BRIDGING THE GAP: METHODOLOGICAL VALIDATION OF A RASCH-BASED DIAGNOSTIC TOOL FOR MATH KNOWLEDGE

PONTE PARA O ENSINO SUPERIOR: VALIDAÇÃO METODOLÓGICA DE INSTRUMENTOS DIAGNÓSTICOS BASEADA EM RASCH PARA O CONHECIMENTO MATEMÁTICO

Henrique Antonio Mendonça Faria
https://orcid.org/0000-0001-6976-6897
São Paulo State University, Unesp
Araraquara, Brazil
henrique.faria@unesp.br

Jorge Manuel Vieira Capela
https://orcid.org/0000-0002-0800-655X
São Paulo State University, Unesp
Araraquara, Brazil
jorge.capela@unesp.br

Marisa Veiga Capela
https://orcid.org/0000-0002-5408-6896
São Paulo State University, Unesp
Araraquara, Brazil
marisa.capela@unesp.br

Fabio Roberto Chavarette
https://orcid.org/0000-0002-1203-7586
São Paulo State University, Unesp
Araraquara, Brazil
fabio.chavarette@unesp.br

Abstract. Many students transitioning from secondary school to higher education face significant challenges with foundational mathematics, impacting academic success. This study investigates the one-parameter Rasch model within Item Response Theory (IRT) as a diagnostic tool to assess mathematical skills and identify weaknesses in this population. The aim of this paper is twofold: first, to conduct a literature review on standardized assessment instruments; second, to exemplify how a Rasch-based diagnostic instrument can assess mathematical proficiency. The literature review examined psychometrics and IRT, highlighting the Rasch model's potential for analysing student responses. Microdata from the 2023 Brazilian National High School Exam (ENEM) for 160 students were used. Student responses to 45 mathematics items were analysed using the Rasch model implemented in R. The analysis included estimating item difficulty, visualising the distribution of student abilities, and comparing Rasch-generated scores with official ENEM scores derived from a three-parameter IRT model. The results suggest that the Rasch model provides valuable insights into mathematical proficiency, even with smaller sample sizes. These findings have implications for developing targeted interventions and support programmes, as identifying weaknesses allows institutions to improve support programme effectiveness and enhance student success and retention.

Keywords: Education; Assessment; Mathematics; Rasch; IRT.

1. INTRODUCTION

Standardized tests are assessment instruments used to measure individuals' skills, knowledge, aptitudes, and other attributes in a uniform and consistent manner. They are designed to ensure that all examinees are evaluated under similar conditions, using the same scoring and interpretation criteria. These tests are widely applied in educational, professional, and psychological contexts, providing an objective means of comparing performance across different individuals or groups. The development, administration, and evaluation of test results fall within a broad area of psychology known as Psychometrics (Gregory, 2010). Tests play a crucial role in assessing and understanding individuals' cognitive, behavioural, and emotional capacities. They are developed with scientific rigour to ensure that their applications are valid, reliable, and fair for the entire evaluated population.

In education, standardized tests provide a uniform measure to assess students' academic performance. They enable comparisons of results among different student groups, schools, and even educational systems, promoting equity and fairness in educational assessment. According to Urbina (2004), these tests serve various purposes, including: (a) Learning assessment – They help determine how much students have learned in relation to the established curriculum, identifying areas of strength and weakness. (b) Educational decision-

making – The results of these tests are often used to make important decisions, such as grade promotion, resource allocation, and the development of school improvement programs. (c) Accountability – They allow schools and teachers to be held accountable for students' performance, encouraging continuous improvement in the quality of education. (d) Admissions and certifications – They are essential in many university admission processes and professional certification programs.

One of the main advantages of standardized tests is their objectivity. They are administered and scored uniform, minimizing the influence of subjective biases. This ensures that the results remain consistent regardless of who administers the test (Urbina, 2004). Another key characteristic is that standardized tests allow for the comparison of results across different individuals or groups over time and across various locations. This is essential for longitudinal research and comparative studies among diverse populations. In the educational context, standardized tests help identify students' individual needs, contributing to the development of personalized pedagogical strategies. They are also used to guide decisions regarding career choices and personal development.

The history of standardized psychological testing dates back to the late 19th and early 20th centuries, marking a significant evolution in assessment methodologies. The historical milestones of this evolution were outlined by Gregory (2010). Francis Galton (1822–1911) was a pioneer in the study of individual differences. He was among the first to develop a battery of tests measuring sensory and motor abilities, laying the foundation for quantifying human capabilities. Wilhelm Wundt (1832–1920), recognized for founding the first psychology laboratory in 1879, used experimental methods to study mental processes, emphasizing objective measurement and the reproducibility of experiments.

Influenced by Galton, James McKeen Cattell (1860–1944) introduced the concept of "mental tests" and highlighted the importance of experimentation and measurement in psychology. His work set the modern agenda for the development of psychological testing. In this research tradition, the intelligence tests developed by Alfred Binet and Théodore Simon in 1905 were the first to systematically assess intelligence. In the educational field, Edward Lee Thorndike and Joseph Rice made significant contributions by developing methods to measure learning efficiency in schools, creating scales that allowed for comparisons of students' performance in specific skills. Thorndike, in 1910, published his writing scale, which was followed by standardized tests designed to assess arithmetic, reading, and spelling skills.

Lewis Terman's revision in 1916 led to the Stanford-Binet test, widely used to measure intelligence in the United States. By the 1920s, objective examinations based on the Army Alpha, a military test, began to be used alongside high school grades for college and university admissions decisions, ultimately leading to the creation of the Scholastic Aptitude Test (SAT).

In this context, the following research question arises: to what extent can a diagnostic tool based on the Rasch model provide valid measures of mathematical knowledge for students transitioning to higher education? The main objective of this study is twofold: first, to conduct a literature review on assessment instruments; second, to exemplify how a Rasch-based diagnostic instrument can assess mathematical proficiency. The one-parameter logistic model was applied to the responses of mathematics items in a sample from the 2023 ENEM examination. The resulting difficulty and ability estimates were compared with those derived from the official three-parameter IRT model. The study demonstrates that identifying item difficulty patterns is relevant so that targeted pedagogical interventions can be adopted in university-level mathematics courses.

2. METHODOLOGY

2.1 SYSTEMATIC LITERATURE REVIEW

This research employed a systematic literature review (SLR) methodology to analyse the scientific literature on the use of standardized mathematics skills tests in selection processes, focusing on their impacts on equity and effectiveness. The review process was guided by the principles outlined in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Page et al., 2021) to ensure a rigorous and transparent approach. Studies were grouped according to specific sets of keywords, resulting in 5 distinct searches. For inclusion in the review, articles had to: (1) be published between 2000 and 2025; (2) be available in peer-reviewed journals; (3) address topics related to mathematics in an empirical context. Exclusion criteria were applied after the initial grouping of results. Articles that did not meet the specific filters (empirical; peer-reviewed; mathematics) were discarded. Furthermore, following the reading of abstracts, the most relevant articles from the initial grouped set for each search descriptor were selected.

Two electronic databases were utilised for conducting the searches: Scopus and Web of Science. After an initial comparative analysis, it was observed that Web of Science contained the majority of articles found in Scopus, as well as additional relevant articles. For this reason, Web of Science was adopted as the primary source for the review. The final search in both databases was conducted on 28 April 2026. Search strategies were developed using specific sets of keywords. Five distinct searches were performed, using different combinations of terms: Search 1: mathematics and skill and logical and reasoning and tests; Search 2: multiple-choice and tests and higher and education and empirical and analysis; Search 3: large-scale and tests and higher and education and mathematics; Search 4: psychometric and theories and item and response and theory and mathematics; Search 5: Rasch and model and item and response and theory and mathematics. For all searches, the following filters were applied: publications between 2000 and 2025 and only articles published in journals. This strategy resulted in an initial total of 58 articles in Scopus and 142 articles in Web of Science.

The article selection process involved reading the abstracts of all articles identified in the initial searches. The main criterion for selection was the relevance of the topic addressed in the article in relation to the objectives of the review. This assessment allowed for the identification of the most pertinent articles among the initial results. After applying the specific filters (empirical; peer-reviewed; mathematics) and relevance assessment, 85 articles from Web of Science considered the most relevant from the initial set were selected, as show in the follow chart of Figure 1. These articles were read in full and comprise the references of the review.

Data were collected regarding the following aspects of the included studies: methodological approaches used in mathematical tests; applied psychometric theories, with emphasis on Item Response Theory and the Rasch Model; empirical results related to large-scale tests in higher education; analyses of multiple-choice tests in educational contexts; assessments of mathematical abilities and logical reasoning. This systematic approach allowed for a comprehensive review of the relevant literature, providing a solid foundation for subsequent analysis. To complement the literature review, secondary data available from the repository of the National Institute of Educational Studies and Research Anísio Teixeira (INEP, 2024) were utilised. These data were incorporated into the analysis of a group of 160 students and results from the National High School Examination (ENEM). The Rasch model in Item Response Theory was applied to this dataset. The results presented in graphs demonstrate the validity of the model for the analysis of reduced datasets, typical of incoming

classes in undergraduate courses in Science, Technology, Engineering and Mathematics (STEM) fields.

The analysis of the collected data was conducted qualitatively, using thematic analysis techniques. Specifically, the thematic analysis was guided by the principles of trustworthiness outlined by Nowell et al. (2017), focusing on credibility, transferability, dependability, and confirmability. It is important to note that this research did not involve the collection of primary data from human subjects. Therefore, there was no need for approval by an ethics committee.

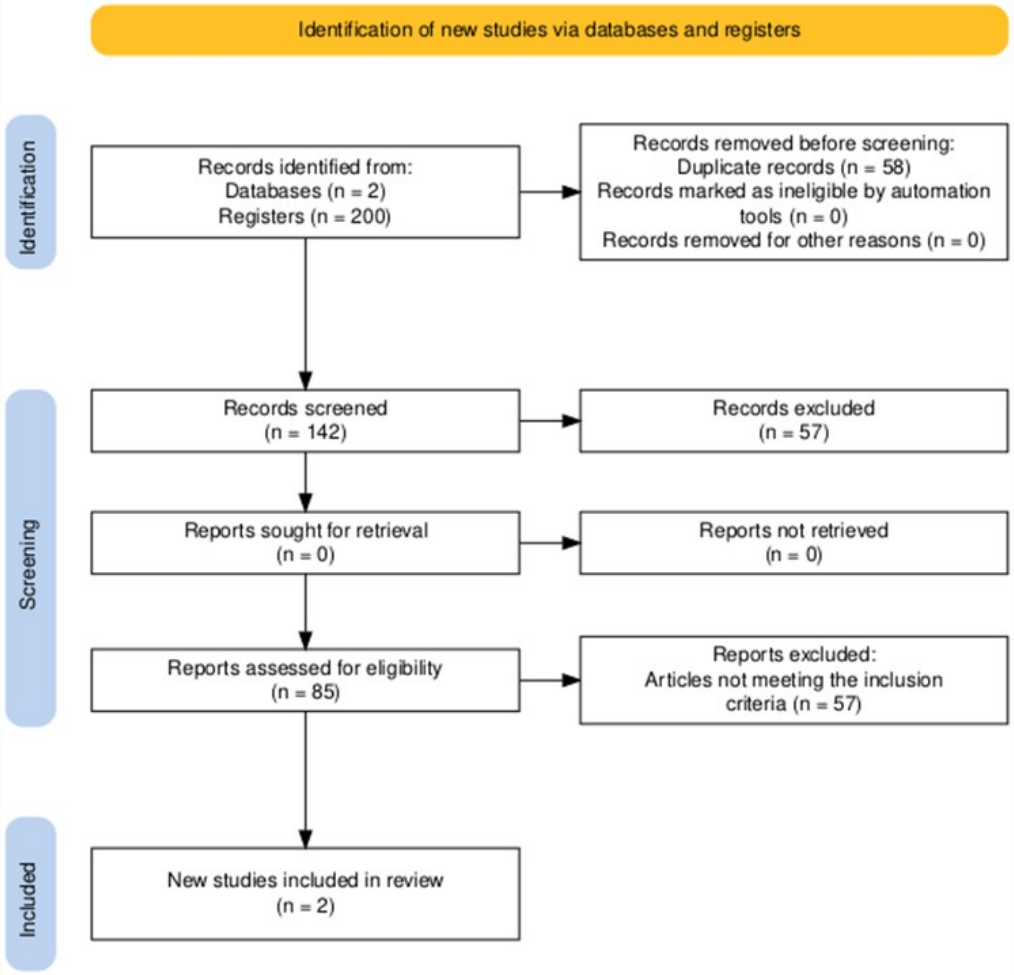

Figure 1. PRISMA graphic. Constructed by author used online PRISMA Flow Diagram in R package. Source: Haddaway et al. (2020).

2.2 RASCH MODEL APPLICATION PROCEDURE

This study adopts a quantitative psychometric paradigm, grounded in the mathematical framework of Item Response Theory (IRT), specifically the one-parameter logistic (Rasch) model (Rasch, 1960). The Rasch model defines the probability of a correct response to an item as a logistic function of the difference between the respondent's ability and the item's difficulty, thereby producing conjoint measurement on a common logit scale. This approach was chosen for its property of specific objectivity, which ensures that item and person parameters can be estimated independently of the particular sample and the specific set of items employed, respectively (Bond & Fox, 2015).

The analytical procedure comprised two main stages. First, item difficulty parameters were estimated via conditional maximum likelihood estimation (CMLE), using the response

matrix of 160 examinees who participated in the 2023 ENEM mathematics examination. Item fit was evaluated through infit and outfit mean-square statistics, with acceptable thresholds set between 0.7 and 1.3 logits, following established guidelines in the Rasch measurement literature (Wright & Stone, 1979; Linacre, 2022). Second, person ability estimates were obtained through maximum likelihood estimation (MLE) and calibrated onto the same logit scale as the item difficulties, allowing for direct comparison between examinee proficiency and item demands.

All estimation procedures were performed using the Winsteps® software package (Linacre, 2022). Model-data fit was assessed at both the item and person levels, and the internal consistency of the instrument was evaluated using the person separation reliability index (Andrich, 1978). The resulting item-person map (Wright map) was examined to evaluate the alignment between the distribution of item difficulties and the range of student abilities, providing substantive evidence for the instrument's diagnostic capacity within the context of higher education transition (Boone, Staver, & Yale, 2014).

3. SYSTEMATIC LITERATURE REVIEW

3.1 MATHEMATICS SKILLS AND LOGICAL REASONING TESTS

Mathematics skills and logical reasoning tests are essential for evaluating individuals' ability to solve quantitative problems and apply logical thinking in various situations (Cresswell & Speelman, 2020). These tests are used in educational and professional contexts to measure proficiency in mathematics and the potential of students or candidates. The tests focus on assessing knowledge and competencies in areas such as arithmetic, algebra, geometry, statistics, and calculus (Tatsuoka, Corter, & Tatsuoka, 2004). The tests can also be subdivided into: quantitative reasoning tests; progress tests in mathematics; logical reasoning tests; abstract reasoning tests; and deductive and inductive reasoning tests.

According to Urbina (2004), **quantitative reasoning tests**, components of the Graduate Record Examination (GRE) and the Graduate Management Admission Test (GMAT) in the United States, measure an individual's ability to interpret and analyse quantitative data, solve complex mathematical problems, and apply mathematical concepts in practical situations (Pickering et al., 2025). They assess skills in areas such as algebra, geometry, arithmetic, and data analysis.

Progress tests in mathematics are used to monitor the development of students' mathematical skills over time. These tests are applied at different educational levels to assess mastery of fundamental mathematical concepts and identify areas that need improvement (Attridge & Inglis, 2013).

Logical reasoning tests are designed to evaluate individuals' ability to think logically and systematically, identify patterns, solve problems, and make logical inferences (Wong, 2018). These tests are used in academic selection processes and professional recruitment.

Abstract reasoning tests measure the ability to understand non-verbal concepts and solve problems using logical thinking (Vecchiato et al., 2013). These tests are frequently used to assess candidates' aptitude for positions that require strong analytical and problem-solving skills. Tests like the American Raven's Progressive Matrices are classic examples of abstract reasoning tests, used in both educational and occupational contexts.

Deductive reasoning tests assess the ability to apply general rules to specific situations to reach logical conclusions, while **inductive reasoning tests** measure the ability to identify patterns and make generalizations from specific data (Cresswell & Speelman, 2020).

In the corporate context, logical reasoning tests are often used as part of the selection process to identify candidates with strong analytical skills (Yu et al., 2022). These tests help companies select individuals capable of solving complex problems and making decisions

based on logic and analysis (Urbina, 2004). Companies frequently use logical reasoning tests as part of their aptitude assessments to ensure candidates possess the skills necessary to perform roles that require critical thinking and problem-solving.

3.2 Multiple-Choice Tests

Multiple-Choice Tests (MCTs) are widely used assessment tools in various educational and professional contexts. They consist of questions or items that provide the examinee with several answer options, of which only one is correct (Haladyna & Rodriguez, 2013). These tests are designed to measure knowledge or skills in a specific area, offering a standardized and objective form of evaluation. Due to their structure, MCTs are particularly effective in assessing factual knowledge, understanding of concepts, and skills in analysis and application.

MCTs are characterized by their objectivity, as the correct answer is predetermined, eliminating subjectivity in the scoring process. This allows for rapid and efficient grading, often carried out automatically through software. Furthermore, the standardization of items facilitates the comparison of results between different groups of examinees, ensuring fairness in the assessment (Zoanetti et al., 2009). Additionally, when designing an MCT, it is crucial to avoid unintentional cues that might help students identify the correct answer without the necessary knowledge. This includes avoiding predictable answer patterns, inconsistent language between alternatives, and the use of absolute terms such as “always” or “never”, which may hint at the correct answer.

Despite their popularity and usefulness, multiple-choice tests have some limitations, according to Haladyna and Rodriguez (2013). One of the most common criticisms is that these tests tend to favour memorization over a deep understanding of concepts. Since the answers are predefined, students may be prompted to remember specific facts and details instead of developing a broader and more critical understanding of the content (Tongchai et al., 2011).

Another limitation of MCTs is the difficulty of creating items that effectively assess higher cognitive skills. While it is possible to design MCTs that address analysis, synthesis, and evaluation, this requires a high level of skill in question writing and a deep understanding of the content to be assessed. Moreover, creating good distractors is a constant challenge, as they must be plausible enough for well-prepared students to consider them as potential correct answers (Slepkov et al., 2021).

MCTs can also be vulnerable to biases, both cultural and contextual. A poorly written item may favour or disadvantage certain groups of examinees, depending on their cultural, socioeconomic, or educational background. This can result in assessments that do not accurately reflect an individual’s skills or knowledge but rather their familiarity with the test format or the specific content of the item (Haladyna and Rodriguez, 2013). However, many of these disadvantages can be avoided through careful test design.

Despite some limitations, multiple-choice tests offer several advantages over essay exams, making them a suitable choice for large-scale assessments. One of the main advantages is the objectivity in grading, which ensures that all examinees are judged based on the same criteria. This is especially important in high-stakes exams, such as entrance exams or professional certification tests, where fairness is critical.

Another significant advantage of MCTs is their breadth. They allow coverage of a wide range of content in a single exam, providing a broad overview of the student’s knowledge across various areas. Additionally, automated grading saves time and resources, allowing results to be made available quickly (Shard et al., 2024).

MCTs are also flexible and can be adapted to assess various levels of difficulty and types of knowledge. They can be used in diagnostic tests, formative and summative assessments,

and in adaptive teaching and learning contexts where immediate feedback is essential for student progress.

The creation of a multiple-choice test is a process that requires care and attention to ensure the validity and reliability of the instrument. Monteiro (2010) emphasizes that the construction of an MCT should begin with a clear definition of the learning objectives that the test intends to assess. From these objectives, items are created to reflect the skills and knowledge that are meant to be measured (Zoanetti et al., 2009).

In the item construction phase, it is important to ensure that the questions are formulated clearly and precisely, avoiding ambiguities that might confuse the examinees. Moreover, the answer choices must be carefully crafted. Incorrect alternatives, known as distractors, play a crucial role, as they must be plausible to prevent students from easily eliminating them and arriving at the correct answer by exclusion. Effective distractors are those that represent common errors or misconceptions about the topic, challenging the student to demonstrate a true understanding of the content (Slepkov et al., 2021).

Another important aspect of creating MCTs is the review and validation of items before their application. This can involve having the items analysed by experts in the content area and conducting pilot tests to identify potential flaws or ambiguities. The review can also help ensure that the items are free from cultural or linguistic biases that might disadvantage or benefit certain groups of students.

In addition to measuring conceptual knowledge in mathematics, MCTs can be used to assess logical reasoning and problem-solving skills. Questions that involve hypothetical scenarios or require students to make inferences from provided data can help assess the students' ability to apply their knowledge in new and unfamiliar contexts (Zhang et al., 2023). However, an MCT's ability to evaluate such skills largely depends on the clarity and precision with which the questions are formulated, as well as the relevance of the answer alternatives (Monteiro, 2010). Including questions that require critical analysis or judgment can increase the test's validity as a measure of complex thinking (Tongchai et al., 2011).

Multiple-choice tests are powerful and versatile tools in educational assessment, capable of measuring a variety of skills and knowledge objectively and efficiently. However, to maximize their effectiveness, it is essential that they are carefully designed, with attention to the clarity of the questions, the plausibility of distractors, and the relevance of the content being assessed. Despite criticisms and limitations, such as the emphasis on memorization and the difficulty in measuring analytical and application skills, multiple-choice tests remain one of the most widely used forms of assessment, due to their numerous advantages in terms of objectivity, efficiency, and flexibility (Shard et al., 2024).

The multiple-choice format is widely used in contemporary psychological and educational tests; however, its construction requires more meticulous attention than other formats due to the complexity involved in formulating the question stems and the response alternatives. To develop these items, there are guidelines aimed at ensuring a systematic and valid construction for the tests. Among these guidelines, those outlined in the work of Haladyna, Downing, and Rodriguez (2002) stand out, as they synthesized over 40 accumulated taxonomies over 54 years, consolidating 36 guidelines.

Despite the significant contributions of these guidelines, some difficulties were identified, such as content overlaps, language imprecision, and an excessive number of guidelines. To overcome these limitations, Moreno, Martínez, and Muñiz (2004) developed a more synthetic set of 12 guidelines for the development of multiple-choice tests (MCTs). This new set seeks to increase the validity of the assessment instruments, emphasizing congruence with the purpose of the evaluation. The guidelines were reorganized and reduced in number, eliminating redundancies and focusing on central aspects.

Among the 12 guidelines proposed by (Moreno, Martínez, & Muñiz, 2004), two are related to the content to be assessed: (1) the content should represent a sample from the specified content in the specification table, avoiding trivial items; and (2) representativeness should guide the level of complexity, specificity, and the memorization or reasoning basis of the item, as well as the best way to express it. Three other guidelines address the expression of content in the items: (3) the main point should be expressed in the statement, and each option should be grammatically consistent with the statement; (4) syntax and grammar structure should be correct, avoiding ambiguities, confusions, and the excessive use of negative expressions; and (5) the semantics should align with the content and the target audience. The remaining guidelines focus on constructing the answer choices: (6) there should only be one correct option, accompanied by plausible alternatives; (7) the correct option should be distributed across different positions; (8) the minimum number of options is three; (9) the options should be presented vertically; (10) the set of options for each item should appear structured; (11) the options should be autonomous, without overlaps or references to others; and (12) no option should stand out from the others in terms of content or appearance.

The creation of effective questions requires attention to specific principles that ensure the validity and reliability of the assessment (Haladyna, Downing, and Rodriguez, 2002). For a multiple-choice question to be considered effective, it is necessary to understand its main components: the stem, which introduces the situation or problem to be solved; the alternatives, which offer a set of possible answers, including the correct answer and distractors (incorrect but plausible answers); and the key, which identifies the correct alternative.

When developing a multiple-choice question in the educational context, it is essential that it aligns with the learning objectives of a course or cycle of studies. Furthermore, both the stem and the alternatives must be clear and concise, avoiding ambiguities or the use of jargon that might confuse students (Trevisan and Amaral, 2016). Another important point is the use of Bloom's Taxonomy, which recommends the formulation of questions that not only assess memorization and comprehension but also address higher levels of cognition, such as application, analysis, synthesis, and evaluation.

To illustrate, consider a question formulated with the stem: "Consider the graph below (not presented here), which shows the function $f(t)=at^2+bt+c$ representing the trajectory of a projectile launched in the Cartesian plane. The projectile reaches its highest point at the instant $t=3$ and touches the ground again at $t=6$. Knowing that the graph intersects the axis y at $(0,2)$, determine the value of the coefficient a". **Alternatives:** $A)\ -1/9;\ B)\ -2/9;\ C)\ -1/6;\ D)\ -2/6;\ E)\ -4/9$. **Correct Answer:** $)\ -1/9$.

This question requires the student to perform a complete analysis of the equation of a parabola and apply concepts of quadratic functions, such as vertex, roots, and the position of the graph in the Cartesian plane. The student needs to interpret the data provided in the graph, identifying that the vertex of the parabola occurs at $t=3$, that the root occurs at $t=6$, and that the point of intersection with the axis occurs at y is $(0,2)$.

The reasoning begins with the identification of the form of the quadratic function: $y(t)=a(t-3)^2+y_v$, where y_v is the height of the vertex. Knowing that the function intersects the axis y at $(0,2)$, the student needs to substitute $t=0$ and $f(0)=2$ to find the relationship between the coefficients. Additionally, the student must use the condition that the function equals zero at $t=6$ to calculate the coefficient a. This type of question is an example of the analysis and synthesis stages of Bloom's Taxonomy, as the student must break down the problem into its components (analysis) and reorganize the information to reach a solution

(synthesis). They need to apply knowledge of quadratic functions in an integrated manner and use advanced calculation and logical reasoning skills to solve the question.

The formulation of effective statements is one of the crucial steps in creating multiple-choice questions. Double negatives should be avoided, as sentences with complicated negations tend to confuse the reader and can undermine the validity of the question. It is preferable to use affirmative statements, which are more direct and easier to understand (Haladyna, Downing, and Rodriguez, 2002). Additionally, contextualizing the questions in situations relevant to the students can improve engagement and the relevance of the assessment, as long as the context is relevant to the content being evaluated.

When developing alternatives, it is recommended to use between three and five options, although the quality of the alternatives is more important than the quantity. Distractors, or incorrect alternatives, should be plausible so that a student with a superficial understanding of the content considers them as possible correct answers. This is crucial for ensuring that the question can effectively discriminate between different levels of comprehension among students. It is also significant to avoid predictable patterns in the position of the correct answer, such as always placing it in the same option. Alternating the position of the correct answer helps ensure that test participants do not answer based on assumptions or guessing strategies.

The analysis and review of questions are essential steps to ensure the validity and reliability of the test. Each question should be reviewed to ensure that it truly measures what it is intended to measure, without the influence of irrelevant factors. Peer review is a recommended practice, as it allows other educators to identify potential ambiguities or interpretation issues that the original author may not have noticed (Haladyna, Downing, and Rodriguez, 2002). After the test is administered, performance analysis of the questions should be conducted to identify questions that exhibited unexpected behaviour, such as all students answering correctly or incorrectly, which could indicate issues in the formulation of the question.

Ethical considerations must also be taken into account when formulating multiple-choice questions. It is essential that the questions are framed impartially, without favouring or disadvantaging any regional or cultural group. This includes attention to the language and context used in the questions (Haladyna, Downing, and Rodriguez, 2002). Furthermore, when using scenarios based on real-life cases, care must be taken to modify details that could reveal confidential information, thus preserving privacy and ethics in assessment.

Creating multiple-choice questions that are valid, reliable, and fair is a challenge that requires practice and continuous revision. By following these guidelines, educators can develop assessments that truly measure students' skills and knowledge, contributing to a more effective teaching and learning process. After the creation stage comes the analysis of the results. For some time now, analysis based solely on the number of correct answers has lost relevance in more advanced MCTs. Nowadays, large-scale tests such as National Assessment of Educational Progress (NAEP) and the *Exame Nacional do Ensino Médio* (ENEM), use statistical methods that go beyond simply counting correct answers and reveal the degree of consistency among the test-takers. This type of analysis was developed based on psychometric theories.

3.3 Large-scale tests in Higher Education

Interest in students' mathematical practices has grown over the last 40 years. These studies have focused on the cognitive skills and strategies used by mathematics experts and adults in informal contexts, i.e., outside the school environment (Monteiro, 2010). From this line of research, an understanding of mathematical practices emerged, outlining the specific behaviours involved in mathematical reasoning and problem-solving. As demonstrated by

Wu et al. (2025), cognitive analysis of mathematical processes provides valuable insights into students' thinking patterns and skill development. Two relevant instruments in the current educational context, the NAEP and the ENEM, present fundamental characteristics consolidated for test development.

3.3.1 National Assessment of Educational Progress

The NAEP is a U.S. instrument used to assess the academic performance of students in various areas, including mathematics (NAEP, 2026). Since 1973, NAEP has provided a national overview of American students' knowledge and skills in this subject. Aiming to track developments in education, NAEP periodically updates its parameters to evaluate what students should know and be able to do in mathematics. The focus is on ensuring that the assessment reflects changes in research, educational policies, and pedagogical practices. Nortvedt (2018) has highlighted how large-scale assessments significantly influence educational policy and curriculum development in mathematics education.

To develop the assessment items, the NAEP involves specialists in mathematics education, researchers, teachers, and other professionals in the field. The purpose is to create an instrument that is accessible to all students, regardless of their socioeconomic, ethnic, cultural, or linguistic background. Research by Arikan et al. (2020) emphasizes the importance of appropriate sampling and measurement techniques in ensuring large-scale assessments are representative and valid. The exam seeks to understand not only the mathematical content acquired by students, but also their ability to apply this knowledge in different contexts. The results of the NAEP are used by educators, policymakers, parents, and the public to understand students' proficiency levels in mathematics (NAEP, 2026). Furthermore, this data serves as a basis for developing strategies to improve education.

The mathematical content assessed by the NAEP covers five main areas: Number Properties and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra. While these areas are distinct, it is important to note that there are often intersections between them. For example, arithmetic calculations are essential in various areas such as Measurement and Geometry.

The NAEP emphasizes the importance of mathematical thinking, including understanding different numerical representations, mental and estimated calculation, and solving problems involving proportions and percentages. The assessment also addresses measurement, focusing on attributes such as length, area, volume, time, and temperature. Additionally, the exam includes questions about Geometry, involving properties of plane and solid figures, as well as geometric transformations.

The area of Data Analysis, Statistics, and Probability is crucial for developing skills in data collection, organization, analysis, and interpretation. Students are assessed on their ability to create graphical representations of data, calculate statistical measures, and understand probabilistic concepts. Finally, Algebra is addressed with questions about algebraic expressions, equations, functions, and relations.

It is important to highlight that the NAEP aims to assess the application of mathematical knowledge in contextualized situations, beyond understanding theoretical concepts. Bolondi et al. (2018) demonstrate that the formulation of mathematical items significantly impacts how students engage with and perform on assessments. The assessment includes items that require students to model real-world situations using mathematical tools, interpret results, and communicate their ideas mathematically. In summary, the NAEP provides a comprehensive view of students' mathematical knowledge, considering both specific content and the application of that knowledge. By evaluating these different dimensions, the exam contributes to understanding students' mathematical performance and provides valuable information for improving mathematics education.

Mathematics education research has undergone a "social turn" (Forzani et al., 2022), characterized by mathematical learning embedded in social activities, including discursive practices. This perspective aligns with Lohbeck et al. (2024) findings on how students evaluate their mathematical abilities through both social and dimensional comparisons. This implies that students use their mathematical knowledge and skills in various social contexts, such as at school, at home, or in games with friends. The 2026 NAEP mathematics framework reflects this broader and more comprehensive view of what it means to know and practice mathematics. For the first time, mathematical practices have been incorporated as a fundamental component of the NAEP assessment. These NAEP mathematical practices were selected after a review of current literature, national and international assessment frameworks, as well as national and state standards in the United States. As Kijima and Lipscy (2024) observe, international testing frameworks have increasingly influenced national assessment policies and educational outcomes. Five main practices stand out: representation; abstraction and generalization; justification and proof; mathematical modelling; and collaboration in mathematics.

Representation involves recognizing, using, creating, interpreting, or translating between representations that are appropriate to the grade level and the mathematical content being assessed. This practice emphasizes students' ability to approach different forms of mathematical representation, such as graphs, tables, equations, and visual models, which are crucial for understanding and communicating mathematical ideas.

Abstraction and generalization refer to the ability to contextualize, identify similarities between cases, items, or problems, and extend reasoning to a broader domain, according to the grade level and mathematical content being assessed. This practice is central to the development of abstract thinking skills, enabling students to identify patterns and generalize mathematical concepts to new situations.

Justification and proof involve creating, evaluating, presenting, or refuting mathematical arguments in ways appropriate to the development and mathematical content. Justifying and proving are essential components of rigorous mathematical thinking, requiring students to support their conclusions with logical and evidence-based arguments.

In *mathematical modelling*, understanding a scenario and identifying a problem to be solved are key steps. Then, mathematizing it allows finding a suitable solution for understanding the phenomenon or process. Finally, verification and feasibility of the model are necessary. Wang et al. (2023) have demonstrated how mathematical modelling competency positively affects student creativity, particularly when supported by curiosity and guided inquiry teaching. This practice emphasizes the application of mathematics in real-world contexts, enabling students to use mathematical tools to solve complex problems and interpret the results in terms of their applicability.

The fifth item, *collaboration in mathematics*, refers to the social endeavour of doing mathematics together with others through discussion and collaborative problem-solving, where ideas are presented, debated, connected, and developed toward a shared solution and understanding. Collaboration in mathematics involves collective thinking between individuals to construct a solution to the issue in a quantitatively appropriate way.

Mathematical practices contribute to mathematical reasoning, but they also depend on other skills, such as conceptual understanding and computational fluency. Similarly, mathematical practices can contribute to conceptual understanding, but they are not interchangeable with it (Forzani et al., 2022). Lin and Chen (2023) emphasize the importance of self-direction in mathematics learning, which encompasses self-efficacy, self-regulation, and intrinsic motivation − factors that significantly support mathematical practice and achievement. Conceptual understanding refers to knowledge of the underlying structures and

relationships in mathematics that transcend the application of familiar algorithms, whereas practices are fluid and responsive to both familiar and unfamiliar problems.

The increased emphasis on mathematical practices is evident in the state and national standards of U.S. assessments. It is now widely accepted that knowing and doing mathematics involves practices such as generalizing, conjecturing, justifying, modelling, problem-solving, communicating, and making sense of concepts. As students engage with and discuss mathematical ideas and issues, both individually and in groups, they participate in these practices and become familiar with the norms of doing mathematics. The inclusion of mathematical practices in the NAEP is not separate from the mathematical content; it signifies a significant shift in the NAEP's mathematics framework, aiming to ensure that the assessment captures a complete and comprehensive picture of students' mathematical knowledge and skills.

3.3.2 National Secondary Education Examination

The National Secondary Education Examination (Exame Nacional do Ensino Médio - ENEM) was created in 1998 in Brazil with the objective of evaluating students' performance at the end of Basic Education, measuring the development of the skills and competencies necessary for the full exercise of citizenship (Andriola, 2011). Initially, the exam had a few numbers of participants, but its importance and reach grew significantly over the years. In 2008, the ENEM reached over four million registrants, and in 2010, this number exceeded 4.6 million (Brasil, 2025). In addition to assessing students' performance, the ENEM became a selection criterion for the University for All Program (ProUni) starting in 2004 and, later, for several Higher Education Institutions (IES) as part of their admission processes.

The Ministry of Education (MEC) in 2010 proposed a reformulation of the ENEM, making it a unified selection instrument for federal IES (Brasil, 2025). The new ENEM began to consist of tests evaluating four knowledge areas: languages, codes, and their technologies (including the essay); human sciences and their technologies; natural sciences and their technologies; and mathematics and its technologies. Each area is made up of 45 multiple-choice questions, applied over two days. The essay must be written in Portuguese in a discursive-argumentative format, addressing topics of social, scientific, cultural, or political order.

One of the main differences between the old and the new ENEM is the structure of the exams. Until 2008, the exam consisted of 63 interdisciplinary items, without direct connection to high school content and without the ability to compare student results over time (Brasil, 2025). The new format allows for this comparison, enabling the creation of historical series that can contribute to broad educational analyses. Hargreaves (2020) notes that such structural changes in large-scale assessments can significantly impact teaching practices and educational priorities. Additionally, the new version of the ENEM exam seeks not only to assess students' knowledge but also their skills and competencies in tackling problem situations, constructing arguments, and developing proposals.

In 2015, there was an expansion of the item bank, aiming for greater diversity in the questions and the style of questions. However, the basic structure of the exam remained the same, with 180 questions and the essay (Brasil, 2025), due to the adoption of the New High School in Brazil, the ENEM may undergo significant changes.

Higher Education Institutions (IES) have the autonomy to decide how to use the ENEM in their selection processes, choosing to use it as a single phase, in combination with their own entrance exam, or as a stage for remaining vacancies (Brasil, 2025). The Unified Selection System (SISU) was created to manage this selection process, allowing candidates to be selected solely based on the grades obtained in the ENEM. Therefore, the exam has

become an evaluation model adapted to the needs of a constantly changing society, focusing on the development of competencies essential for the contemporary citizen.

The ENEM aims to assess complex competencies, such as mastery of languages, understanding of phenomena, problem-solving, argument construction, and proposal development. Bittmann (2022) has shown that academic competencies and educational aspirations co-develop, with improvements in one area fostering advancement in the other. These competencies should be developed throughout Basic Education and are essential for the formation of future university students, who will need to apply them in generating new scientific knowledge, proposing solutions to social problems, and promoting technological innovations.

The ENEM uses principles from Applied Mathematics, Advanced Statistics, Computer Science, and Cognitive Psychology to ground and analyse its exams. Wang (2020) highlights how the integration of data science with large-scale assessments can enhance our understanding of educational contexts and student performance. The items of the exams are designed based on rigorous pedagogical procedures and analysed both quantitatively and qualitatively to ensure their quality. The quantitative analyses are based on Item Response Theory (IRT). The items in the ENEM are designed to assess skills and competencies, regardless of the student's formal content knowledge. The questions present information that allows candidates to interpret, infer, deduce, and solve problems based on the provided data set (Brasil, 2012). This model moves away from a focus solely on knowledge of school content, as is the case in traditional entrance exams. The goal is to assess what the student is capable of doing with the information presented, encouraging the development of broader competencies.

In summary, the ENEM is an important tool for evaluating the competencies and skills of students at the end of basic education, as well as being a fundamental instrument for access to higher education in Brazil. Gambini et al. (2024) demonstrate the value of such assessments in predicting university performance, underscoring their importance in educational transitions. Following its reformulation, the exam has taken on a central role in educational opportunities, contributing to academic mobility and the restructuring of the Brazilian high school curricula. Fischbach et al. (2013) have similarly found that proficiency scores on large-scale assessments can effectively predict educational outcomes, reinforcing the value of these evaluation instruments.

3.4 Psychometric theories

To ensure the effectiveness of tests, it is essential to understand the theories that underpin their construction and analysis. Although the historical evolution of tests has been addressed, the importance of the theories that enable their construction is often not fully recognized. The statistical theories of testing help estimate essential psychometric properties, such as the reliability and validity of tests, ensuring that decisions based on these instruments are appropriate and accurate. Without these theories, it would be impossible to guarantee the accuracy and relevance of the tests, which could significantly disadvantage or benefit the individuals being assessed (Santos et al., 2021).

Test construction is guided by psychometric theories that inform their development. The Classical Test Theory (CTT) and Item Response Theory (IRT) are two primary approaches in this field. CTT, which developed from the pioneering work of Charles Spearman in the early 20th century, is based on the idea that the score obtained on a test is composed of a true score and a measurement error (Muñiz, 2010). Spearman's model assumes that the true score is the average of scores obtained from an infinite number of test administrations, and that the measurement error is independent of the true score and across different test administrations

(Martins et al., 2022). These assumptions are fundamental for estimating the reliability and validity of tests, and they form the foundation of classical psychometrics.

An example that helps in understanding these two analytical approaches can be observed in the field of Physics, as shown when developing assessments for visualization skills (Munns et al., 2023). The concept of spatial scale refers to the size of the objects with which we interact, ranging from micrometers in microorganisms to light-years in astronomical distances. This concept is essential for understanding the relationship between the microscopic and macroscopic worlds. Recent scientific advances have allowed for a better understanding of this relationship. For example, the volume of a balloon, which is a macroscopic property, depends on the microscopic characteristics of the gas particles that inflate it. Understanding the influence of scales is crucial for analysing physical problems and others. The type of gas constituent, whether lighter or heavier, reveals the microscopic characteristic, while its volume reveals the macroscopic characteristic. Both influence the balloon's ability to float.

Similarly, in Psychology and Education, Modern Psychometrics uses both approaches for test evaluation. Classical Test Theory (CTT) focuses on the properties of the test as a whole, using a macroscopic scale. Item Response Theory (IRT), on the other hand, focuses on the study of individual items to make precise measurements, employing a microscopic approach. The main difference between IRT and CTT lies in their approaches: IRT analyses individual items, while CTT considers the test as a unit, a distinction essential for developing accurate measurement instruments (Bulut et al., 2025). Both are valid for evaluating test results, each with your qualities and specificities.

3.4.1 Item Response Theory (IRT)

The CTT is widely used and has become consolidated over the years with established formulas and methods for evaluating the accuracy of tests. However, the classical approach has limitations. It does not ensure the invariance of measurements with respect to the instrument used and relies on samples to estimate psychometric properties, which can compromise the precision of results (Kieftenbeld et al., 2011). These issues led to the development of IRT, which emerged to address these concerns and provide a more robust and consistent measurement.

IRT, which began to be formalized in the 1960s, introduces more complex and detailed models, such as the Rasch logistic model, which provides familiar measurement properties that facilitate interpretation (Madison et al., 2024). This theory addresses some limitations of CTT by ensuring the invariance of estimates of individuals' abilities and the properties of items, difficulty and discrimination, regardless of the instrument or sample used. IRT provides a more refined approach to measuring psychological variables, allowing for a more accurate estimation of the reliability and validity of tests in different contexts and populations.

IRT represents a significant advancement in psychometrics, offering new concepts and tools that complement and expand upon CTT methods, particularly when analysing differential item functioning across diverse populations (Gonçalves et al., 2013). However, CTT is still widely used and remains relevant in psychological practice. Both approaches coexist and are applied according to the situation and the specific needs of psychological assessment.

Theoretical models do not emerge suddenly or linearly. IRT, while addressing the problems of CTT, does not contradict its fundamental conclusions but adds assumptions that allow it to answer questions not addressed by CTT. The same is true in the example of the gas inflating a balloon. The development of IRT was gradual, as reviewed by Andriola (2009), starting over 50 years ago. L. L. Thurstone, in 1925, introduced curves associating

age and the proportion of correct answers, precursors to modern item characteristic curves. M. W. Richardson, in 1936, formulated the item information function. G. A. Ferguson, in 1942, developed concepts equivalent to the difficulty parameter of IRT. P. F. Lazarsfeld, in 1950, introduced the concept of latent traits.

F. M. Lord, in 1952, formulated the main concepts of IRT. Between 1957 and 1958, A. Birnbaum proposed logistic models with greater mathematical ease. G. Rasch, in 1960, introduced the one-parameter logistic model, emphasizing the principles of convergence and separability. In 1968, F. M. Lord and M. R. Novick published a work that marked the end of the theoretical phase of IRT. The worldwide acceptance of IRT only occurred two decades later, with the publication of F. M. Lord's work in 1980 on the practical applications of IRT, which continue to evolve with advanced statistical methods like semi-parametric approaches (Liu, 2025).

The terms "latent attribute" and "construct" are used to refer to the psychosocial characteristics of interest for measurement, a concept fundamental to educational assessment development (Schadl & Lindmeier, 2025). The definition of the construct is directly related to issues of validity. Validity refers to the degree to which inferences made from test scores correspond to the defined "construct", which is particularly important when evaluating specific interests, such as in mathematics (Wei et al., 2014). For example, in large-scale achievement tests, such as mathematics tests, the construct may be based on the mathematics curriculum. Item Response Theory (IRT) helps to link the construct to the types of inferences that can be made from the test scores.

In Item Response Theory (IRT), the construct has a specific meaning. While in Classical Test Theory (CTT) inferences are mainly made based on the total score obtained on a test, in IRT the focus is on measuring the level of a latent attribute in each individual, an attribute that cannot be directly observed. For example, if the construct to be measured is proficiency in geometry, the test items are designed to assess different aspects of that specific knowledge, which requires innovative item formats to capture all dimensions (Moon et al., 2020).

IRT assumes that the level of the latent attribute is what determines an individual's responses to the items, not the other way around. This differs from approaches where the total score directly depends on the specific items used. For example, in a mathematics test composed of algebra and geometry questions, the average score from these questions does not directly define the student's level of ability in mathematics, requiring specialized assessment instruments to evaluate concept readiness (Peralta et al., 2020). In IRT, it is possible to alter the items in the test, as long as they continue to measure the same construct, and still maintain a consistent assessment of the latent attribute level, such as mathematical proficiency.

The IRT model presented is one-dimensional, meaning that a set of items measures a single latent attribute, though testing this assumption requires sophisticated analysis as demonstrated in cognitive assessment development (Primi et al., 2016). If a test contains multiple latent attributes, it is considered multidimensional. In practice, it is difficult to find items that test exactly the same "construct". For example, solving mathematical problems may involve both reading and mathematical skills, making the test multidimensional.

3.4.2 Main Logistic Models of IRT

The simplest model for (IRT) is the one-parameter logistic model, proposed by G. Rasch in 1960, which is particularly useful for psychometric analysis of achievement scales (Bellini et al., 2019). This model calculates the probability of answering an item correctly based on its difficulty (parameter b). The two-parameter logistic model, developed by A. Birnbaum between 1957 and 1968, includes discrimination (parameter a) in addition to difficulty (b). The three-parameter model, also by Birnbaum, adds a parameter for the probability of

guessing the correct answer (c). Finally, the four-parameter model, suggested by Barton and Lord in 1981, includes an additional parameter (Yi) for situations where individuals with high proficiency do not answer the item correctly due to specific circumstances, which requires sophisticated validation methods (Stites et al., 2019).

The assumption of unidimensionality in Item Response Theory (IRT) indicates that the complexity of solving a problem should be caused by a single latent variable (θ), a concept essential in the development of career interest surveys like the STEM (Kier et al., 2014). Unidimensionality proposes a functional relationship between θ and the responses to the items, represented by the item characteristic curve. The local independence of items states that the response to one item should not influence the responses to the other items, which is mathematically expressed as the product of the probabilities of answering each item correctly, a property heavily studied in imputation methods for educational assessment (Thomas, 2000).

3.5 Fundamentals of the Rasch Model

The Rasch model is based on two main assumptions. First, it assumes that the attribute being measured can be represented in a single dimension, on which both individuals and items are positioned. According to Prieto (2003), the probability of a person answering an item correctly is determined by the difference between the person's ability level and the difficulty of the item. Madison et al. (2024) demonstrated that this one-parameter approach provides familiar measurement properties, including sum score sufficiency and invariant item and person ordering. The model uses a logistic function to model this relationship, where the probability of a correct response is expressed by equation (1):

$$\ln\left(\frac{P_{is}}{1-P_{is}}\right)=\theta_s-\beta_i \qquad (1)$$

Equation (1) indicates that the logarithm of the ratio between the probability (P_{is}) of individual *i* answering item *s* correctly and the probability of error ($1-P_{is}$) in an item is a function of the difference between the person's ability level (θ_s) and the item difficulty (β_i).

From Equation (1) it is obtained that the probability of success in the Rasch model is a sigmoid function, also known as the logistic function, as presented in Equation (2):

$$P_{is}=\frac{e^{(\theta_s-\beta_i)}}{1+e^{(\theta_s-\beta_i)}} \qquad (2)$$

The Rasch model predicts that the probability of a correct answer will be 50% when the person's ability is equal to the item difficulty. When the ability is higher, the probability of success increases, and when it is lower, the probability decreases. Liu and McKeough (2005) applied this principle to analyse the developmental sequence of energy concepts using item difficulty estimates from Rasch modelling of Trends in International Mathematics and Science Study (TIMSS) data.

Susongko et al. (2020) underscore the importance of a comprehensive validity framework, grounded in Messick's theory, when evaluating the quality of items in assessment instruments, particularly when used in conjunction with Rasch modelling. Messick's approach encompasses six principal facets of validity: content, substantive, structural, generalisability, external, and consequential. Each of these aspects provides a distinct lens through which the validity of an instrument can be scrutinised, ensuring that the assessment is not only relevant and representative of the construct it purports to measure (content validity), but also that student responses genuinely reflect their knowledge and skills (substantive validity).

Within Rasch modelling, this validity framework is operationalised through various statistical analyses. Structural validity, for example, is evaluated by examining the fit of items

to the Rasch model, ensuring that items function consistently and measure a single latent construct. Toribio and Albert (2011) illustrated the effectiveness of various discrepancy measures for detecting misfitted items in IRT models through posterior predictive model checking. External validity is investigated through the instrument's relationship with other relevant measures, whilst consequential validity considers the social and ethical implications of using the instrument. Li and Wu (2023) demonstrated this by employing the Bookmark method with Rasch modelling to establish standards for mathematics achievement tests, showing how generalisability theory can evaluate the precision of standard setting results.

The most commonly used scale for interpreting the parameters is the logit scale, which is the natural logarithm of the odds ratio on the left-hand side of Equation (1), offering an intuitive interpretation of the item and student parameters. To apply the Rasch model, it is necessary to estimate the item and person parameters. When the item parameters are known, it is possible to estimate the individual parameters using conditional estimation methods. Otherwise, joint estimation is used, where both item and person parameters are estimated simultaneously. Bartolucci, Bacci, and Gnaldi (2014) described this process in detail in their R package for multidimensional latent class item response models, which implements maximum likelihood estimation via an Expectation-Maximization algorithm. The most common method for this estimation is maximum likelihood, which seeks the parameters that maximize the likelihood of the observed responses.

The precision of the estimates is measured by the standard error of measurement, which is inversely proportional to the test information function. The information function is the sum of the item information functions, and the information of an item varies along the ability continuum, being maximum when the item difficulty corresponds to the individual's ability level. Robitzsch (2023) explored how different analytical decisions in scaling educational large-scale assessment data can impact parameter estimation, demonstrating that model uncertainty can have substantial effects on country-level estimates.

The Rasch model presents some advantages over other models applied in Item Response Theory (IRT). The following characteristics stand out:

Joint measurement: the parameters of individuals and items are expressed in the same units and located on the same continuum, allowing for a more realistic and detailed analysis. The interpretation of scores does not depend on group norms, but rather on the probability of a person correctly solving certain items. Wilson (2023) extensively discusses this property in his book on constructing measures, emphasizing how the Rasch model provides a framework for developing high-quality instruments across various disciplines.

Specific objectivity: comparisons between individuals and items are independent, which ensures that the difference between two people or between two items does not depend on the characteristics of the sample. Bansilal (2015) demonstrated this principle in a study of mathematics teachers' proficiency, where teacher ability and item difficulty were represented on the same scale to provide more informative and fair assessment.

Interval properties: the logit scale has interval properties, which allows for the use of rigorous parametric analyses, ensuring the invariance of differential scores along the continuum.

Specificity of the standard error of measurement: the Rasch Model allows for quantifying the information provided by a test at different points along the continuum and adjusting the difficulty of the items to increase the precision of measurements. Cho et al. (2010) leveraged this property in their latent transition analysis with a mixture Rasch model, which permitted within-class variability on the latent variable for more effective measurement of treatment effects.

To ensure that the advantages of the Rasch model are effective, it is essential that the data are consistent with the model's assumptions. The presence of responses that deviate from the

expected can indicate that the parameters of subjects and items have no theoretical meaning. The fit analysis can be performed using statistical analysis based on residuals, which help to identify items and individuals that do not fit the model. Hos and Saraç (2025) demonstrated this by comparing different mixture models for PISA mathematics data, finding that incorporating response times as a covariate improved model fit and helped identify distinct ability and non-ability latent classes.

3.5.1 The Rasch Model and its application in Item Response Theory

The Rasch Model and the Rasch Model within the IRT approach present significant differences that affect their implementation, interpretation, and sample requirements. Whilst related, they represent distinct perspectives in psychometric measurement, with important implications for researchers and assessors.

The Rasch Model, developed by Georg Rasch in 1960, is often considered a distinct measurement paradigm, with a central focus on the principle of "specific objectivity" and measurement invariance. As Boone and Noltemeyer (2017) highlight, proponents of the Rasch tradition consider this model philosophically distinct from other IRT models, emphasising that data should fit the model, not vice versa. According to Andrich (2004), in the pure Rasch model, items that do not fit the model requirements are typically revised or removed, rather than adopting a more complex model. This approach assumes all items have equal discriminative power and is characterised by its single parameter (item difficulty), resulting in comparatively smaller sample requirements, generally around 100-250 participants (Linacre, 2012).

In contrast, within the Item Response Theory approach, the Rasch Model is classified as a one-parameter logistic model (1PL), considered merely a special case within a broader family of models. As explained by Embretson and Reise (2013) and De Ayala (2022), IRT encompasses more complex models, such as the two-parameter model (2PL, which adds item discrimination) and the three-parameter model (3PL, which adds a guessing parameter). In this perspective, if data do not fit the Rasch model, researchers typically consider more complex alternative models that might better explain response patterns. This flexibility comes at the cost of significantly larger sample requirements, with recent studies suggesting that samples of 500-1,000 participants or more are necessary for stable parameter estimates in more complex IRT models (Jiang et al., 2016; Golino et al., 2020).

The practical implications of these differences are substantial for researchers. As noted by Wind (2017) and Demars (2018), the choice between the pure Rasch approach and the IRT approach should be guided not only by statistical considerations but also by measurement objectives, practical constraints, and theoretical assumptions. For instance, when measurement invariance and direct interpretability are prioritised, the traditional Rasch approach may be preferable. In contrast, when the primary goal is to maximise statistical fit or when there are theoretical reasons to expect items to vary in discrimination, the broader IRT approach may be more appropriate. Recent research by Huggins-Manley (2018) demonstrates that these methodological choices can significantly impact assessment results and interpretations, particularly in educational contexts where important decisions are based on the outcomes.

4. RESULTS AND DISCUSSION

To validate the applicability of the proposed methodology for students transitioning from secondary school to higher education, we used microdata from the 2023 ENEM examination as a case study. Since ENEM is the primary gateway examination for Brazilian universities, the sample directly represents the target population of incoming first-year students. This

alignment ensures that the diagnostic insights derived from the Rasch-based analysis are contextually relevant for designing targeted pedagogical interventions at the university level.

Knowledge assessment, often performed through instruments with multiple-choice items (MC), is positive, but it has limitations. Students may get MC items right by chance or without mastering the subject, while others, with partial knowledge, may be harmed. Given these limitations, Item Response Theory (IRT) emerges as an alternative for analysis. IRT allows analysing students' responses beyond the simple number of correct answers, considering the item difficulty.

The present study aimed to develop a standardized test to assess the mathematical skills of students entering a higher education course, considering a limited number of students, which can vary from one to four classes. For this purpose, the analysis with IRT, using the one-parameter Rasch model, is useful. This approach allows for an efficient analysis of students'' proficiency and the difficulty of test questions.

In order to exemplify and compare the applicability of this proposal, Microdata from the 2023 National High School Exam (ENEM) were selected (INEP, 2024). The sample consisted of the results of the blue Mathematics test of 160 students from the municipality of Araraquara. The answers to the 45 questions were transformed into a binary matrix, in which the value 0 represents a wrong answer and the value 1 indicates a correct answer. This binary format is essential for the application of the Rasch model, which assumes that the probability of getting an item right depends on both the student's ability and the difficulty of the question.

Regarding the sample size of 160 respondents, it is important to consider the principles of parameter estimation within the Rasch model. While Rasch modelling can be effectively employed with smaller samples compared to other statistical techniques, the appropriateness of the sample size is contingent on the number of parameters being estimated (Bond & Fox, 2015). Our sample size was determined to be adequate based on recommendations for item calibration stability, with a participant-to-item ratio exceeding the suggested minimum of 10:1, ensuring sufficient precision in estimating item difficulties and person abilities (Reise, Bonifay, & Haviland, 2013). This ratio supports the validity of our findings (Susongko et al., 2024).

Furthermore, the Rasch model's sample-free item calibration and person-free ability of estimation contributes to the generalizability of our results, as item parameters are not dependent on the specific sample used (Boone, Staver, & Yale, 2014). Additionally, the rigorous quality control measures applied in our study, including item fit statistics and differential item functioning analysis, further validate the appropriateness of our sample size by ensuring the stability and consistency of item parameters across different subgroups.

For data analysis, the R 3.3.0 (R Core Team, 2024) software was used with the following packages: eRm, ggplot2, ltm. This software allows the application of the Rasch model and the graphical visualization of results accessible to most educators. The first step consisted of estimating the model parameters, which resulted in determining the difficulty of each question, presented in Figure 2.

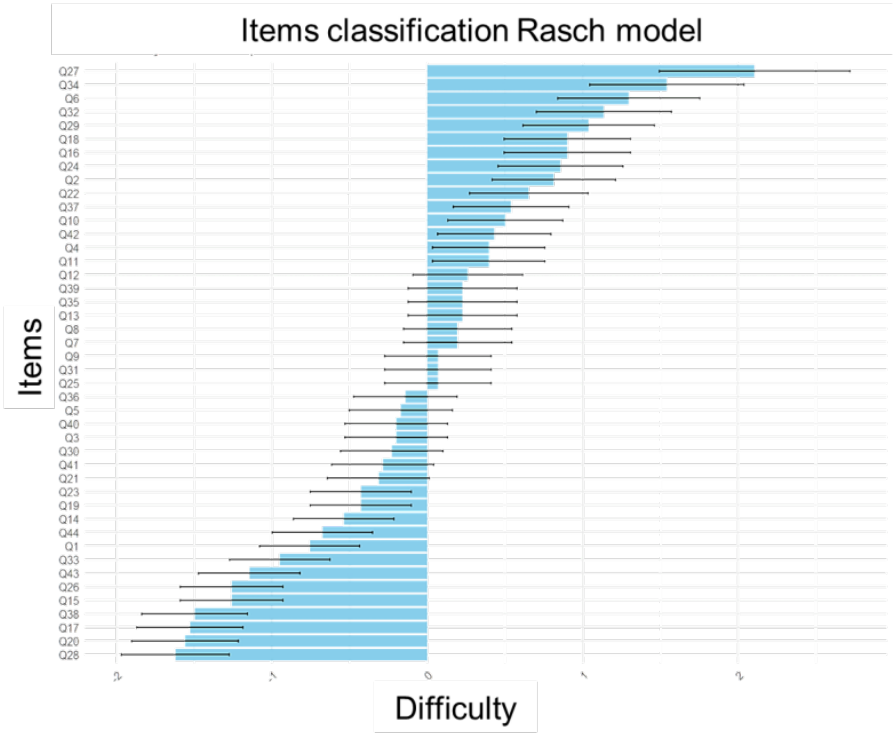

Figure 2. Estimated difficulty of each question on the 2023 ENEM maths test. Sample universe of 160 students from the municipality of Araraquara-SP. Results of the application of the Rasch model in the R software. Source: INEP Microdata (2024).

These values represented in an item difficulty graph allow identifying which questions presented a greater or lesser degree of complexity for the evaluated students. The difficulty scale ranges from -2 to +2, with the easiest items located at the left end of the scale and increasing from left to right, reaching the highest difficulty value at the right end of the difficulty axis.

To address the need for a more profound understanding of item difficulty levels, we have developed Table 1, which presents a detailed classification of each item in the instrument. This table organises the items into five categories of difficulty: very easy, easy, moderate, difficult, and very difficult, based on the logit values obtained through Rasch analysis. This categorisation enables a substantive interpretation of the measured construct, clearly identifying which aspects of knowledge or skill present greater challenges for the participants.

The logit values were converted into qualitative categories following a metric criterion of statistical significance based on the Least Significant Difference (LSD), estimated at 0.8 logit according to the measurement uncertainty benchmarks proposed by Ekstrand et al. (2022). Consequently, values below −1.2 logits are considered ‘very easy’; values between −1.2 and −0.4 (excluding the upper limit) are classified as ‘easy’; the central interval between −0.4 and +0.4 represents ‘moderate’ difficulty, which corresponds to the region of maximum precision of the instrument; values between +0.4 and +1.2 are ‘difficult’; and values equal to or greater than +1.2 logits are classified as ‘very difficult’. This classification provides a pedagogical interpretation of the results, allowing educators to identify which specific competencies represent greater difficulty for students and, consequently, which areas warrant greater attention in the teaching-learning process.

Table 1. Item difficulty levels and their substantive interpretations.

Item	Item description	Difficulty value (logit)	Difficulty category	Substantive interpretation
1	Assesses basic, easily accessible concepts.	-2	Very Easy	Most participants master basic skills.
2	Measures fundamental, straightforward competencies.	-1	Easy	Demonstrates introductory domain knowledge.
3	Evaluates intermediate proficiency and application.	0	Moderate	Requires more than basic knowledge.
4	Requires advanced analysis and interpretation.	+1	Difficult	Integrates concepts with complex processing.
5	Assesses higher-level, integrated skills.	+2	Very Difficult	Represents the most advanced challenges.

Source: by the authors (2026).

Next, a histogram was created representing the distribution of abilities of the 160 students based on their answers to the items, shown in Figure 3. The scale used ranged from -1 to 4, with students classified with values close to -1 being considered less skilled, while those with values close to 4 demonstrated greater proficiency in the test. This graph made it possible to visualize the dispersion of abilities within the sample, identifying groups of students with different levels of mathematical knowledge.



Figure 3. Distribution of the ability of 160 students from the municipality of Araraquara on the 2023 ENEM maths test. Results of the application of the Rasch model in the R software.

Source: INEP Microdata (2024).

Subsequently, a graph of the score distribution according to the Rasch model was generated, Figure 4. This visualization allows verifying the overall distribution of scores obtained by the students and how these scores behave within the analysed sample universe. The students' scores are the result of the application of the one-parameter model and may differ from the score calculated by INEP. This may occur because the agency uses the three-parameter Birnbaum model.

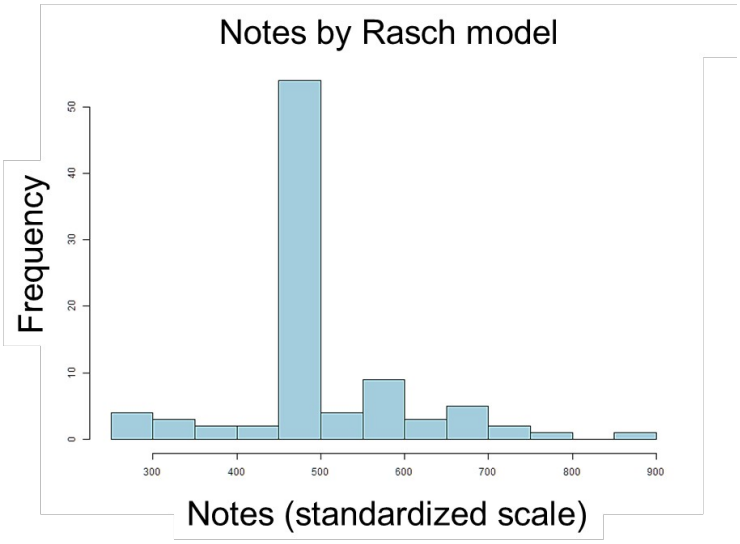

Figure 4. Distribution of scores according to the Rasch model. Results of the application of the Rasch model in the R software. Source: INEP Microdata (2024).

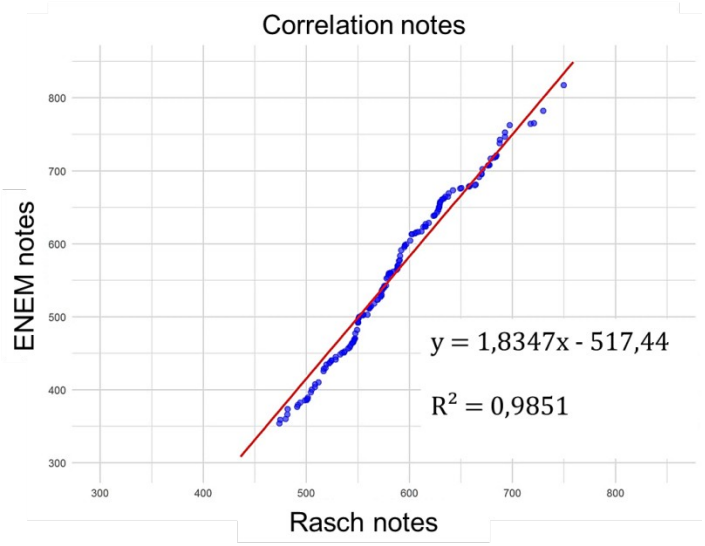

Figure 5. Correlation between the scores of 160 students from the municipality of Araraquara obtained by the Rasch model in IRT and the official scores of the 2023 ENEM obtained by a three-parameter model. Results obtained in the R software.

Source: by the authors (2025).

The correlation presented in Figure 4 reveals a linear trend between the scores calculated by the ENEM model and the Rasch model. For scores around 600, both models are equivalent. However, for scores below 550, the Rasch model overestimates the score. This means that the same test-taker obtains, for example, a score of 500 by the Rasch model and a score close to 400 by the ENEM model. This behaviour occurs because the three-parameter model used in the ENEM takes into account other discriminating items for the final score calculation.

Although the one-parameter Rasch model is a simpler version within IRT, it has proven suitable for the analysis of multiple-choice tests in contexts where the number of students evaluated is between one and two hundred. In this way, the proposed approach can be used as an efficient assessment tool for students entering higher education courses, allowing for an objective and standardized analysis of student performance in a reduced sample universe.

5. CONCLUSIONS

The literature review conducted as part of this study confirmed that standardized tests play a crucial role in evaluating academic performance and supporting educational decision-making. Their objectivity, standardization, and ability to compare results across different groups remain their primary advantages, as they provide quantitative data that can identify areas of proficiency and deficiency, reshaping pedagogical practice and the development of educational policies.

On the other hand, the literature review also revealed limitations of standardized tests. The emphasis on specific skills and the objective nature of the questions can restrict the assessment of more complex skills, such as creativity and critical thinking. Furthermore, the validity and reliability of tests can be questioned, especially when they are used to making high-stakes decisions, such as selection for higher education programs. Equity in the application of standardized tests is another crucial point to be considered. Factors such as the socioeconomic, cultural and linguistic context of students can influence test performance, leading to unfair results. It is essential that tests be culturally adapted and that the different realities of students be considered when interpreting results.

The historical evolution of standardized tests demonstrates the constant search for more accurate assessment instruments. However, much remains to be done to ensure that these tests are used appropriately. It is necessary to invest in research that investigates the long-term impacts of standardized tests, as well as in assessment practices that complement traditional tests, such as formative assessment and portfolio assessment. Item Response Theory (IRT) has brought significant changes to psychological and educational assessment, focusing the analysis on individual items rather than the test as a whole. One-dimensional IRT models are widely used, but multidimensional IRT, which deals with polytomous data, is becoming increasingly relevant. In Brazil, IRT is used by institutions such as INEP to provide results for ENEM and other large-scale assessments, in addition to state education secretariats. However, there is a need for more training and research in the area. Future developments may lead to more significant use of multidimensional IRT models. In line with the second objective of this paper, to exemplify how a Rasch-based diagnostic instrument can assess mathematical proficiency, the present study demonstrated that it is possible to use simpler models, such as the one-parameter Rasch model, for analysis of standardized instruments in reduced sample universes, as illustrated by the ENEM 2023 microdata case study with 160 students.

It is essential to acknowledge the limitations of this case study that impact the interpretation and generalisation of results. Firstly, the sample was restricted to 160 students from a single geographical region (Araraquara, Brazil), which may limit the applicability of findings to broader or culturally diverse educational contexts. The exclusive use of ENEM 2023 data provides only a specific temporal snapshot, failing to capture longitudinal variations in mathematical skill's development. Methodologically, whilst the one-parameter Rasch Model demonstrated utility, it does not consider factors such as discrimination and guessing, which two- or three-parameter IRT models could capture. Furthermore, uncontrolled external variables such as socioeconomic background, prior education quality, and motivational factors may have influenced the observed outcomes. The absence of cross-validation with other diagnostic instruments also represents a limitation that future studies could address. These considerations are essential for properly contextualising the contributions of this work within the broader field of STEM educational assessment.

This research offers insights into mathematical assessment in STEM education, though several limitations exist. The cross-sectional design prevents observation of longitudinal developments in assessment practices. Data collection faced potential response biases, with participants holding stronger opinions potentially overrepresented. Despite sampling efforts,

certain institutional contexts remained underrepresented, limiting generalisation across all educational settings. Additionally, the self-reported nature of practices may reflect aspirational rather than actual implementation. These constraints suggest our findings should be contextualised accordingly; the assessment frameworks identified should be considered emerging models requiring validation through further research employing diverse methodological approaches, including observational and longitudinal studies to better capture the dynamic nature of STEM assessment practices.

Based on the diagnostic approach exemplified in this study, future research could investigate the impact of using different types of standardized tests on higher education student performance, and analyze the relationship between Rasch-based diagnostic scores and other indicators of academic success — such as attendance, engagement, and retention. The development of assessment instruments that combine Rasch diagnostics with complementary forms of assessment, such as formative assessment and portfolio evaluation, represents another strategy of interest. Additionally, future studies could explore the influence of cultural and socioeconomic context on performance in diagnostic tools, as well as investigate the long-term impact of such assessments on student learning trajectories across different populations and educational contexts in higher education.

FUNDING STATEMENT

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

DATA AVAILABILITY STATEMENT

The secondary data used in this study are available in the *Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira* (INEP) repository: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem.

REFERENCES

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, 42(1), I-7.

Andriola, W. B. (2009). Modern psychometrics: Characteristics and trends. *Estudos em Avaliação Educacional*, 20(43), 319–340.

Andriola, W. B. (2011). Twelve reasons favoring the adoption of the National High School Exam (ENEM) by Federal Institutions of Higher Education. Ensaio: *Avaliação e Políticas Públicas em Educação*, 19(70), 107–125.

Arikan, S., Özer, F., Seker, V., & Ertas, G. (2020). The importance of sample weights and plausible values in large-scale assessments. *Journal of Measurement and Evaluation in Education and Psychology*, 11(1), 43-60.

Attridge, N., & Inglis, M. (2013). Advanced mathematical study and the development of conditional reasoning skills. *PLOS ONE*, 8(7), e69399.

Bansilal, S. (2015). A Rasch analysis of a Grade 12 test written by mathematics teachers. *South African Journal of Science*, 111(5-6), 1-9.

Bartolucci, F., Bacci, S., & Gnaldi, M. (2014). MultiLCIRT: An R package for multidimensional latent class item response models. *Computational Statistics & Data Analysis*, 71, 971-985.

Bellini, D., Crescentini, A., Zanolla, G., Cubico, S., Favretto, G., Faccincani, L., Ardolino, P., & Gianesini, G. (2019). Mathematical Competence Scale (MCS) for primary school: The psychometric properties and the validation of an instrument to enhance the sustainability of talents development through the numeracy skills assessment. *Sustainability*, 11(9), 2569.

Bittmann, F. (2022). Investigating the co-development of academic competencies and educational aspirations in German primary education. *Frontiers in Education*, 7, 923361.

Bolondi, G., Branchetti, L., & Giberti, C. (2018). A quantitative methodology for analyzing the impact of the formulation of a mathematical item on students learning assessment. *Studies in Educational Evaluation*, 58, 37-50.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model*: fundamental measurement in the human sciences. New York: Taylor & Francis.

Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Cogent Education*, 4(1), 1416898.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht: Springer.

Brasil. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). History. Retrieved February 10, 2025, from https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem/historico.

Brasil. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). (2012). Technical note: Item Response Theory. Retrieved January 12, 2025, from https://download.inep.gov.br/educacao_basica/enem/nota_tecnica/2011/nota_tecnica_tri_enem_18012012.pdf.

Bulut, H. C., Bulut, O., & Clelland, A. (2025). A psychometric network analysis approach for detecting item wording effects in self-report measures across subgroups. *Field Methods*, 37(1), 20-36.

Cho, S.-J., Cohen, A. S., Kim, S.-H., & Bottge, B. (2010). Latent Transition Analysis With a Mixture Item Response Theory Measurement Model. *Applied Psychological Measurement*, 34(7), 483-504.

Cresswell, C., & Speelman, C. P. (2020). Does mathematics training lead to better logical thinking and reasoning? A cross-sectional assessment from students to professors. *PLOS ONE*, 15(7), e0236153.

De Ayala, R. J. (2022). *The theory and practice of item response theory* (2nd ed.). New York: Guilford Press.

DeMars, C. (2018). *Classical test theory and item response theory*. In The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development (pp. 49-73). Wiley Blackwell.

Ekstrand, J., Westergren, A., Årestedt, K. et al. (2022). Transformation of Rasch model logits for enhanced interpretability. *BMC Medical Research Methodology*, v. 22, n. 332.

Embretson, S. E., & Reise, S. P. (2013). Item response theory for psychologists. New York: Taylor & Francis.

Fischbach, A., Keller, U., Preckel, F., & Brunner, M. (2013). PISA proficiency scores predict educational outcomes. *Learning and Individual Differences*, 24, 63-72.

Forzani, E., Afflerbach, P., Aguirre, S., Brynelson, N., Cervetti, G., Cho, B. Y., ... & Uccelli, P. (2022). *Advances and missed opportunities in the development of the 2026 NAEP Reading Framework.* Literacy Research: Theory, Method, and Practice, 71(1), 153–189.

Gambini, A., Desimoni, M., & Ferretti, F. (2024). Predictive tools for university performance: An explorative study. *International Journal of Mathematical Education in Science and Technology*, 55(3), 691-717.

Golino, H., Shi, D., Christensen, A. P., Garrido, L. E., Nieto, M. D., Sadana, R., & Thiyagarajan, J. A. (2020). Investigating the performance of exploratory graph analysis and traditional techniques to identify the number of latent factors: A simulation and tutorial. *Psychological Methods*, 25(3), 292-320.

Gonçalves, F. B., Gamerman, D., & Soares, T. M. (2013). Simultaneous multifactor DIF analysis and detection in Item Response Theory. *Computational Statistics & Data Analysis*, 59, 144-160.

Gregory, R. J. (2010). *Psychological testing*: History, principles, and applications (6th ed.). New York: Pearson.

Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. Campbell Systematic Reviews, 18(2), e1230.

Haladyna, T.M., & Rodriguez, M.C. (2013). *Developing and Validating Test Items*. New York: Taylor & Francis.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309–333.

Hargreaves, A. (2020). Large-scale assessments and their effects: The case of mid-stakes tests in Ontario. *Journal of Educational Change*, 21(3), 393-420.

Hos, H. Y., & Saraç, M. U. (2025). How Does Incorporating the Response Times into Mixture Modelling Influence the Identification of Latent Classes for Mathematics Literacy Framework in PISA 2022? *Education and Science*, 50(1), 129-146.

Huggins-Manley, A. C. (2018). *Developments in latent variable measurement models*. In V. Aryadoust & M. Raquel (Eds.), Quantitative data analysis for language assessment: Fundamental techniques (pp. 48-72). New York: Routledge.

INEP - *Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira*. (2024). ENEM 2023 Microdata. Retrieved February 15, 2025, from https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem.

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology*, 7, 109.

Kieftenbeld, V., Natesan, P., & Eddy, C. (2011). An Item Response Theory analysis of the Mathematics Teaching Efficacy Beliefs Instrument. *Journal of Psychoeducational Assessment*, 29(5), 443-454.

Kier, M. W., Blanchard, M. R., Osborne, J. W., & Albert, J. L. (2014). The development of the STEM Career Interest Survey (STEM-CIS). *Research in Science Education*, 44(3), 461-481.

Kijima, R., & Lipscy, P. Y. (2024). The politics of international testing. *The Review of International Organizations*, 19(1), 1-31.

Li, G., & Wu, Y. (2023). Application of the Bookmark method: setting standard for the ninth-grade mathematics achievement test in China. *Current Psychology*, 42(33), 28941-28952.

Lin, C. H., & Chen, C. H. (2023). Development, testing, and application of a mathematics learning scale of self-direction. *Frontiers in Psychology*, 14, 1145442.

Linacre, J. M. (2012). A user's guide to WINSTEPS MINISTEP Rasch-model computer programs. Beaverton, Oregon: Winsteps.com.

Liu, C. W. (2025). Semi-parametric item response theory with O'Sullivan splines for item responses and response time. *Applied Psychological Measurement*, 49(4-5), 224-242.

Liu, X. F., & McKeough, A. (2005). Developmental growth in students' concept of energy: Analysis of selected items from the TIMSS database. *Journal of Research in Science Teaching*, 42(5), 493-517.

Lohbeck, A., Kocaj, A., Schneider, R., & Jansen, M. (2024). Testing the generalized internal/external frame of reference model among students with and without special educational needs in regular and special education schools. *Learning and Individual Differences*, 111, 102427.

Madison, M. J., Wind, S. A., Maas, L., Yamaguchi, K., & Haab, S. (2024). A One-Parameter Diagnostic Classification Model with Familiar Measurement Properties. *Journal of Educational Measurement*, 61(3), 408-431.

Martins, P. S. R., Barbosa-Pereira, D., Valgas-Costa, M., & Mansur-Alves, M. (2022). Item analysis of the Child Neuropsychological Assessment Test (TENI): Classical test theory and item response theory. *Applied Neuropsychology-Child*, 11(3), 339-349.

Monteiro, A. F. (2010). *Multiple choice test*: Construction, application and evaluation. Revista Brasileira de Educação, 15(44), 324–341.

Moon, J. A., Sinharay, S., Keehner, M., & Katz, I. R. (2020). Investigating technology-enhanced item formats using cognitive and item response theory approaches. *International Journal of Testing*, 20(2), 122-145.

Moreno, R., Martínez, R. J., & Muñiz, J. (2004). Guidelines for the construction of multiple choice items. *Psicothema*, 16(3), 490–497.

Muñiz, J. (2010). Test theories: Classical theory and item response theory. *Papeles del Psicólogo*, 31(1), 57–66.

Munns, M. E., He, C., Topete, A., & Hegarty, M. (2023). Visualizing cross-sections of 3D objects: Developing efficient measures using item response theory. *Journal of Intelligence*, 11(11), 205.

NAEP - *National Assessment of Educational Progress*. (2026). Mathematics framework for the 2026 National Assessment of Educational Progress. National Assessment Governing Board (U.S. Department of Education). Avaliable in: https://www.nagb.gov/naep-subject-areas/mathematics/2026-naep-mathematics-framework.html. Accessed: 01 fev. 2026.

Nortvedt, G. A. (2018). Policy impact of PISA on mathematics education: The case of Norway. *European Journal of Psychology of Education*, 33(3), 427-444.

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, 16(1), 1609406917733847.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372.

Peralta, Y., Kohli, N., Strom, A., Duranczyk, I., Mesa, V., & Watkins, L. (2020). A psychometric analysis of the Algebra and Precalculus Concept Readiness Assessment for Community Colleges. *Journal of Psychoeducational Assessment*, 38(3), 321-336.

Pickering, J., Attridge, N., Inglis, M., & Morsanyi, K. (2025). Numeracy, logical reasoning and real-world decision making. *Research in Mathematics Education*, 1-17.

Prieto, G., & Delgado, A. (2003). Analysis of a test using the Rasch model. Psicothema, 15(1), 94–100.

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the Cognitive Reflection Test applying Item Response Theory (IRT). *Journal of Behavioral Decision Making*, 29(5), 453-469.

R Core Team (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. Avaliable in: https://www.R-project.org Accessed: 18 fev. 2024.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological data in the 21st century. *Journal of Personality Assessment, 95(2), 129-140*.

Robitzsch, A. (2023). Exploring the Multiverse of Analytical Decisions in Scaling Educational Large-Scale Assessment Data: A Specification Curve Analysis for PISA 2018 Mathematics Data. *European Journal of Investigation in Health, Psychology and Education*, *12*(7), 731-754.

Santos, O., Stefanovska-Petkovska, M., Virgolino, A., Miranda, A. C., Costa, J., Fernandes, E., Cardoso, S., & Carneiro, A. V. (2021). Functional health literacy: Psychometric properties of the newest vital sign for Portuguese adolescents (NVS-PTeen). *Nutrients*, 13(3), 790.

Schadl, C., & Lindmeier, A. (2025). Preparing for digital learning monitoring in the fraction context: Assessment of students' prior knowledge according to evidence-based cognitive models. *International Journal of Science and Mathematics Education*, *23*(7), 2199-2224.

Shard, Kumar, D., & Koul, S. (2024). Investigating students' perception about LMS-based online examination practices. *International Journal of Information and Learning Technology*, 41(1), 36-60.

Slepkov, A. D., Van Bussel, M. L., Fitze, K. M., & Burr, W. S. (2021). A modified reliability index for multiple-choice tests: A synthesis of item and test indices. *Sage Open*, *11*(2), 21582440211016838.

Stites, N. A., Douglas, K. A., Evenhouse, D., Berger, E., Deboer, J., & Rhoads, J. F. (2019). A validation and differential item functioning (DIF) study of an abbreviated dynamics concept inventory. *International Journal of Engineering Education*, 35(2), 491-509.

Susongko, P., Wahab, N. B. A., Arfiani, Y., & Kusuma, M. (2024). Validation and Implementation of 3-Dimensional Scientific Literacy Test (Lisa3D Test): Measuring Scientific Literacy for Senior High School Students based on Scientific Reasoning, Scientific Inquiry, and Nature of Science. *Journal Pendidikan IPA Indonesia*, 13(3).

Susongko, P., Yuenyong, C., Kusuma, M., & Arfiani, Y. (2024). Developing the Nature of Science (NOS) Test Using COVID-19 Theme. *Pegem Journal of Education and Instruction*, 14(2), 57-66.

Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal*, 41(4), 901-926.

Thomas, N. (2000). Assessing model sensitivity of the imputation methods used in the national assessment of educational progress. *Journal of Educational and Behavioral Statistics*, 25(4), 351-371.

Tongchai, A., Sharma, M. D., Johnston, I. D., Arayathanitkul, K., & Soankwan, C. (2011). Consistency of students' conceptions of wave propagation: Findings from a conceptual survey in mechanical waves. *Physical Review Special Topics-Physics Education Research*, 7(2), 020101.

Toribio, S. G., & Albert, J. H. (2011). Discrepancy measures for item fit analysis in item response theory. *Journal of Statistical Computation and Simulation*, 81(10), 1345-1360.

Trevisan, A. L., & Amaral, R. G. do. (2016). The revised Bloom Taxonomy applied to assessment: A study of mathematics written tests. *Ciência & Educação (Bauru)*, 22(2), 451–464.

Urbina, S. (2004). *Essentials of psychological testing*. John Wiley & Sons, Inc.

Vecchiato, G., Susac, A., Margeti, S., Fallani, F. D., Maglione, A. G., Supek, S., Planinic, M., & Babiloni, F. (2013). High-resolution EEG analysis of power spectral density maps and coherence networks in a proportional reasoning task. *Brain Topography*, 26(2), 303-314.

Wang, T., Zhang, L., Xie, Z., & Liu, J. (2023). How does mathematical modeling competency affect the creativity of middle school students? The roles of curiosity and guided inquiry teaching. *Frontiers in Psychology*, 13, 1044580.

Wang, Z. (2020). When large-scale assessments meet data science: The big-fish-little-pond effect in fourth- and eighth-grade mathematics across nations. *Frontiers in Psychology*, 11, 579545.

Wei, T., Chesnut, S. R., Barnard-Brak, L., Stevens, T., & Olivárez, A. (2014). Evaluating the Mathematics Interest Inventory using Item Response Theory: Differential item functioning across gender and ethnicities. *Journal of Psychoeducational Assessment*, 32(8), 747-761.

Wilson, M. (2023). *Constructing Measures*: An Item Response Modeling Approach. New York: Taylor & Francis.

Wind, S. A. (2017). An instructional module on Rasch measurement theory. *Educational Measurement*, 36(2), 50-66.

Wong, T. T. Y. (2018). Is conditional reasoning related to mathematical problem solving? *Developmental Science*, 21(5), e12644.

Wu, X., Li, N., Wu, R., & Liu, H. (2025). Cognitive analysis and path construction of Chinese students' mathematics cognitive process based on CDA. *Scientific Reports*, 15(1), 4397.

Yu, M., Cui, J., Wang, L., Gao, X., Cui, Z., & Zhou, X. (2022). Spatial processing rather than logical reasoning was found to be critical for mathematical problem-solving. *Learning and Individual Differences*, 100, 102230.

Zhang, W.-X., Hsu, Y.-S., Ho, Y.-T., & Liu, C.-C. (2023). Effects of a SSI-Based Learning Module on Students' Decision-Making Abilities. *Journal of Research in Education Sciences*, 68(1), 167-198.

Zoanetti, N., Griffin, P., Beaves, M., & Wallace, E. M. (2009). Rasch scaling procedures for informing development of a valid Fetal Surveillance Education Program multiple-choice assessment. *BMC Medical Education*, 9(1), 20.