ISSN: 2317-0840

Data Reduction in Item Response Theory Models

Renan Barufaldi Bueno[1†], Marcelo Andrade da Silva[1]

[1] *Universidade de São Paulo (USP), Escola Superior de Agricultura "Luiz de Queiroz" (ESALQ), Departamento de Ciências Exatas, Piracicaba - SP.*

Abstract: *One of the main classes of statistical models established in psychometric literature is Item Response Theory (IRT), which plays a crucial role in constructing scales of latent traits and allows for the creation of measurement instruments that are more precise and adapted to respondents' characteristics. The primary objective of this paper is to analyze, through simulation studies, the possible consequences of reducing response categories on the parameter estimates of IRT models, as this process can simplify the estimation phase and is used in some research. After analyzing the obtained results, it is concluded that this reduction mechanism provides satisfactory computational optimization. However, it also results in the loss of information, as the quality of IRT parameter estimates is negatively affected, directly interfering with the results of applied studies. Finally, this analysis aims to contribute significantly to psychometric literature, becoming capable of promoting new research in this statistical domain, as well as providing a solid foundation for future methodological developments.*

Keywords: *IRT models; simulation; reduction of categories.*

Introduction

Item Response Theory (IRT) represents a category of well-established statistical models in the literature, which estimate unobservable characteristics and provide insights into the quality of items used on a specific measurement scale. According to Andrade et al. (2000), IRT was first employed in Brazil to analyze the data of National Basic Education System (SAEB - *Sistema de Avaliação da Educação Básica*, in portuguese) data in 1995. Based on the results obtained from SAEB, subsequent large-scale assessments, including the São Paulo State School Performance Assessment System (SARESP - *Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo*, in portuguese), were designed and implemented for IRT-based analysis.

In the context of polytomous items, several researchers choose to reduce the number of response categories using IRT models, aiming to decrease computational time, among other reasons. For instance, in a study inspired by Vincenzi et al. (2018), this procedure was employed to investigate sustainability perception among residents of the Paraná Basin III. Initially, the collected data had six response categories (strongly disagree, disagree, slightly disagree, slightly agree, agree, and strongly agree). However, to fit the graded response model proposed by Samejima (1969), researchers grouped the categories, reducing them to three.

Furthermore, reducing response categories may aim to dichotomize items, a procedure converting polytomous items into dichotomous ones, widely adopted in various studies (da Silva et al., 2018; Fragoso and Cúri, 2013). This reduction can offer advantages such as decreased computational time and adaptation to simplified models due to fewer parameters to estimate. However, it can also have disadvantages, including potential information loss, directly impacting the estimation and precision of parameters associated with items and individuals.

[†]Autor correspondente: renanbbueno.rb@usp.br

Manuscrito recebido em: 30/07/2024
Manuscrito revisado em: 22/11/2024
Manuscrito aceito em: 26/11/2024

This research aims to obtain detailed information and conclusions regarding the potential consequences of reducing response categories and using Item Response Theory (IRT) models through simulation studies. The analysis will employ IRT models to simulate and verify the quality of parameter estimates for items and individuals, analyzing results to optimize computational execution time and preserve essential information. Thus, the dissemination of this project aims to stimulate advancements in research fields related to psychometrics and the application of IRT.

Finally, this article is organized as follows. The "Methods" section describes the software used in the study's development and details the IRT models considered in study, based on their parameters obtained through a simulation study employing three statistical measures for their evaluation. The "Results and Discussion" section presents the obtained results, including tables with the respective values of the statistical measures for all parameters of each model in different scenarios. Additionally, the differences between the values obtained with and without the reduction of response categories in the Graded Response Model (GR model) are demonstrated through graphs. This section also includes an application of the study to a real dataset on sustainability obtained by Vincenzi et al. (2018). Moreover, some conclusions and observations are presented in the "Conclusion" section.

Methods

This section presents: (1) a description of the software used, including its characteristics and functionalities; (2) IRT models used in this study; (3) concepts of simulation studies.

Softwares

In this project, the R language (R Core Team, 2023) was utilized as the foundation for the studies. This language is extensively employed in statistical contexts, enabling data analysis, manipulation, and visualization, thereby streamlining various stages of computational processing. Moreover, its versatility is supported by a large active community that provides numerous statistical packages, enhancing essential commands. Specifically, the *tidyverse* package (Wickham *et al.*, 2019), developed to facilitate data manipulation, the *mirt* package (Chalmers, 2012), designed to handle data in major IRT models, and the *geobr* (Pereira et al., 2021) package, programmed to enable the creation of maps, were explored in this project. Finally, it is noteworthy that this language was utilized through the RStudio Team (2022) program.

IRT models

IRT models are statistical models that relate a set of observable variables to a set of unobservable variables, typically referred to as latent variables or traits (Rao; Sinharay, 2007). These models aim to estimate latent traits such as proficiency in educational assessments, consumer satisfaction with a product or service, and measures of depressive symptoms. Practically, individuals in the population of interest, such as students, consumers, patients, among others, are subjected to a measuring instrument like a test, questionnaire, or clinical assessment. These measuring instruments consist of items proposed by experts to evaluate the latent traits of interest in individuals.

Three IRT models were studied and applied in the development of this project. Two of them, known as the two-parameter logistic model (2PL model) and the three-parameter logistic model (3PL model) (Lord, 1952; Birnbaum, 1968), are suitable for dichotomous items, which have only two response categories. The other model used is known as the graded response model (GR model) (Samejima, 1969), which is used for polytomous items with ordered response categories.

Model parameters were determined through classical inference methods implemented in the *mirt* package (Chalmers, 2012), a widely used R-based tool for statistical analysis (R Core Team, 2023). This robust and reliable approach allowed for a detailed and precise analysis of the parameters, ensuring high-quality and reliable results for interpreting the models in question.

Two-Parameter Logistic (2PL model)

The unidimensional two-parameter logistic model, introduced by Lord (1952) and further developed by Birnbaum (1968), remains the most widely used model for dichotomous data in psychometrics. This model is given by

$$P(Y_{ij} = 1|\theta_i, \xi_j) = \frac{1}{1 + e^{-D_{a_j}(\theta_i - b_j)}}$$

where $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, J$, with n representing the number of individuals and J the number of items. Furthermore, Y_{ij} is a dichotomous variable taking the value 1 when individual i responds correctly or 0 when responding incorrectly to item j; θ_i corresponds to the latent trait value of individual i; $\xi_j = (a_j, b_j)'$ is the parameter vector for item j; $P(Y_{ij} = 1|\theta_i, \xi_j)$ represents the probability of individual i answering item j correctly given θ_i, referred to as the Item Response Function (IRF); a_j is the discrimination power of item j proportional to the slope of the Item Characteristic Curve (ICC) at b_j; b_j refers to the difficulty parameter of item j; and finally, D is a constant scaling factor set to 1 (1.7 is used when aiming for logistic function results similar to the normal ogive function).

Three-Parameter Logistic (3PL model)

The unidimensional three-parameter logistic model, developed by Lord (1952) and Birnbaum (1968), shares the core structure of the 2PL model but introduces an additional parameter to account for the possibility of individuals answering items correctly by chance, even without sufficient ability. This possibility is statistically represented by the parameter c. This model is given by

$$P(Y_{ij} = 1|\theta_i, \xi_j) = c_j + (1 - c_j)\frac{1}{1 + e^{-D_{a_j}(\theta_i - b_j)}},$$

where $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, J$, with n representing the number of individuals and J the number of items.

Graded Response Model (GR model)

The graded response model proposed by Samejima (1969) is considered an extension of the 2PL model, as it is used to handle polytomous items. This model was also defined as a difference model by Thissen and Steinberg (1986), due to the response probabilities being modeled by the difference obtained between adjacent cumulative probabilities.

According to the model, the probability P_{ijk} of respondent i choosing category k on item j is modeled based on cumulative probabilities (P^+_{ijk}), defined as

$$P^+_{ijk} = P(Y_{ij} \leq k|\theta_i, \xi_j) = \begin{cases} L(\eta_{ijk}), & \text{if } k \in \{1, 2, ..., m_j - 1\} \\ 1, & \text{if } k = m_j \end{cases}, \tag{1}$$

where $L(\cdot)$ describes the item response function often used in the cumulative logistic distribution.

In the GR model, there are m_j response categories for a given item j, and therefore, there are $m_j - 1$ difficulty parameters (b) for each item. The latent predictor η_{ijk} is defined as

$$\eta_{ijk} = a_j(b_{jk} - \theta_i) = b^\star_{jk} - a_j\theta_i, \tag{2}$$

where a_j is the discrimination parameter for item j, and b_{jk} is the difficulty parameter for response category k of item j.

As mentioned earlier, P_{ijk} is obtained as the difference between adjacent cumulative probabilities, that is,

$$P_{ijk} = P(Y_{ij} = k|\theta_i, \xi_j) = \begin{cases} P^+_{ijk}, & \text{if } k = 1 \\ P^+_{ijk} - P^+_{ij[k-1]}, & \text{if } 2 \leq k < m_j \\ 1 - P^+_{ij[k-1]}, & \text{if } k = m_j \end{cases}. \quad (3)$$

In IRT modeling, assessing the discrimination and difficulty parameters is crucial for understanding the characteristics and effectiveness of items in differentiating respondents. The discrimination parameters represent the degree to which an item's response curve is sensitive to changes in the respondent's latent trait. Thus, higher discrimination parameters indicate that an item is more effective in this differentiation process. Meanwhile, difficulty parameters represent the location of an item's response curve on the latent trait scale, meaning that items with higher difficulty parameters require a higher latent trait value to be scored. Consequently, by examining both parameters, researchers can gain essential insights into the quality and suitability of the items.

Simulation studies

According to Ehrlich (1991), simulation represents a method used to study the behavior of a system by developing a mathematical model that faithfully expresses its aspects. Thus, by modifying this model, it is possible to investigate the effects on the system. In IRT, a simulation study involves simulating the parameters of the used models, employing probability distributions to generate response matrices.

The simulation study is fundamental as it can be applied in various different circumstances, providing a flexible approach. Moreover, it is essential for understanding the behavior of the parameters from the development of response matrices, allowing the construction of different scenarios for comparisons with reality and the development of measures to analyze the quality of the results.

Thus, to compare the performance of parameter estimation of the models and evaluate their recovery, some appropriate statistics were considered. The first statistic used was the Root Mean Square Error (*RMSE*), a measure of precision defined as

$$\text{RMSE} = \sqrt{\frac{1}{20}\sum_{r=1}^{20}(\hat{\pi}_{lr} - \pi_l)^2},$$

where π_l is a parameter of the item or individual with l being a convenient index for this parameter, and $\hat{\pi}_{lr}$ is its respective estimate obtained in replicate r, $r = 1, 2, ..., 20$. This measure allows us to evaluate the accuracy of the estimate found from fitting the models to simulated data, such that lower RMSE values indicate more precise estimates.

The second statistic used was bias, responsible for calculating the degree of bias present in the estimation of the studied parameters. Generally, values close to 0 are desired, as they represent parameters with little bias. This measure is given by

$$\text{Bias} = \hat{\bar{\pi}}_l - \pi_l,$$

where $\hat{\bar{\pi}}_l = \frac{1}{20}\sum_{r=1}^{20}\hat{\pi}_{lr}$.

Finally, the third statistic used was the Pearson correlation coefficient, denoted by r, which assesses the correlation between two linear variables. This coefficient ranges from -1 to

1, indicating a perfect (positive or negative) correlation as it approaches the extremes, and less linear dependence as the value approaches 0. This measure is calculated by

$$r = \frac{\sum_{r=1}^{20}(\hat{\pi}_r - \hat{\bar{\pi}})(\pi_r - \bar{\pi})}{\sqrt{\sum_{r=1}^{20}(\hat{\pi}_r - \hat{\bar{\pi}})^2} \cdot \sqrt{\sum_{r=1}^{20}(\pi_r - \bar{\pi})^2}} = \frac{\text{cov}(\hat{\pi}, \pi)}{\sqrt{\text{var}(\hat{\pi}) \cdot \text{var}(\pi)}},$$

where $\hat{\pi_1}, \hat{\pi_2}, ..., \hat{\pi_{20}}$ and $\pi_1, \pi_2, ..., \pi_{20}$ correspond to the measured values of both variables, and $\hat{\bar{\pi}} = \frac{1}{20}\sum_{r=1}^{20}\hat{\pi}_r$ and $\bar{\pi} = \frac{1}{20}\sum_{r=1}^{20}\pi_r$ refer to their arithmetic means.

Results and Discussion

Below, the results obtained regarding the following psychometric measures will be demonstrated: Root Mean Square Error (RMSE), bias, and average correlation, respectively. These values refer to the discrimination (a), difficulty (b), guessing (c) parameters for 3PL model and the latent variable (θ), calculated according to each of the pre-established scenarios, which vary in terms of sample size (n), items (J), and response categories (m) in the case of GR model.

The processing to obtain these results was conducted using R Core Team (2023), following the steps outlined below: (1) randomly generated parameters a, b, c, and θ; (2) conducted a simulation, adapting them to the specific formulas of each studied model and randomly generating the response matrix; (3) directly estimated these parameters through the mentioned computational program, resulting in estimated values; (4) compared the types of values obtained; (5) these values were then stored in data frames; (6) which were subsequently inserted into formulas for the statistical measures RMSE, bias, and correlation, aiming to obtain results associated with the different predetermined scenarios.

Additionally, in order to conduct a comprehensive Exploratory Data Analysis (EDA), descriptive tables were created. This approach aims to provide a clearer and more focused understanding in interpreting the results obtained, thereby facilitating the formulation of hypotheses and conducting more robust analyses. Thus, these tables offer an organized and detailed view of the data, contributing to a deeper and more effective exploration.

Presented below are Tables 1, 2, and 3, containing the results obtained regarding the values of the statistical measures mentioned earlier and applied to different scenarios.

Table 1: Root Mean Square Error (RMSE), bias, and average correlation for the following scenarios corresponding to 2PL model.

Scenarios	RMSE			Bias			Correlation		
	a	b	θ	a	b	θ	a	b	θ
$n = 100; J = 10$	0.6181	0.5205	0.5241	0.0538	0.1860	0.1365	0.6236	0.9457	0.8381
$n = 500; J = 10$	0.2157	0.1208	0.5361	0.0499	-0.0009	0.01095	0.8613	0.9926	0.8356
$n = 1000; J = 10$	0.1370	0.0887	0.5071	-0.0106	-0.0164	-0.0113	0.9533	0.9953	0.8489
$n = 100; J = 20$	0.4215	0.3703	0.3934	0.0212	0.1560	0.1367	0.6988	0.9346	0.9184
$n = 500; J = 20$	0.2013	0.1062	0.3768	0.0419	0.0102	0.0108	0.8895	0.9945	0.9252
$n = 1000; J = 20$	0.1200	0.0843	0.3885	0.0125	-0.0121	-0.0115	0.9697	0.9949	0.9148

Source: from the authors (2024).

Through the analysis of Table 1, it is noticeable that from scenario 1 to 3, there is a significant increase in the number of individuals, which in turn results in more suitable RMSE and correlation values for parameters a and b, while θ remains nearly constant. Regarding bias, there is a general improvement in the values for all three measures analyzed.

However, when comparing scenarios 1 and 4, 2 and 5, 3 and 6, it can be concluded that maintaining the number of individuals while increasing the number of items leads to better estimation of individual parameters (θ) for the three studied statistical measures. Additionally, there is also a modest improvement in the estimation of parameters a and b.

Table 2: Root Mean Square Error (RMSE), bias, and average correlation for the following scenarios corresponding to 3PL model.

Scenarios	RMSE				Bias				Correlation			
	a	b	c	θ	a	b	c	θ	a	b	c	θ
$n = 100; J = 10$	0.2096	0.1875	0.1984	0.2880	0.0340	0.0289	0.0412	0.0020	0.8977	0.9382	0.9147	0.9128
$n = 500; J = 10$	2.1288	0.4470	0.1565	0.6672	0.7200	-0.0126	-0.0136	0.0108	0.2273	0.8464	0.1635	0.7255
$n = 1000; J = 10$	0.5935	0.4103	0.1321	0.5929	0.1175	-0.0708	-0.0172	-0.0114	0.6570	0.8711	0.1493	0.7819
$n = 100; J = 20$	7.5456	1.1414	0.2009	0.5375	3.1673	0.1684	-0.0045	0.1332	0.0585	0.6357	0.1615	0.8303
$n = 500; J = 20$	0.9081	0.3911	0.1532	0.4710	0.3533	0.02782	0.0056	0.0108	0.5462	0.8963	0.2339	0.8788
$n = 1000; J = 20$	0.3962	0.4326	0.1411	0.4698	0.0675	-0.0024	-0.0037	-0.0114	0.7404	0.8397	0.2189	0.8710
$n = 5000; J = 30$	0.1422	0.1886	0.0657	0.3962	0.0367	-0.0087	-0.0046	0.0034	0.9456	0.9634	0.5958	0.9149

Source: from the authors (2024).

By examining Table 2, it is observed that from scenarios 1 to 3 and 4 to 6, despite an increase in the number of individuals, there is no clear improvement in the estimation of item parameters (a, b, and c) as expected, showing significant instability particularly in scenarios with a low number of individuals.

Furthermore, when comparing scenarios 1 and 4, where the number of individuals was kept at 100 and the number of items increased from 10 to 20, it is generally noted that there is no improvement in the estimation process for any of the parameters, neither for individuals, where a considerable improvement was expected in this case, nor for items, confirming the instability mentioned in the previous paragraph.

Nevertheless, it is important to note that in scenario 7, which includes a large number of individuals (5000) and items (30), the parameter estimation is much more adequate statistically. Therefore, it is concluded that the use of the 3PL model in scenarios with a low number of individuals is impractical due to its instability, thus recommended only for circumstances similar to the cited scenario.

By analyzing Table 3, it is evident that in scenarios 1 to 3 and 4 to 6, with 3 response categories, as well as in scenarios 7 to 9 and 10 to 12, with 4 categories, despite the constant number of items, an increase in the number of individuals leads to improved estimation of item parameters (a and b) across the three psychometric measures evaluated. However, values for the latent variable (θ) do not significantly differ under these circumstances, showing better results in scenarios with a greater number of items. This is supported by the analysis of scenarios 1 and 4; 2 and 5; 3 and 6; 7 and 10; 13 and 16, among others.

Moreover, when evaluating results from scenarios with identical numbers of individuals and items but varying response categories, such as 1, 7, and 13 or 2, 8, and 14, it is inferred that changes in the value of m minimally affect bias and correlation results overall, with a more pronounced impact on the RMSE values, which decrease as the number of response categories increases.

Therefore, it is crucial to emphasize that the accuracy in estimating parameters a, b, and θ is intrinsically linked to the number of individuals, items, and response categories. This process is most effective in scenarios where these quantities are substantial, meeting expectations in broader contexts. Thus, the GR model is not recommended for estimation in scenarios with 100 individuals, as even with variations in m, bias values for θ remain high, deviating significantly from 0 and becoming statistically inadequate.

Table 3: Root Mean Square Error (RMSE), bias, and average correlation for the following scenarios corresponding to GR model.

Scenarios	RMSE			Bias			Correlation		
	a	b	θ	a	b	θ	a	b	θ
$n = 100; J = 10; m = 3$	0.4848	0.3908	0.4419	0.0537	0.1409	0.1366	0.7556	0.9651	0.8927
$n = 500; J = 10; m = 3$	0.1638	0.1296	0.4380	0.0502	0.0165	0.0110	0.9203	0.9927	0.8973
$n = 1000; J = 10; m = 3$	0.1026	0.1061	0.4362	-0.0042	-0.0019	-0.0112	0.9743	0.9957	0.8930
$n = 100; J = 20; m = 3$	0.3104	0.4214	0.3472	0.0030	0.1597	0.1368	0.8317	0.9456	0.9406
$n = 500; J = 20; m = 3$	0.1433	0.1153	0.3000	0.0523	0.0191	0.0109	0.9459	0.9952	0.9538
$n = 1000; J = 20; m = 3$	0.0972	0.0998	0.3257	-0.0040	-0.0142	-0.0120	0.9796	0.9959	0.9422
$n = 100; J = 10; m = 4$	0.3407	0.3049	0.4326	-0.0189	0.0907	0.1533	0.7755	0.9548	0.9210
$n = 500; J = 10; m = 4$	0.1396	0.1036	0.4022	0.0432	0.0260	0.0096	0.9452	0.9892	0.9155
$n = 1000; J = 10; m = 4$	0.0904	0.0846	0.4063	-0.0047	-0.0097	-0.0099	0.9797	0.9939	0.9085
$n = 100; J = 20; m = 4$	0.2707	0.3308	0.3812	-0.0232	0.0935	0.1814	0.8573	0.9233	0.9428
$n = 500; J = 20; m = 4$	0.1222	0.0963	0.2764	0.0371	0.0277	0.0099	0.9563	0.9924	0.9620
$n = 1000; J = 20; m = 4$	0.0761	0.0915	0.2981	-0.0045	-0.0211	-0.0104	0.9879	0.9922	0.9533
$n = 100; J = 10; m = 6$	0.3115	0.3097	0.4002	-0.0313	0.0350	0.1667	0.8092	0.8619	0.9288
$n = 500; J = 10; m = 6$	0.1217	0.1107	0.3770	0.0234	0.0278	0.0080	0.9539	0.9698	0.9225
$n = 1000; J = 10; m = 6$	0.0886	0.1130	0.3854	-0.0112	-0.0275	-0.0117	0.9813	0.9713	0.9173
$n = 100; J = 20; m = 6$	0.2545	0.4392	0.5073	-0.0333	0.0328	0.2319	0.8600	0.7840	0.9208
$n = 500; J = 20; m = 6$	0.1146	0.1120	0.2451	0.03800	0.0391	0.0092	0.9618	0.9715	0.9673
$n = 1000; J = 20; m = 6$	0.0778	0.1134	0.2744	-0.0108	-0.0304	-0.0097	0.9876	0.9668	0.9594

Source: from the authors (2024).

Reduction of response categories

The following Table 4 presents the results of three statistical measures for each psychometric parameter, examined across various scenarios during the process of reducing response categories in GR model. In these scenarios, response categories were reduced from four to two and from six to three. Notably, when reduced to two categories, 2PL model was used to obtain the values of the studied measures.

The data in Table 4 demonstrates that decreasing response categories compared to the last twelve scenarios in Table 3 results in notably higher RMSE and lower correlations, indicating reduced data fidelity. This effect was pronounced primarily in the estimation of the discrimination parameter (a) and latent variable (θ). However, regarding bias, no clear pattern that could be identified was observed.

Besides that, it is crucial to note that this trend described for the examined statistical measures becomes more noticeable in scenarios with only 100 individuals, as in these cases, due to the smaller population sample, the impact of reduction is greater, further highlighting the effects on estimates and consequently on information loss.

Table 4: Root Mean Square Error (RMSE), bias, and mean correlation for different scenarios with response category reduction in GR model.

Scenarios	RMSE			Bias			Correlation		
	a	b	θ	a	b	θ	a	b	θ
$n = 100; J = 10; m = 2$	0.6784	0.4403	0.5776	0.0006	-0.0948	0.00003	0.4915	0.3467	0.7434
$n = 500; J = 10; m = 2$	0.2375	0.1052	0.5685	-0.0157	0.0054	0.0001	0.8259	0.9646	0.7604
$n = 1000; J = 10; m = 2$	0.1327	0.1316	0.5124	0.0346	0.0359	0.0001	0.9474	0.9646	0.8209
$n = 100; J = 20; m = 2$	0.5664	0.3847	0.4921	0.0618	-0.1371	0.0001	0.6368	0.3198	0.8365
$n = 500; J = 20; m = 2$	0.2058	0.0933	0.4491	-0.0022	0.0195	0.0001	0.8698	0.9610	0.8703
$n = 1000; J = 20; m = 2$	0.1428	0.0947	0.4744	-0.0180	-0.0084	-0.0001	0.9531	0.9577	0.8506
$n = 100; J = 10; m = 3$	0.5645	0.4024	0.5688	0.0263	-0.0102	-0.0001	0.3475	0.8782	0.7702
$n = 500; J = 10; m = 3$	0.1694	0.1209	0.5193	-0.0247	0.0032	-0.00004	0.9105	0.9870	0.8144
$n = 1000; J = 10; m = 3$	0.1266	0.1178	0.5404	0.0629	-0.0017	0.0004	0.9660	0.9897	0.7938
$n = 100; J = 20; m = 3$	0.4543	0.9190	0.4410	-0.0004	-0.0161	-0.0001	0.6251	0.8594	0.8740
$n = 500; J = 20; m = 3$	0.1782	0.1173	0.3877	-0.0169	0.0058	-0.0001	0.8968	0.9874	0.9081
$n = 1000; J = 20; m = 3$	0.1169	0.1353	0.4111	-0.0001	0.0011	0.0004	0.9697	0.9810	0.8941

Source: from the authors (2024).

Exploratory analysis of simulation study

Below, graphs will be presented comparing the values of statistical measures for each parameter of IRT, with or without reduction of response categories in specific scenarios.

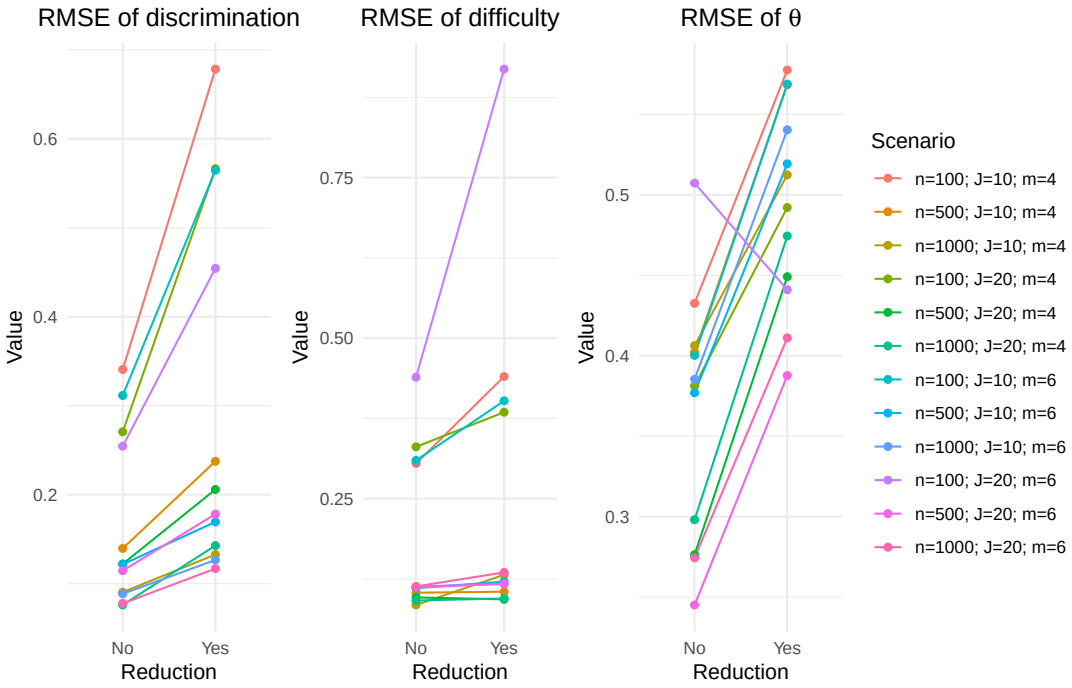
Through the analysis of Figure 1, it is possible to observe a generalized increase in RMSE values for the discrimination parameter when response categories are reduced, highlighting the expected trend of this statistical error increasing with reduction. Furthermore, it is also noted that steeper lines, indicating more pronounced increases in RMSE, correspond to scenarios 01, 04, 07, and 10, which include only 100 individuals (a less representative population sample), a significant factor contributing to this undesirable variation.

Furthermore, there is a notable tendency for RMSE values for difficulty parameters to increase as the number of response categories decreases. This observation suggests that higher values of this psychometric measure are again obtained in scenarios with only 100 individuals, due to the small sample size, statistically confirming the expected result. Specifically, scenario 10 stands out for presenting the highest (worst) RMSE value, showing a particularly pronounced increase from approximately 0.45 to 0.90.

Finally, it is noticeable that, in general, the RMSE results for the latent variable (θ) tend to increase as the number of response categories is reduced. However, scenario 10 stands out as an exception, where there was a decrease in the RMSE value after the reduction, from approximately 0.50 to 0.45. This phenomenon can be attributed to the small sample size (100 individuals), which causes inconsistencies in the generation of values.

When analyzing Figure 2, it can be stated that there was no clear pattern in the estimation of bias values for the discrimination parameter. In some scenarios, such as 03, 04, 06, and 09, there was an increase in values, while in other scenarios there was a decrease with the reduction of response categories. Additionally, some results were relatively far from 0, such as in scenarios 04 and 09, which reached values close to 0.06, indicating a higher degree of bias for this parameter in these cases.

Figure 1: Comparison of RMSE results for the parameters of discrimination (a) and difficulty (b), in addition to the latent variable (θ), with and without reduction of response categories in different scenarios.



Source: from the authors (2024).

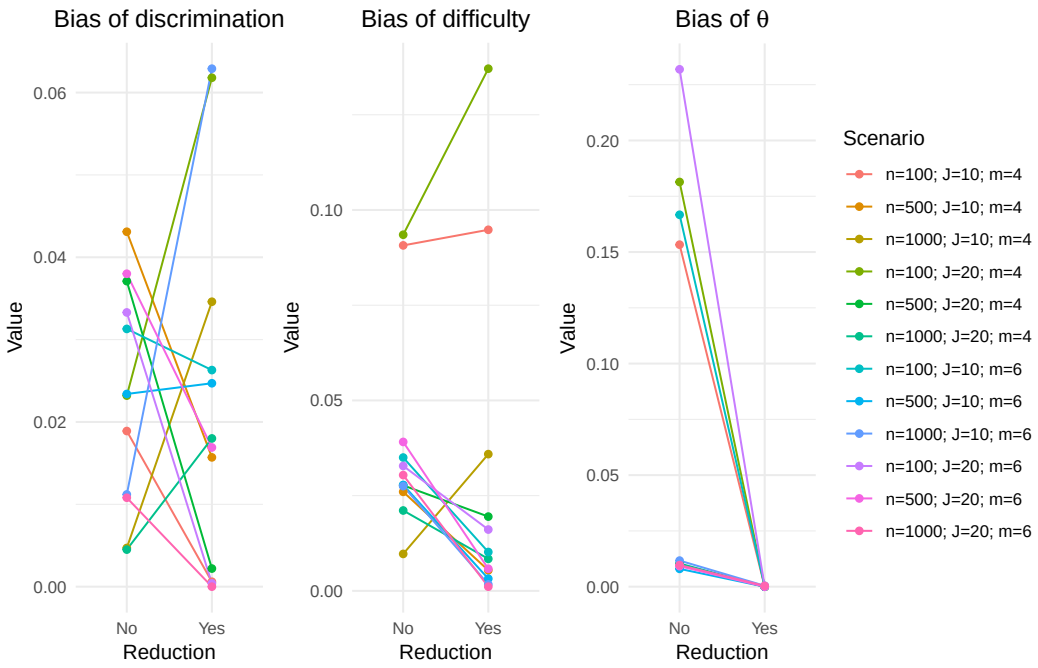
Moreover, there is a clear trend of decreasing bias values for the difficulty parameters as response categories are reduced, deviating from expectations and reaching values close to 0. However, some scenarios, such as 01, 03, and 04, showed a sharp increase in results. Specifically, scenarios 01 and 04 reached extremely high values, around 0.10 and 0.15, respectively. These deviations can be attributed to the small sample size in both cases ($n = 100$), which can lead to biased parameter estimates, yet in this analysis, they align with expectations.

Ultimately, there is a noticeable pattern of decreasing bias values for the latent variable (θ) as response categories are reduced, reaching values very close to 0. Although this is statistically desirable, it deviates from expectations, as according to the applied theory, reducing response categories should result in values farther from 0. Moreover, scenarios 01, 04, 07, and 10, corresponding to those with only 100 individuals, stand out. As previously explained, this results in greater variations shown in the graph due to inconsistencies in the estimation process.

Figure 3 shows decreased correlations for the discrimination parameter across all scenarios due to reduced response categories, as expected. Scenarios 01, 04, 07, and 10 experienced the largest drops, again due to the low sample size ($n = 100$), which intensifies the variations caused by the reduction in categories. Ultimately, it is worth highlighting scenario 07, where the correlation decreased from approximately 0.80 to 0.35, which is undesirable in a statistical context.

Additionally, it is evident that in certain scenarios, such as 01 and 04, the correlation between difficulty parameters decreased as the response categories were reduced, aligning with theoretical expectations. However, it is important to note that in other scenarios, such as 07 and 10, there was a slight increase in correlation after this reduction, making it impossible to establish a standardized behavior for the scenarios.

Figure 2: Comparison between the bias results of the discrimination (a) and difficulty (b) parameters, in addition to the latent variable (θ) depending on the reduction or not of the response categories in different scenarios.



Source: from the authors (2024).

Finally, it is observed that all scenarios followed the expected trend for the latent variable (θ), with the correlation decreasing as the response categories were reduced. In all cases, this reduction in the statistical measure was considerable, with particular emphasis on scenarios 01 and 10, which recorded the largest and smallest variations, respectively, going from approximately 0.925 to 0.750 and from 0.925 to 0.875.

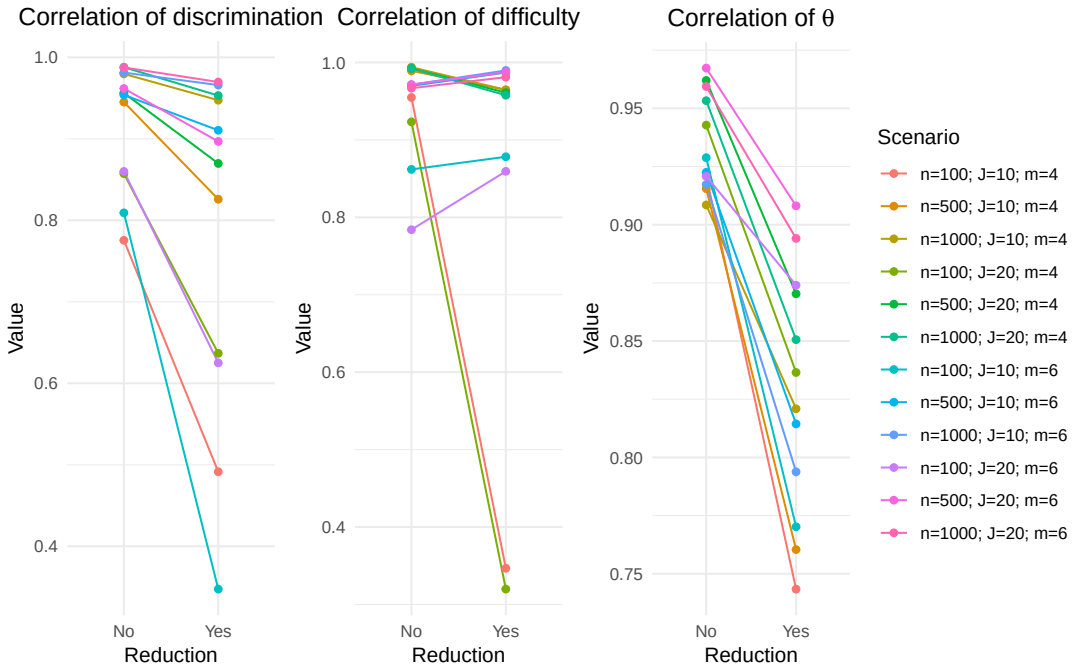
Application with sustainability perception data

The concept of sustainability has received immense attention and priority in various student research projects. This is due to the recent neglect of anthropogenic actions towards environmental conservation. Sustainability involves maintaining a mutual balance between commercial production and the preservation of the natural ecosystem to ensure the smooth functioning of all existing ecological and vital relationships.

Due to the importance of this topic, developing a scale to measure the levels of perception of sustainability in specific population segments becomes essential. Specifically, the approach proposed in Vincenzi et al. (2018) utilizes Item Response Theory (IRT) to estimate unobservable characteristics, providing information about the quality of the items used in the measurement scale of interest.

In this study, a dataset collected from 52 items proposed by Vincenzi et al. (2018) was used to evaluate the perception of sustainability of 2,519 individuals located in the southwestern region of the state of Paraná, in the Paraná III Basin, highlighted in Figure 4. The items were equally divided into four blocks, and by combining two blocks, six different questionnaires were generated, each applied to approximately 420 individuals. Each item has six response categories: *strongly disagree*, *disagree*, *somewhat disagree*, *somewhat agree*, *agree*, and *strongly agree*. The questionnaires were administered in the three cities of the Paraná III Basin with more than 100,000 inhabitants: Cascavel, Foz do Iguaçu, and Toledo, with 852, 833, and 834 respondents, respectively.

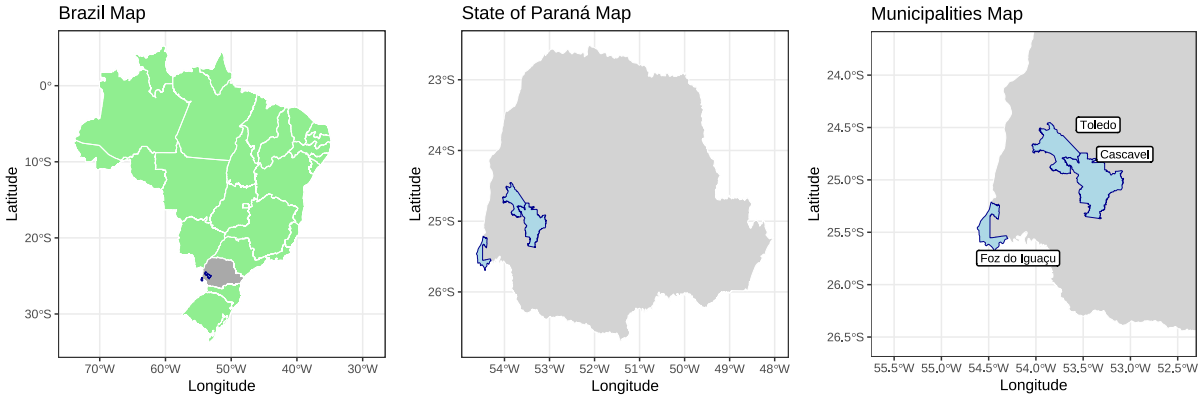
Figure 3: Comparison between the results of the correlation of the discrimination (a) and difficulty (b) parameters, in addition to the latent variable (θ) depending on the reduction or not of the response categories in different scenarios.



Source: from the authors (2024).

From this, it becomes feasible to apply the current study using the sustainability perception data obtained by Vincenzi et al. (2018) and descriptively explored by Bueno et al. (2024), aiming to relate the estimated IRT parameter values with the six response categories of the items and, subsequently, with only three, due to the reduction. This analysis, in turn, allows for the verification of the correlation between the estimates obtained in each of the analyzed scenarios.

Figure 4: Location of the study region.



Source: Vincenzi et al. (2018).

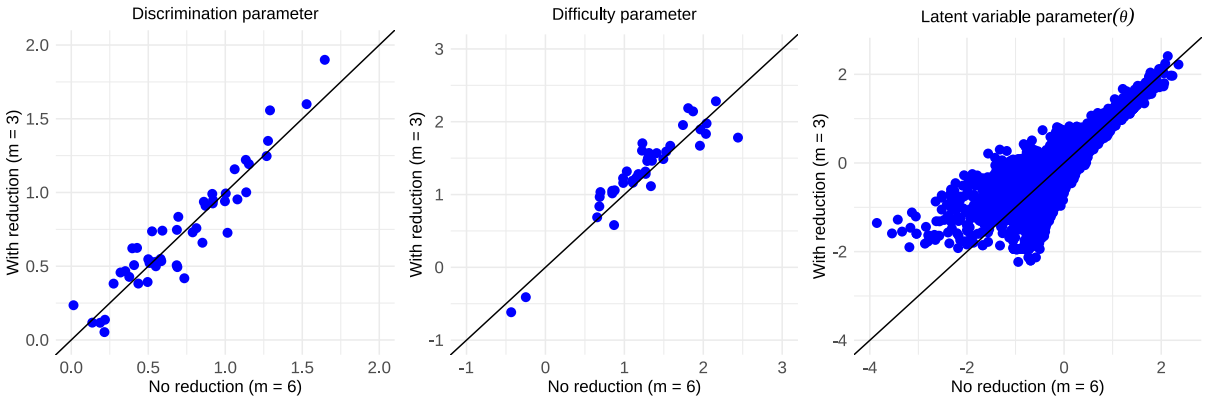
The analysis of Figure 5 reveals that the values related to the item discrimination parameter concerning the number of response categories generally follow the line $y = x$, indicating similarity between the parameter estimates in the two study conditions.

To create the scatter plot associated with the difficulty parameters (b), it was necessary to exclude items with a discrimination parameter lower than 0.25, as these produced completely

disconnected results. By interpreting this graph, it can be generally stated that the values of b in both scenarios were close to the trend line. However, it is also noted that a few items recorded values that deviate from the expected, distancing themselves further from the line $y = x$.

The study of the scatter plot referring to the latent variable (θ) shows that the sustainability perception indices exhibited significant variability for values below 1, while values above 1 resulted in less variation, following the trend of the straight line $y = x$. Additionally, it is noted that the limits of the x-axis, without the reduction in the number of response categories, range from -4 to 2, whereas the y-axis, with the reduction applied, ranges from -2 to 2, having a significantly smaller interval.

Figure 5: Dispersion of values related to the discrimination (a) and difficulty (b) parameters, in addition to the latent variable (θ) with or without reducing the response categories.



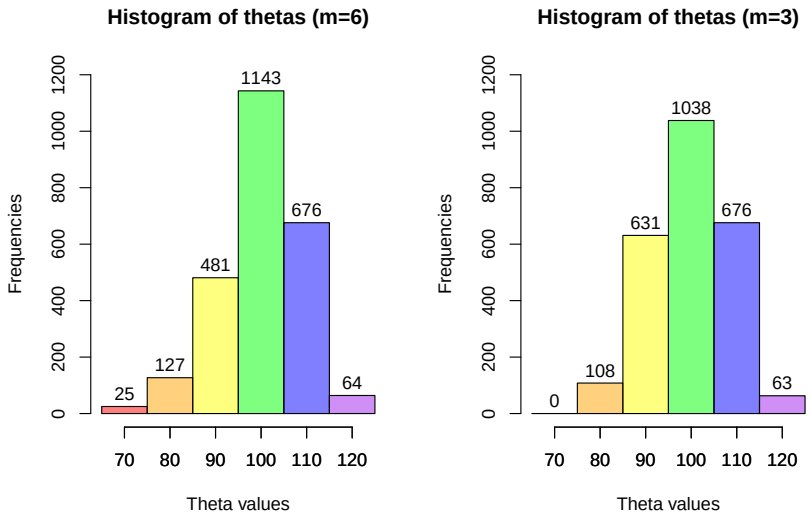
Source: from the authors (2024).

Furthermore, the sustainability perception indices were applied using the scale in Vincenzi et al. (2018), where the original results of θ, which typically range from -3 to 5, were converted to values from 70 to 150 in increments of 10. From this, it was observed that there were no individuals in the range of 130 to 150, allowing for the upper limit of the x-axis to be reduced to 125 in the histograms subsequently developed, in order to facilitate the comparison of latent variable classes obtained with and without the reduction of response categories.

Upon interpreting Figure 6, it is observed that in both histograms, the number of respondents in the last two classes was practically similar. On the other hand, there is a greater discrepancy in the number of participants in the 90 and 100 classes, mainly. This occurred due to the reduction of response categories, which resulted in the translocation of individuals from lower classes to successive ones, due to the decreased discriminative capacity.

Finally, Table 5 was constructed, a two-way table aimed at facilitating the verification of the frequency of individuals in each of the classes corresponding to the latent variable (θ), with and without reduction in response categories. Through its analysis, it is observed that most individuals are located along the highlighted main diagonal, while the table's extremes show an absence of individuals, following the expected trend. However, it is noteworthy that there are some exceptions, such as the intersections between classes $85 \vdash 95$ and $95 \vdash 105$, with elevated values (271 and 158). This indicates that the reduction in response categories is shifting individuals from the expected class to adjacent ones, demonstrating that information loss has altered estimates of sustainability perception.

Figure 6: Scale of values related to the latent variable (θ) distributed by frequency of individuals.



Source: from the authors (2024).

Overall, based on the analysis of the obtained results, it is inferred that changes occurred in the parameter values across all scenarios due to the application of the reduction technique. However, in some cases, these changes were more pronounced than in others, indicating a loss of information. Additionally, regarding the application of reduction discussed in the aforementioned article, it was observed that the discrimination (a) and difficulty (b) parameters followed the trend of the reference line $y = x$, while the latent variable (θ) deviated from expectations for values below 1. Finally, through the interpretation of histograms and the contingency table, a clear shift in the number of respondents within central categories was noted, primarily demonstrating that the applied technique caused a relocation of individuals from their original category to an adjacent one due to the loss of information.

Table 5: A crossing between the classes related to the latent variable (θ) scale with and without the reduction of response categories.

No reduction	With reduction						Total
	$65 \vdash 75$	$75 \vdash 85$	$85 \vdash 95$	$95 \vdash 105$	$105 \vdash 115$	$115 \vdash 125$	
$65 \vdash 75$	**0**	12	13	0	0	0	25
$75 \vdash 85$	0	**20**	93	14	0	0	127
$85 \vdash 95$	0	68	**254**	158	1	0	481
$95 \vdash 105$	0	8	271	**788**	76	0	1143
$105 \vdash 115$	0	0	0	78	**583**	15	676
$115 \vdash 125$	0	0	0	0	16	**48**	64
Total	0	108	631	1038	676	63	2516

Source: from the authors (2024).

Amid growing global concerns about climate change, studies such as this one, which analyze data related to sustainability, become indispensable for supporting analyses and guiding future decisions. This article, in particular, highlights the risks associated with reducing response categories, emphasizing their implications for the interpretation of research results. Furthermore, this study is directly aligned with the 17 Sustainable Development Goals (SDGs) established by the United Nations (UN) in 2015, with a particular focus on Goal 11: Sustainable Cities

and Communities. In this sense, the research contributes significantly to achieving targets 11 and 11.a, which aim to make cities more sustainable and to foster positive economic, social, and environmental relationships between different areas, respectively.

Beyond its application to sustainability perception data, this study demonstrates considerable potential for use in various other fields, provided that questionnaires containing polytomous items are available. An example is presented by Moreira et al. (2024), who explored pig producers' perceptions regarding the use of technologies aimed at animal welfare in pig farming. This research was conducted through questionnaires administered to producers in different cities across Brazil, addressing aspects such as property infrastructure, technology adoption, knowledge about animal welfare, and interaction with academic institutions.

Therefore, it can be asserted that reducing response categories, which effectively decreases data density, leads to a loss of information. Nevertheless, it is important to investigate whether there are circumstances under which applying this technique might be advantageous. This possibility, in turn, could inspire the development of new studies aimed at creating a metric capable of quantifying this impact and establishing thresholds that facilitate result interpretation, thereby assisting in the final decision-making process.

Final remarks

In this paper, we conducted a simulation study to determine the potential consequences of reducing response categories in the parameter estimation process via IRT models. We analyzed RMSE, bias, and correlation values obtained in different scenarios for each psychometric parameter through tables. Additionally, by reducing response categories in the simulation study, we were able to perform an exploratory analysis using graphs comparing the different statistical measures for each IRT parameter across scenarios, with and without reduction. Finally, we applied this methodology to sustainability perception data from the article Vincenzi et al. (2018), aiming to investigate the effects of reduction on IRT parameters in real-world data.

Therefore, concerning human resource development, this work provided means for the scientific and investigative development of the student, stimulating their potential through new theoretical and practical knowledge in the fields of statistics and psychometrics, as well as enabling the development of new studies related to this theme.

References

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C. Teoria da Resposta ao Item: conceitos e aplicações. *ABE, São Paulo*, 2000.

BIRNBAUM, A. Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*, 1968.

BUENO, R. B. et al. Analysis of environmental sustainability perception in Paraná Basin III residents via Item Response Theory. *Brazilian Journal of Biometrics*, v. 42, n. 2, p. 158-170, 2024.

CHALMERS, R. P. mirt: A multidimensional item response theory package for the R environment. *Journal of statistical Software*, v. 48, p. 1-29, 2012.

MOREIRA, M.d.R et al. The Perception of Brazilian Livestock Regarding the Use of Precision Livestock Farming for Animal Welfare. *Agriculture*, v. 14, n. 8, p. 1-17, 2024.

DA SILVA, M. A. et al. Estimating the DINA model parameters using the No-U-Turn Sampler. *Biometrical Journal*, v. 60, n. 2, p. 352-368, 2018.

EHRLICH, P. J. Pesquisa operacional: curso introdutório. *Atlas*, 1991.

FRAGOSO, T. M.; CÚRI, M. Improving psychometric assessment of the Beck Depression Inventory using multidimensional item response theory. *Biometrical Journal*, v. 55, n. 4, p. 527-540, 2013.

LORD, Frederic. A theory of test scores. *Psychometric monographs*, 1952.

PEREIRA, R. H. M. et al. geobr: download official spatial data sets of Brazil. R package version 1.6.1, 2021.

R CORE TEAM. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vienna, Austria, 2023.

RAO, C. R.; SINHARAY, S. (Ed.). *Psychometrics. Elsevier*, 2007.

SAMEJIMA, F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*, 1969.

TEAM, RStudio. *RStudio: integrated development for R 2022.* RStudio, PBC: Boston, MA, USA, 2022.

THISSEN, D.; STEINBERG, L. A taxonomy of item response models. *Psychometrika*, v. 51, n. 4, p. 567-577, 1986.

VINCENZI, S. L. et al. Assessment of environmental sustainability perception through item response theory: A case study in Brazil. *Journal of Cleaner Production*, v. 170, p. 1369-1386, 2018.

WICKHAM, H. *et al.* Welcome to the Tidyverse. *Journal of open source software*, v. 4, n. 43, p. 1686, 2019.