STATISTICAL COMPETENCIES IN THE TRAINING OF MATHEMATICS TEACHERS IN 2017 ENADE: AN IRT APPLICATION

Sandra Cristina Martini Rostirola, Elisa Henning, and Ivanete Zuchi Siple
Santa Catarina State University, Brazil
sandra.rostirola@ifc.edu.br

The National Students' Performance Examination [Exame Nacional de Desempenho dos Estudantes–Enade] is a large-scale assessment instrument for Brazilian higher education programs whose results, along with other quality indicators, allow for an evaluation of the country's higher education. This work analyzed 2017 Enade questions for the mathematics degree program regarding the statistical competencies of future educators. The quantitative approach methodology allowed analyzing 10,869 participants' responses through Item Response Theory (IRT) using the Three-Parameter Logistic Model. The results indicate evidence of weakness in questions related to probability and statistics regarding t levels of difficulty and discrimination in addition to reflecting discrepancies between the statistical content in the official test descriptors and those found in the questions.

INTRODUCTION

The National Students' Performance Examination [Exame Nacional de Desempenho dos Estudantes–Enade] is one of the assessment instruments that allows measuring the quality of Brazilian higher education, both for traditional classrooms and for distance education. Enade evaluates the performance of students in the last period of the undergraduate programs regarding the syllabus defined in each program's curriculum and the development of competencies and skills required for a deeper general and professional training. (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [INEP], 2022). This evaluation has been carried out annually since 2004, and the included educational areas change each year in a rotation system. In 2017, it included the areas of Engineering, Information Technology, and Teaching Degrees, including the degree in mathematics. The mathematics degree program aims to help the development of mathematics educators' competencies, including curriculum, evaluation, methodological aspects, and the selection and production of teaching materials.

Pedagogical knowledge, ability to solve problems, and ability to establish mathematical models and conjectures for content areas such as statistics are included among the teaching competencies. The latter is the main interest in this study because understanding of statistics structure and concepts allows one to understand and be part of the social environment in which he or she is inserted, with information interpretation, data analysis, and decision making providing an opportunity to experience citizenship. To measure the future mathematics teacher's competencies, Enade is based on a competency matrix that guides the skills to be assessed through the examination.

This study analyzed the 2017 Enade's questions for the mathematics degree program and future educators' statistical competencies, measuring the quality of the test and of each item (question). The 2017 Enade examination consisted of 40 questions, five of which were essays and 35 of which were multiple-choice questions with five alternatives—one correct answer and four incorrect distractors. From that set of questions, two—items 20 and 21—comprise this study's analysis base because they are related to statistics. This work aims to assess the statistical competencies of the mathematics degree program's students who took the 2017 Enade. It should be noted that the mathematical model employed in Enade is Classical Test Theory (CTT). However, for the purpose of a wider scope, the items were analyzed using Item Response Theory (IRT).

METHODS

The adopted methodology is the quantitative approach from a documental perspective using Enade's 2017 test for the mathematics teaching degree program to analyze two questions (items) related to the contents of probability and statistics, as described by the Reference Curriculum in the Summary Report (INEP, 2018). The test consisted of 40 questions, five essays and 35 multiple-choice questions. The data regarding the answer pattern from 10,869 participants were collected through microdata from INEP. All participants who answered at least one item from the multiple-choice questions were included.

Classical Test Theory (CTT), probability models of Item Response Theory (IRT), and the Nominal Response Model (NRM) were applied to the response pattern set of the 35 objective items.

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org ©2022 ISI/IASE

Data processing was aided by the RStudio interface for the statistical software program R (R Core Team, 2021), with special mention for packages mirt (Chalmers, 2012) and psych (Revelle, 2021) using the maximum marginal likelihood method.

An exploratory analysis of the items through difficulty and discrimination indices by Point Biserial Correlation Coefficient (BCC) was performed for the CTT. The IRT models were applied from a three-parameter logistic model—M3LP—with estimates of (a) discrimination, (b) difficulty, and (c) pseudo-guessing, as well as measurement of the latent trait with a higher-than-60% probability of correct answer in the computational scale (0, 1).

Later, an analysis was performed through CTT and parameters estimated through NRM, acquiring measurements related to the probability of a participant choosing a distractor or the correct answer, besides enabling the test's edumetric analysis that contributes to the improvement of curricular and pedagogical terms.

DATA ANALYSIS

As previously explained, the 2017 Enade evaluation system proposes a matrix of competencies associated with certain contents, with the definition of some general skills (the same for all programs) and some specifically for the teaching of mathematics. In the case of the specified items, for measuring purposes, the ability in probability and statistics was defined as the latent trait (θ), with no damage to the unidimensionality and local independence of each item. With the parametrization of the 35 objective items, we obtained the parameters of (a) discrimination, (b) difficulty, and (c) pseudo-guessing related to the IRT, accompanied by their respective Standard Error (SE), and complemented by the Difficulty Index (PI) and the Discrimination Index by Point-Biserial Correlation Coefficient (BCC) proposed by the CTT, which are shown in Table 1.

Table 1. Parameters estimated through IRT and CTT for items 20 and 21

	IRT Parameters				
Item	a (SE)	b (SE)	c (SE)	PI	BCC
20	3.6 (0.342)	1.85(0.039)	0.20 (0.005)	0.21	0.34
21	0.78 (0.57)	10.68(17.532)	0.29(0.005)	0.09	0.12

Baker (2001) suggests the values of 0.7 (moderate discrimination) to 1.7 (strong discrimination) for the discrimination parameter. Hence, item 20 has a very high discrimination parameter (3.6). Considering the average for parameter (a) for the 35 items (M = 1.57 and SD = 1.09) and lower and higher values of 0.31 and 4.3, respectively, we see that item 20 was one of those with the highest discrimination in the test. Item 21's discrimination parameter, however, is considered moderate (0.78).

For parameter (b), which measures the test difficulty, Andrade, Tavares, and Valle (2000) indicate values between -2 and +2 as good estimates because those are considered typical. The lowest and highest values ranged from -1.08 to 10.68 (M = 1.8 and SD = 2.13). Item 20 is within the theoretical interval and can be classified as difficult. However, item 21 has a very high difficulty parameter, which makes it the most difficult question on the test. Parameter (c), which indicates the probability of a correct answer by chance (pseudo-guessing), is within the expected values for five-alternative questions, according to what Nojosa (2001) defined, with values in the range from 0.1 to 0.3 considered acceptable.

Item 20 has high-discrimination and average-difficulty characteristics, fulfilling the measurement objective. Item 21, however, has values that classify it as an inadequate-quality question. Figures 1(a) and 1(b) show the logistical curve of items 20 and 21, evidencing the weakness mentioned regarding item 21, which nearly does not have a discrimination angle.

To complement the CTT, the difficulty index was used, which is calculated with the proportion of correct answers. According to that parameter, item 20 is considered difficult, and item 21 is considered very difficult (Vilarinhos, 2015). The discrimination index shows that an item is able to set apart respondents with high and low scores. According to the classification by Ebel (1954), cited by Piton-Gonçalves and Almeida (2018), both items need to be revised in that regard. The point-biserial correlation for item 21 (0.12) caused it to be removed from the final grade computation because of the system adopted for Enade. Both items were analyzed to assess the proficiency level (that is, each individual's aptitude or latent trait) required for an answer in a scale (0, 1), which are indicated in Table 2 along with each item's discrimination and correct answer probability.

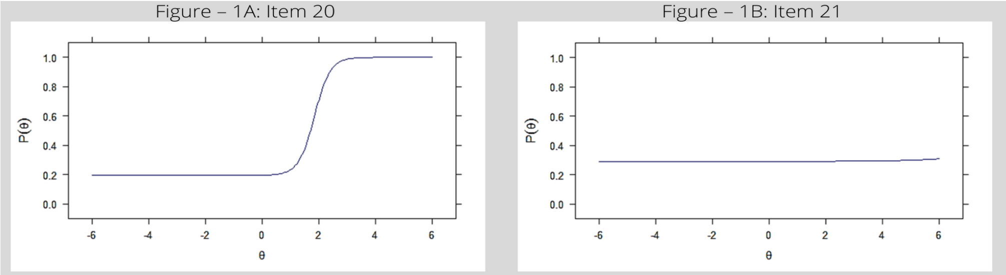


Figure 1. (a) Item 20 and (b) Item 21

So, for item 20, the participant must have a 2.0 skill for a higher-than-60% chance of selecting the correct answer. Item 21 has no latent trait compatible with the participants at the scale (0,1). As a whole, the test measures latent traits from -1 to 3.8 more effectively, as Figure 2 shows.

Table 2. Skill scale (0,1)

Items	Skill Scale, range										
	a	b	-4.0	-3.0	-2.0	-1.0	0.0	1.0	2.0	3.0	4.0
20	3.6	1.85	0.00	0.00	0.00	0.00	0.00	0.05	0.63	0.98	1.00
21	0.78	10.69	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01

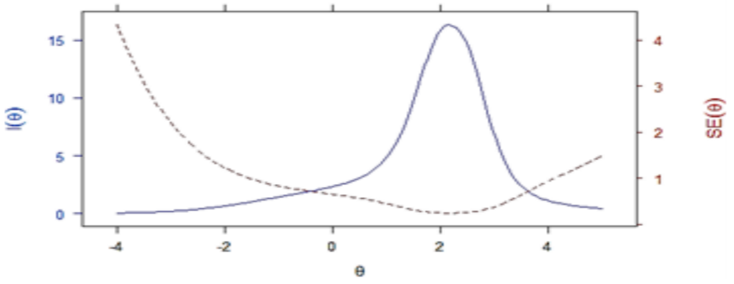
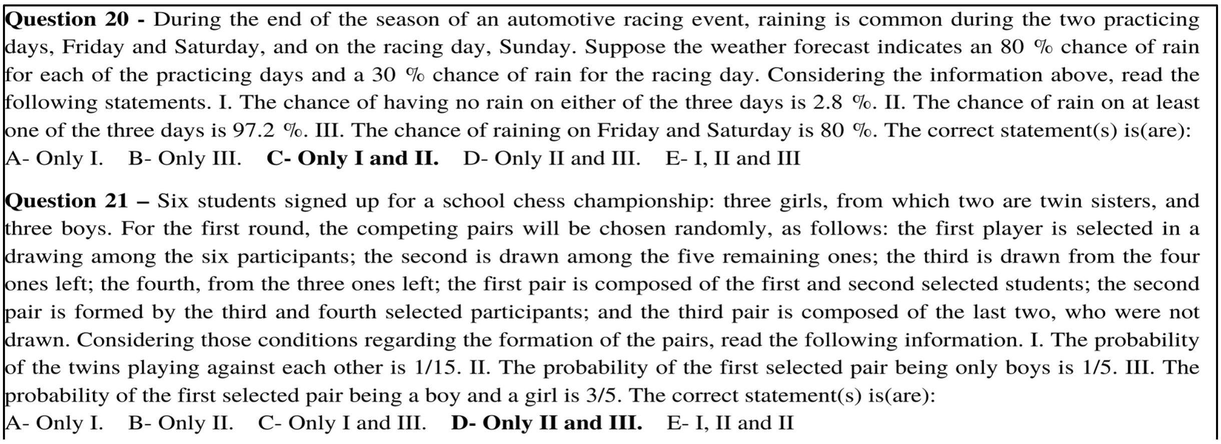


Figure 2. Test information curve

Analyzing the distractors for items 20 and 21 (Figure 3), we see that, for item 20, the biserial coefficient holds positive for the correct answer and, for the NRM, the distractor parameters show a reduced probability of being chosen in relation to the correct answer. Item 21, however, has distractors B and E with parameters 0.09 and 0.13, whereas the correct answer is 0.004. When answering the questions, the first calculations lead to three fractions equal to 1/15. If the respondent does not interpret the need to add those fractions, option E is mistakenly chosen. Such discrepancy can be observed in the proportion of 30 % of participants choosing it. Still, when analyzed through the CTT, this item presented negative biserials for all alternatives, with the highest corresponding to the correct answer (-0.01), too close to distractors B and E (-0.07 and -0.05, respectively). All this information allows for the conclusion that item 21 should have been edited and adjusted. Considering those aspects, item 20 is considered of moderate capacity to measure the skill in probability and statistics latent trait, whereas item 21 does not allow for a precise measurement, be it through an IRT, CTT, or even an NRM approach.

CONCLUSIONS

According to the evaluation of parameters through IRT and CTT, item 20 presents an average quality. Item 21 has evidence of fragility such as the high difficulty parameter (b) and associated standard error. Moreover, its Point Biserial Correlation Coefficient is too low, which classifies it as inefficient.



Question 20 - During the end of the season of an automotive racing event, raining is common during the two practicing days, Friday and Saturday, and on the racing day, Sunday. Suppose the weather forecast indicates an 80 % chance of rain for each of the practicing days and a 30 % chance of rain for the racing day. Considering the information above, read the following statements. I. The chance of having no rain on either of the three days is 2.8 %. II. The chance of rain on at least one of the three days is 97.2 %. III. The chance of raining on Friday and Saturday is 80 %. The correct statement(s) is(are):
A- Only I. B- Only III. **C- Only I and II.** D- Only II and III. E- I, II and III

Question 21 – Six students signed up for a school chess championship: three girls, from which two are twin sisters, and three boys. For the first round, the competing pairs will be chosen randomly, as follows: the first player is selected in a drawing among the six participants; the second is drawn among the five remaining ones; the third is drawn from the four ones left; the fourth, from the three ones left; the first pair is composed of the first and second selected students; the second pair is formed by the third and fourth selected participants; and the third pair is composed of the last two, who were not drawn. Considering those conditions regarding the formation of the pairs, read the following information. I. The probability of the twins playing against each other is 1/15. II. The probability of the first selected pair being only boys is 1/5. III. The probability of the first selected pair being a boy and a girl is 3/5. The correct statement(s) is(are):
A- Only I. B- Only II. C- Only I and III. **D- Only II and III.** E- I, II and II

Figure 3. Item 20 and Item 21

For that reason, that item was not used by INEP when calculating final scores. Thus, considering that those two items aimed at evaluating the skills in probability and statistics, the evaluation is harmed by the removal of one item. Items 1 and 10 also included statistical contents, but they were not indicated in the competencies matrix used when the test was developed.

The data indicate that studies involving the assessment of skills related to statistics are justified because the institutionalized test presents evidence that it needs improvement in the scope of assessment and deeper edumetric analysis for more precise and accurate skill measurement. Statistics are indispensable in the teaching degree program because they give the opportunity to develop competencies for future professional experiences and for the development of the citizen.

REFERENCES

Andrade, D. F., Tavares, H. R., & Valle, R. C. (2000). *Teoria da resposta ao item: Conceitos e aplicações*. Associação Brasileira de Estatística.

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). Eric Clearinghouse on Assessment and Evaluation.

Chalmers, R., P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2018). *Relatório síntese de área: Matemática*. Ministério da Educação. Retrieved July 16, 2021, from https://download.inep.gov.br/educacao_superior/Enade/relatorio_sintese/2017/Matematica.pdf

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2022). *Exame nacional de desempenho dos estudantes (Enade).* Ministério da Educação. Retrieved September 16, 2021, from https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exameseducacionais/enade

Nojosa, R T. (2001). *Modelos multidimensionais para a teoria de resposta ao item*. [Dissertação de Mestrado, Universidade Federal de Pernambuco]. Repositório da Universidade Federal de Pernambuco. https://repositorio.bc.ufg.br/tede/handle/tede/4058

Piton-Gonçalves, J., & Almeida, A. M. (2018). Análise da dificuldade e da discriminação de itens de matemática do ENEM. *REMAT: Revista Eletrônica da Matemática, 4*(2), 38–53. https://doi.org/10.35819/remat2018v4i2id3060

R Core Team. (2021). *R: A language and environment for statistical computing* (Version 4.1) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W. (2021). *psych: Procedures for personality and psychological research* (Version 2.1.9) [Computer software]. Northwestern University. https://CRAN.R-project.org/package=psych Version = 2.1.9

Vilarinho, A. P. L. (2015*). Uma proposta de análise de desempenho dos estudantes e de valorização da primeira fase da OBMEP*. [Dissertação de Mestrado, Universidade de Brasília]. Repositório da Universidade de Brasília. https://repositorio.unb.br/handle/10482/19335