Open Access Library Journal
2019, Volume 6, e5242
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Item Response Theory Modeling of High School Students' Behavior in a High-Stakes Exam

Helen Gomes[1], Raul Matsushita[1], Sergio Da Silva[2*]

[1]Department of Statistics, University of Brasilia, Brasilia, Brazil
[2]Department of Economics, Federal University of Santa Catarina, Florianopolis, Brazil
Email: *professorsergiodasilva@gmail.com

How to cite this paper: Gomes, H., Matsushita, R. and Da Silva, S. (2019) Item Response Theory Modeling of High School Students' Behavior in a High-Stakes Exam. *Open Access Library Journal*, **6**: e5242.
https://doi.org/10.4236/oalib.1105242

Received: February 21, 2019
Accepted: February 25, 2019
Published: February 28, 2019

Abstract

We put forward a model based on item response theory that highlights the role of latent features called "proficiency" and "propensity". The model is adjusted to data from the decisions made in a high-stakes exam taken by 10,822 Brazilian high school students. Our model aims to recover information regarding the role the latent features (proficiency and propensity) play in a decision. We find that the decision of responding or not and also the decision of responding correctly or not in a group of items can be described by a two-dimensional logistic model, even if there are imperfections from an item-by-item adjustment. Not only proficiency, but also refraining from responding is found to depend on both the characteristics of the items and the latent features of the students. In particular, the least proficient students prefer to leave an item blank, rather than respond it incorrectly. There is a negative linear correlation between scoring in the exam and propensity, and scoring and proficiency are positively correlated although nonlinear.

Subject Area

Statistics

Keywords

Psychometrics, Item Response Theory, Student Behavior, High-Stakes Exams

1. Introduction

Admissions to higher education institutions in Brazil are traditionally made through an entrance exam called the "vestibular". However, since 1995 a three-stage evaluation process has become adopted by some universities as an alternative to the vestibular. Here, we consider one such an evaluation taken by the University of Brasi-

lia called the "Serial Evaluation Program", or PAS. In particular, we take recently publicly available data for the third stage of PAS for the years 2006 to 2008 and focus on the exam given on 7 December 2008 for 10,822 last-year high school participants.

The third stage of PAS's exam involves two sections: 1) a foreign language section; 2) a general knowledge section that considers Portuguese, math, physics, biology, chemistry, the arts, philosophy, geography, history, literature and sociology. Here, we concentrate on the second section and its true or false questions composed of 100 items. The dataset is available at Figshare (https://doi.org/10.6084/m9.figshare.5882377.v1).

This is a high-stakes environment for the applicants [1] [2] [3] because the exam payoff means entering a top university. Under such circumstances, participants are expected to behave strategically [4].

This work considers item response theory [5] to model the participants' behavior. In psychometrics, item response theory (IRT) constitutes a set of methodologies that allow the estimation of intangible individual characteristics (or latent features), such as intelligence, personality traits, emotional states, proficiency and risk taking [5].

In particular, we postulate here the probability of a high-school participant to correctly answer an item on the exam depends on both the intrinsic characteristic of such an item, such as its degree of difficulty, and the participant's proficiency on the subject the item refers to. Acting strategically, the participant may also either provide an incorrect response or leave the question blank. Here, leaving the question blank is strategically better because answering incorrectly is a loss. When facing difficult questions, we also assume the participant makes a decision taking into account both intrinsic difficulty and intuitive latent features that we call "propensity".

Leaving a question blank may also mean a participant's low proficiency regarding the item as well as the propensity to avoid a loss accruing from answering incorrectly. Our model aims to recover information regarding the roles the latent features of proficiency and propensity play in a decision.

Section 2 introduces a model of proficiency and propensity based on item response theory. Section 3 analyzes the data using the model and shows the results found. And Section 4 concludes the study.

2. An IRT Model of Proficiency and Propensity

Consider a group of n participants who take part in an exam made up of I items. Let R_{ij} be a dummy variable that takes on value 1 if individual j does answer item i (where $1 \le j \le n; 1 \le i \le I$), or takes on value 0 if individual j does not answer item i. In addition, let U_{ij} be another dummy such that $U_{ij} = 1$ if item i is correctly answered by individual j, and $U_{ij} = 0$ if item i is not correctly answered or left blank by individual j.

Figure 1 displays the possible paths for result U_{ij}. First, individual j decides

whether or not to answer item *i*. If individual *j* decides to answer, his or her answer may end up correct or incorrect. If he or she decides not to answer, then $U_{ij}=0$. Thus, if $R_{ij}=0$ then $U_{ij}=0$ with probability 1.

Table 1 shows the joint probability distribution of variables R_{ij} and U_{ij}. Their conditional probabilities are:

$$\pi_{00}=P\left[R_{ij}=0,U_{ij}=0\right]=P\left[R_{ij}=0\,|\,U_{ij}=0\right]\cdot P\left[U_{ij}=0\right] \quad (1)$$

$$\pi_{01}=P\left[R_{ij}=1,U_{ij}=0\right]=P\left[R_{ij}=1\,|\,U_{ij}=0\right]\cdot P\left[U_{ij}=0\right] \quad (2)$$

$$\pi_{11}=P\left[R_{ij}=1,U_{ij}=1\right]=P\left[R_{ij}=1\,|\,U_{ij}=1\right]\cdot P\left[U_{ij}=1\right]=P\left[U_{ij}=1\right]. \quad (3)$$

The probabilities *P* in (1), (2) and (3) are obviously related to *i* and *j*, but subscripts have been omitted for notational convenience.

Setting $0^0\equiv 0$, the joint probability distribution can be written as

$$P\left[R_{ij}=r,U_{ij}=u\right]=\left\{P\left[R_{ij}=r\,|\,U_{ij}=u\right]\right\}^{1-u}P\left[U_{ij}=u\right], \quad (4)$$

where $r,u\in\{0,1\}$.

The conditional probabilities in (2) and (3) capture the trade-off faced by individual *j* of either responding to item *i* incorrectly or leaving the item blank. These two possibilities refer to the event $U_{ij}=0$. However, treating missing data as incorrect is the least desirable way to account for missing-not-at-random responses in large-scale surveys [6], because a participant tends to leave blank those items he or she considered difficult. To be in control, the participant manages to pick those items that match his or her proficiency. Incorrect answers and nonresponses have the same payoff, but considering nonresponses as the same as incorrect answers bias proficiency estimates [6] [7].

To remedy this deficiency, we consider the approach initiated by Knott *et al.*

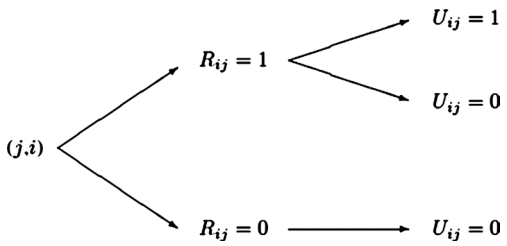


Figure 1. Possible paths for result U_{ij}.

Table 1. Joint distribution between R_{ij} (*j* does respond = 1; does not = 0) and U_{ij} (correct = 1; incorrect = 0).

		R_{ij}		Total
		0	1	
U_{ij}	0	π_{00}	π_{01}	$\pi_{0\cdot}$
	1	0	π_{11}	π_{11}
	Total	π_{00}	$\pi_{\cdot 1}$	1

[8] and Albanese and Knott [9], and followed by many [10]-[16]. We introduce in our model a bivariate latent feature (θ_1, θ_2), where θ_1 is what we called propensity in the previous section, and θ_2 is proficiency. Propensity precisely means responses to the items convey information regarding the participants who are more prone to answer incorrectly rather than not to answer. Our strategy of modeling allows us to incorporate nonresponses explicitly into an analysis. In particular, we consider the terms in Equation (4) to be described by two-parameter logistic equations [8] [9]:

$$P\left[R_{ij} = r \mid U_{ij} = 0\right] = \frac{\exp\left[ra_{1i}\left(\theta_{1j} - b_{1i}\right)\right]}{1 + \exp\left[a_{1i}\left(\theta_{1j} - b_{1i}\right)\right]} \tag{5}$$

$$P\left[U_{ij} = u\right] = \frac{\exp\left[ua_{2i}\left(\theta_{2j} - b_{2i}\right)\right]}{1 + \exp\left[a_{2i}\left(\theta_{2j} - b_{2i}\right)\right]} \tag{6}$$

where a_{1i} and a_{2i} are parameters related to the discriminating power of a participant, and b_{1i} and b_{2i} are difficulty parameters related to item i [17] [18]. Here, subscript 1 refers to propensity, while subscript 2 refers to proficiency. The latent feature θ_{2j} is the proficiency of participant j, and θ_{1j} is the propensity of participant j to answer incorrectly.

Equations (5) and (6) give a precise meaning to the latent features. Propensity is defined exactly by Equation (5), while proficiency is defined by Equation (6). Propensity is related to the conditional probability of an incorrect response against the nonresponse option. Thus, propensity refers to making a mistake by choosing the incorrect response rather than making a mistake by leaving an item blank. Of note, a risk is involved while choosing, and thus risk taking is implicitly considered in propensity.

By considering an item incorrect, a participant 1) may provide an incorrect response or 2) may leave the item blank. A high propensity means the participant picks the former. Because propensity is defined based on a probability conditional to the space of incorrect items, here the correct decision is not to answer.

Propensity and proficiency latent features are usually considered in models of "nonignorable nonresponses" [6] [7]. Here, we consider a two-dimensional IRT model to deal with such nonignorable nonresponses in tests with dichotomous items. While the propensity dimension provides information about omitted behavior, the proficiency dimension is related to a candidate's ability.

Considering Equations (1)-(3), the latent variables θ_{1j} and θ_{2j} refer to the logit functions:

$$\eta_{1ij} = \ln\frac{\pi_{01}}{\pi_{00}} = a_{1i}\left(\theta_{1j} - b_{1i}\right) \tag{7}$$

$$\eta_{2ij} = \ln\frac{\pi_{11}}{\pi_{00} + \pi_{01}} = a_{2i}\left(\theta_{2j} - b_{2i}\right). \tag{8}$$

Substituting the one-dimensional logistic Equations (5) and (6) into (4) yields

the bidimensional model:

$$P\left[R_{ij}=r,U_{ij}=u\right]=\frac{\exp\left[r(1-u)\left(\eta_{1ij}+u\eta_{2ij}\right)\right]}{1+\exp\left[\eta_{2ij}\right]+(1-u)\exp\left[\eta_{1ij}\right]\left\{1+\exp\left[\eta_{2ij}\right]\right\}}. \quad (9)$$

This IRT model of proficiency and propensity is noncompensatory [18], in that the low proficiency of participant j in answering an item correctly, θ_{2j}, cannot be compensated by his or her propensity, θ_{1j}. We estimate the item-related parameters— a_{1i}, a_{2i}, b_{1i}, b_{2i}—by maximum likelihood, whereas the latent features— θ_{1j}, θ_{2j}—are estimated by the expected a posteriori method [17] [18]. All the scripts were built using the *R* language (https://cran.r-project.org/).

In item response theory, the items are usually evaluated taking into account their adhesion to an adjusted model [19]. In particular, for the joint distribution in **Table 1** of an item i its chi-square statistics are given by

$$\chi_i^2=n\left[\frac{\left(\hat{\pi}_{11,i}-\tilde{\pi}_{11,i}\right)^2}{\hat{\pi}_{11,i}}+\frac{\left(\hat{\pi}_{01,i}-\tilde{\pi}_{01,i}\right)^2}{\hat{\pi}_{01,i}}+\frac{\left(\hat{\pi}_{00,i}-\tilde{\pi}_{00,i}\right)^2}{\hat{\pi}_{00,i}}\right], \quad (10)$$

where

$$\hat{\pi}_{00,i}=1-\hat{\pi}_{11,i}-\hat{\pi}_{01,i},$$

$$\hat{\pi}_{11,i}=\sum_{j=1}^{n}\hat{P}\left(U_{ij}=1\right)/n,$$

$$\hat{\pi}_{01,i}=\sum_{j=1}^{n}\hat{P}\left(R_{ij}=1\,|\,U_{ij}=0\right)\cdot\hat{P}\left(U_{ij}=0\right)/n$$

are the aggregates of the estimates of the probabilities in model (9), and $\tilde{\pi}_{11,i}$, $\tilde{\pi}_{01,i}$, $\tilde{\pi}_{00,i}=1-\tilde{\pi}_{11,i}-\tilde{\pi}_{01,i}$ are the corresponding empirical frequencies, that is, the ratio between the number of occurrences and the total number of participants. Under the null hypothesis that the model fits the joint distribution in **Table 1**, the chi-square statistics (10) have 2 degrees of freedom as they depend on two random variables.

In particular, to assess the similarity between the expected and observed fractions π_{kk} in a set of I items, for either $k=0$ (nonresponses) or $k=1$ (correct responses), we consider the Pearson correlation measures

$$\rho_k=\frac{\sum_{i=1}^{I}\left[\hat{\pi}_{kk,i}-m\left(\hat{\pi}_{kk,i}\right)\right]\left[\tilde{\pi}_{kk,i}-m\left(\tilde{\pi}_{kk,i}\right)\right]}{\sqrt{\sum_{i=1}^{I}\left[\hat{\pi}_{kk,i}-m\left(\hat{\pi}_{kk,i}\right)\right]^2\sum_{i=1}^{I}\left[\tilde{\pi}_{kk,i}-m\left(\tilde{\pi}_{kk,i}\right)\right]^2}}, \quad (11)$$

where $m\left(\pi_{kk,i}\right)=\sum_{i=1}^{I}\pi_{kk,i}/I$.

Next, we show the analysis of data and the results from model (9).

3. Results

Figure 2(a) shows a funnel-shaped dispersion between the total of unanswered items, T_n, and the total of correct responses, T_c, for a participant. As T_n rises, the variability of T_c plummets. **Figure 2(b)** shows a triangle-shaped dispersion between the total of unanswered items, T_n, and the total of incorrect responses, T_w. While the distributions of T_c and T_w are roughly sinusoid, the distribu-

tion of T_n reveals the concentration of zeros (only 12.5 percent of participants responded all the items). The variability of correct and incorrect responses for the participants who did not leave items blank is high, thus suggesting they are likely to present a larger propensity θ_1.

Figure 3 shows the percentage of nonresponses for each item considering the four groups of disciplines, as in Table 2. As can be seen in Figure 3, the disciplines in groups II and III of Table 2 show more nonresponses (p-value = 0.0005; Kruskal-Wallis test, d.f. = 3). For this reason, model (9) will consider such a fact.

Regarding the percentage of incorrect responses relative to all incorrect responses for the groups, that is,

$$P_{wt} = \frac{\pi_{01}}{\pi_{00} + \pi_{01}} \times 100, \tag{12}$$

Figure 4 reveals absence of pattern (p-value = 0.16; Kruskal-Wallis test, d.f. = 3). Figure 5 shows the dispersion of the empirical values of conditional probabilities

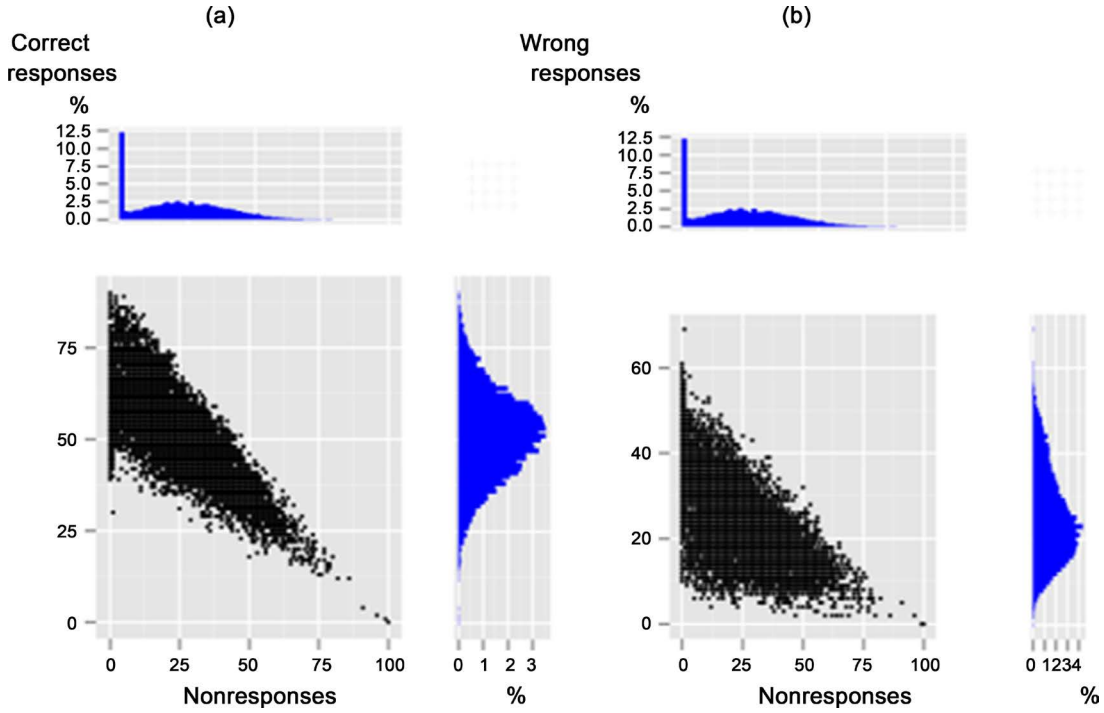


Figure 2. (a) Dispersion between the total of unanswered items and the total of correct responses; (b) dispersion between the total of unanswered items and the total of incorrect responses. Their respective marginal distributions are also shown.

Table 2. Groups of disciplines in the exam.

Group	Disciplines	Number of items
I	Portuguese, philosophy, geography, history, sociology	24
II	Physics, math	24
III	Chemistry, biology	25
IV	Literature, arts	27

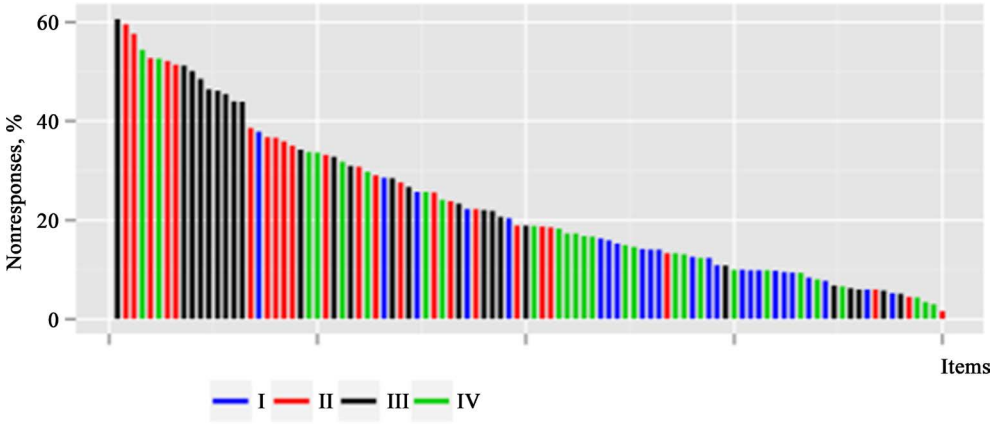


Figure 3. Percentage of nonresponses, by group. Groups II and III show more nonresponses.

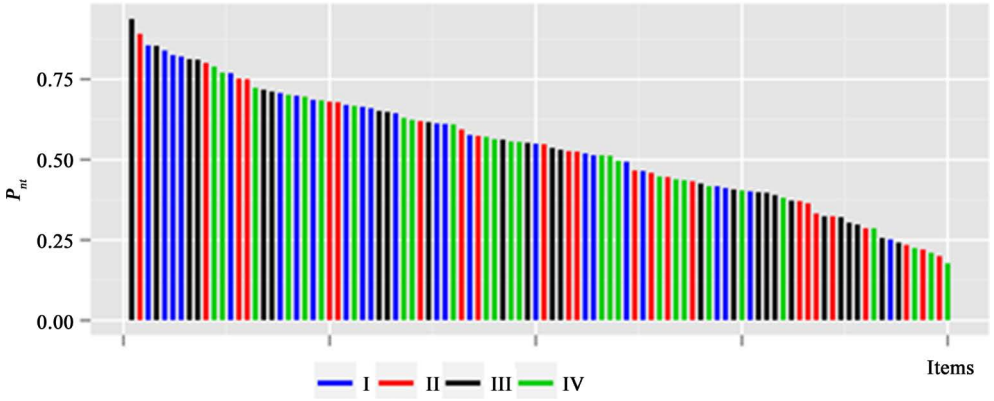


Figure 4. The percentage of incorrect responses relative to all incorrect responses for the groups does not have a pattern.

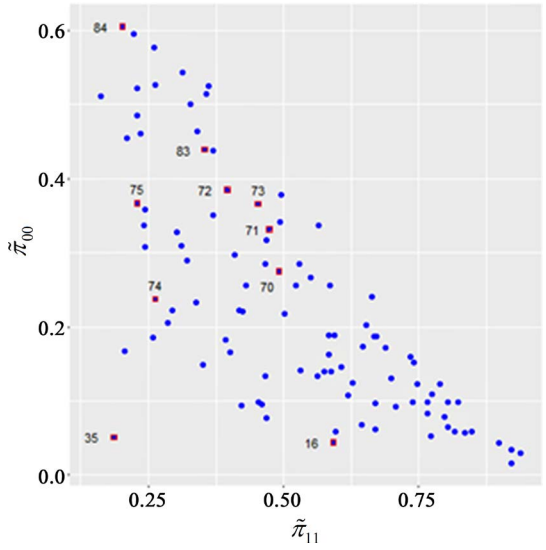


Figure 5. Dispersion of the empirical values of $\tilde{\pi}_{00}$ and $\tilde{\pi}_{11}$.

π_{00} and π_{11}. Nonresponses $\tilde{\pi}_{00}$ are expected to drop as correct responses $\tilde{\pi}_{11}$ increase, as seen, for instance, for items 70, 71, 72, 73, 83 and 84 (highlighted in **Figure 5**). However, for the other items highlighted, this did not oc-

cur and incorrect responses were more common than nonresponses. For instance, item 35 presented only 5.1 percent of nonresponses and 76.3 percent of incorrect responses (and 18.6 percent of correct responses).

Tables 3-6 show the parameter estimates of the items from marginal maximum likelihood. They also show the observed percentage frequencies along with the expected percentage frequencies from our adjusted model. The χ^2 statistics and their corresponding *p*-values are also shown.

Figure 6 shows the dispersion between the discriminating power parameter and the difficulty parameter regarding propensity, by group. Most responses to the items allow a reasonable discriminating power, that is, $a_1 > 1$, and present a

Table 3. Results for Group I.

Item	Parameter estimate				Frequency, %						Fit	
					Expected			Observed				
	a_1	b_1	a_2	b_2	π_{00}	π_{01}	π_{11}	π_{00}	π_{01}	π_{11}	$\chi^2_{(2)}$	*p*-value
11	1.43	−0.74	0.87	−0.42	13.6	28.5	57.9	14.0	28.5	57.5	1.74	**0.4183**
12	1.54	−0.84	1.08	−0.81	9.7	22.6	67.6	9.7	23.4	66.8	3.89	**0.1431**
13	1.29	−1.54	0.77	0.27	9.4	45.4	45.2	9.9	45.0	45.1	2.41	**0.2997**
23	1.85	−0.20	0.79	−2.02	8.9	10	81	9.9	9.6	80.4	13.42	0.0012
24	1.70	−0.62	0.47	−3.80	5.1	9.9	85	5.9	9.4	84.7	15.69	0.0004
25	1.75	−0.54	0.52	−0.69	14.8	26.7	58.5	16.3	25.5	58.3	21.62	0.0000
26	1.84	−0.63	1.51	−0.50	12.6	23.7	63.7	12.5	24.8	62.7	6.89	**0.0319**
27	2.15	−0.38	1.52	−1.10	9.6	12.4	78	9.9	13.4	76.7	12.42	0.0020
28	2.11	−0.02	0.61	−1.86	13.2	12.3	74.5	15.3	10.6	74.1	58.92	0.0000
30	1.71	0.72	0.70	0.03	35.8	14.7	49.5	37.9	12.8	49.3	45.13	0.0000
31	1.69	−1.08	1.08	−1.39	5.2	16.6	78.2	5.2	17.5	77.3	6.44	**0.0399**
32	2.27	−0.55	1.49	−1.11	8.3	13.7	78	8.3	15.0	76.7	15.20	0.0005
36	1.39	−0.73	1.08	−0.42	13.5	27.2	59.3	14.0	27.2	58.7	2.82	**0.2440**
37	1.91	−0.26	0.65	−2.07	9.9	12.3	77.8	10.9	11.8	77.3	13.48	0.0012
38	1.77	−0.29	0.60	−1.95	10.8	14.1	75.1	12.3	13.0	74.7	29.69	0.0000
39	1.92	0.06	0.86	−0.85	18.6	15.7	65.7	20.4	14.5	65.1	27.95	0.0000
45	1.59	−0.87	0.55	−0.25	12.9	34	53.2	14.2	32.8	53.0	16.96	0.0002
52	1.56	−0.27	0.76	0.41	24.8	32.2	43	25.7	31.3	43.0	6.85	**0.0326**
62	1.56	0.16	0.87	−1.36	14.7	11.2	74.1	15.9	10.7	73.4	13.56	0.0011
67	1.22	−1.74	0.43	0.77	8.6	49.3	42	9.3	48.6	42.0	7.48	0.0238
89	1.08	−1.95	0.56	0.23	7.6	45.6	46.8	7.7	45.5	46.8	0.32	**0.8515**
90	1.26	−1.57	0.65	0.27	9.2	44.9	45.9	9.5	44.7	45.8	0.95	**0.6222**
114	1.45	−0.83	0.27	3.31	20.2	50.4	29.4	22.2	48.4	29.3	28.72	0.0000
115	1.54	−0.06	0.23	0.57	25.9	27.4	46.7	28.6	24.8	46.6	55.73	0.0000

Table 4. Results for Group II.

Item	Parameter estimate				Frequency, %						Fit	
					Expected			Observed				
	a_1	b_1	a_2	b_2	π_{00}	π_{01}	π_{11}	π_{00}	π_{01}	π_{11}	$\chi^2_{(2)}$	p-value
15	1.10	−1.48	0.11	−22.14	1.6	6.1	92.2	1.5	6.2	92.2	**1.00**	**0.6075**
16	0.88	−2.77	0.38	−1.01	4.4	36.3	59.3	4.4	36.3	59.2	**0.14**	**0.9339**
17	0.94	−1.50	0.43	0.34	13	40.5	46.5	13.3	40.2	46.5	**1.20**	**0.5476**
18	1.09	0.07	0.67	0.89	34.6	28.8	36.6	35.0	28.1	36.9	**2.71**	**0.2585**
55	1.70	−0.42	0.75	−0.50	18.7	22.8	58.5	18.8	22.8	58.4	**0.12**	**0.9421**
56	1.87	0.80	0.94	0.74	49.9	14.9	35.2	51.4	12.9	35.8	39.33	0.0000
57	1.40	−1.14	0.68	1.69	18.1	56.4	25.5	18.5	55.6	25.9	**2.98**	**0.2257**
59	1.64	−0.47	0.77	1.66	30.4	45.8	23.8	30.8	44.9	24.4	**4.07**	**0.1305**
60	1.62	−0.25	0.65	1.89	34.8	41.2	24	35.9	39.7	24.4	9.98	0.0068
64	1.31	−0.92	0.44	−3.49	5.9	12.3	81.8	5.9	12.5	81.6	**0.56**	**0.7542**
65	1.36	0.00	0.52	−1.45	18.1	14.6	67.3	18.6	14.2	67.1	**2.95**	**0.2287**
66	1.09	−0.68	0.70	0.53	22.5	36	41.5	22.2	36.1	41.7	**0.76**	**0.6849**
70	2.19	−0.12	0.91	0.05	28	23	48.9	27.5	23.4	49.1	**1.45**	**0.4834**
71	2.37	0.05	0.74	0.17	31.3	21.5	47.2	33.2	19.5	47.3	33.99	0.0000
72	2.27	0.14	0.85	0.58	37.5	23.4	39.1	38.5	21.9	39.6	13.99	0.0009
73	2.58	0.16	0.79	0.28	35	19.9	45.1	36.6	18.1	45.3	26.67	0.0000
74	2.15	−0.74	0.78	1.49	22.6	51.7	25.6	23.7	50.0	26.2	13.08	0.0014
75	2.14	−0.27	0.73	1.85	35.2	42.5	22.3	36.7	40.4	22.9	19.53	0.0001
76	1.82	−0.18	0.74	−0.13	24.6	23.2	52.2	25.5	22.3	52.2	8.49	0.0144
77	2.12	0.38	1.56	1.10	53.1	25.5	21.4	52.1	24.9	22.9	14.02	0.0009
78	2.32	0.69	1.68	0.91	58.2	17.2	24.6	57.6	16.2	26.2	17.93	0.0001
79	2.23	0.47	0.85	1.40	51.7	22.7	25.5	52.7	21.1	26.2	17.00	0.0002
94	1.42	−0.42	0.68	1.21	28.9	39.3	31.8	28.9	38.9	32.2	**1.14**	**0.5668**
102	1.84	0.76	0.97	1.52	58.7	19.9	21.4	59.5	18.2	22.3	21.55	0.0000

low degree of difficulty ($b_1 < 0$). This result suggests those with lower propensity to respond incorrectly prefer not to respond (that is, they give nonresponses).

Figure 7 shows the dispersion between the discriminating power parameter and the difficulty parameter regarding proficiency, by group. Now, responses to the items allow less power to discriminate the most proficient participants, apart from items 20, 77, 78, 82, 83, 95, 96, 118 and 119, where $a_2 > 1$ and $b_2 > 0$.

To illustrate this, first consider items 83 and 84 of group III, whose responses are "correct". The parameter estimates for item 83 were $a_1 = 2.35$, $b_1 = 0.27$, $a_2 = 1.02$ and $b_2 = 0.72$. For item 84, they were $a_1 = 2.40$, $b_1 = 0.61$, $a_2 = 0.71$ and $b_2 = 2.15$. Thus, as for propensity, responses to the items allow

Table 5. Results for Group III.

Item	Parameter estimate				Frequency, %						Fit	
					Expected			Observed				
	a_1	b_1	a_2	b_2	π_{00}	π_{01}	π_{11}	π_{00}	π_{01}	π_{11}	$\chi^2_{(2)}$	p-value
19	1.17	−1.09	0.40	−1.26	10.8	27.1	62.1	10.8	27.3	61.9	**0.27**	**0.8730**
20	1.57	0.18	1.16	1.27	45.8	31.9	22.3	46.0	30.4	23.5	14.87	0.0006
21	1.86	0.11	0.84	−0.15	28.7	18.4	52.8	28.5	18.7	52.8	**0.80**	**0.6700**
22	1.27	−1.43	0.28	−2.61	6.7	26.3	67	6.2	26.8	67.0	**5.74**	**0.0568**
35	0.79	−3.74	0.43	3.58	5.1	76.6	18.3	5.1	76.3	18.6	**0.46**	**0.7933**
82	1.74	−0.38	1.00	0.00	23	26.9	50.1	21.8	28.0	50.2	12.12	0.0023
83	2.35	0.27	1.02	0.72	43	22.3	34.7	43.9	20.6	35.4	19.15	0.0001
84	2.40	0.61	0.71	2.15	58.1	22.3	19.6	60.5	19.3	20.2	63.90	0.0000
86	2.01	0.00	0.50	−0.43	25.1	19.9	55.1	26.7	18.3	55.0	24.59	0.0000
87	1.99	0.05	0.64	2.25	43.7	35.8	20.5	45.4	33.6	21.0	23.43	0.0000
88	1.63	−0.31	0.86	1.13	32	38.5	29.5	32.8	37.0	30.2	10.13	0.0063
91	2.27	0.31	0.97	0.66	42.4	21.2	36.4	43.8	19.1	37.1	30.13	0.0000
92	2.31	0.16	0.77	2.37	49.7	34.7	15.6	51.2	32.6	16.3	21.48	0.0000
93	2.67	0.28	0.71	1.04	43.4	23.1	33.5	46.3	19.7	34.0	83.34	0.0000
97	1.58	−1.27	0.33	−1.83	7.3	28.2	64.5	6.7	28.8	64.4	**6.71**	**0.0348**
98	1.51	−0.39	0.42	−0.95	17.6	22.9	59.5	18.8	21.8	59.4	15.17	0.0005
99	1.94	0.64	0.80	1.03	49.3	18.5	32.2	50.1	17.2	32.8	13.14	0.0014
100	2.15	−0.48	0.43	−3.93	6.8	9.5	83.7	5.8	10.8	83.5	34.05	0.0000
101	1.30	−1.75	0.19	−2.11	5.8	34.5	59.7	5.9	34.4	59.6	**0.29**	**0.8641**
103	1.91	0.33	0.57	0.05	32.9	17.7	49.4	34.2	16.4	49.4	16.17	0.0003
104	1.40	−0.70	0.39	1.78	22.6	43.7	33.6	23.3	42.9	33.8	**3.97**	**0.1372**
111	0.98	−1.19	0.20	4.67	19.8	51.8	28.4	20.6	50.9	28.5	**4.94**	**0.0848**
118	1.90	−0.36	1.06	0.93	31.0	38.9	30.2	30.9	38.0	31.1	**5.33**	**0.0696**
119	1.96	0.23	1.14	1.33	48.3	30	21.7	48.4	28.7	22.9	13.76	0.0010
120	1.89	−0.58	0.83	0.43	22.1	35.8	42.1	22.0	35.5	42.5	**0.58**	**0.7474**

for good discriminating power and have positive difficulty parameters. This suggests the responses to the items convey information regarding the participants who are more prone to respond incorrectly rather than not to respond.

Table 7 compares the expected joint percentage distribution of R_{ij} and U_{ij} from our model (9) with its empirical joint distribution (in parentheses). There is poor adjustment to model (9) for items 83 (χ^2 = 19.15; p-value < 0.001) and 84 (χ^2 = 63.90; p-value < 0.001), if taken in isolation. However, if considered together with the other items from Group III, items 83 and 84 do not deviate significantly from the expected lines in Figure 9 and Figure 10.

Table 6. Results for Group IV.

Item	Parameter estimate				Frequency, %						Fit	
					Expected			Observed				
	a_1	b_1	a_2	b_2	π_{00}	π_{01}	π_{11}	π_{00}	π_{01}	π_{11}	$\chi^2_{(2)}$	p-value
33	1.81	−0.42	0.77	−3.51	3.4	3.9	92.7	3.5	4.3	92.2	**5.40**	**0.0673**
34	1.43	−0.87	0.43	−2.12	8.9	20.1	71	9.2	20.0	70.8	**1.49**	**0.4751**
40	1.43	−0.85	0.30	1.51	18.1	42.8	39.1	18.2	42.6	39.2	**0.22**	**0.8957**
41	1.60	−0.62	0.58	−2.52	7.5	12.3	80.2	7.9	12.4	79.7	**2.81**	**0.2450**
42	1.34	−1.01	0.37	1.11	16.1	43.9	40	16.6	43.3	40.1	**2.87**	**0.2378**
43	1.48	−0.63	0.42	−1.07	14.1	25.1	60.8	14.6	24.8	60.6	**2.55**	**0.2790**
44	1.09	−1.52	0.36	3.87	16	63.6	20.4	16.7	62.7	20.6	**4.78**	**0.0915**
46	1.78	−0.72	0.45	−3.27	6.5	12.9	80.6	6.5	13.1	80.4	**0.64**	**0.7249**
47	1.88	−0.11	0.66	−2.52	9.2	7.9	82.9	9.9	7.7	82.3	**6.20**	**0.0449**
48	1.95	−0.04	0.63	−1.36	16.8	13.9	69.3	17.2	13.9	68.8	**1.45**	**0.4831**
49	1.64	−0.12	0.52	0.74	29.1	30	40.8	29.8	29.3	41.0	**3.71**	**0.1564**
50	1.24	−1.34	0.23	2.75	13.8	51.3	35	15.0	50.0	35.0	13.78	0.0010
51	1.60	−0.82	0.36	−0.73	13	30.6	56.4	13.3	30.4	56.4	**0.75**	**0.6859**
61	1.94	0.46	0.78	−0.98	23	10.1	66.9	24.0	9.6	66.3	**7.71**	**0.0211**
68	1.67	0.16	0.67	−0.57	24.4	16.8	58.9	25.7	15.7	58.6	14.35	0.0008
80	1.38	−0.30	0.42	2.81	33	43.1	23.9	33.7	42.2	24.1	**3.34**	**0.1878**
85	1.46	−0.46	0.64	−1.44	12.5	17	70.5	13.1	16.9	70.0	**3.67**	**0.1595**
95	2.67	0.73	1.02	0.94	53.4	16.3	30.3	54.4	14.4	31.2	31.10	0.0000
96	2.54	0.81	1.13	0.63	50.8	13.8	35.4	52.5	11.3	36.2	69.79	0.0000
105	1.64	−0.22	0.56	−1.15	16.9	18.2	64.9	17.3	18.1	64.6	**1.09**	**0.5787**
106	2.43	0.67	0.87	−0.35	33	10.2	56.8	33.8	9.8	56.5	**4.45**	**0.1081**
109	1.96	−0.48	0.64	−3.63	4.2	5.5	90.3	4.4	5.8	89.8	**2.34**	**0.3106**
110	1.81	−0.04	0.87	−1.74	11.6	8.6	79.8	12.3	8.8	78.8	**6.34**	**0.0420**
112	1.94	−0.56	0.55	−2.02	9.8	15.9	74.3	9.8	16.3	73.9	**1.19**	**0.5513**
113	1.89	−0.31	1.04	−3.05	2.9	2.5	94.6	3.0	3.2	93.9	13.88	0.0010
116	2.01	−0.07	0.80	−0.99	17.9	14.7	67.4	18.8	14.5	66.8	**5.81**	**0.0548**
117	1.96	0.10	0.73	0.20	30.7	22.6	46.7	31.8	21.5	46.7	9.94	0.0070

As another example, consider items 70-75 from group II, where item 70 is "incorrect" and the remaining are "correct." Parameter estimates for these items are presented in previous Table 4. As for proficiency, such items have moderate discriminating power ($0.73 \le a_2 \le 0.91$) and positive difficulty ($0 \le b_2 \le 1.85$). Regarding propensity, the items show high discriminating power ($2.14 \le a_1 \le 2.58$) and the difficulty parameters are $-0.74 \le b_1 \le 0.16$.

Table 8 shows the expected joint distributions from model (9) and the empirical

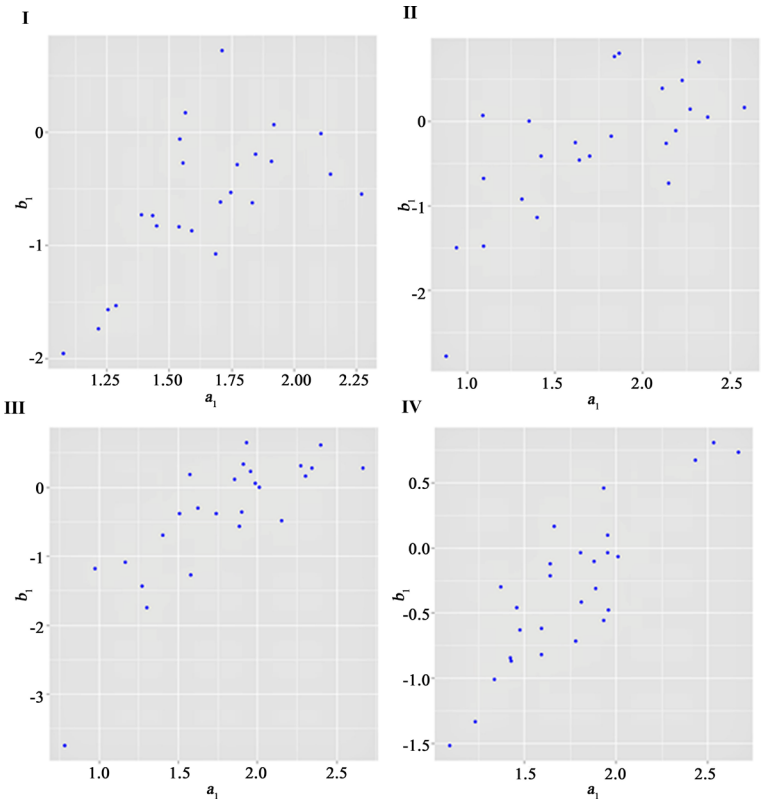


Figure 6. Dispersion between the discriminating power parameter and the difficulty parameter regarding propensity, by group.

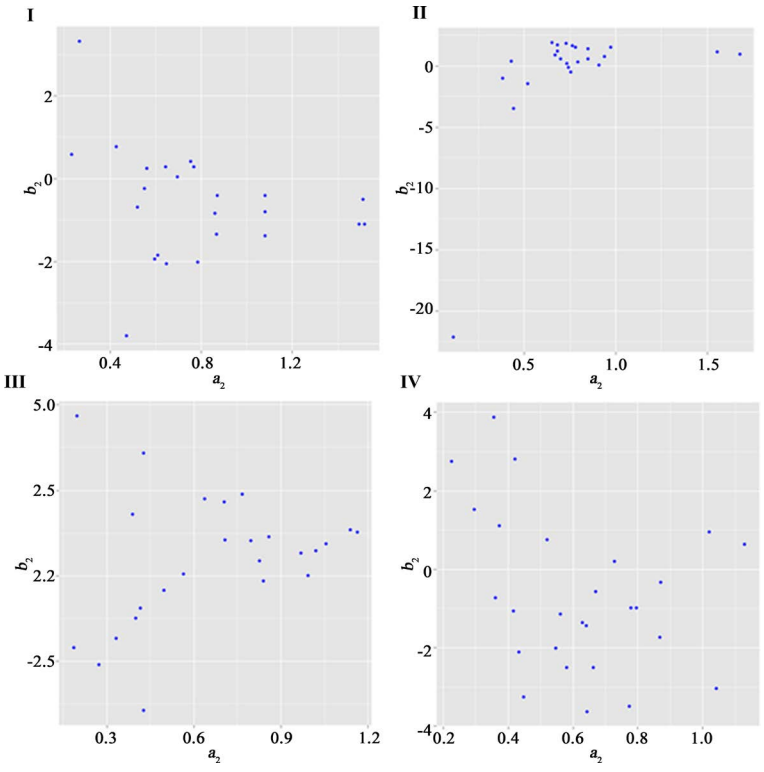


Figure 7. Dispersion between the discriminating power parameter and the difficulty parameter regarding proficiency, by group.

Table 7. Joint percentage distribution of R_{ij} (responded = 1; didn't respond = 0) and U_{ij} (correct = 1; incorrect = 0): Expected from model (9) and empirical (in parentheses). Items 83 and 84.

Item 83		R_{ij}		Item 84	R_{ij}	
		0	1		0	1
U_{ij}	0	43.0% (43.9%)	22.3% (20.6%)	0	58.1% (60.5%)	22.3% (19.3%)
	1	0	34.7% (35.4%)	1	0	19.6% (20.2%)

Table 8. Joint percentage distribution of R_{ij} (responded = 1; didn't respond = 0) and U_{ij} (correct = 1; incorrect = 0): Expected from model (9) and empirical (in parentheses). Items 70 - 75.

Item 70		R_{ij}		Item 71	R_{ij}	
		0	1		0	1
U_{ij}	0	28.0% (27.5%)	23.0% (23.4%)	0	31.3% (33.2%)	21.5% (19.5%)
	1	0	48.9% (49.1%)	1	0	47.2% (47.3%)
Item 72		R_{ij}		Item 73	R_{ij}	
		0	1		0	1
U_{ij}	0	37.5% (38.5%)	23.4% (21.9%)	0	35.0% (36.6%)	19.9% (18.1%)
	1	0	39.1% (39.6%)	1	0	45.1% (45.3%)
Item 74		R_{ij}		Item 75	R_{ij}	
		0	1		0	1
U_{ij}	0	22.6% (23.7%)	51.7% (50.0%)	0	35.2% (36.7%)	42.5% (40.4%)
	1	0	25.6% (26.2%)	1	0	22.3% (22.9%)

ones. Again, apart from item 70, the items did not appear to adhere to the model ($\chi^2 > 13$; p-value < 0.002). However, both fractions of observed correct responses ($\tilde{\pi}_{11}$) and observed nonresponses ($\tilde{\pi}_{00}$) fall near their corresponding expected lines given by model (9) (**Figure 9** and **Figure 10**).

Figure 8 summarizes the χ^2 statistics distances between the observed distributions and those expected from model (9) for the items, by group. Horizontal dashed lines divide those 52 items for which the model is better adjusted (9 items from Group I; 12 from II; 10 from III; and 21 from IV). For all the 52 items beneath the lines, $\chi^2 < 9.21$ with p-values less than 1 percent. However, perhaps apart from the Group I items in **Figure 9**, there are more than 52 items whose observed frequencies $\tilde{\pi}_{00}$ and $\tilde{\pi}_{11}$ are similar to the expected frequencies

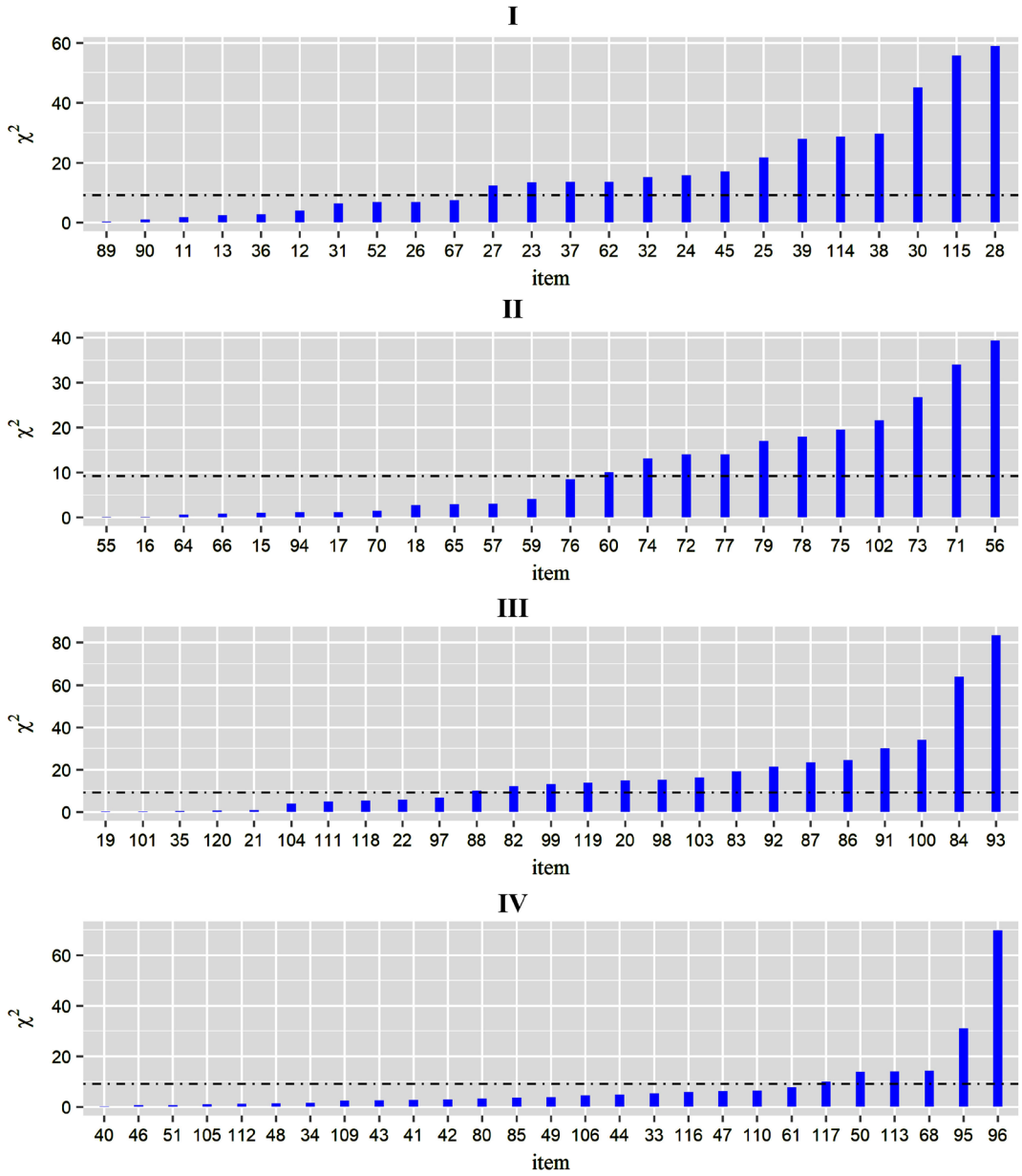


Figure 8. χ^2 distances (with d.f. = 2) between the observed distributions and those expected from model (9) for the items, by group. The items whose distances are statistically null fall below the horizontal dashed lines (critical value of 9.21 at the significance level of 1 percent), and thus are well adjusted to model (9).

from the model, $\hat{\pi}_{00}$ and $\hat{\pi}_{11}$, with $\rho > 0.995$ (Figure 9 and Figure 10).

A slightly different picture emerges from the Group I nonresponses (Figure 9), where the fractions of nonresponses, $\tilde{\pi}_{00}$, fall above the expected ones, $\hat{\pi}_{00}$.

Figure 11 shows the joint distribution between θ_1 and θ_2, by group. It suggests the existence of at least two types of participants. The clusters of dots at the top refer to the participants who do not leave items blank ($T_n = 0$). Overall, less proficient participants (low θ_2) are less likely to respond incorrectly and then leave an item blank (low θ_1). However, proficiency changes over after a threshold and then propensity tends to the modal region of the distributions.

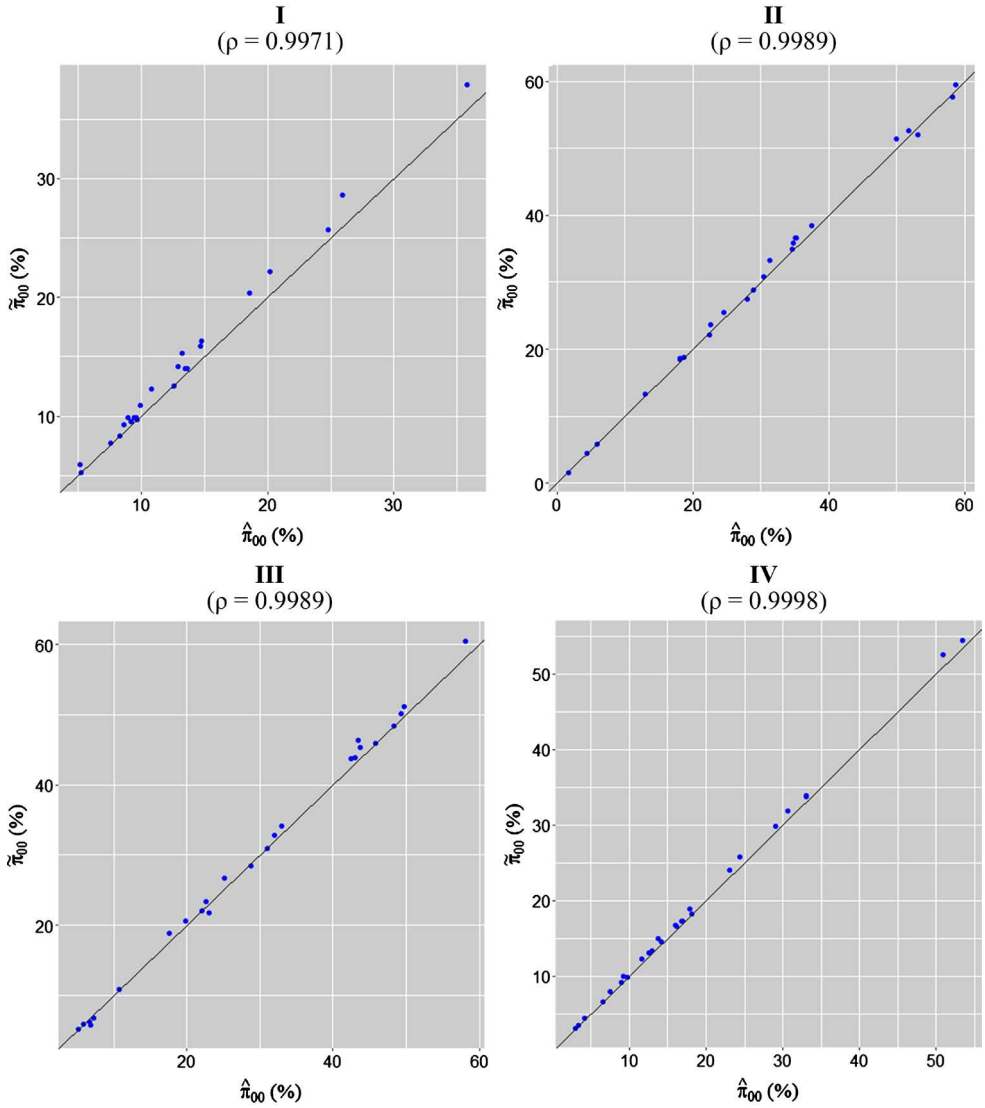


Figure 9. Dispersion between the observed $\tilde{\pi}_{00}$ and the expected $\hat{\pi}_{00}$ frequencies of nonresponses, by group (corresponding Pearson correlation coefficients ρ in parentheses).

Figure 12 and Figure 13 show the relationship between score, *S*, and the latent variables θ_1 and θ_2, by group. Score and propensity present low negative linear correlation (Figure 12). Solid lines show conditional mean values $S \mid \theta_1$ that are adjusted nonparametrically using the LOESS method. From a threshold (say, $\theta_1 > 1$), participants with higher propensities score lower. However, before this threshold is reached, expected scores lie on a plateau around which dispersions are funnel shaped. Moreover, and as expected, participants who are more proficient tend to lie above the solid line.

Figure 13 shows scoring and proficiency are positively correlated and nonlinear. Participants who are more proficient tend to score more and at a higher intensity (slope) than that of those who score lower. As expected, participants with higher propensities tend to lie below the solid line. For a given level of proficiency, θ_2, participants with higher propensities tend to score lower. However,

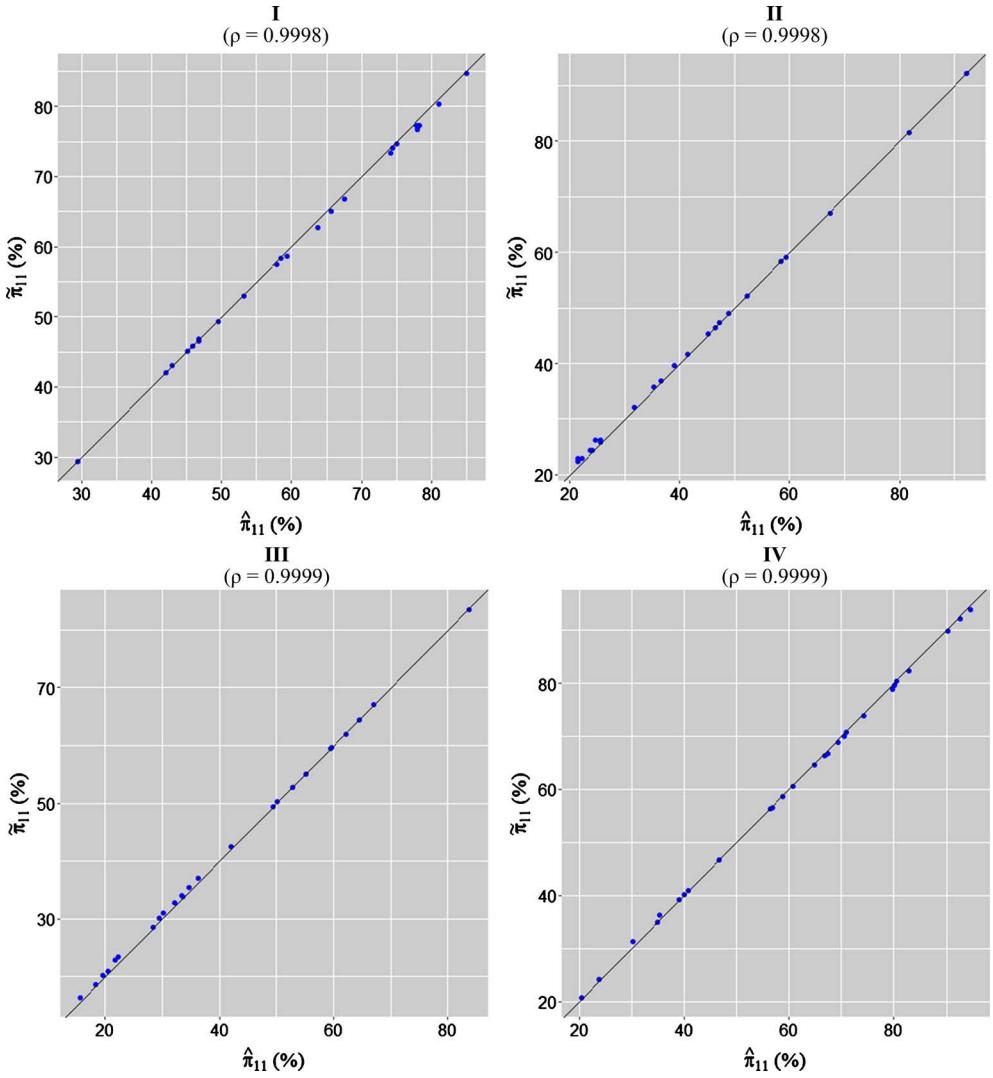


Figure 10. Dispersion between the observed $\tilde{\pi}_{11}$ and the expected $\hat{\pi}_{11}$ frequencies of correct answers, by group (corresponding Pearson correlation coefficients ρ in parentheses).

as θ_2 rises, the dispersion of S lessens, and this dampens the effect of θ_1. Yet, as θ_2 is reduced, θ_1 impacts S more, and $S \mid \theta_2$ tends to flatten.

4. Conclusions

This work considers item response theory to model 10,822 Brazilian high school students' behavior in a high-stakes exam that may enable them to enter a top university. We put forward a model based on item response theory that highlights the role of latent features that we call "proficiency" and "propensity".

The key strategic decision of a participant is to either risk an incorrect response or leave the question blank. Leaving the question blank is strategically better, because responding incorrectly is a loss. A participant then decides by taking into account both intrinsic difficulty and the latent feature of propensity.

Leaving a question blank may also reflect the participant's low proficiency regarding the item as well as the propensity to avoid the loss accruing from responding

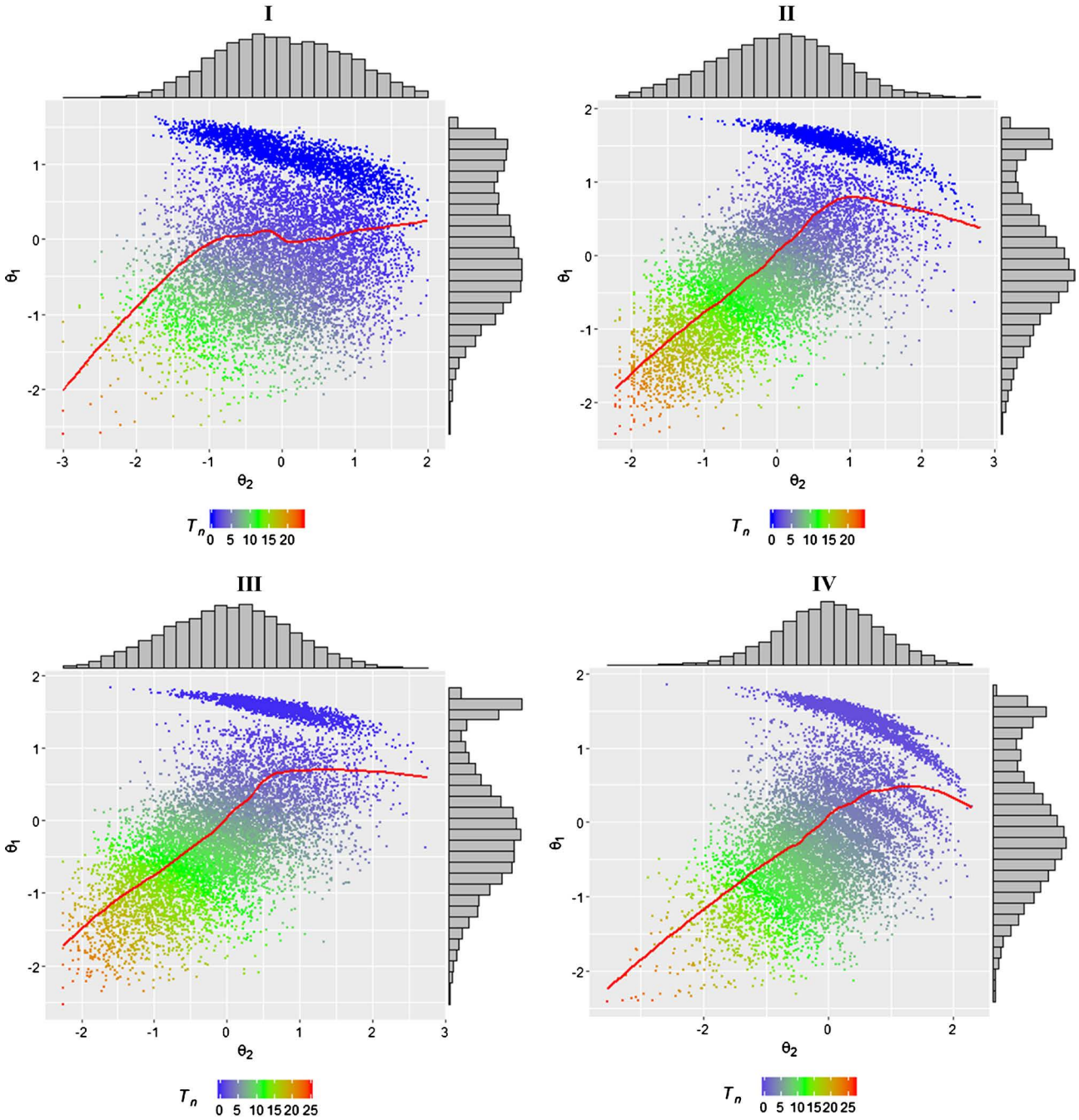


Figure 11. Dispersion between the latent features θ_1 and θ_2, by group. Solid red lines show conditional mean values $\theta_1 \mid \theta_2$ that are adjusted nonparametrically using the LOESS method, and T_n shows the total of unanswered items.

incorrectly. Our model aims to recover information regarding the role the latent features—proficiency and propensity—play in a decision.

In the model we set, propensity is defined exactly by Equation (5), while proficiency is defined by Equation (6). Propensity means propensity to respond incorrectly rather than not to respond. And (low) proficiency of responding correctly cannot be compensated by the propensity of responding incorrectly.

We estimate by maximum likelihood (using the language *R*) the parameters

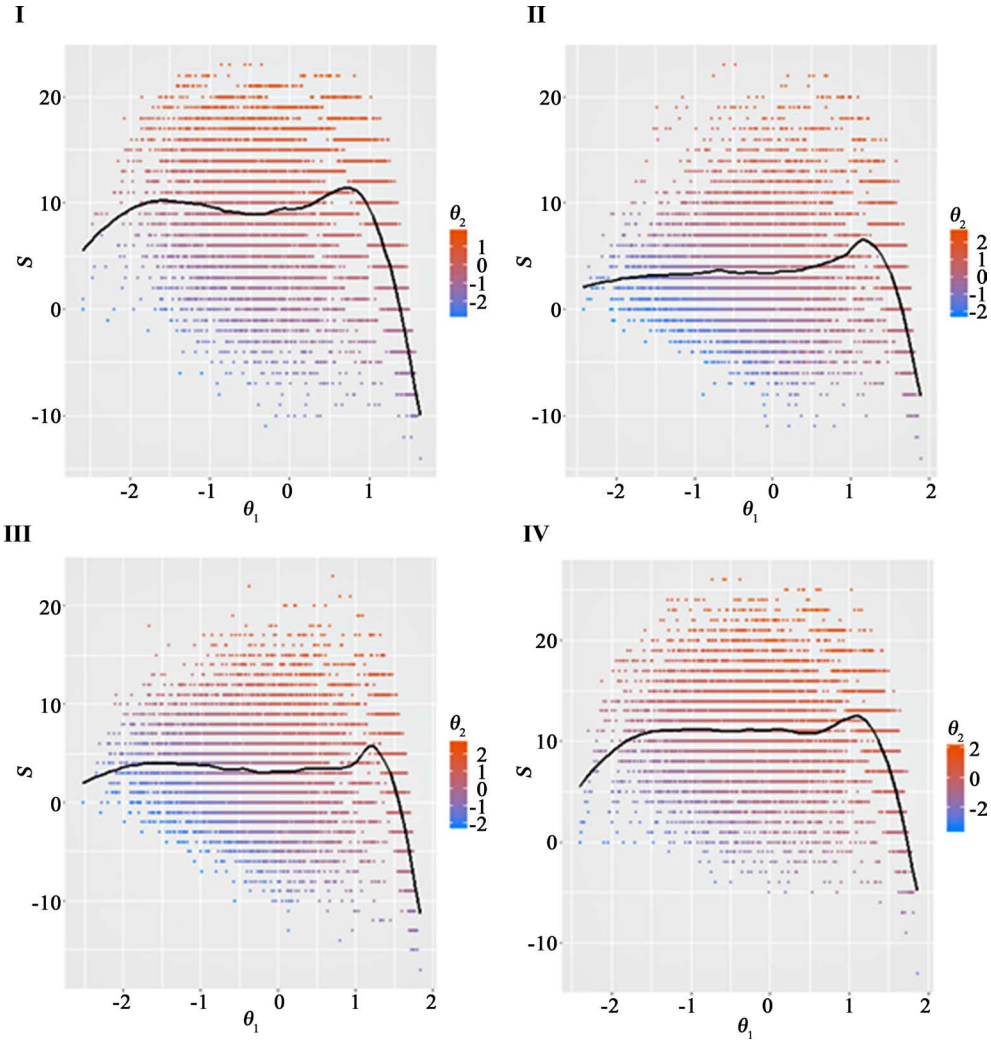
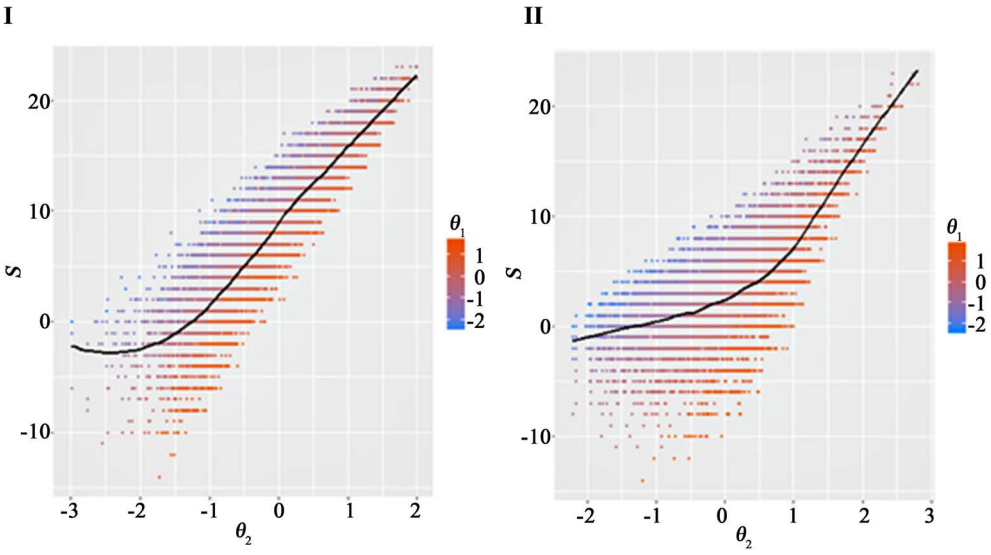


Figure 12. Dispersion between score S and propensity θ_1, by group. For each group, the linear correlations between S and θ_1 are, respectively, −0.20, −0.03, −0.17 and −0.14. Solid lines show conditional mean values $S \mid \theta_1$ that are adjusted nonparametrically using the LOESS method.



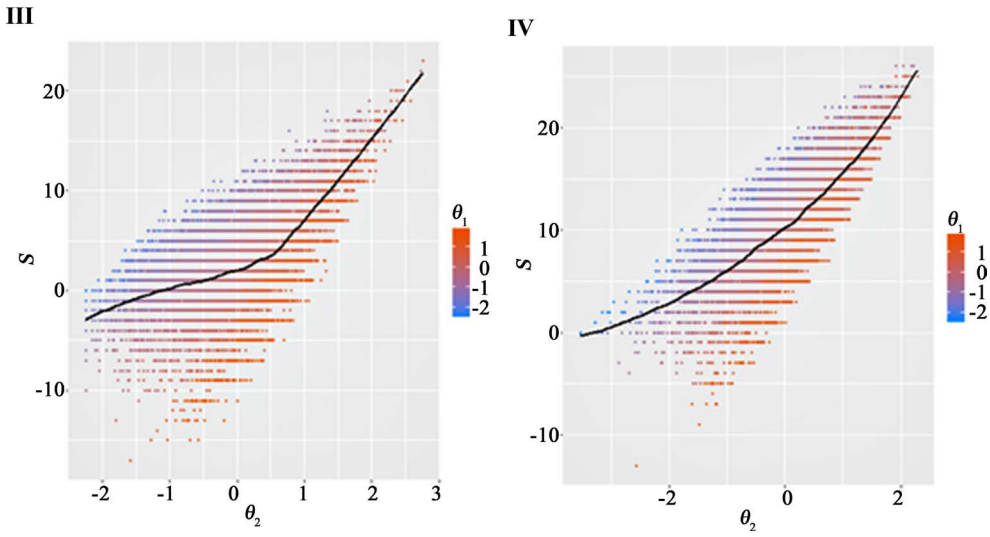


Figure 13. Dispersion between score S and proficiency θ_2, by group. For each group, the correlations between S and θ_2 are, respectively, 0.91, 0.67, 0.60 and 0.76. Solid lines show conditional mean values $S \mid \theta_2$ that are adjusted nonparametrically using the LOESS method.

related to the discriminating power of a participant and the parameters of difficulty related to an item. Proficiency and propensity are estimated by the expected a posteriori method.

Based on the chi-squared distances, 52 items out of 100 proved to be a good fit to the model. For each group, the overall adhesion of data to our adjusted model was evaluated by the Pearson correlation coefficient (ρ). Both responding correctly and propensity showed a strong agreement with the adjusted model (**Figure 9** and **Figure 10**), with $\rho > 0.995$.

This suggests the decision of responding or not and also the decision of responding correctly or not in a group of items can be described by a two-dimensional logistic model, even if there are imperfections coming from an item-by-item adjustment.

Refraining from responding is found to depend on both the characteristics of the items and the latent features of the participants. In particular, the least proficient participants prefer to leave an item blank rather than respond it incorrectly.

Scoring on the exam and propensity present a low negative linear correlation. However, scoring and proficiency are positively correlated although nonlinear. Thus, for a given level of proficiency, after a threshold is reached, students with higher propensities score lower.

Acknowledgements

We acknowledge financial support from Cebraspe, CNPq and Capes.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this pa-

per.

References

[1] Klein, S.P. and Hamilton, L. (1999) Large-Scale Testing: Current Practices and New Directions. *Rand Education*. https://www.rand.org/content/dam/rand/pubs/issue_papers/2006/IP182.pdf

[2] Hamilton, L.S., Stecher, B.M. and Klein, S.P. (2002) Making Sense of Test-Based Accountability in Education. *Rand Education*. https://www.rand.org/content/dam/rand/pubs/monograph_reports/2002/MR1554.pdf

[3] Abdelfattah, F.A. (2007) Response Latency Effects on Classical and Item Response Theory Parameters Using Different Scoring Procedures. PhD Thesis, Ohio University, Athens, OH.

[4] Lievens, F., Sackett, P.R. and Buyse, T (2009) The Effects of Response Instructions on Situational Judgment Test Performance and Validity in a High-Stakes Context. *Journal of Applied Psychology*, **94**, 1095-1101. https://doi.org/10.1037/a0014628

[5] Baker, F.B. (2001) The Basics of Item Response Theory. ERIC Clearinghouse on Assessment and Evaluation, College Park, MD.

[6] Rose, N., von Davier, M. and Xu, X. (2010) Modeling Nonignorable Missing Data with Item Response Theory (IRT). Technical Report, ETS, Princeton.

[7] Bertoli-Barsotti, L. and Punzo, A. (2013) Rasch Analysis for Binary Data with Nonignorable Nonresponses, *Psicologica*, **34**, 97-123.

[8] Knott, M., Albanese, M. and Galbraith, J. (1990) Scoring Attitudes to Abortion. *The Statistician*, **40**, 217-223. https://doi.org/10.2307/2348494

[9] Albanese, M. and Knott, M. (1992) TWOMISS: A Computer Program for Fitting a One- or Two-Factor Logit-Probit Latent Variable Model to Binary Data When Observations May Be Missing. LSE Technical Report, London.

[10] Knott, M. and Tzamourani, P. (1997) Fitting a Latent Trait Model for Missing Observations to Racial Prejudice Data. In: Rost, J. and Langeheine, R., Eds., *Applications of Latent Trait and Latent Class Models in the Social Sciences*, Waxmann, Munster, 244-252.

[11] Bartholomew, D.J., de Menezes, L.M. and Tzamourani, P. (1997) Latent Trait Class of Models Applied to Survey Data. In: In: Rost, J. and Langeheine, R., Eds., *Applications of Latent Trait and Latent Class Models in the Social Sciences*, Waxmann, Munster, 219-232.

[12] O'Muircheartaigh, C. and Moustaki, I. (1996) Item Non-Response in Attitude Scales: A Latent Variable Approach. *Proceedings of the American Statistical Association*, Section of Survey Research Methods, 938-943.

[13] O'Muircheartaigh, C. and Moustaki, I. (1999) Symmetric Pattern Models: A Latent Variable Approach to Item Non-Response in Attitude Scales. *Journal of the Royal Statistical Society A*, **162**, 177-194. https://doi.org/10.1111/1467-985X.00129

[14] Moustaki, I. and Knott, M. (2000) Weighting for Item Non-Response in Attitude Scales by Using Latent Variable Models with Covariates. *Journal of the Royal Statistical Society A*, **163**, 445-459. https://doi.org/10.1111/1467-985X.00177

[15] Moustaki, I. and O'Muircheartaigh, C. (2000) A One Dimension Latent Trait Model to Infer Attitude from Nonresponse for Nominal Data, *Statistica*, **60**, 259-276.

[16] Moustaki, I. and O'Muircheartaigh, C. (2002) Locating "Don't Know", "No An-

swer" and Middle Alternatives on an Attitude Scale: A Latent Variable Approach. In: Marcoulides, G.A. and Moustaki, I., Eds., *Latent Variable and Latent Structure Models*, Lawrence Erlbaum Associates, London, 15-40.

[17] Andrade, D.F. and Tavares, H.R. (2005) Item Response Theory for Longitudinal Data: Population Parameter Estimation. *Journal of Multivariate Analysis*, **95**, 1-22. https://doi.org/10.1016/j.jmva.2004.07.005

[18] Reckase, M.D. (2009) Multidimensional Item Response Theory. Springer, New York. https://doi.org/10.1007/978-0-387-89976-3

[19] Hambleton, R.K., Swaminathan, H. and Rogers, H.J. (1991) Fundamentals of Item Response Theory. Sage, London.