Scientific Research Publishing

Creative Education, 2025, 16(6), 932-946
https://www.scirp.org/journal/ce
ISSN Online: 2151-4771
ISSN Print: 2151-4755

Assessing the Quality of Examination Questions in Medical Education: A Classical Test and Item Response Theory Approach in a Morphology Course

Ângela Tavares Paes*, Danielle Tamashiro Duarte, Natália Oliveira Feitosa, Marcella M. Ceratti, Felipe Prieto Siqueira, Pedro Afonso Liberato, Carlos Augusto Cardim de Oliveira

Faculdade Israelita de Ciências da Saúde Albert Einstein, Hospital Israelita Albert Einstein, São Paulo, Brazil
Email: *angela.tpaes@einstein.br

How to cite this paper: Paes, Â. T., Duarte, D. T., Feitosa, N. O., Ceratti, M. M., Siqueira, F. P., Liberato, P. A., & de Oliveira, C. A. C. (2025). Assessing the Quality of Examination Questions in Medical Education: A Classical Test and Item Response Theory Approach in a Morphology Course. *Creative Education, 16,* 932-946.
https://doi.org/10.4236/ce.2025.166058

Received: May 26, 2025
Accepted: June 27, 2025
Published: June 30, 2025

Abstract

Introduction: One of the main challenges in educational courses in medical schools is the development of tests which are able to measure the proficiency of the students enrolled in an accurate manner. Despite the advances in the educational system, conventional methods of testing, such as the sum of points from multiple choice questions, are still commonplace and professors seldom focus on the quality of the questions contained within the examination. **Objectives:** The purpose of this study is to foster a discussion on the quality of exams in medical education, using a psychometric analysis of tests from a Morphology course as an example. **Methods:** Four examinations (tests) from the Morphology course in the first year of a private medical school were analyzed. The questions therein were all in a multiple-choice model, assessing basic concepts of anatomy, radiology and pathology. Techniques from Classical Test Theory (CTT) and Item Response Theory (IRT) using the mirt package were applied to evaluate the effectiveness of the questions in measuring students' learning outcomes. **Results:** In the analyzed examinations, a discrepancy in the distribution of question difficulty was observed, with a predominance of easy questions, contrasted by a scarce or entirely absent presence of difficult ones. Additionally, a considerable proportion of questions exhibited a high likelihood of being answered correctly by chance. Overall, the discrimination rates of the examination questions were low (below 20%), with none approaching 100%. Moreover, a great number of questions displayed discrimination rate near zero, a value which disfavors the differentiation of students with both greater and lesser proficiency in the subjects examined. **Conclusions:** Although CTT and IRT are well-established and effective methods for assessing the quality of test

questions, they are seldom utilized to guide educators in the development of more equitable examinations. These methods should be employed to assist professors in enhancing item development and the creation of more effective assessments, ensuring a better balance between easy, moderate, and difficult questions.

Keywords

Educational Evaluation, Academic Success, Medical Education, Quality of Tests, Proficiency Measure, Classical Test Theory (CTT), Item Response Theory (IRT)

1. Introduction

There is a plethora of discussions in medical education regarding pedagogical approaches, educational platforms, technologies, and the educators' professional development. However, there is a notable absence of discourse surrounding the evaluation systems themselves, with low-quality assessments often used as the basis for determining whether students have acquired sufficient knowledge to pass a course.

There are several studies focusing on measuring students' performances and their ability to recall what was taught, but there still does not seem to be a consensus on the best method of measurement (Zgheib et al., 2011; Zeferino & Passeri, 2007). Regardless of the subject or analytical method used, numerous examinations suffer from poor quality, often exhibiting issues such as low discrimination rates, a high percentage of correct answers due to chance (random success), an uneven distribution of difficulty levels, among other limitations (De Champlain, 2010). Knowledge is not only the pure and simple transference of information, but a solid construction that includes critical thinking, professional posture and ethics in the construction of ideas and actions that can be useful and profitable to people (Farias et al., 2015). In this context of necessary improvement, it is essential to employ tools that assess whether the tests administered throughout the course effectively measure students' proficiency.

One of the challenges faced by the teaching staff in a medical educational institution is the design of assessments, which include selecting the most relevant topics, formulating well-structured questions, determining the appropriate format (multiple-choice or open-ended), and deciding whether the test will be administered individually or in groups (Rosso & Taglieber, 1992). Throughout the semesters, professors gradually modify the design of their questions according to their perceptions and feedback from the students themselves. Questions are reformulated in order to maintain the clarity of the texts, evaluate the depth of the content requested and to ponder whether it succeeded in measuring the knowledge of the students.

The lack of quality in a question may arise from various factors, such as ambig-

uous wording, conceptual errors which stem from shortcomings in content delivery during lectures, or differences in the interpretation of the items. Moreover, there are also questions with a low difficulty rate, which themselves fail to differentiate students with greater or lesser proficiency in the subjects covered. Furthermore, the possibility of answering a question correctly by chance can also mask to what extent a student truly knows about a subject (Vendramini et al., 2004; Andrade et al., 2000). It is also expected that an examination could provide a reliable way to quantify the student's knowledge regarding a discipline, or regarding a branch of study as a whole. The classical way of obtaining test scores (grades) is merely by the sum of points of correct answers, with the complexity and relative weight of each question given subjectively by the teacher. This means that the grade could not accurately and/or precisely exhibit the student's knowledge (Andrade et al., 2000; Pasquali, 2009; Thorpe & Favia, 2012).

The evaluation of items has received the attention of researchers within the educational fields for several years. The two most used analysis methodologies are the Classical Test Theory (CTT) and the Item Response Theory (IRT). CTT has certain limitations, including its inherent restriction by merely comparing a student's scores with those obtained on the same test or similar assessments. Another significant limitation is that the results of the analysis may vary depending on the sample, meaning they are influenced by the group of respondents rather than solely by the test itself. Additionally, a notable drawback of CTT is its scoring system. In CTT, the final grade is determined by the total sum of points. Consequently, in tests where all questions carry the same weight, two students who answer the same number of questions correctly will receive identical scores, even if their levels of skill and knowledge differ. The simple summation of points fails to differentiate students who provide more coherent responses or correctly answer more complex questions from the rest of the students (Pasquali, 2009; Thorpe & Favia, 2012).

CTT emphasizes the total score of the test, whereas IRT centers on the individual items (questions). A student's proficiency is assessed using a statistical model that incorporates three key parameters: the discrimination index of the items, the difficulty level, and the probability of answering correctly by chance (Andrade et al., 2000). Once these parameters are estimated (in a process known as "calibration"), it becomes possible to calculate the proficiency of individuals across different tests (Hair Jr. et al., 2005). This represents a significant advantage, particularly in fields such as medical sciences, which aim to advance educational technologies while also integrating and evolving the accumulated knowledge of medical students throughout the period in which they are enrolled (Costa, 2010).

Both CTT and IRT offer critical indicators that contribute to the development of assessments that effectively align with the taught content and yield grades that more accurately reflect students' proficiency. The widespread adoption of these methodologies by educators could provide valuable insights to assist in the creation of more reliable tests for evaluating student learning.

The aim of this paper is to promote a discussion on a fairer and more accurate

approach to the formulation of examination questions, as opposed to the methods currently employed. This will be demonstrated through the application of CTT and IRT in the analysis of multiple-choice questions from tests administered to students in a one-semester Morphology course. The methodology of the present study, albeit restricted to one subject, aims to support the discourse among any subjects in medical education.

2. Methods

2.1. Participants and Ethical Considerations

This study is part of a project that involves all first-year disciplines of medical graduation at *Faculdade Israelita de Ciências da Saúde Albert Einstein*, a private school of medicine in São Paulo, Brazil. The main project was approved by the Research Ethics Committee under number 51413821.6.0000.0071. In this study, only examinations containing multiple-choice questions of the Morphology discipline were considered. The examinations were applied to students in the first semester of 2019. A total of 4 theoretical tests (containing a total of 125 questions) were analyzed.

The choice of this subject is due to the fact that the university has already undergone a three-year adaptation phase. As a result, the professors had the opportunity to adjust their assessments based on their experience with previous cohorts of students.

The data was extracted from the Canvas platform (Hospital Israelita Albert Einstein, 2022) with the consent of the professor responsible for the course. Only the grades of the 45 students who signed an informed consent form were analyzed. At no point during the project were the names of students or faculty members disclosed.

2.2. Dataset

Four assessments from the Morphology I course, administered during the first semester of 2019, were analyzed. The tests contained 25, 25, 25, and 50 multiple-choice questions, respectively, covering fundamental concepts in anatomy, radiology, and pathology.

For each test (1, 2, 3, and 4), a database was created in Excel, with each question assigned a numerical identifier within the respective test (e.g., "test1_item1"). This variable was recorded in two formats: polytomous, indicating the letter corresponding to the correct answer, and dichotomous, denoting whether the answer was correct or incorrect. Additionally, a final grade, representing the total points obtained on each test, was also included.

2.3. Data Analysis and Theoretical Approach

For the purposes of data analysis, participants' responses to the multiple-choice questions—each comprising a single correct alternative—were examined. Following the identification of the correct responses, the responses were converted into

dichotomous items, coded as 0 (incorrect) or 1 (correct).

The psychometric properties were analyzed by both CTT and IRT. Within the CTT framework, the following indices were calculated: item difficulty (proportion of correct responses), discrimination index, biserial correlation, and Cronbach's alpha. For the IRT analysis, a three-parameter logistic (3PL) model was employed, in which item parameters include difficulty, discrimination, and guessing.

All statistical analyses were conducted using the R statistical software (R Core Team, 2024) and the mirt package for Multidimensional Item Response Theory, also within the R environment (Chalmers, 2012).

2.3.1. Classical Test Theory (CTT)

In the CTT model, item difficulty is determined by the proportion of correct responses and can be used to classify items into five categories: very easy (more than 90% correct responses), easy (70% - 90%), moderate (30% - 70%), difficult (10% - 30%), and very difficult (less than 10%) (Vilarinho, 2015). In addition to the difficulty parameter, the discrimination parameter is also employed to assess item quality. Items with discrimination values below 20% may require revision, as they demonstrate limited ability to distinguish between higher- and lower-performing individuals. Conversely, values equal to or greater than 40% are indicative of high-quality items with strong discriminatory power (Condé, 2001; Pasquali, 2009; Vilarinho, 2015).

2.3.2. Item Response Theory (IRT)

In the context of IRT, the 3PL model for an assessment comprising I items and n students is represented by Equation (1):

$$P\left(U_{ij}=1\middle|\theta_j\right)=c_i+\frac{1-c_i}{1+\exp\left[-a_i\left(\theta_j-b_i\right)\right]} \tag{1}$$

with $i=1, 2, \ldots, I$ and, $j=1, 2, \ldots, n$, in which:

U_{ij} is a dichotomous variable which is equal to 1 when student j answers question i correctly and equal to 0 when student j answers question i incorrectly.

θ_j represents the ability/latent trait of the j-th individual (i.e., the individual's grade).

b_i represents the difficulty of question i, in the same scale of the ability. It represents a threshold parameter for the necessary ability to answer the question correctly.

a_i represents the discrimination of the question i (i.e., the rate of change in the probability of answering question i correctly in regards to the ability; the slope of the probability function). Moreover, the value of the parameter is proportional to the derivative of the probability function $P(U_{ij}=1|\theta_j)$ at $\theta_j=b_i$.

c_i represents the casual probability of answering correctly question i (i.e., the probability of guessing the correct answer). Moreover, the value of the parameter is the asymptotic minimum of the probability function $P(U_{ij}=1|\theta_j)$.

$P(U_{ij}=1|\theta_j)$ is the probability of the j-th individual with ability θ_j answering

question *i* correctly.

These parameters are used to generate a probability function for each item, represented graphically by the Item Characteristic Curve (ICC) (Andrade et al., 2000; Costa, 2010; Gyamfi & Acquaye, 2023). An example of such a curve is presented in **Figure 1**:

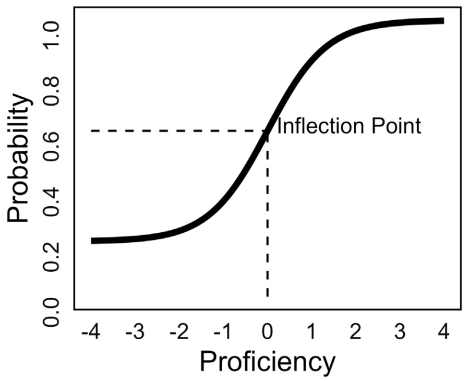

Figure 1. Item Characteristic Curve (ICC) for a test item with a discrimination parameter of 1.5, a difficulty parameter of 0, and a guessing parameter of 0.2. From Wells (2021).

3. Results

The distribution of student scores across the four assessments is presented in **Figure 2**. Scores ranged from 54 to 100, with the majority exceeding 70% (the minimum threshold required to pass the course). Greater variability was observed in Test 4, which consisted of 50 items, in contrast to the 25 items included in the other assessments. The larger number of items positively influenced the test's reliability, as reflected by the highest Cronbach's alpha coefficient observed in Test 4 (0.815).

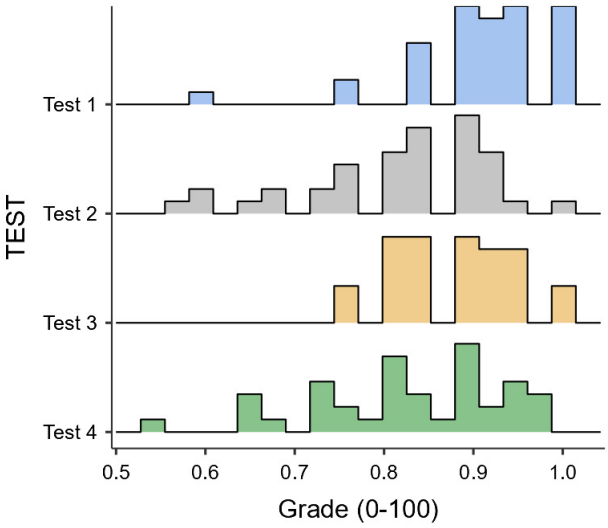

Figure 2. Distribution of test scores based on Classical Test Theory (CTT), represented as the sum of item scores on a 0 - 100 scale.

Table 1. Summary of key CTT metrics for the four assessments.

	Test 1 (25 questions)	Test 2 (25 questions)	Test 3 (25 questions)	Test 4 (50 questions)
Overall Cronbach's alpha	0.5898	0.5590	0.3675	0.8150
N questions with 100% correct answers	5 (20%)	5 (20%)	8 (32%)	7 (14%)
Biserial correlation (BC)				
Number of questions with BC ≥ 0.5	6	2	1	6
Number of questions with BC < 0.5	14	23	16	37
Facility Index (FI)				
Very easy (FI > 0.9)	17	13	14	19
Easy (FI between 0.7 and 0.9)	7	9	7	24
Moderate (FI between 0.3 and 0.7)	1	3	4	7
Difficult/Very difficult (FI < 0.3)	-	-	-	-
Discrimination index (DI)				
Good discrimination rate (DI ≥ 0.4)	3	3	2	12
Good discrimination rate, with room for improvement (ID between 0.3 and 0.4)	2	5	6	5
Moderate discrimination rate (between 0.2 - 0.3)	4	5	2	7
Poor discrimination rate (DI < 0.2)	16	12	15	26

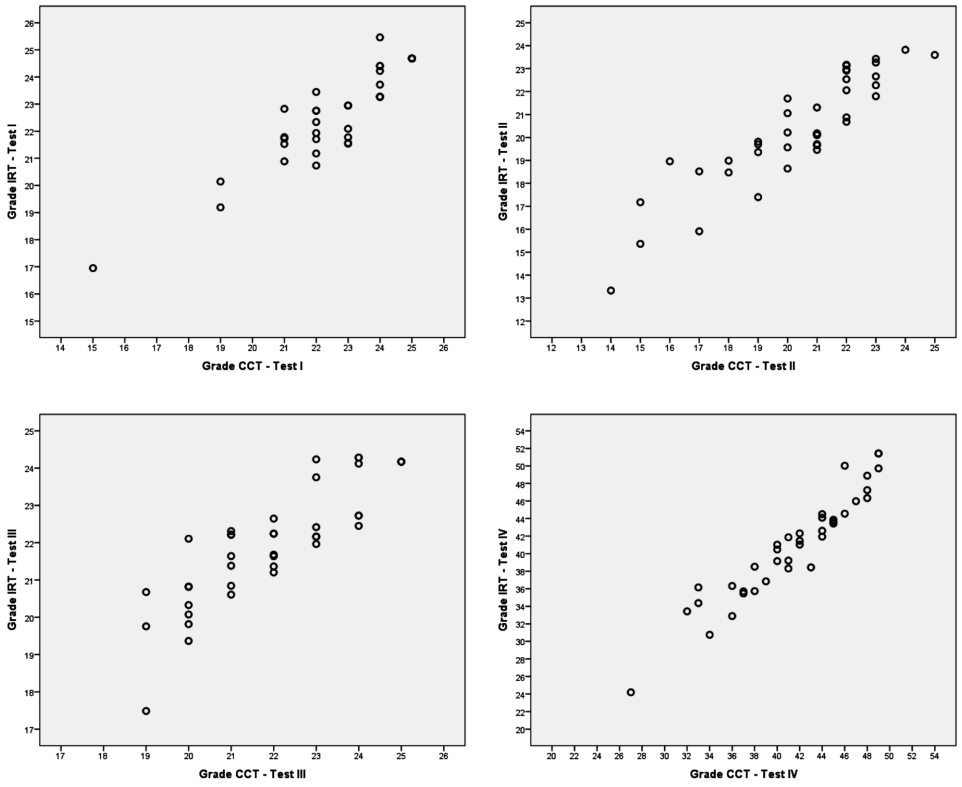

Figure 3. Scatterplot of student scores comparing IRT and CTT methods across the four assessments. Pearson's r correlation coefficients are: 0.913 (Test 1), 0.899 (Test 2), 0.855 (Test 3), and 0.951 (Test 4).

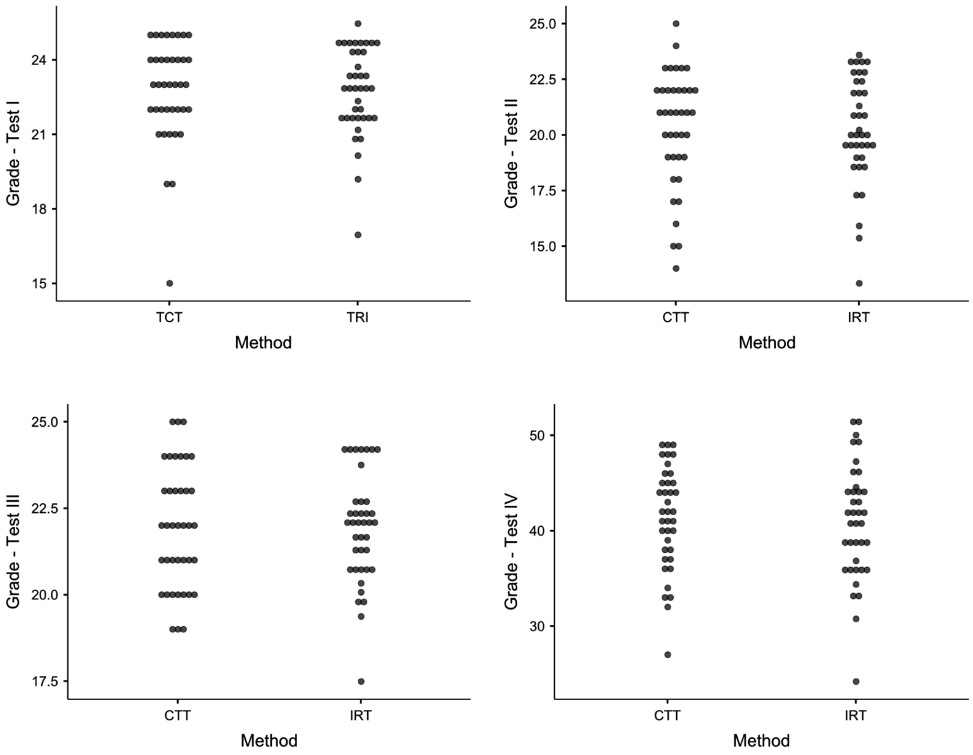

Figure 4. Scatterplots of grades by CTT and IRT.

Table 1 presents a summary of key CTT metrics, including the difficulty (or facility) index, Cronbach's alpha coefficient, biserial correlation coefficient, and discrimination index. Based on the biserial correlation values, no item approached the ideal threshold of 1.0, suggesting a weak relationship between individual item performance and overall test scores. According to the established difficulty classification by Condé (2001), all assessments contained a substantial proportion of items classified as "very easy," including some with a 100% correct response rate. Notably, no items were categorized as "difficult" or "very difficult," which deviates from the recommended structure of a well-balanced test—ideally comprising a mix of easy, moderate, and difficult items.

When comparing the scores derived from CTT and IRT, a strong correlation is observed (Figure 3). However, IRT demonstrates superior discriminatory capacity in distinguishing among students of varying performance levels (Figure 4).

Table 2. CTT and IRT metrics for Test 1.

Question*	CTT-related coefficients				IRT parameters		
	Cronbach's alpha if item excluded (overall 0.5898)	Biserial correlation	Facility rate	Discrimination rate	Discrimination parameter	Difficulty parameter	Guessing parameter
1	0.59	0.29	0.87	0.125	0.270	−7.18	0.020
2	0.54	0.54	0.85	0.375	1.600	−1.48	0.002
3			1.00	0			
4	0.60	0.03	0.95	0.063	−0.720	4.02	0.190

Continued

5	0.57	0.37	0.90	0.188	0.450	−5.00	0.010
6	0.55	0.53	0.92	0.250	1.930	−1.93	0.002
7			1.00	0			
8	0.55	0.53	0.92	0.250	2.610	−1.71	0.001
9			1.00	0			
10	0.59	0.13	0.97	0.125	5.340	−0.51	0.910
11	0.60	0.14	0.95	0.125	0.250	−10.40	0.290
12	0.63	0.19	0.74	0.188	6.630	1.09	0.700
13			1.00	0			
14	0.59	0.29	0.87	0.375	7.690	−0.26	0.680
15	0.55	0.62	0.97	0.125	8.500	−2.23	0.002
16	0.59	0.14	0.95	0.125	0.120	−23.67	0.160
17	0.58	0.43	0.59	0.500	1.460	−0.33	0.001
18			1.00	0			
19	0.60	0.11	0.92	0.125	9.940	−0.04	0.840
20	0.54	0.58	0.92	0.250	3.600	−1.55	0.001
21	0.58	0.30	0.97	0.125	10.630	−0.95	0.840
22	0.60	0.06	0.97	0	0.030	−129.25	0.150
23	0.54	0.56	0.79	0.563	1.470	−1.24	0.001
24	0.55	0.48	0.92	0.250	1.660	−2.10	0.003
25	0.58	0.41	0.76	0.625	11.500	0.43	0.650

*Items with 100% correct responses are highlighted; for these items, certain coefficients could not be calculated.

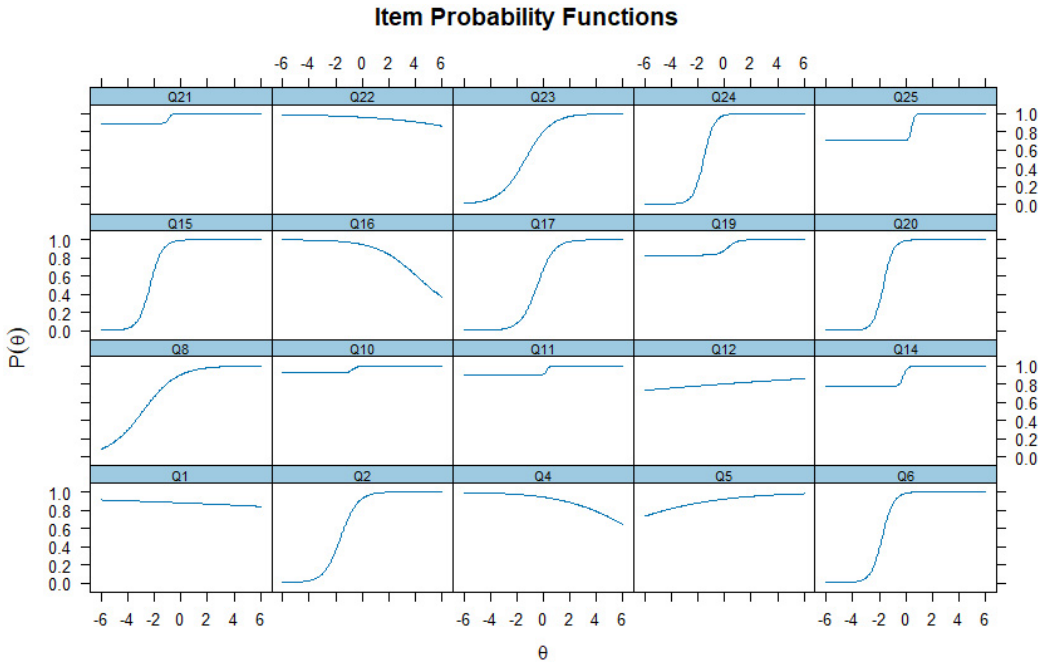

Figure 5. Item Characteristic Curves (ICCs) for several questions in Test 1, based on the Item Response Theory (IRT) 3PL model. Items with 100% correct responses are unable to be displayed.

Table 2 and Figure 5 present the Item Characteristic Curves (ICCs) derived from the 3PL IRT model fitted to Test 1. Based on these results, items 2, 6, 8, 17, 23, and 24 exhibited the most favorable combination of parameters: low guessing probability, high discrimination, and a clear positive relationship between student proficiency and the probability of a correct response (Baker & Kim, 2017). These items demonstrate greater capacity to accurately assess student knowledge and should be prioritized for inclusion in item banks. However, in contrast, the remaining items showed limited discriminatory power, with similar probabilities of correct responses among students regardless of their proficiency levels.

Table 3. CTT and IRT metrics for Test 4.

Question*	CTT-related coefficients			IRT parameters			
	Cronbach's alpha if item excluded (overall 0.8150)	Biserial correlation	Facility rate	Discrimination rate	Discrimination parameter	Difficulty parameter	Guessing parameter
1			1.00	0			
2	0.7979	0.706	0.76	0.750	6.78	−0.84	0.128
3			1.00	0			
4	0.8158	0.106	0.95	0.083	0.16	−17.06	0.154
5	0.8109	0.446	0.97	0.083	5.16	−2.49	0.017
6	0.8129	0.291	0.97	0.083	4.25	−1.31	0.832
7	0.8094	0.426	0.92	0.167	1.50	−2.45	0.021
8	0.8089	0.484	0.95	0.167	5.16	−1.96	0.014
9			1.00	0			
10	0.8139	0.224	0.92	0.167	2.51	−0.07	0.851
11	0.8168	−0.018	0.97	0	−0.17	19.87	0.144
12			1.00	0			
13	0.8106	0.368	0.81	0.333	1.63	−0.56	0.523
14	0.8231	0.006	0.76	0.083	−0.18	4.67	0.197
15	0.8104	0.372	0.84	0.333	6.78	−0.31	0.631
16			1.0	0			
17	0.8077	0.499	0.92	0.250	4.95	−1.72	0.015
18	0.8097	0.395	0.87	0.250	1.41	−1.95	0.039
19	0.8102	0.380	0.81	0.417	8.56	−0.29	0.577
20	0.8160	0.043	0.97	0	0.25	−14.11	0.147
21	0.8183	0.184	0.31	0.167	5.55	1.38	0.230
22	0.8134	0.239	0.95	0.083	0.77	−3.74	0.325
23	0.8096	0.418	0.52	0.583	4.23	0.74	0.371
24	0.8072	0.470	0.81	0.417	1.31	−1.65	0.020
25	0.8025	0.583	0.71	0.667	1.89	−0.83	0.067
26	0.8106	0.365	0.89	0.167	1.06	−2.64	0.029
27	0.8079	0.449	0.74	0.500	5.87	−0.12	0.470
28	0.8024	0.584	0.68	0.667	1.61	−0.81	0.017

Continued

29	0.8150	0.205	0.87	0.250	4.51	0.02	0.757
30	0.8133	0.259	0.97	0.083	5.17	−0.73	0.920
31	0.8112	0.344	0.84	0.250	0.96	−2.13	0.048
32	0.8171	0.163	0.81	0.083	0.58	−2.71	0.034
33	0.8145	0.236	0.84	0.167	2.36	0.63	0.777
34	0.8174	0.116	0.87	0	0.32	−5.31	0.169
35	0.8130	0.336	0.47	0.333	5.98	0.79	0.314
36	0.8120	0.358	0.42	0.583	3.93	1.08	0.303
37	0.8186	0.173	0.31	0.250	2.27	1.79	0.255
38	0.8026	0.589	0.76	0.500	3.60	−0.40	0.422
39	0.8142	0.234	0.87	0.250	0.82	0.02	0.757
40	0.8151	0.227	0.81	0.250	8.59	0.05	0.650
41	0.8122	0.300	0.89	0.333	8.59	−0.65	0.675
42	0.8122	0.306	0.95	0.167	6.12	−1.13	0.730
43	0.8026	0.616	0.84	0.500	7.19	−0.96	0.334
44			1.0	0			
45			1.0	0			
46	0.8104	0.379	0.76	0.500	1.83	−0.05	0.541
47	0.8056	0.169	0.76	0.083	0.21	−5.13	0.076
48	0.8097	0.408	0.65	0.417	0.73	−0.90	0.044
49	0.8114	0.336	0.87	0.167	0.54	−3.90	0.042
50	0.8050	0.556	0.87	0.333	1.57	−1.89	0.021

*Items with 100% correct responses are highlighted; for these items, certain coefficients could not be calculated.

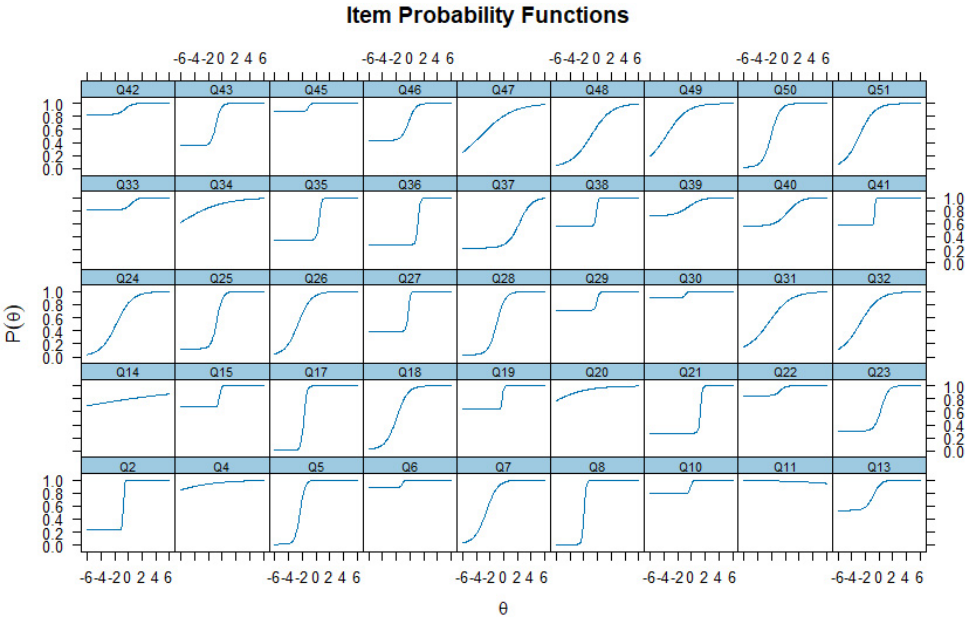

Figure 6. Item Characteristic Curves (ICCs) for several questions in Test 4, based on the Item Response Theory (IRT) 3PL model. Items with 100% correct responses are unable to be displayed.

The same analysis can be made regarding Test 4 in both Table 3 and Figure 6. Regarding all IRT parameters (difficulty, discrimination and guessing) shown by the ICC, questions 2, 5, 7, 8, 17, 18, 24, 25, 26, 28, 31, 32, 48, 50, 51 are the most adequate ones.

4. Discussion

Among the four assessments, Test 4 demonstrated the most balanced distribution of item difficulty, with an appropriate mix of easy, moderate, and difficult questions. This improved distribution may be attributed to the larger number of items, as Test 4 was the only exam to include 50 questions—twice as many as the other tests.

When evaluating the items with the highest CTT discrimination indices, a consistent pattern emerged: all tests featured at least two highly discriminating questions related to histology. Moreover, histology items frequently appeared among the most difficult questions according to the CTT difficulty index. Notably, two of the most challenging items across all four tests involved the identification of anatomical structures in images.

A comparison between highly discriminating and difficult items revealed that the most difficult questions were rarely the most discriminating. This is likely because items that are excessively difficult and answered correctly by very few students fail to distinguish effectively between different levels of proficiency.

In general, items that required straightforward recall tended to be rated as more difficult under CTT analysis but appeared less difficult based on the IRT curves. Some exceptions to this trend were observed, which may be explained by the repetition of certain key questions during morphology courses. These items are often included in assessments due to their relevance to medical residency exams, future coursework, or clinical practice, making them familiar and easier for students. Despite their lower difficulty, their inclusion remains justified, as they represent essential knowledge for medical education.

Interestingly, not all items identified as high-quality by the IRT model—whether due to strong discrimination, appropriate difficulty, or low guessing probability—were similarly rated by CTT. This discrepancy arises because CTT evaluates each parameter in isolation, while IRT integrates difficulty, discrimination, and guessing into a single item characteristic curve. As a result, IRT provides a more comprehensive assessment of item performance, making it easier to determine whether an item is suitable for evaluating student knowledge and worth retaining in a test item bank.

Regarding student scores, the IRT-based grading appears to offer a fairer evaluation, as it better differentiates between students with higher proficiency and those who may have answered more items correctly by chance. This supports the use of IRT for more precise and informative assessment in medical education.

4.1. Limitations

The sample size of this study was low due to the limitation to only one class and

the small number of students who agreed to participate in the study. In the future, a larger study, with more classes and more disciplines involved, may give a better idea about the benefits of IRT for student assessment and test calibration. Additional analyses have been conducted with other classes and subjects, although limited to the data that was available by the ethics committee of the institution. Nevertheless, the numerical results presented in our article are merely illustrative and intended to highlight the importance of encouraging educators to examine the quality of the questions they develop. Moreover, there was no aim in drawing inferences with the results, all of which reflect a very particular case.

Another limitation of this study is that some questions are repeated from previous years. If the students in the class had access to these questions previously, maybe they were pointed out as easy, not because of the intrinsic difficulty of the question, but because they were questions already known by the students.

A few aspects of the study design may have influenced the findings. The sample size was relatively small, as it only included students from a single class, with participation based on voluntary consent. While the results provide meaningful preliminary insights, future research involving larger and more diverse student groups—across multiple classes and subject areas—may offer a more comprehensive understanding of the usefulness of IRT for assessment and test calibration.

Furthermore, the use of the Morphology course as the basis for the study design served solely as an example. This course has unique characteristics that set it apart from other contexts. Nevertheless, it illustrates the widespread issue of poorly constructed exam questions and underscores how Item Response Theory (IRT) can more effectively differentiate student performance. These considerations are highly relevant and applicable across various disciplines and fields of study.

In addition, some assessment items were reused from previous years. Although this practice is common in educational settings, prior exposure to these questions could have affected how students perceived their difficulty. As a result, certain items may have appeared easier due to familiarity rather than a lower level of challenge. Still, these items are often retained in assessments due to their relevance to foundational medical knowledge; hailed as indispensable.

4.2. Supporting the Application of CTT and IRT

Both approaches are well-supported by the literature in education (Thorpe & Favia, 2012; Pasquali, 2009; Vilarinho, 2015; Andrade et al., 2000), and their implementation depends on the educator's goals and available resources. CTT, being a simpler method, is often easier to apply. For instance, the Canvas platform (used in the institution where this study was conducted) automatically generates CTT-related metrics for multiple-choice questions.

In contrast, applying IRT involves more complex procedures, particularly in estimating item parameters and generating Item Information Curves (IICs), which require the use of statistical software. In this study, the *mirt* package (Chalmers, 2012) was used for such analyses, but the R statistical environment offers many

other specific packages for IRT analysis, such as *Rirt*, *Birtr*, *shortIRT*, *cacIRT*, *Cirt*, *D3mirt*, *eirm*, *em IRT*, *Hirt*, *irtoys*, *irtQ*, *irtrees*, *KernSmoothIRT* and *pcIRT*. Each package includes a reference manual with detailed guidance, a general description, example applications, and additional bibliography.

Additionally, reading other scientific articles that apply IRT can help educators become more familiar with the terminology and typical formats of presenting results. Recent studies have demonstrated the effective application of IRT in educational contexts, as shown in the studies by Salele et al. (2025) and Esomonu et al. (2025). Although these examples are not about medical education, they offer some orientations for educators seeking to enhance their assessment practices.

5. Conclusion

Although medical education has advanced significantly in recent years, there remains a gap in research focused on the effectiveness of test evaluation methods. More detailed quantitative analyses of student performance data can offer valuable insights for educators—supporting improvements not only in instructional approaches but also in the design and calibration of assessment tools. The integration of methodologies such as CTT and IRT has the potential to enhance the quality of evaluations, as demonstrated in large-scale educational assessments such as Brazil's ENEM (National High School Exam) (Condé, 2001).

The findings of this study support prior hypotheses regarding the limited efficiency of multiple-choice questions when used without careful validation. Nonetheless, this issue can be mitigated through the application of established psychometric techniques. Both CTT and IRT offer useful frameworks for identifying and refining test items, allowing for the creation of calibrated question banks that facilitate the development of more effective and reliable assessments.

By adopting these approaches, educators can design exams with improved item congruence and discriminatory power, leading to a more accurate measurement of students' true learning and knowledge acquisition.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Andrade, D. F., Tavares, H. R., & Valle, R. C. (2000). *Teoria de Resposta ao Item: Conceitos e Aplicações.* Associação Brasileira de Estatística.

Baker, F. B., & Kim, S.-H. (2017). *The Basics of Item Response Theory Using R.* Springer. https://doi.org/10.1007/978-3-319-54205-8

Chalmers, R. P. (2012). MIRT: A Multidimensional Item Response Theory Package for The*R* Environment. *Journal of Statistical Software, 48,* 1-29. https://doi.org/10.18637/jss.v048.i06

Condé, F. N. (2001). *Análise Empírica de Itens [Technical Report].* Instituto Nacional de Estudos e Pesquisas Educacionais—DAEB/INEP/MEC.

Costa, M. C. (2010). *Aplicando a Teoria de Resposta ao Item a Dados Psicométricos.*

Universidade Federal do Rio de Janeiro.

De Champlain, A. F. (2010). A Primer on Classical Test Theory and Item Response Theory for Assessments in Medical Education. *Medical Education, 44,* 109-117. https://doi.org/10.1111/j.1365-2923.2009.03425.x

Esomonu, N. P. M., & Anayo, O. I. (2025). *Development, Standardization and Bench Mark of Mathematics Proficiency Test for Senior Secondary School Students Using Item Response Theory.* UNIZIK.

Farias, P. A. M. d., Martin, A. L. d. A. R., & Cristo, C. S. (2015). Aprendizagem Ativa na Educação em Saúde: Percurso Histórico e Aplicações. *Revista Brasileira de Educação Médica, 39,* 143-150. https://doi.org/10.1590/1981-52712015v39n1e00602014

Gyamfi, A., & Acquaye, R. (2023). Parameters and Models of Item Response Theory (IRT): A Review of Literature. *Acta Educationis Generalis, 13,* 68-78. https://doi.org/10.2478/atd-2023-0022

Hair Jr., J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (2005). *Análise Multivariada de Dados.* Bookman.

Hospital Israelita Albert Einstein (2022). *Plataforma Canvas.* FICSAE. https://einstein.instructure.com/login/canvas

Pasquali, L. (2009). *Psicometria: Teoria dos Testes na Psicologia e Educação.* Vozes.

R Core Team (2024). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Rosso, A. J., & Taglieber, J. E. (1992). Métodos Ativos e Atividades de ensino. *Perspectiva, 10,* 37-46.

Salele, N., Shahadat Hossain Khan, M., Hasan, M., & Ali, S. (2025). Advancing Four-Tier Diagnostic Assessments: A Novel Approach to Mapping Engineering Students' Conceptual Understanding in Microwave Engineering Course. *IEEE Access, 13,* 59886-59910. https://doi.org/10.1109/access.2025.3555432

Thorpe, G. L., & Favia, A. (2012). *Data Analysis Using Item Response Theory Methodology: An Introduction to Selected Programs and Applications [Technical Paper No. 20].* Psychology Faculty Scholarship, University of Maine.

Vendramini, C. M. M., Silva, M. C. D., & Canale, M. (2004). Análise de itens de uma prova de raciocínio estatístico. *Psicologia em Estudo, 9,* 487-498. https://doi.org/10.1590/s1413-73722004000300017

Vilarinho, A. P. (2015). *Uma Proposta de Análise de Desempenho dos Estudantes e de Valorização da Primeira Fase da OBMEP.* Master's Thesis, Universidade de Brasília.

Wells, C. S. (2021). Item Response Theory. In C. S. Wells (Ed.), *Assessing Measurement Invariance for Applied Research* (pp. 108-160). Cambridge University Press. https://doi.org/10.1017/9781108750561.004

Zeferino, A. M., & Passeri, S. M. (2007). Avaliação da Aprendizagem do Estudante. *ABEM, 3,* 39-43.

Zgheib, N. K., Simaan, J. A., & Sabra, R. (2011). Using Team-Based Learning to Teach Clinical Pharmacology in Medical School: Student Satisfaction and Improved Performance. *The Journal of Clinical Pharmacology, 51,* 1101-1111. https://doi.org/10.1177/0091270010375428