Biometrical Journal **55** (2013) 4, 527–540 DOI: 10.1002/bimj.201200197

Improving psychometric assessment of the Beck Depression Inventory using Multidimensional Item Response Theory

Tiago M. Fragoso*[,1] and **Mariana Cúri**[2]

[1] Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo 05508-090, Brazil
[2] Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos 13560-970, Brazil

Received 02 October 2012; revised 19 January 2013; accepted 31 January 2013

We studied the latent factor structure of the Beck Depression Inventory (BDI) under the light of Multidimensional Item Response Theory models. Under a Bayesian Markov chain Monte Carlo setting, we chose the most adequate model, estimated its parameters and verified its fit to the data. An evaluation of the inventory in terms of the assumed dimensions seems to agree with previous investigations in the factor structure of the BDI present in the literature. Cognitive and somatic-affective latent traits were identified in the analysis making possible the interpretation of symptom evolution along these dimensions, in terms of probability of their appearance.

Keywords: Beck Depression Inventory; Item Response models; Psychometrics.

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

Item Response Theory (IRT) models are most frequently employed in educational settings to model the interaction between a multiple choice item and a respondent in terms of the probability of an answer which varies in function of a latent trait belonging to the individual, and item parameters inherent to the item answered.

Such models can be of use to psychometric applications due to its natural formulation, only shifting the interpretation of a correct or incorrect response to the item to a present or not present behavior or symptom, according to the latent factor influencing the answers. Most studies in this field are composed of a variety of Classical Test Theory, factor analysis, discriminant analysis, and other statistical tools (check Wang et al., 2005, for a typical analysis) and would be enhanced by a single interpretable theoretical framework, as IRT.

A central assumption in most IRT applications, however, is that only one latent trait significantly influences the respondent. Such assumption is not always warranted. Multiple analysis of Beck Depression Inventory (BDI) applications hint at a separate second dimension to depression (Steer et al., 1999; Cohen, 2008), which calls for an extension of said unidimensional IRT models like the one used in Castro et al. (2010).

Adding up to the need to expand IRT models to measure more than one latent trait, the quality of parameter estimates is at stake when additional dimensions are ignored due to the lack of conditional independence (i.e., independence between answered items given a subject with fixed latent trait), which

*Corresponding author: e-mail: fragoso@ime.usp.br, Phone: +55-11-3091-6129, Fax: +55-11-3091-6130 is paramount in obtaining item and individual parameter estimates that are more precise and less biased. A complete study in a similar setting can be found in Finch (2011).

However, with the additional interpretation given by extra dimensions, there is a correspondent increase in model complexity. The multiple latent traits might add up together to affect the response probability, effectively compensating a deficiency in one trait with a surplus on other. Models with this behavior are called compensatory models. Alternatively, each trait may act separately, requiring an elevated trait level on every dimension in order to obtain a high response probability. Such models are called noncompensatory.

Compensatory Multidimensional Item Response Theory (MIRT) models are most commonly applied and available in popular software for IRT data analysis, such as NOHARM (Fraser, 1988) and TESTFACT (Wilson et al., 1987). Noncompensatory models are less popular, but estimation procedures were presented in Bolt and Lall (2003) using Bayesian methods.

In this work, we analyzed a dataset with the responses of 1111 college students to the 21 items of BDI. We employed uni- and multidimensional IRT models. The compensatory nature of the traits or lack thereof is a premise not naturally present in most practical settings, so both models were fit to the data and selected using the deviance information criterion (DIC, Gelman et al., 2004) as a model choice criteria. The DIC was also used to select between an unidimensional or multidimensional IRT models, deeming the model with the smaller DIC most appropriate to the data. Although there is a rich literature in dimension choice for MIRT models, there is plenty of psychological literature indicating more than one dimension to depression (Vanheule et al., 2008). Also, a likelihood ratio test using marginal likelihood estimates significantly rejects one dimensional models in favor of two dimensions (p-value ≤ 0.01) confirming the psychological findings.

With the most appropriate model, we then obtain point estimates to item and individual parameters and check model fit by a posterior predictive check using the obtained a posteriori density and a test score statistic similarly as in Beguin and Glas (2001) and Sinharay et al. (2006).

This work is structured as follows. Throughout the next session, we describe the application of the BDI used for the modeling (Section 2.1), fundamental models of MIRT (Section 2.2.1), Bayesian estimation and model choice (Section 2.2.2), and model fit (Section 2.2.3). We then present the results of the modeling in Section 3 and perform the classification and interpretation of said results in Section 4. Finally, Section 5 presents the conclusion of this study.

2 Methods

2.1 Description of the dataset

We employed a dataset of an application of the Portuguese version of the BDI to 1111 college students gently given by Dr. Teng Chei-Tung from the Hospital das Clínicas, Faculdade de Medicina, Universidade de São Paulo. The patients responded to the 21 items composing of four possible answers regarding the intensity of a symptom of depression such as guilt or insomnia in a scale of 0, indicating absence of said symptom, to 3, indicating a constant presence of the symptom evaluated in the past two weeks.

Differently from Castro et al. (2010), we did not take the responses in the 0–3 scale. In order to employ MIRT models to the data, we dichotomized the dataset to 0 if the respondent indicated an absence of the symptom, and 1 if the answer was 1, 2, or 3 in the former scale. The rationale employed is that the answer is taken as an indication of the symptom's presence.

We provide a translated version of the Portuguese language BDI applied to our sample in Dr. Cúri's homepage.[1]

[1] http://www.icmc.usp.br/~mcuri

2.2 Statistical analysis

Being still a novel procedure in many applications, we take the length to explain the statistical tools used in MIRT data analysis and model estimation in a bit more depth. A more complete reference can be found in Bolt and Kim (2007).

2.2.1 Multidimensional item response models

Extending IRT models to allow for other latent traits carries an extra assumption regarding the interaction among latent traits resulting in the response probability. Models that assume that a low value of one latent trait can be compensated by a higher value in another are called compensatory models, whereas models where the opposite assumption is made are regarded as noncompensatory.

In order to choose the most adequate model for the data, we employ a compensatory model, the Multidimensional Logistic 2 Parameter (ML2P) model (McKinley and Reckase, 1980) where the presence of the symptom evaluated in item i, $i = 1, \ldots, 21$, is indicated by respondent j, $j = 1, \ldots, 1111$, $(X_{ij} = 1)$ with probability

$$P(X_{ij} = 1 | \boldsymbol{\theta}_j, \boldsymbol{a}_i, b_i) = \frac{1}{1 + \exp\left(-\sum_{k=1}^{p} a_{ki}\theta_{kj} + b_i\right)}, \tag{1}$$

where the p latent traits for the N-th individual are represented by the vector $\boldsymbol{\theta}_j = (\theta_{1j}, \ldots, \theta_{pj})$, $j = 1, \ldots, N$, and $k = 1, \ldots, p$. The item parameters are a vector of p discrimination parameters, one for each dimension evaluated, $\boldsymbol{a}_i = (a_{1i}, \ldots, a_{pi})$, and a difficulty parameter for the item i, b_i.

Being only slightly similar to a regression model, it is appropriate to explain the model parameters in light of the application at hand. The latent traits represent the aspects of the latent construct being measured, which we assume independent and influencing all subjects.

As for the discrimination vector, it shares a similar interpretation with the factor loadings obtained in factor analysis models more commonly used in Classical Test Theory. Each component of the vector represents the “loading” of each latent trait on the response probability. Items with a high discrimination on a latent trait are more likely answered in different ways by individuals with different latent traits.

The difficulty parameter can be interpreted as the severity of the symptom described in the item, for instance, if the difficulty of BDI’s ninth item (regarding suicidal thoughts) is high, it means that it takes an elevated value of the latent traits in order to obtain an endorsement of the individual to the existence of the symptom (Thomas, 2010).

In order to evaluate the MIRT models against the premise of unidimensionality, we fit the model assuming $p = 1$, resulting in the unidimensional logistic 2 parameter model, and $p = 2$, implying that two latent traits influence the answers.

We also employed the Non Compensatory Logistic 2 Parameter model (Sympson, 1977; Whitely, 1980) that models the probability of a response indicating presence of the symptom evaluated in the i-th item by the j-th individual by the probability

$$P(X_{ij} = 1 | \boldsymbol{\theta}_j, \boldsymbol{a}_i, \boldsymbol{b}_i) = \prod_{k=1}^{p} \frac{1}{1 + \exp\left(-a_{ki}\theta_{jk} + b_{ik}\right)}, \tag{2}$$

where the parameters are defined in a similar fashion as in (1), the only distinction being that we now deal with as many difficulty parameters as there are dimensions, since the item is assumed to evaluate each dimension of the latent trait separately, implying a severity of each aspect of depression required to give an affirmative answer.

The noncompensatory model differs from the model (1) in the fundamental assumption that a high level in both latent traits are required in order to obtain an affirmative response from the individual.

Due to the additive nature of the compensatory model (1), a lower level in one latent trait can be “compensated” by a higher level in another, obtaining the same response probability of a more balanced individual, hence the denomination.

Albeit enriching in terms of interpretation to the latent traits, the interpretation of the item parameters are not as straightforward as the unidimensional case due to the multitude of discrimination parameters and a difficulty parameter that is not in the same scale as the latent trait.

For the ML2P, Reckase (1996) proposes some summary statistics in order to obtain similar conclusions when evaluating unidimensional and multidimensional items alike. The discriminatory power of the i-th item is defined as

$$MDISC_i = \sqrt{\sum_{k=1}^{p} a_{ki}^2}, \tag{3}$$

which encompasses all the dimensions evaluated and gives a measure of the discrimination across all latent traits assumed. When $p = 1$, we simply have the discrimination parameter for the sole latent trait, in analogy with the unidimensional models.

A modification to the item difficulty in order to put it in the latent trait scale is given by

$$MDIFF_i = \frac{-b_i}{MDISC_i}. \tag{4}$$

Note that, for $p = 1$, we simply rewrite (1) in terms of the distance between the difficulty and the latent trait, which is the most common formulation of the unidimensional IRT models, in contrast with the linear model formulation usually encountered in the literature for MIRT models.

2.2.2 *Parameter estimation*

To obtain estimates for the item and individual parameters, we used a joint Bayesian estimation through Markov chain Monte Carlo (MCMC) sampling as discussed in Patz and Junker (1999). In this estimation paradigm, we assume a priori probability distributions for the parameter we aim to estimate and obtain possible values of said parameters according the probability distribution of said parameters given the observed data. Such distribution is called a posteriori distribution and summary statistics of the sampled values can be used to obtain point estimate and credibility intervals for the model parameters.

We specified the a priori as follows: for the latent traits, we assumed the frequent standard normal (zero mean and unit variance) frequently employed in the literature, as for each component of the discrimination vector across all items, we assumed a log-normal distribution of location parameter 1 and scale parameter 0.5 and a standard normal distribution for the difficulty parameters for every item as well.

Note that in most IRT applications, the latent traits are assumed as drawn from a standard normal distributions for reasons of identifiability, that is, a scale change for the latent traits would produce the same likelihood, producing an array of technical and practical problems. We assume a standard distribution for the normal as a reasonable assumption that the latent traits are a priori independent among each other and are homoscedastic in the population. The reasons for this premise are that other identification procedures like ad hoc restraints on item parameters like the ones adopted in Bolt and Lall (2003) imply assumptions about the items we did not want to carry over to our analysis of the item parameters. More general modelings can allow for correlated traits like the one presented in Fu et al. (2009).

As point estimates for the parameters, we calculated the sample mean for the adequate samples obtained from the MCMC procedure, with its fine tuning parameters (autocorrelation and burn-in) determined by the Raftery and Lewis (1995) and Geweke (1992) procedures. Ninety-five percent credibility intervals were obtained by taking the sample 0.025 and 0.975 quantiles.

Table 1 Deviance Information Criterion (DIC) for the three possible models.

	UL2P	ML2P	MN2P
DIC	24,645	24,023	41,610

For model choice, we employed the DIC as exposed in Gelman et al. (2004) with the model with the smallest DIC value being selected as most adequate. Estimation of the DIC index was performed in the R Statistical Software (R Development Core Team, 2005) handling the samples obtained from the (Spiegelhalter et al., 2003) software with the coda package (Plummer et al., 2006). The R scripts used in our analysis can be obtained at Dr. Cúri's webpage[2] or from request from the authors. The reader further interested in the fitting of IRT models using R and BUGS software is also directed to Li and Baser (2012) for a more complete tutorial.

2.2.3 Model fit

Under the Bayesian MCMC framework, we can evaluate the model fit to some desirable aspects of the dataset using posteriori predictive checks (Gelman et al., 2004). This technique allows us to specify an aspect of the dataset we find desirable to measure model fit and summarize it into a predictive index, which is called a p-value because it is a proportion.

We used as test statistic the difference between the predicted scores under the model and the observed scores in the sample as adopted by Beguin and Glas (2001)

$$T(\boldsymbol{X}_r) = \sum_{k=0}^{Q} \frac{\left(N_k^{(r)} - f^{(r)}(k)\right)^2}{f^{(r)}(k)}, \tag{5}$$

where $N_k^{(r)}$ represents the observed frequency of the k-th score ($k = 0, \ldots, 21$) under the r-th dataset obtained by using the r-th MCMC sample to simulate the responses whereas $f^{(r)}(k)$ represents the frequency of said score expected by the model on the r-th simulated dataset using the sample.

Calculating the frequency the value the statistic (5) gets over the samples is lower than its value calculated on the original dataset, one can obtain a number similar to the p-value where extreme values (i.e., too close to 0 or 1) indicate poor model fit and average values indicate adequate model fit. We can also compute the distribution for the scores using the simulated datasets and use it as a graphical check of model fit. A more detailed exploration of the technique in MIRT settings is given in Beguin and Glas (2001).

3 Results

The models were fitted to the data using the MCMC sampling algorithms available in the WinBUGS (Spiegelhalter et al., 2003) software. We then took 105,000 samples, discarding the first 5000 as burn-in and took every 50 iteration as thinning, resulting in an effective size of around 2000. The procedures mentioned in Section 2.2.2 indicated convergence of the chain.

We then calculated the DIC for the unidimensional model (UL2P), the multidimensional compensatory model (ML2P) and the multidimensional noncompensatory model (MN2P). The results can be found in Table 1. The lowest DIC is the one from the ML2P, therefore, we adopted it as the most adequate model for the data.

We computed the sample means and standard deviations to obtain item parameter estimations as displayed in Table 2 and in analogy with factor analysis approaches, we computed standardized

[2] http://www.icmc.usp.br/∼mcuri

Table 2 Point estimates and standard deviations for the BDI item parameters.

Item	a_1		a_2		b	
	Mean	(std)	Mean	(std)	Mean	(std)
1	1.14	(0.13)	0.70	(0.14)	0.76	(0.09)
2	1.06	(0.13)	0.97	(0.15)	−0.69	(0.09)
3	2.64	(0.27)	0.71	(0.23)	−1.58	(0.16)
4	1.09	(0.16)	1.46	(0.19)	0.11	(0.09)
5	2.43	(0.25)	0.46	(0.20)	−1.40	(0.14)
6	1.31	(0.13)	0.34	(0.13)	−0.89	(0.09)
7	3.09	(0.40)	0.43	(0.23)	0.47	(0.14)
8	1.45	(0.14)	0.30	(0.13)	0.79	(0.09)
9	1.22	(0.16)	0.96	(0.18)	−2.49	(0.15)
10	0.80	(0.12)	0.86	(0.13)	−1.17	(0.09)
11	0.59	(0.12)	1.03	(0.13)	0.05	(0.07)
12	0.70	(0.13)	1.17	(0.16)	−0.73	(0.09)
13	0.99	(0.14)	1.16	(0.15)	−0.61	(0.09)
14	0.63	(0.10)	0.64	(0.11)	−0.73	(0.07)
15	1.12	(0.15)	1.30	(0.16)	−0.18	(0.09)
16	0.32	(0.11)	1.21	(0.17)	−0.02	(0.08)
17	0.53	(0.14)	1.40	(0.19)	0.58	(0.09)
18	0.47	(0.11)	0.73	(0.11)	−1.08	(0.08)
19	0.15	(0.08)	0.58	(0.13)	−2.26	(0.12)
20	0.50	(0.09)	0.59	(0.11)	−0.82	(0.07)
21	0.36	(0.10)	0.74	(0.12)	−1.33	(0.09)

a_k: Item discrimination on the k-th latent trait $k = 1, 2$.
b : Item difficulty.
std: Standard deviation.

discriminations for both traits in every item by dividing the estimated discrimination by the item discriminating power 3, obtaining a "loading" of sorts which we use to classify items between traits using an established cut-off point. In the present work, standardized discriminations higher than 0.5 were considered important and used for classification.

In order to evaluate model fit, we calculated the test statistic 5 for the observed data and for the 2000 datasets simulated using the MCMC samples and calculated the frequency of the event consisting of the value of the statistic being lower than the value calculated for the real dataset. The obtained p-value was 0.13, indicating an adequate capability of predicting test scores, which is also corroborated by the test score mean frequency and 95% credibility intervals (Fig. 1).

Regarding the latent trait estimates, we obtained 2222 values, 2 for each patient. Summary statistics of the estimates resulted in a sample mean of approximately 0 for both dimensions, variances of 0.62 and 0.71 respectively and a correlation between latent traits of 0.25.

4 Discussion

As a mean of comparison with Castro et al. (2010) which evaluated the Brazilian version of the BDI using an unidimensional IRT model, we fit the same model to the dataset as well as a compensatory

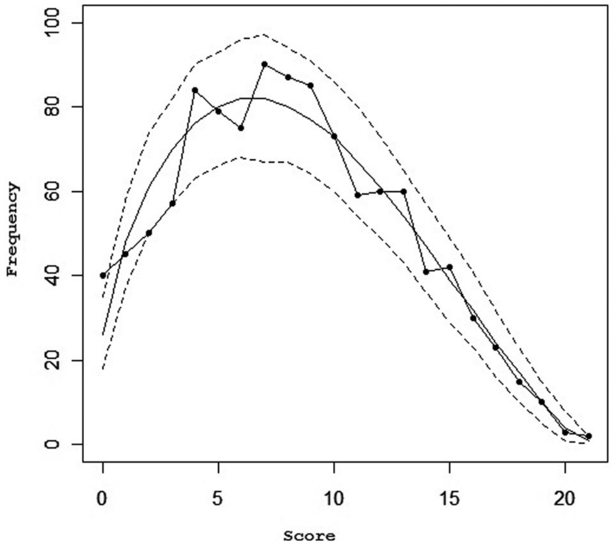

Figure 1 Observed scores (connected dots), expected scores (full line), and 95% credibility interval (dashed line).

and noncompensatory two-dimensional model in order to evaluate certain assumptions in light of the model fit, which we measure by the DIC.

The results as expressed in Table 1 point out that the compensatory two dimensional model is the most adequate for the data, having a lower value of the DIC than the unidimensional model, therefore indicating two factors influencing individual responses. A lower value for the DIC in the compensatory model and a much higher one on the noncompensatory model also indicate that both aspects of depression interact among each other, compensating a deficiency in one dimension with a surplus in another.

Fitting the compensatory model to the data, we then obtain point estimates and standard errors for both discrimination parameters and difficulties, as one can see in Table 2. We observe that the difficulty parameters are overall very low and negative values, indicating that, after the scaling shown in (4), we have high difficulties consistent with the purpose of the BDI to evaluate depressed individuals instead of the average population (which we center in 0).

The latent trait estimates can be interpreted looking at the respondents' response pattern. For instance, the individual with the higher score in the first dimension, individual 322, with estimates $\hat{\boldsymbol{\theta}}_{322} = (2.02, 0.33)$, representing a high level on the first latent trait dimension and an average value for the second. Using Table 4, we can conclude the individual must have the symptoms represented by items 1,4, 7, and 8 with high probability. Indeed, the individual has positive responses for items 1, 7, and 8. For the second dimension, we can see in Table 5 that items 4, 11, 12, 15, 16, and 17 are expected with high probability for individuals with high latent trait values on the second latent trait. The individual presents only symptoms indicated by items 16 and 17, the remaining symptoms expected to appear following the increase of depression associated with the second dimension.

As for the discrimination parameters, the best way to interpret might be scaling over the overall discriminative power $MDISC_i$ as shown in (3), wielding a number between 0 and 1 analogous with factor analysis loadings. In a similar fashion, we used the fixed cut-off to classify items. The standardized values are shown in Table 3 where one can classify the items relative to each dimension.

The first conclusion we reached is that the BDI has indeed a multidimensional structure, as indicated by items having significant loadings in both dimensions (items 1, 2, 4, 9, 10, 12 through 15, 18, and 20),

Table 3 Significant standardized discriminations for the BDI items (cut-off = 0.5).

Item	a_1^*	a_2^*
1	0.85	0.52
2	0.74	0.68
3	0.97	
4	0.60	0.80
5	0.98	
6	0.97	
7	0.99	
8	0.98	
9	0.79	0.62
10	0.69	0.72
11		0.87
12	0.51	0.86
13	0.65	0.76
14	0.70	0.71
15	0.65	0.76
16		0.97
17		0.94
18	0.54	0.84
19		0.97
20	0.65	0.76
21		0.90

a_k^*: Item standardized discrimination on the k-th latent trait $k = 1, 2$.

which is a common conclusion in many factor analysis studies, as reported in Vanheule et al. (2008). The solid advantage of MIRT modeling in this context is that one can associate the items with both associated latent traits, perhaps characterizing the items as indicators of more nonspecific aspects of depression.

By evaluating the items which were reduced to one dimension by the cut-off, we can also derive similar interpretations for the latent traits as the ones observed in the literature. Items 3 (sense of failure), and 5 through 8 (guilty feelings, sense of punishment, self-dislike, and self-accusations) seem strongly related with the first dimension of the latent trait, indicating the cognitive dimension pointed in Steer et al. (1999). Whereas items 11 (irritability), 16 (sleep disturbance), 17 (fatigability), 19 (weight loss), and 21 (loss of libido) were associated with the second dimension, often regarded as the "somatic-affective" dimension.

Our findings are resonant with a factor analysis performed in the same dataset of healthy Brazilian college students studied in Gorenstein et al. (1999), which performed multiple factor analyses agreeing with most of our findings in their two-factor modelings and agree in the first factor with their three-factor approach with our second dimension being associated with their second and third factor.

Similarly to Cohen (2008), we aimed to interpret the BDI items in terms of its dimensions to clarify its characteristics in terms of the symptoms evaluated in the inventory. To this purpose, we used the items' discrimination in both dimensions and used it to evaluate which items are more sensible to one dimension more than another. A graphical representation can be observed in Fig. 2 where we classify the items relative to the two dimensions adopted (θ_1 and θ_2) using the values in Table 3.

Another way to visualize the item parameters is through a vector plot proposed in Reckase (1996), in which we plot the items as vectors scattered in the plane in positions relative to their normalized

Table 4 Anchor items in dimension 1, for fixed values for dimension 2.

	θ_1		
θ_2	-2	0	2
-2		5. Guilty 9. Suicidal	2. Pessimism 7. Dislike 8. Self-accusation 10. Crying 14. Body image
0		2. Pessimism 3. Failure 5. Guilty 6. Punishment 10. Crying 12. Social withd. 13. Indecisiveness 14. Body image	1. Sadness 4. Insatisfaction 7. Dislike 8. Self-accusation
2	2. Pessimism 4. Insatisfaction 9. Suicidal 10. Crying 13. Indecisiveness 14. Body image 15. Work	1. Sadness 3. Failure 5. Guilty 6. Punishment	8. Self-accusation

difficulty 4 and angles relative to both axes depending of their normalized discriminations. Such plot is shown in Fig. 3, albeit in more real settings and complex tests it can get really cluttered and hard to read.

We concluded that the inventory evaluates both assumed dimensions adequately and the positions of the vectors in the positive quadrant imply that the items aim to evaluate above average values of depression, which is consistent with the intent of the test. The vectors closer to the θ_1 axis (items 3, 6, 5, and 8) and those closer to the θ_2 axis (items 17, 19, 21, for example) are also consistent to the interpretation of dimensions previously discussed.

To focus on the evolution of depression symptoms, we considered four values for each of the two BDI dimensions, denoted by θ_{jk}, for $j = 0, 1, 2, 3$ and $k = 1, 2$: a very low value, $\theta_{0k} = -4$, a low value, $\theta_{1k} = -2$, an intermediate value, $\theta_{2k} = 0$, and a high value, $\theta_{3k} = 2$. In this way, we defined 16 intensities for depression combining four levels for the first dimension with four levels for the second one. One item i is called an anchor item (Beaton and Allen, 1992) for dimension 1 in level θ_{j1}, for $j = 1, 2, 3$, if:

(1) $P(X_{i,j} = 1|\theta_{j1}, \theta_{j',2}) \geq 0, 6$
(2) $P(X_{i,j-1} = 1|\theta_{j-1,1}, \theta_{j',2}) < 0, 5$
(3) $P(X_{i,j} = 1|\theta_{j1}, \theta_{j',2}) - P(X_{i,j-1} = 1|\theta_{j-1,1}, \theta_{j',2}) \geq 0, 3,$

where $\theta_{j',2}$ is fixed at -2, 0, or 2, and $P(X_{i,j} = 1|\theta_{j1}, \theta_{j',2})$ is the probability of the presence of symptom i, specified by the adopted model (1). In other words, an anchor item has an abrupt increase in its probability of presence in a specific level of the latent trait, characterizing that trait level in terms of

Table 5 “Anchor” items in dimension 2, for fixed values for dimension 1.

	θ_1		
θ_2	−2	0	2
−2		9. Suicidal	2. Pessimism 10. Crying 14. Body image 18. Appetite 21. Libido
0	21. Libido	2. Pessimism 10. Crying 12. Social withd. 13. Indecisiveness 14. Body image 18. Appetite 21. Libido	4. Insatisfaction 11. Irritability 12. Social withd. 15. Work 16. Sleep 17. Fatigability
2	2. Pessimism 4. Insatisfaction 10. Crying 11. Irritability 12. Social withd. 13. Indecisiveness 14. Body image 15. Work 16. Sleep 17. Fatigability	1. Sadness 4. Insatisfaction 11. Irritability 17. Fatigability	

symptoms that are most likely to manifest in patients with the specified levels of depression. We can define an “anchor” item for dimension 2 in an identical manner.

The thresholds are chosen in terms of the response probability, the exact values inherited from educational settings (Beaton and Allen, 1992; Harraway and Andrade, 2006). For an item to be an anchor for a trait level, the response probability at the anchor level must be high (as enforced by the first item in the definition), the probability at the preceding level must be relatively low (second item), and the anchor item must mark a transition from a low probability to a high response probability (third item).

In Tables 4 and 5, we present the anchor items of dimensions 1 and 2 for levels −2, 0, and 2 of the latent traits. Note that all items, except items 19 (weight loss) and 20 (somatic preoccupation), are anchors for some level of at least one dimension. Those items more associated to the cognitive trait (3, 5, 6, 7, and 8; see Fig. 2) are anchor only in this dimension, whereas the items more related to the somatic-affective trait (11, 16, 17, and 21; see Fig. 2) are anchor only in this trait dimension. These facts corroborate the previous interpretation and definition of cognitive and somatic-affective dimensions. The nonidentification of items 19 and 20 as anchor for any level of the dimensions may be justified by the low estimates of discrimination parameters.

In order to illustrate the usefulness of Tables 4 and 5, suppose a subject with an intermediate level for dimension 1 (equal to 0, for instance) and a low level for dimension 2 (equal to −2). A worsening in cognitive aspects may be remarked by symptoms as pessimism, dislike, self-accusation, crying, and

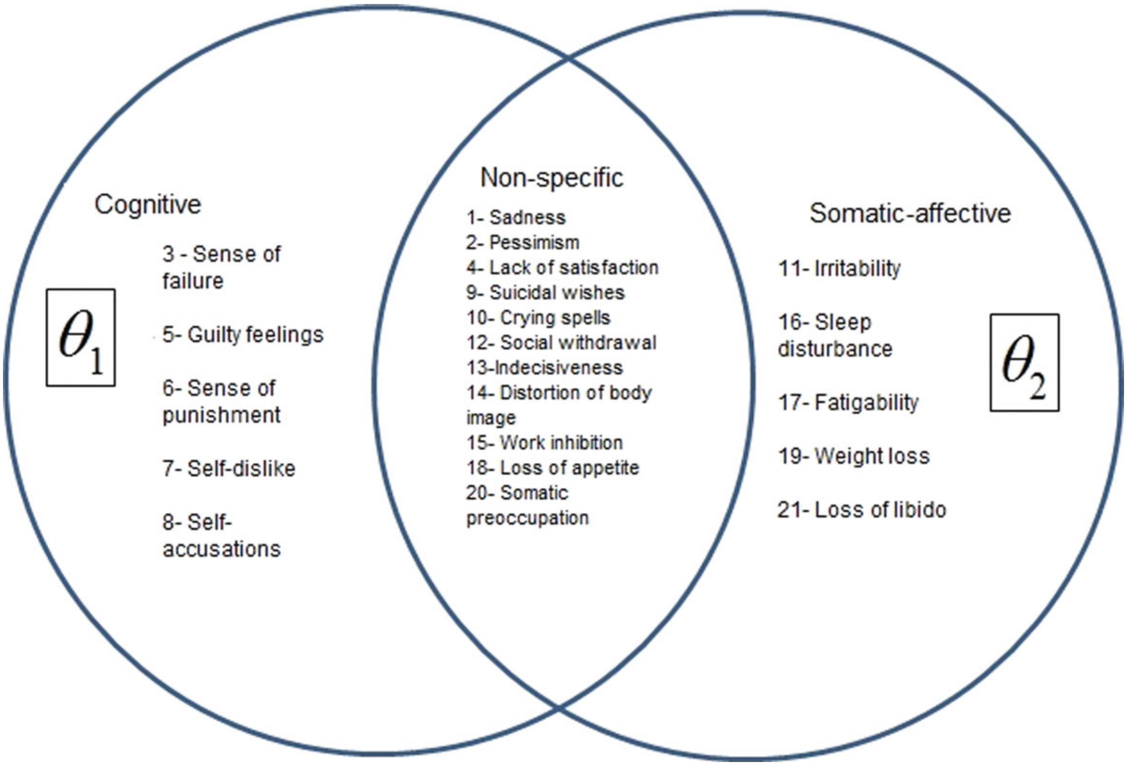

Figure 2 Venn diagram for the BDI-II items.

body image, while a worsening in somatic-affective aspects probably leads to the presence of pessimism, crying, social withdraw, indecisiveness, body image, appetite, and libido symptoms.

As a final remark, our model proved itself capable of modeling the expected behavior of the score, which may provide consistent conclusions with Classical Test Theory methods. In Fig. 1, we note the observed scores appearing in the 95% credibility bands, indicating a good predicting capability of the model, as summarized by a not too extreme posterior predictive checking p-value of 0.13. In the same way, we represent in Fig. 4 the boxplots of latent trait estimates categorized by groups of nondepressed (BDI total score 0–15), dysphoria (BDI total score 16–20), and depressed subjects (BDI total score 21–63), as suggested by Kendall et al. (1987). For both dimensions, we can see that the latent trait estimates increase as expected from group nondepressed through depressed one, being more distinguishable between nondepressed and dysphoria groups.

5 Conclusions

MIRT models can prove themselves as a valuable modeling tool for psychometric data, confirming conclusions obtained under different techniques, like the factor analysis performed by Steer et al. (1999) to verify the number of underlying factors to depression or the item classification performed by Cohen (2008) under a single class of models and estimation paradigms.

Such possibility is extremely empowering for the analysis, as a unified theoretical framework may make certain questions way easier to ask. For instance, the interaction between the latent aspects of depression can be translated as a different model, the existence of more sociably acceptable behavior

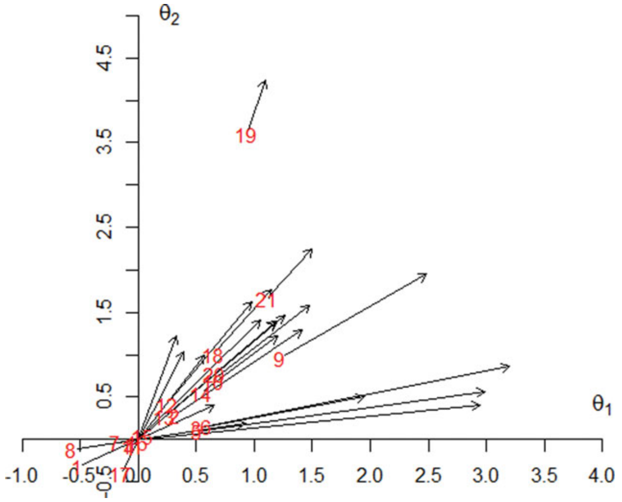

Figure 3 Vector plot for the BDI-II items.

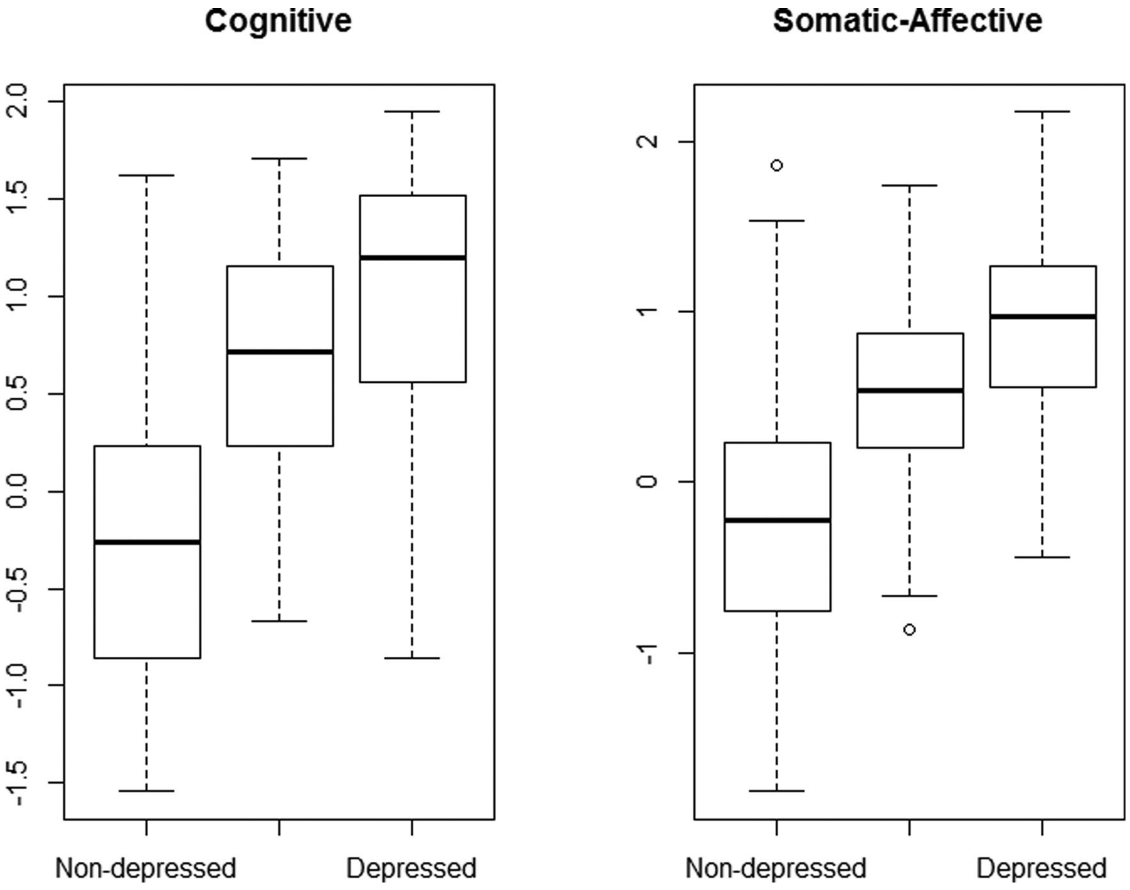

Figure 4 Boxplots for latent trait estimates of subjects classified as nondepressed, dysforia, or depressed using Kendall criteria.

in answering some questions can be interpreted in the context of the guessing parameter in the three-parameter model.

With an interpretable parameter like the latent traits, one can also proceed with the analysis using the obtained estimates as independent variables in regressions over predictors such as sex, income, literacy, and other sensible indicators, using the IRT model we fitted as a first step in the analyses (Fox and Glas, 2001).

The Bayesian estimation methods used are also very well discussed in IRT models (Bock et al., 1988; Patz and Junker, 1999; Beguin and Glas, 2001, to cite a few) and may provide powerful methods for most of the possibilities raised in this section.

Acknowledgments The authors would like to thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and the Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) for the financial support through student grants and the Instituto de Ciências Matemáticas e de Computação of the Universidade de São Paulo - São Carlos (ICMC-USP-São Carlos).

Conflict of interest
The authors have declared no conflict of interests.

References

Beaton, A. E. and Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics* **17**, 191–204.

Beguin, A. A. and Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis in multidimensional IRT models. *Psychometrika* **66**, 541–562.

Bock, R. D., Gibbons, R. and Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement* **12**, 261–280.

Bolt, D. M. and Kim, J. S. (2007). *Estimating Item Response Models using Markov Chain Monte Carlo Methods*. National Council on Measurement in Education, Edimonton.

Bolt, D. M. and Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement* **27**, 395–414.

Castro, S. M. J., Trentini, C. and Riboldi, J. (2010). Teoria da resposta ao item aplicada ao inventório de depressão de beck. *Revista Brasileira de Epidemiologia* **13**, 487–501.

Cohen, A. (2008). The underlying structure of the Beck Depression Inventory II: a multidimensional scaling approach. *Journal of Research in Personality* **42**, 779–786.

Finch, H. (2011). Multidimensional item response theory parameter estimation with nonsimple structure items. *Applied Psychological Measurement* **35**, 67–82.

Fox, J. P. and Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using gibbs sampling. *Psychometrika* **66**, 271–288.

Fraser, C. (1988). Noharm ii: a fortran program for fitting unidimensitonal and multidimensional normal ogive models of latent trait theory. Technical report, The University of New England, Armisdale, Australia.

Fu, Z., Tao, J. and Shi, N. (2009). Bayesian estimation on the multidimensional three-parameter logistic model. *Journal of Statistical Computation and Simulation* **79**, 819–835.

Gelman, A., Carlin, J. B. and Rubin, D. B. (2004). *Bayesian Data Analysis* (2nd ed.). Chapman and Hall/CRC, Boca Raton, Florida.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: Bernardo, J. M., Berger, J. O., Dawid, A. P. and S. A. F. M. (Eds.), *Bayesian Statistics 4*. Clarendon Press, Oxford.

Gorenstein, C., Andrade, L., Vieira Filho, A. H. G., Tung, C. T. and Artes, R. (1999). Psychometric properties of the Portuguese version of the beck depression inventory on brazilian college students. *Journal of Clinical Psychology* **55**, 553–562.

Harraway, J. A. and Andrade, D. F. (2006). An item response analysis of statistics use in the workplace. In *Proceedings of 7th International Conference on Teaching Statistics - ICOTS 7*, International Association for Statistical Education, Salvador, Brazil.

Kendall, P. C., Hollon, S. D., Beck, A. T., Hammen, C. L. and Ingram, R. E. (1987). Issues and recommendations regarding the use of the Beck Depression Inventory. *Cognitive Therapy and Research* **11**, 289–299.

Li, Y. and Baser, R. (2012). Using r and winbugs to fit a generalized partial credit model for developing and evaluating patient-reported outcomes assessments. *Statistics in Medicine* **31**, 2010–2026. Available at http://dx.doi.org/10.1002/sim.4475

McKinley, R. L. and Reckase, M. D. (1980). The use of the general rasch model with multidimensional item response data Technical report, American College Testing, Iowa City, IA.

Patz, R. J. and Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response theory models. *Journal of Educational and Behavioral Statistics* **24**, 146–178.

Plummer, M., Best, N., Cowles, K. and Vines, K. (2006). Coda: convergence diagnosis and output analysis for MCMC. *R News* **6**, 7–11. Available at http://CRAN.R-project.org/doc/Rnews/

R Development Core Team (2005). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E. and Lewis, S. M. (1995). The number of iterations, convergence diagnostics and generic metropolis algorithms. In: Gilks, D. S. W. R. and Richardson, S. (Eds.), *Practical Markov Chain Monte Carlo*, Chapman and Hall, Londres.

Reckase, M. D. (1996). A linear logistic multidimensional model. In: van der Linden, W. J. and Hambleton, R. K. (Eds.), *Handbook of Modern Item Response Theory*. Springer, New York, pp. 271–286.

Sinharay, S., Stern, H. S. and Johnson, M. S. (2006). Posterior predictive assessment in item response theory. *Applied Psychological Measurement* **30**, 298–321.

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2003). *WinBUGS User Manual Version 1.4*. MRC Biostatistics Unit, Cambridge.

Steer, R. A., Ball, R., Ranieri, W. F. and Beck, A. T. (1999). Dimensions of the Beck Depression Inventory-II in clinically depressed outpatients. *Journal of Clinical Psychology* **55**, 117–128.

Sympson, J. (1977). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference*. University of Minnesota, Minneapolis, pp. 82–98.

Thomas, M. L. (2010). The value of item response theory in clinical assessment: a review. *Assessment* **18**, 291–307.

Vanheule, S., Desmet, M., Groenvynck, H., Rosseel, Y. and Fontaine, J. (2008). The factor structure of the Beck Depression Inventory-II: an evaluation. *Assessment* **15**, 177–187.

Wang, Y. P., Andrade, L. H. and Gorenstein, C. (2005). Validation of the Beck Depression Inventory for a Portuguese-speaking Chinese community in brazil. *Brazilian Journal of Medical and Biology Research* **38**, 399–408.

Whitely, S. (1980). *Measuring aptitude processes with multicomponent latent trait models*. Technical report. University of Kansas, Lawrence.

Wilson, D., Wood, R. and Gibbons, R. D. (1987). *TESTFACT: test scoring, item statistics and item factor analysis*. Scientific Software, Mooresville.