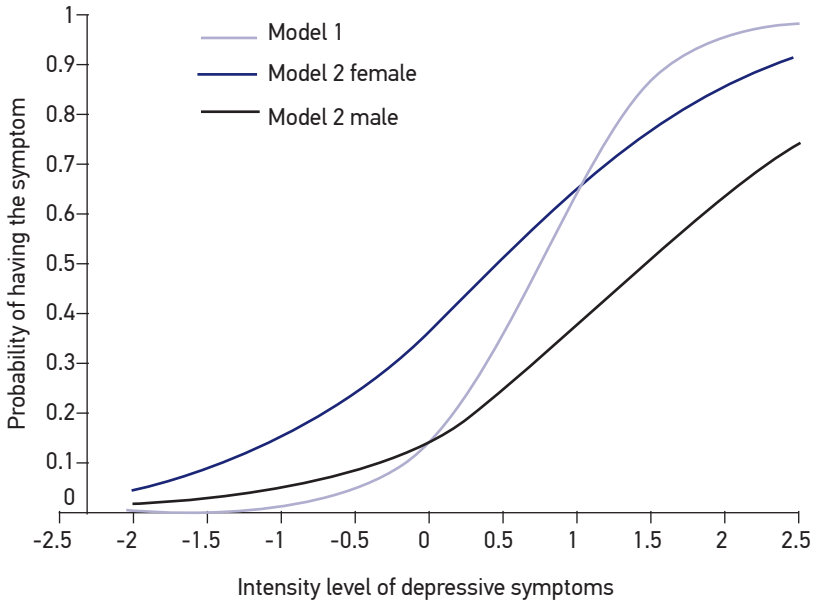

Figure 1. Item Characteristic Curve (ICC) for the symptom loss of libido (item 21) according to the 2-parameter logistic model (1) and for the Embarrassing Item model (2) for male and female gender.

The depressive symptoms levels estimated under the IRT models are in the same symptom severity scale estimated for each BDI item; therefore, they are comparable. The 95th percentiles of the depressive symptoms intensity levels are 1.598 and 1.593 for models 1 and 2, respectively. From the 201 individuals with depressive symptoms severity higher than 95th percentile for each model, 194 are classified equally by both models. The characteristics (Table 3) of this group show that almost 80% are derived from the psychiatric group, approximately 68% of them are women, most of them (over 58%) do not have a partner and they are, on average, 37 years old.

The depressive symptoms intensity estimatives obtained according to models 1 and 2 present high association, with correlation coefficient equal to 0.99.

DISCUSSION

We used the two-parameter logistic model (model 1) in order to compare it to the model for embarrassing items (model 2) because both of them include the parameters for discrimination

Table 3. Description of individuals with high level of depressive symptoms, estimated as a value above the 95th percentile.

Variable	Psychiatric	Clinical	Overall population	Total
Model 1	n = 157	n = 18	n = 26	n = 201
Age (n = 200)				
Mean	39.18	34.67	29.76	37.60
Standard deviation	12.5	13.26	13.83	13.08
Minimum	15	18	18	15
Maximum	75	64	67	75
School education (n = 194)				
Less than 5 years	54	9	12	38.1%
Complete Elementary School	43	3	5	25.9%
Complete High School	43	5	5	26.9%
Complete College Degree	11	0	4	7.7%
Marital status (n = 199)				
Single	45	8	19	36.2%
Married	74	8	1	41.7%
Separated, divorced or widow(er)	36	2	6	22.1%
Gender (n = 201)				
Male	36	11	16	31.3%
Female	121	7	10	68.7 %
Model 2	n = 155	n = 17	n = 29	n = 201
Age (n = 200)				
Mean	39.03	34.12	28.64	37.15
Standard deviation	12.46	13.45	13.48	13.16
Minimum	15	18	15	15
Maximum	75	64	67	75
School education (n = 194)				
Less than 5 years	53	8	13	37.6%
Complete Elementary School	43	3	6	26.4%
Complete High School	43	5	5	26.9%
Complete College Degree	10	0	4	7.1%
Marital status (n = 199)				
Single	47	8	22	38.7%
Married	72	7	1	40.2%
Separated, divorced or widow(er)	34	2	6	21.1%
Gender (n = 201)				
Male	38	10	18	32.8%
Female	117	7	11	67.2 %

$\hat{\theta}$: Latent trait (depressive symptoms level) estimated from the sample data.
$\hat{\theta}$ =1.598 for model 1 and $\hat{\theta}$=1.593 for model 2.

and depressive symptoms severity. Other studies have already used the 2-parameter logistic model for psychiatric data: Schaeffer[16], in 1988, adjusted this model to the response for 11 depression symptoms for which there are 4 categories of answers ("never", "once up until now", "relatively often" and "many times") and Kessler et al.[17] used it in building 2 scales (one of them with 10 items and the other one with 6) on mental health.

The findings regarding model 1, as for the presence of DIF in eight BDI items, show that men and women with the same depressive symptoms intensity responded differently to the items sadness, feeling of failure, dissatisfaction, guilt, punishment, crying, fatigability and loss of libido. Several studies[4-6,18-24] corroborate these findings; however, the different functioning (DIF) of the item crying in relation to gender is what is observed in most of them. A good part of the studies which show the gender difference in relation to crying emphasizes that women tend to cry more often than men[5,6,21]. This may be another reflex of the well known tendency of women crying more easily and intensely than men in a variety of anguishing situations rather than being an indicator of gender difference in the prevalence of depression[18]. This conclusion suggests crying as a response for anguish is, mostly, determined by gender; therefore, men and women with the same intensity level of depressive symptoms will probably not answer to the item crying in the same way, which is confirmed in this study. Originally, the BDI scale has four categories, considering that, specially on crying, the higher importance category states that the individual lost their ability to cry, even if they feel like it, while the first three categories determine an increase in the number of times they are used to crying. Of all the men who got a 1 in the dichotomous scale, over half of them answered category 3, the same occurring when observing only men in the group of 5% higher estimated levels for the depressive symptoms intensity, showing they are serious candidate to a positive diagnosis on depression. This loss of the capacity of crying by men is also present in the study by Hammen and Padesk[4], in which the BDI is worked in its original scale.

When comparing the result found for models 1 and 2 in relation to the discrimination on depressive symptoms by the items, it is possible to notice that, considering items with values of $a_i \geq 1$[8,10] as having reasonable discrimination, the same 19 items in the two models are in this category, except only for loss of weight and self reproaching. In the study of Cúri et al.[8], in which a three-parameter logistic model was adjusted, only the loss of appetite had an estimative below this cutoff point, however, the symptom of weight loss is very close to this region. On the other hand, the most discriminated symptoms, feeling of failure and dissatisfaction, are the ones present in models 1 and 2 and in the one adjusted by Cúri et al.[8], showing that these are important symptoms in the discrimination of population for the intensity of their depressive symptoms.

A result shown in model 2 was the greater severity on the symptom loss of libido for men rather than for women, since there is a higher probability of its occurrence in

higher levels of depressive symptoms intensity for men than for women. The importance of loss of libido for men is shown in several studies. In a clinical randomized trial on the sexual effects (such as improvement in loss of libido and erectile dysfunction) of testosterone replacement in men diagnosed with deeper depression[25], the authors intended to verify whether the treatment would be efficient in this population the same way it is in general population. However, the testosterone replacement did not have the expected known effect, indicating that maybe the problem was the condition of the depression in the target population.

The groups formed by the 5% of individuals with higher estimative of depressive symptoms intensity (latent trait being measured), obtained from models 1 and 2, evidences female superiority in the psychiatric group, since over 75% of these groups is formed by women. These data are consistent to the evidence that depression is twice to three times more common among adolescent and adult women than it is among adolescent and adult men[26], because these women have higher levels of depressive symptoms intensity, being strong candidates for having a positive depression diagnosis.

It is important to emphasize that models 1 and 2 track basically the same individuals as belonging to these groups with the highest estimatives on intensity of depressive symptoms. From 201, only 7 women and 7 men had disagreeing classifications, considering that model 1 tracks more women above the 95th percentile and model 2 tracks more men above their respective 95th percentile. These differences seem to occur due to the fact that the intensity levels of the estimated depressive symptoms for these individuals are at the limits of their respective 95th percentile.

CONCLUSION

Two IRT models were adjusted to the dichotomous BDI data: the 2-parameter logistic model (model 1) and the IRT model for embarrassing items (model 2), which includes the presence of DIF items.

The results found in models 1 and 2 are quite similar, especially in the case of the estimatives on the intensity of depressive symptoms for each individual, proved by the high correlation between the IRT scores. Despite that, model 2 is still better, since it shows the differences in the severity of depressive symptoms in the evaluated groups, bringing, this way, more information to the researcher on the studied population. The use of a model with wider reach in terms of target population may be a very useful alternative also in the clinical field, where the existence of validated models may contribute in the identification of individuals as potentially depressed.

A limitation of this work is that it consists of an empiric comparison, being necessary a broader study, using, for example, simulated data.

Still, as commented by Cúri et al.[8], it is necessary the extension of model 2 to items with ordinal responses, since, as well as the BDI, countless instrument of psychiatric measures have items of ordinal responses, and their transformation in dichotomous items (absence or presence, for example) do not make complete usage of the available information, and possibly producing inconsistent results.

REFERENCES

1. Teresi JA, Fleishman JA. Differential item functioning and health assessment. Qual Life Res 2007; 16(Suppl 1): 33-42.
2. Embretson SE, Reise SP. Item Response Theory for Psychologists. New Jersey: Lawrence Erlbaum Associates; 2000.
3. Lord F. Applications of item response theory to practical testing problems. Hillsdale: Routledge; 1980.
4. Hammen CL, Padesky CA. Sex differences in the expression of depressive responses on the Beck Depression Inventory. J Abnorm Psychol 1977; 86(6): 609-14.
5. Santor D, Ramsay J, Zuroff D. Nonparametric item analyses of the Beck Depression Inventory: evaluating gender item bias and response option weights. Psychol Assess 1994; 6: 255-70.
6. Salokangas RK, Vaahtera K, Pacriev S, Sohlman B, Lehtinen V. Gender differences in depressive symptoms. An artefact caused by measurement instruments? J Affect Disord 2002; 68(2-3): 215-20.
7. Beck AT, Steer RA. Beck Depression Inventory. Manual. San Antonio, TX: Psychological Corporation; 1993.
8. Cúri M, Singer JM, Andrade DF. A model for psychiatric questionnaires with embarrassing items. Stat Methods Med Res 2001; 20(5): 451-70.
9. Cunha JA. Manual da versão em português das Escalas Beck. São Paulo: Casa do Psicólogo; 2001.
10. Andrade DF, Tavares HR, Valle RC. Teoria da Resposta ao Item: conceitos e aplicações. In: Anais do 14º SINAPE; 2000 jul 28; Caxambu (MG).
11. Teresi JA, Ocepek-Welikson K, Kleinman M, Cook KF, Crane PK, Gibbons LE, et al. Evaluating measurement equivalence using the item response theory log-likehood ratio (IRTLR) method to assess differential item functioning (DIF): applications (with illustrations) to measure of physical functioning ability and general distress. Qual Life Res 2007; 16(Suppl 1): 43-68.
12. Thissen D. Dave Thissen's Front Page. Disponível em www.unc.edu/~dthissen/dl.html. (Acessado em 26 de julho de 2008).
13. Samejima F. Estimation of latent ability using a response pattern of graded scores. Madison (WI): Psychometric Society; 1969.
14. Lunn DJ, Thomas A, Best N, Spiegelhalter D. Winbugs – a Bayesian modeling framework: concepts, structure, and extensibility. Stat Comput 2000; 10: 325-37.
15. Castro SMJ, Trentini C, Riboldi J. Teoria da resposta ao item aplicada ao Inventário de Depressão Beck. Rev Bras Epidemiol 2010; 13(3): 487-501.
16. Schaeffer NC. An application of item response theory to the measurement of depression. Sociol Methodol 1988; 18: 271-307.
17. Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SLT, et al. Short screening scales to monitor population prevalence and trends in non-specific psychological distress. Psychological Medicine 2002; 32: 959-76.
18. Romans SE, Tyas J, Cohen MM, Silverstone T. Gender differences in the symptoms of major depressive disorder. J Nerv Ment Dis 2007; 195(11): 905-11.
19. Stommel M, Given BA, Given CW, Kalaian HA, Schulz R, McCorkle R. Gender bias in the measurement properties of the Center for Epidemiologic Studies Depression Scale (CES-D). Psychiatry Res 1993; 49(3): 239-50.
20. Wilhelm K, Parker G, Asghari A. Sex differences in the experience of depressed mood state over fifteen years. Soc Psychiatry Psychiatr Epidemiol 1998; 33(1): 16-20.
21. Carter JD, Joyce PR, Mulder RT, Luty SE, McKenzie J. Gender differences in the presentation of depressed outpatients: a comparison of descriptive variables. J Affect Disord 2000; 61(1-2): 59-67.
22. Gelin M, Zumbo B. Differential item functioning results may change depending on how an item is scored: an ilustration with the Center for Epidemiologic Studies Depression Scale. Educ Psychol Meas 2003; 63: 65-74.
23. Wenzel A, Steer RA, Beck AT. Are there any gender differences in frequency of self-reported somatic symptoms of depression? J Affect Disord 2005; 89(1-3): 177-81.

24. Angst J, Gamma A, Gastpar M, Lepine JP, Mendlewicz J, Tylee A; Depression Research in European Society Study. Gender differences in depression. Epidemiological findings from the European DEPRES I and II studies. Eur Arch Psychiatry Clin Neurosci 2002; 252(5): 201-9.

25. Seidman SN, Roose SP. The sexual effects of testosterone replacement in depressed men: randomized, placebo-controlled clinical trial. J Sex Marital Ther 2006; 32(3): 267-73.

26. Beyer JL, Nash J, Shelton R, Loosen PT. Transtorno depressivo maior. In: Jorge MR. Manual diagnóstico e estatístico de transtornos mentais. 4ª edição. Porto Alegre: Artmed; 2000. p. 288-324.

Received on: 12/13/2012
Final version presented on: 03/29/2013
Accepted on: 06/05/2013