em teia
Revista de Educação Matemática e Tecnológica Iberoamericana

UNIVERSIDADE FEDERAL DE PERNAMBUCO

ISSN 2177-9309

Pós-Graduação em Educação Matemática e Tecnológica
EDUMATEC

Planning and analyzing the quality of a descriptive statistics assessment in health science courses

Planejamento e análise da qualidade de uma avaliação de estatística descritiva em cursos da área da saúde

Alessandra Aparecida da Silva Menezes
Doctor of Science – Public Health/Epidemiology
Universidade Federal de São Paulo – São Paulo – Brazil
aasmenezes@unifesp.br
https://orcid.org/0000-0002-3102-4748

Nathália Petraconi
Doctor of Science – Neurology and Neuroscience
D'Or Institute for Research and Education – Rio de Janeiro – Brazil
nathaliapetraconi@hotmail.com
https://orcid.org/0000-0002-3515-542X

Camila Bertini Martins
Doctor of Science – Statistics
Universidade Federal de São Paulo – São Paulo – Brazil
cb.martins@unifesp.br
https://orcid.org/0000-0002-8252-8815

Abstract

Traditional statistics education in the health sciences often relies on transmissive, decontextualized methods that prioritize rote memorization over statistical thinking, the investigative process essential for problem-solving and decision-making. To bridge these gaps, we developed and evaluated an assessment tool comprising seven multiple-choice questions and one short-essay question, grounded in the GAISE guidelines and centered on a real-world public health problem. We enrolled twenty students from a health-related course at a public university in São Paulo. To establish validity evidence (across cognitive, instructional, and inferential dimensions), we assessed students' perceptions of difficulty and calculated difficulty and discrimination indices for each item. We found strong alignment with classroom content and the ability to identify conceptual misconceptions and discriminate between varying levels of students' knowledge. Analysis of the essay responses and student reports revealed that synthesis capacity and the visual complexity of graphs act as cognitive barriers, independent of a student's underlying conceptual understanding. While the small sample size limits the findings to this

course context, this study contributes to the continuous improvement of the teaching-learning process by offering a practical framework for aligning statistics assessment with evidence-based pedagogical practices.

Keywords: Learning Assessment; Descriptive Statistics; Health Education; Validity Evidence; Classroom.

Resumo

O ensino tradicional de Estatística nas ciências da saúde frequentemente baseia-se em métodos transmissivos e descontextualizados que privilegiam a memorização mecânica em detrimento do pensamento estatístico — o processo investigativo essencial à resolução de problemas e à tomada de decisões. Para superar essas lacunas, desenvolvemos e avaliamos um instrumento de avaliação composto por sete questões de múltipla escolha e uma questão dissertativa curta, fundamentado nas diretrizes GAISE e centrado em um problema real de saúde pública. Participaram do estudo vinte estudantes de um curso da área da saúde de uma universidade pública de São Paulo. Para estabelecer evidências de validade (nas dimensões cognitiva, instrucional e inferencial), avaliamos a percepção dos estudantes sobre a dificuldade das questões e calculamos os índices de dificuldade e discriminação de cada item. O instrumento apresentou forte alinhamento com o conteúdo das aulas, além de capacidade para identificar equívocos conceituais e discriminar entre diferentes níveis de conhecimento. A análise das respostas dissertativas e dos relatos dos participantes indicou que a capacidade de síntese e a complexidade visual dos gráficos funcionam como barreiras cognitivas, independentemente da compreensão conceitual prévia do estudante. Embora o tamanho reduzido da amostra limite as conclusões ao contexto deste curso, o estudo contribui para a melhoria contínua do processo de ensino-aprendizagem, ao oferecer uma estrutura prática para alinhar a avaliação em Estatística a práticas pedagógicas baseadas em evidências.

Palavras-Chave: Avaliação de Aprendizagem; Estatística Descritiva; Educação em Saúde; Evidências de Validade de Instrumento; Sala de Aula.

INTRODUCTION

A solid understanding and application of statistical concepts have become increasingly important in the training of health professionals. Statistics underpin scientific production and inform evidence-based clinical decisions and public policy (Groth, 2021). Skills such as interpreting graphs, understanding summary measures, and communicating numerical data are essential competencies for informed clinical decision-making (Groth, 2021; Zikmund-Fisher; Thorpe; Fagerlin, 2025). However, statistics education in health sciences has traditionally relied on transmissive, decontextualized approaches detached from clinical and epidemiological practice, contributing to students' aversion and a lack of interest (Pereira; Dufranc; Villagra, 2019). This approach, often centered on the mechanical execution of calculations and memorization of procedures, tends to neglect

statistics as an investigative process for problem-solving and decision-making (Ben-Zvi, 2011; Burrill; Pfannkuch, 2024).

When assessment aligns with the curriculum, teaching methods, and student needs, it can effectively address these limitations. In this context, assessment becomes a continuous process for collecting and interpreting evidence, supporting both formative and summative purposes (Pellegrino; DiBello; Goldman, 2016). Assessment strategies can be diverse, incorporating hands-on projects, oral presentations, and group discussions. When well-designed, these activities offer rich opportunities for students to demonstrate their understanding, promoting a more authentic and meaningful evaluative experience (Pellegrino; DiBello; Goldman, 2016).

In health education, multiple-choice questions (MCQs) are widely used due to their objectivity, scalability, and broad content coverage (Parekh et al., 2024). However, recent research indicates that errors in MCQs formulation remain frequent (Ali; Zahra, 2024). MCQs also have inherent limitations, such as the potential for random correct answers, limited insight into students' reasoning and problem-solving abilities, and reduced capacity for evaluating communication skills. Despite that, both MCQs and alternative assessment types have strengths and weaknesses (Mee et al., 2024; Parekh et al., 2024).

MCQs can play a formative role if they are constructed according to clear, verifiable quality criteria (Elgadal; Mariod, 2021). In the field of biology, for example, Hicks (2021) used paired MCQs to demonstrate the need to foster conceptual understanding rather than formula memorizing. Additionally, MCQs may support learning by encouraging retrieval practice, which enhance long-term retention (Nair; Feroze, 2023).

Conversely, flaws in MCQ construction can compromise the interpretation of results (Rezigalla et al., 2024). Recognizing the complexity of designing effective assessments, initiatives such as the Guidelines for Assessment and Instruction in Statistics Education (GAISE) (Perrett, 2024) and the Levels of Conceptual Understanding in Statistics project (LOCUS Project, 2025) provide worked examples and commentary to illustrate how to craft effective questions, thereby translating theoretical principles into classroom practice.

Concerns regarding conceptual assessment in statistics have led to the development of alternative methods, such as student project, graph critiques, and short essays (Ben-Zvi, 2011; Burrill; Pfannkuch, 2024). While the GAISE (Perrett, 2024) and LOCUS (2025) provide foundational principles, the health context entails specific competencies beyond those addressed in generic statistics education.

Recent discourse reveals critical distinction: health science students must not only compute or select correct statistical procedures but also explain how these concepts underpin evidence-based recommendations and risk communication (Aggarwal, 2018; Groth, 2021; Zikmund-Fisher; Thorpe; Fagerlin, 2025). This deeper understanding requires assessment formats specifically designed to elicit student reasoning. Nevertheless, a recent review highlights a scarcity of studies dedicated to evaluating statistical concepts within health education (Pereira; Dufranc; Villagra, 2019).

In addition, concerns regarding the quality of MCQ-based evaluations in health programs are more commonly addressed in large-scale examinations than in daily classroom settings (Hamamoto Filho; Bicudo, 2020). Standard statistical techniques for establishing quality may also be constrained in small-scale contexts, such as individual classrooms (Belov; Lüdtke; Ulitzsch, 2025; Tavakol; Dennick, 2011). However, for exploratory purposes, difficulty indices (the percentage of correct responses) and discrimination indices (which reflect the ability to differentiate between students who have mastered the concept and those who have not) can yield valuable performance data and provide formative insights to inform teaching practice (Elgadal; Mariod, 2021).

In contemporary literature, validity is not regarded as an intrinsic property of a test but rather as an argument supported by evidence. Pellegrino, DiBello, and Goldman (2016) proposed an interpretive framework that clarifies the claims made about an assessment and the evidence used to support them. This framework comprises three complementary components: cognitive validity, which examines whether an instrument accesses relevant domain knowledge; instructional validity, which evaluates the alignment between the instrument, the curriculum, and learning opportunities; and inferential validity, which concerns the accuracy of inferences drawn from student performance to inform diagnostic and pedagogical decision-making.

In this study, we present the development and quality analysis of an assessment instrument composed of MCQs and of a short essay, designed for both summative and formative use in an introductory, health-oriented statistics course. This mixed format combines the objectivity of MCQs with the capacity of the essay component to assess higher-order reasoning and communication skills. Our quality argument is structured around Pellegrino, DiBello, and Goldman's (2016) cognitive, instructional, and inferential validity dimensions, supplemented by an analysis of students' reflections on perceived challenges, and item-level difficulty and discrimination indices tailored to the classroom context.

METHODS

Setting and participants

This study is part of the project "Research in Statistics Teaching," approved by the institution's Research Ethics Committee (CAAE no. 74235523.6.0000.5505). We enrolled twenty students in a required biostatistics course (60 class hours) within a health sciences curriculum at a public university in São Paulo, Brazil, in 2022. The course was conducted by an interdisciplinary pedagogical team and offered during the initial years of training. The curriculum sought to balance theoretical and practical instruction through a diverse set of activities designed to foster statistical skills and development throughout the training path. All course activities were graded and contributed to the final marks for student approval.

For data presentation and anonymity, students are identified hereinafter by fictitious names. The corresponding author will provide all data and materials upon request, in accordance with ethical guidelines and institutional regulations on participant privacy.

The evaluation activity

We designed the assessment activity in accordance with the theoretical and methodological framework outlined in the previous section. Guided by the GAISE guidelines (2024) and the concept of cognitive validity (Pellegrino; DiBello; Goldman, 2016), we developed the MCQs to assess five core competencies:

- Recognize problems where the statistics investigative process is applicable.

- Solving problems using the statistical investigative process.
- Interpreting the insights and limitations provided by graphical data.
- Recognizing and explaining the role of variability within the field of statistics.
- Recognizing and explaining the role of randomness in study planning and the formulation of conclusions.

To ensure instructional validity (Pellegrino; DiBello; Goldman, 2016), the MCQs were restricted to topics covered during the course: basic sampling concepts, summary measures, frequency distributions, graphs, and tables. These topics were taught with the explicit objective that students would both understand the underlying theory and know how to apply them to practical problem-solving.

Seven MCQs, each with four response options and a single correct answer, addressed a public-health case study of excessive sound-pressure levels during indoor fitness classes at a fictional gym. In this scenario, the gym manager conducted a study to address the issue, combining a brief survey of clients and staff regarding perceived loudness with sound-pressure measurements collected over several sessions in one week. An introductory text framed the items, outlining the health risks of excessive sound pressure in enclosed spaces and summarizing relevant Brazilian technical standards.

The MCQs followed the manager's step-by-step investigation, requiring students to apply statistical concepts to study design, data analysis, and interpretation. A simulated dataset supported the descriptive statistics, figures, and tables used in constructing the items. Distractors (incorrect answer options) were designed so that each MCQ included at least two plausible alternatives, based on common misconceptions related to the statistical concepts addressed (e.g., extrapolating conclusions from a sample to a population without considering representativeness). Each of the seven MCQs was weighted equally (1.43 points each), for a total of ten points. Table 1 summarizes the content areas and skills assessed by each MCQ.

Table 1 – Contents and skills assessed by multiple-choice questions, São Paulo, Brazil, 2022

Question	Content covered	Skill required of the subject
Q01	Summary measures for qualitative variables;	Understand that non-numerical data collected for statistical problem-solving have a distribution that

Question	Content covered	Skill required of the subject
	comparison of distributions; bar chart.	can be described by absolute and relative frequencies. Use frequency distributions and bar graphs to compare and interpret differences between two or more distinct qualitative datasets.
Q02	Summary measures for quantitative variables; comparison of distributions; One-dimensional scatter plots.	Understand that numerical data can be described by measures of central tendency (median and/or mean) and variability (interquartile range and/or mean absolute deviation). Select and use statistics appropriate to the distribution format, translated into one-dimensional scatter plots, to compare and interpret differences between numerical data sets.
Q03; Q06	Population-sample relationship.	Understand that, under certain conditions (such as representativeness and random sampling), statistics allow for inferences about a population to be drawn from a sample.
Q04	Relationship between two quantitative variables; Two-dimensional scatter plots.	Employ and interpret scatter plots to investigate patterns of association between two distinct numerical variables.
Q05	Summary measures for quantitative variables; comparison of distributions; box plot.	Understand that numerical data can be described by measures of position and dispersion. Use distribution-appropriate statistics, translated into box plots, to compare and interpret differences between two or more distinct numerical data sets.
Q07	Summary measures for bivariate qualitative data; relationship between two qualitative variables; 2 x 2 contingency tables.	Understand that association patterns in bivariate categorical data can be identified by displaying absolute and relative frequencies in contingency tables. Use relative frequencies to describe and interpret associations and trends between variables.

Source: Prepared by the author.

To stimulate higher-level critical thinking, we included a short essay alongside the MCQs. To answer this question, students were required to use information from the instrument's introduction and the preceding MCQs to identify and discuss the manager's solution. We assessed the essay using the rubric (a scoring guide with criteria and performance levels) in Table 2, which operationalizes the Gal's (2004) model of statistical literacy. This model comprises four interdependent components: literacy skills (the ability to read and interpret data displays and statistical language), statistical knowledge (understanding of statistical concepts and procedures), context knowledge (familiarity with the specific domain), and critical stance (the ability to question data, evaluate reasoning, and consider limitations).

We mapped these four components onto two criteria with distinct performance levels: (a) identification of solution (0–1), which assesses literacy skills and context knowledge; and (b) use of evidence and critical thinking in justification (0–2), which evaluates statistical knowledge, context knowledge, and critical stance. This operationalization captures students' progression from insufficient understanding to satisfactory or excellent performance, the latter characterized by the critical evaluation of data limitations. The essay score was rescaled to a maximum of ten points to align with the total MCQ score, and the final score was calculated as the average of the two components.

Table 2 – Rubric for evaluation of the short essay operationalizing with the Gal's (2004) statistical literacy model, São Paulo, Brazil, 2022

Criterion	Gal's Components	Performance level descriptor
Identification of solution	Literacy skills; Context knowledge	0 – Insufficient: Fails to identify the solution or gives an irrelevant/confused response. Shows no evidence of comprehension of the problem context. 1 – Satisfactory: Correctly identifies the manager's solution, demonstrating a clear understanding of the problem along with adequate literacy and context knowledge.
Use of evidence and critical thinking in the justification	Statistical knowledge; Context knowledge; Critical stance	0 – Insufficient: No evidence or data used; reasoning is vague or contradictory. Fails to recognize data limitations or validity concerns. 1 – Satisfactory: Incorporates some data to support conclusions, though organization or logic may be weak; minor inconsistencies are present. Demonstrates limited awareness of data limitations. 2 – Excellent: Provides clear, coherent, and persuasive argumentation supported by multiple data points, well-integrated into the conclusion. Recognizes limitations in data collection or scope (e.g., "measurements limited to one week"); questions the validity of data sources or acknowledges alternative interpretations. Demonstrates a high level of critical evidence evaluation.

Source: Prepared by the author.

The activity was hosted in the institution's Virtual Learning Environment (VLE) for three consecutive days, during which students were encouraged to provide honest

responses. During this period, students were permitted to review and modify their previously submitted responses. Initially, the first MCQ (Q01) was published with an error: the prompt lacked clear instructions (e.g., “check the correct alternative”). After four students had responded and one alerted the team to the issue, the form was corrected to ensure all subsequent students access Q01 with an error-free statement. This procedural issue is methodologically significant, as it may have introduced noise into the item’s difficulty index.

Student feedback prompts

After the activity, participants were invited to answer two open-ended questions:

- “Which questions did you find the most difficult?” and
- “What were the main challenges, and what can be done to overcome them?”

We systematically catalogued and examined the identified challenges and recommendations to determine the proportion of students who perceived each MCQ as particularly challenging. Quotations were lightly edited for clarity, with any clarifications or omissions indicated in square brackets.

Quantitative analysis

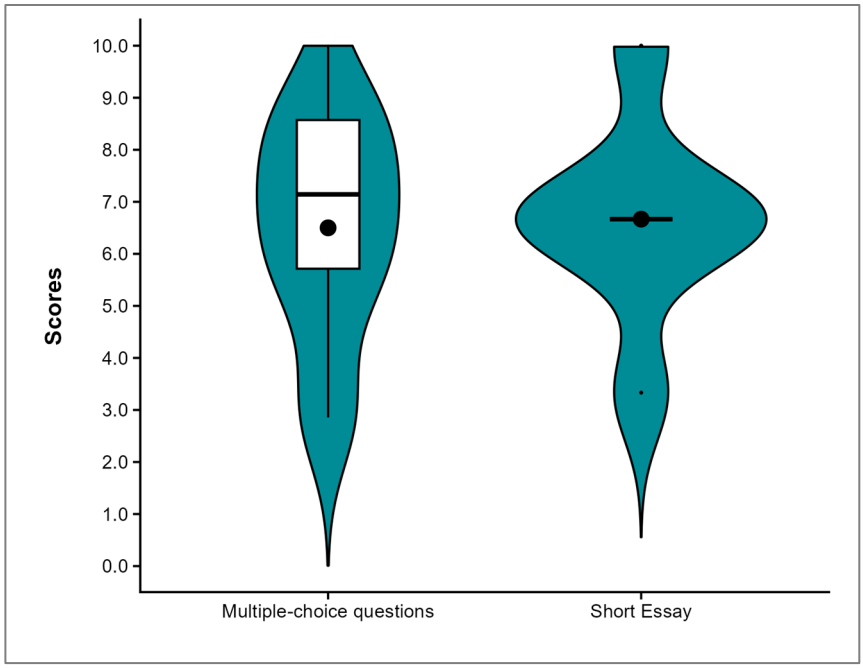
We used R 4.5.0 to compute the difficulty and discrimination indices for each MCQ, the latter via point-biserial correlation (a statistical measure of the relationship between a dichotomous variable (correct/incorrect) and a continuous variable (total score)) (Rizopoulos, 2007). Given the sample size constraints (Belov; Lüdtke; Ulitzsch, 2025), we used difficulty and discrimination indices for formative and diagnostic purposes within classroom contexts rather than as evidence of generalizable psychometric properties. The classification of these indices was based on the National Student Performance Exam (ENADE), a standardized assessment administered for undergraduate heath programs in Brazil (Brazil, 2024), which we treated as preliminary and exploratory. These calculations illustrate how educators can analyze their classroom assessments to identify student learning patterns and guide instructional modifications. To examine the relationship between performance on the short essay and the total MCQ score, we compared the median and mean essay scores across tertiles of the MCQ results.

Feedback to the students

After grading the assignment, students received their individual grades and a document detailing the rationale behind the distractors via VLE. During a subsequent face-to-face session, we expanded the feedback by discussing the short essay and the two open-ended questions. We also presented the difficulty and discrimination indices as practical examples of how statistics can be applied to educational assessment.

RESULTS

Figure 1 – Distribution of total MCQ score and score of the short essay (N = 20), São Paulo, Brazil, 2022



Source: Prepared by the author.

The distribution of scores for both assessment components are presented in Figure 1. Total MCQ score was generally high, with a median of 7.1 and an interquartile range of 5.7 to 8.6, although a few students scored markedly lower. The short essay revealed a similar central tendency (quartiles of 6.7), but exhibited less variability, characterized by a tighter concentration around the quartiles and the presence of a low outlier.

While most students successfully identified the solution implemented in the case, performance varied more significantly regarding the quality of argumentation. A few students failed to provide relevant data to support their conclusions [“Reduce the noise

during classes."(Avery)], yielding low scores. Several responses made only superficial reference to the data, resulting in intermediate scores, whereas the strongest answers integrated multiple pieces of evidence in a coherent and persuasive manner:

> The most likely solution identified by the manager was to reduce the ambient noise level, since it was almost always above the regulated limit (50 dB, with a median of approximately 70 dB), and most clients and staff considered it inadequate (only 1 out of 25 staff and 1 out of 20 students did not complain about the volume). However, the survey conducted with clients showed that only 52% were satisfied with the new rules. (Morgan)

Despite the greater homogeneity of the short essay regarding the MCQs' total score, the mean of the former increased monotonically across tertiles of the latter (6.2, 6.7, and 7.2).

The analyses of the MCQs indicated that most of them were classified as "easy", according to the difficulty index, while all items demonstrated "good" or "very good" ability to discriminate among students with different levels of knowledge. To facilitate comparison, the data in Tables 3 and 4 are presented in descending order of difficulty, based either on the statistical index (Table 3) or on the students' perception (Table 4).

Table 3 presents the percentage of students who selected each answer, alongside the difficulty and discrimination indices for each question. Q01 and Q06 had the lowest percentage of correct answers, identifying them as the most statistically challenging items. While many MCQs were rated as "easy" or "moderately difficult", all demonstrated satisfactory capacity to distinguish between students with varying levels of knowledge.

Nonetheless, it is important to note that, given the small sample size (N=20), these indices should be interpreted as exploratory illustrations of how educators can diagnose item performance in a classroom context, rather than as robust evidence of generalizable psychometric properties. Additionally, the difficulty observed in Q01 may have been influenced by the procedural issue described at the end of "The evaluation activity" subsection.

Table 3 – Distribution of MCQ options, item difficulty, and discrimination (N = 20), São Paulo, Brazil, 2022

	% of students in each Answer Option

Question	A	B	C	D	Difficulty Index	Discrimination Index
Q01	0	60	40#	0	0.40 (hard)	0.47 (very good)
Q06	20	20	50#	10	0.50 (medium)	0.52 (very good)
Q04	0	25	60#	15	0.60 (medium)	0.31 (good)
Q07	70#	0	30	0	0.70 (easy)	0.48 (very good)
Q05	25	0	0	75#	0.75 (easy)	0.38 (good)
Q03	0	80#	5	15	0.80 (easy)	0.63 (very good)
Q02	0	80#	5	15	0.80 (easy)	0.37 (good)

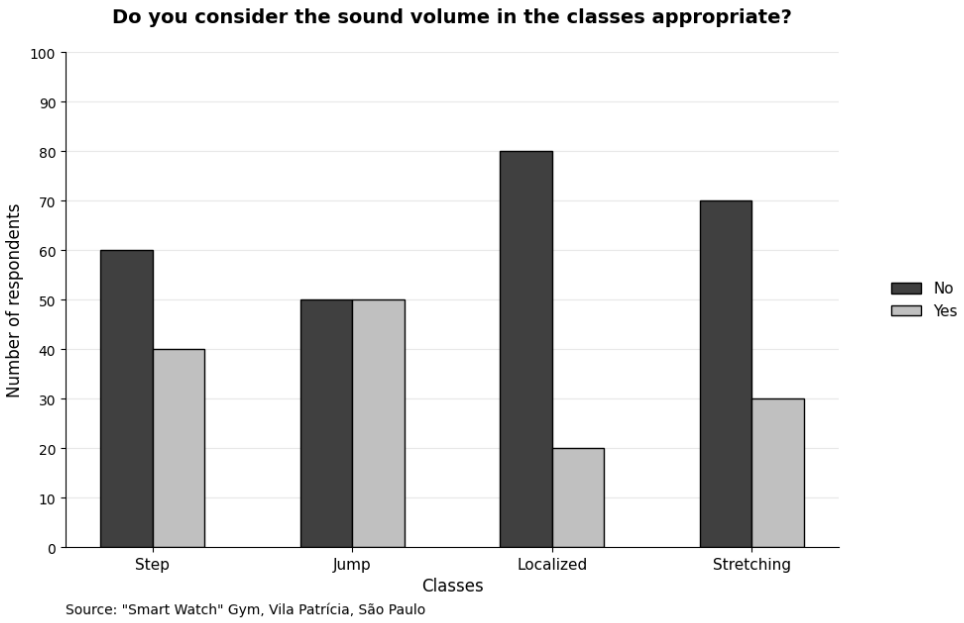
Note: Correct options are marked with #. Difficulty = proportion correct. Discrimination = point–biserial. Given small sample size, these indices are presented for classroom-level diagnostic and formative use only. Interpretation beyond this specific context is not warranted.

Source: Prepared by the author.

For each item, we examined which misconception-focused distractors were or were not selected (see Table 3). Figures 2 and 3 illustrate Q01 and Q07, respectively; along with Q05, these questions contained two distractors not selected by any student.

Figure 2 – Example question with two unused distractors (A and D) and one distractor (B) suggesting misinterpretation of variability in bar charts

Q01 - The gym offers step, jump, localized, and stretching classes, exercise modalities with 100 clients in each group. In January 2022, each of these clients was asked whether, overall, they considered the sound volume appropriate for those classes. The graph below shows the results of the survey. In which group were the answers more consistent (homogeneous)?



A. Step
B. Jump
C. Localized
D. Stretching

Source: Prepared by the author.

Figure 3 – Example of a question with two distractors (B and D) not chosen by any student (N = 20), São Paulo, Brazil, 2022

Q07 - The results of the survey conducted among the gym's clients at the end of April 2022 are recorded in the table below. The manager believes that satisfaction with the new rules for using the fitness rooms was higher among women. Which of the following options provides the best evidence to support the gym manager's assumption?

Survey participants' satisfaction with the adjustments to sound pressure levels in fitness classes, "Smart Watch" Gym, Vila Patrícia, São Paulo, Apr./2022

	Satisfied	Dissatisfied	Total
Women	63	34	97
Men	20	43	63
Total	83	77	160

Source: "Smart Watch" Gym, Vila Patrícia, São Paulo

A. Women are 76% (63/83) of those satisfied and 44% (34/77) of those dissatisfied.
B. Men are a minority in the survey (63/160, 39%).
C. Women make up 61% (97/160) of the clients, and 52% (83/160) of the clients are satisfied with the changes.
D. A higher participation of women normally exists in all surveys.

Source: Prepared by the author.

Eighteen (90%) of the twenty students answered the open-ended question regarding the perception of difficulty in the MCQs. Table 4 presents these perceptions, listing the questions in descending order of frequency of indication as more difficult. MCQs Q04 and Q05, which involved the interpretation of two-dimensional scatterplots and box plots, respectively, were the most mentioned as challenging.

Tables 3 and 4 reveal gaps between actual performance and perceived difficulty. For example: Q01 had a 60% error rate, but only 28% of students found it difficult. Q06 had a 50% error rate, but 22% found it difficult. In contrast, 50% of students perceived Q04 as difficult, yet 60% answered correctly. Similarly, 39% found Q05 difficult, although 75% answered correctly.

Table 4 – Student-perceived difficulty by question and visualization (N = 18), São Paulo, Brazil, 2022

Question	Data visualization	n	%
Q04	Two-dimensional scatter diagram	9	50
Q05	Box plots	7	39
Q01	Bar Chart	5	28
Q02	One-dimensional scatter diagram	4	22
Q06	Did not use data visualization	4	22
Q07	Table 2 x 2	4	22
Q03	Did not use data visualization	0	0

Note: n = number of students; % = percentage of students who indicated the question as one of the most difficult.
Source: Prepared by the author.

Regarding question Q01, students who responded before the correction reported greater difficulty, even though they were later given the opportunity to redo it with the corrected statement: “Mainly the first [question]... which did not have the command [instruction]” (Taylor). Setting this case aside, other participants reported difficulty selecting the correct alternative when multiple options appeared plausible, as noted in one comment: “Question number 7 seemed to present two possible alternatives, as both [sets of] data presented were correct” (Casey). The main challenges reported by the students involved interpreting various types of graphs, especially box plots and both one-dimensional and two-dimensional scatter plots, in addition to the need to recall and simultaneously apply statistical concepts. These challenges were illustrated in the students’ comments: “The question about the box plot, dispersion [...]” (Taylor); “The questions with the scatter plot, as they are more difficult graphs to interpret” (Quinn); “The ones with dot plots – I find other types of graphs and tables easier to interpret” (Morgan); and “Question 2 because of the type of graph it presents [one-dimensional dispersion]” (Cameron).

Other participants mentioned challenges related to decision-making and argumentation in the short essay, “[the more difficult was] the last question, which was very subjective, since it was not clear which area to turn to solve the problem.” (Riley); “The final question, I confess that I found it difficult to align the data from the previous questions to formulate a solution ...” (Casey).

Students’ suggestions for overcoming these difficulties included rereading the prompts: “I read and reread the questions” (Alex); and the revision of the class material: “[...] To overcome them, I will review the contents addressed in these two questions in order to improve the graphic interpretation” (Jordan). In addition, many students highlighted the importance of more practice in graphical interpretation: “Interpreting scatter plots and box plots was the most difficult; by training this interpretation, one can improve” (Taylor).

Finally, the adoption of practical teaching methods, such as the discussion of clinical cases at the end of classes, was suggested as an effective way to fix the concepts studied. One student exemplified this idea by sharing their experience:

> There are several concepts that need to be memorized and applied at once. A teaching method that I found very interesting was that of the cardio teacher, who, at the end of all classes, gave a clinical case for us to discuss and apply the concepts studied. I find it easier to fix the matter that way. In biostatistics, we did the exercise in only one class, but I would like them all to be like this.(Morgan)

DISCUSSION

This study describes the process of designing an assessment that integrates both MCQs and a short essay within a public health context to evaluate students' conceptual understanding of statistics. The development process emphasized rigorous item construction and addressed multiple quality dimensions at various stages before and after administration.

The decision to incorporate a diverse mix of assessment formats—MCQs, a short essay, and open-ended perception items—was pedagogically intentional. While MCQs facilitate broad content coverage and objective scoring, the essay component fosters integrative reasoning and communication skills (Perrett, 2024). Moreover, the open-ended prompts helped verify the alignment between the intended pedagogy and students' interpretations. This multimethod approach is consistent with current recommendations to employ varied assessment strategies to obtain a more comprehensive understanding of student learning.

Contextualizing questions in clinical or public health scenarios engages learners by linking abstract statistical concepts to practical decision-making. In this study, each item formed part of a health-investigation narrative, requiring students to interpret data within authentic scenarios (e.g., comparing groups with a bar-chart or evaluating survey results in a 2×2 contingency table). This contextualization scaffolds the statistical knowledge necessary for clinical and public health decisions (Aggarwal, 2018; Groth, 2021; Zikmund-Fisher; Thorpe; Fagerlin, 2025) and promotes transfer of classroom knowledge to applied problems. The approach aligns with health-education strategies that

emphasize conceptual understanding and interpretation over rote calculation (Hicks et al., 2021; Pereira; Dufranc; Villagra, 2019).

For formative purposes, we designed MCQs to target key conceptual competencies from the GAISE (Perrett, 2024) and statistical literacy in health (Aggarwal, 2018; Groth, 2021; Zikmund-Fisher; Thorpe; Fagerlin, 2025). All assessment content aligned with course topics and pedagogy, supporting instructional validity. Beyond scoring correctness, we: examined which incorrect answer options students selected to identify misconceptions; evaluated short essay performance; and analyzed student feedback to open-ended prompts. This triangulation approach provides stronger evidence of conceptual understanding than correctness alone, reinforcing inferential validity (Pellegrino; DiBello; Goldman, 2016).

Operationalizing the Gal's (2004) statistical literacy model through the rubric revealed how these instructional components manifested in student performance. Most students demonstrated satisfactory proficiency in identifying solutions and using evidence. However, their reasoning about critical evaluation of data revealed differentiated levels of sophistication. Morgan's response exemplified advanced reasoning: explicitly acknowledging that measurements were limited to one week and proposing that longer sampling would strengthen conclusions. In contrast, most other students correctly cited data to support their conclusions but did not question measurement scope, sampling generalizability, or alternative interpretations. These patterns indicate that while students can apply statistical knowledge in familiar classroom contexts, developing critical stance, the ability to question data validity and acknowledge analytical limitations, requires more explicit and sustained instruction in health statistics education.

Despite these promising formative insights, the assessment-quality analysis must be interpreted within the constraints of this specific study design. The assessment was administered to a small sample from a single cohort as an unsupervised online activity, which inherently limits the generalizability of observed patterns. Furthermore, certain quantitative results, such as difficulty and discrimination indices, are not definitive parameters of item quality; these indices may differ substantially in larger, more diverse student groups, or in different administrative contexts. Accordingly, the quantitative

findings presented here should be understood as exploratory illustrations of how teachers can analyze classroom assessments locally, rather than as robust evidence of item properties generalizable beyond this specific context.

Analysis of difficulty indices indicated that most MCQs were easy or moderately challenging, consistent with students' prior learning and supportive of instructional validity (Pellegrino; DiBello; Goldman, 2016). An overrepresentation of difficult items would have suggested misalignment between assessment and instruction. Despite the lack of generalizability, all items exhibited satisfactory discrimination, suggesting that higher-performing students were more likely to answer correctly than lower performers in this cohort of students. The two items with the lowest proportions of correct answers still indicated discriminability, pointing that well-constructed MCQs can be both challenging and diagnostic.

At the item level, we examined how distractors mapped onto common statistical misunderstandings. In our study, one distractor that misrepresented variability in a bar graph was frequently chosen (Q01, option B). However, given the procedural issue reported, it is important to recognize that a high percentage of distractor choice responses might indicate confusion about the task itself rather than a misunderstanding of statistical concepts. Conversely, addressing these procedural issues openly increases inferential validity, since it helps us discern when difficulties with the assessment tool may obscure a more accurate understanding of the real cognitive obstacles to learning.

We also observed unselected distractors (e.g., Q07, option D). Their non-selection implies that they were either implausible or failed to capture the intended misconception. Although unselected distractors do not reduce an item's scored performance, they signal opportunities to strengthen item design: each distractor should be plausible enough to attract students who hold the targeted misconception or are unsure about the concept.

These findings align with previous research advocating for item analysis to enhance classroom assessments (Elgadal; Mariod, 2021). Furthermore, they highlight the formative importance of item analysis in educational settings. Through this process, we were able to pinpoint content areas and cognitive skills requiring further development, which led us to revisit the concept of variability with students and revise items whose distractors did not perform as intended.

Student feedback highlighted difficulty with data visualization, but this perception only partially aligned with difficulty indices. As noted by Zaidi (2018), discrepancies between quantitative results and students' perceptions may arise from the structure of MCQs or from students' self-confidence, even when grounded in misconceptions. Another possibility is that students may have selected the correct answer through visual pattern recognition or process of elimination, thereby activating only partial understanding rather than deep conceptual understanding.

According to Duval's theory (2012), conceptual understanding of mathematical objects requires coordinating multiple semiotic representation registers—distinct systems of representation such as graphical displays, geometric figures, algebraic expressions, and natural language. A student may comprehend the abstract concept of "variability" yet struggle to decode a box plot's visual structure, which demands simultaneously coordinating five summary statistics (minimum, Q1, median, Q3, maximum) within a compact spatial arrangement. Two-dimensional scatter plots compound this difficulty: students must navigate Cartesian coordinate systems while extracting association patterns. The management of semiotic registers has been identified as a primary difficulty in mathematics education, necessitating explicit and scaffolded practice (Duval, 2012; Ferretti; Gambini; Spagnolo, 2024).

In our course, these results pointed out the need to strengthen instruction on reading and interpreting visual data displays. As an immediate pedagogical response, we incorporated more scaffolded practice with graph interpretation, progressively layering tasks from simpler plots (e.g., one-dimensional dot plots) to box plots comparing groups, to build fluency. Students themselves recommended "training this interpretation," and some reported plans to re-read questions and review class materials, indicating that the assessment served a reflective function.

One concrete student suggestion was to conclude classes with brief case discussions focused on data interpretation. This recommendation aligns with research on Case-Based Learning (CBL) in medical and health-care fields (McLean, 2016; Sistermans, 2020). CBL is defined as an inquiry-structured learning experience utilizing live or simulated patient cases to solve or examine a clinical problem, with guidance from a teacher and adherence to stated learning objectives (McLean, 2016; Sistermans, 2020).

In health education, mathematical and statistical knowledge must inform the design of case-based learning scenarios. Authentic cases should engage students with fundamental statistical concepts (e.g., variability, uncertainty, stochastic phenomena) rather than isolated computational procedures (Aggarwal, 2018; Groth, 2021; Zikmund-Fisher; Thorpe; Fagerlin, 2025). These cases should bridge statistical representations and contextual understanding by connecting data patterns to geographic, temporal, and epidemiological contexts. Implementation approaches are diverse, ranging from small group discussions and role-play activities to multidisciplinary teamwork and online debates (McLean, 2016; Sistermans, 2020). The current study demonstrates one approach: using a contextualized public health scenario with multiple-choice questions and a short essay to assess statistical concept understanding.

As we state above, the study has limitations. First, the sample consisted of twenty students from a single program and institution, which limits the generalizability of these results to other educational contexts or disciplines. Second, the difficulty and discrimination indices are sample-dependent, and an item categorized as “easy” in our class might behave differently with a different cohort. Third, the assessment content focused on a specific curriculum and context of a biostatistics course. Finaly, we also acknowledge a minor procedural issue during administration, the initial error in Q01’s prompt, which could have affected the students’ performance or perceptions of that item.

At this stage, our objective was not to develop a psychometrically validated instrument for large-scale application. We focused on refining MCQs, widely used to assess the understanding of statistical concepts in health, which, according to our experience and the literature (Ali; Zahra, 2024; Parekh et al., 2024; Pereira; Dufranc; Villagra, 2019), remains problematic. The study’s strength, therefore, lies in presenting a concrete and reproducible process for developing and evaluating MCQs integrated into best pedagogical practices. Regarding the use of fictitious data, we acknowledge that real data may be more motivating for students. However, real data may increase cognitive complexity (Phadke; Beckman; Morgan, 2024), and simulations offer diverse didactic situations that facilitate learning through the comparison and confrontation of results.

Grounded in a culture of continuous feedback that promotes student self-assessment and ongoing improvement (Boud; Dawson, 2023), this study aimed to

document the development of a contextualized assessment in public health, analyzing its items and student reflections. Thus, our study may contribute to practical insights for statistics educators in health programs seeking to refine their assessments. We encourage further research and the application of these ideas in other contexts. By continuing to iterate and study our assessments, we not only improve our own teaching practices but also contribute significantly to the knowledge of statistics education in the health professions. This study, with its classroom-based insights, is a step in this direction. Future work should incorporate pre-post measures to examine learning gains and retention.

CONCLUSION

The combined assessment structure — MCQs, short essay, item-level analysis, and student feedback — enabled us to evaluate student understanding of statistical concepts in health contexts and make timely instructional improvements. Although these findings are context-specific to this course, the reproducibility of this multifaceted assessment framework, grounded in established pedagogical principles, offers insights on how to improve teaching statistics education across health programs, providing both immediate insights into student learning and a practical, evidence-based model for continuous improvement of the teaching-learning process.

ACKNOWLEDGMENTS

We sincerely thank the students who participated in this study for their engagement and valuable reflections. We also express our gratitude to Professor Leandro F. M. Rezende for his thoughtful reading and constructive feedback on the manuscript.

REFERENCES

AGGARWAL, Rakesh. Statistical Literacy for Healthcare Professionals: Why is It Important? **Annals of Cardiac Anaesthesia**, v. 21, n. 4, p. 349, dez. 2018. Disponível

em: https://journals.lww.com/aoca/fulltext/2018/21040/statistical_literacy_for_healthcare_professionals_.1.aspx

ALI, Kamran; ZAHRA, Daniel. Ten tips for effective use and quality assurance of multiple-choice questions in knowledge-based assessments. **European Journal of Dental Education: Official Journal of the Association for Dental Education in Europe**, v. 28, n. 2, p. 655–662, maio 2024. Disponível em: https://pubmed.ncbi.nlm.nih.gov/38282273/

BELOV, Dmitry I.; LÜDTKE, Oliver; ULITZSCH, Esther. A supervised learning approach to estimating IRT models in small samples. **British Journal of Mathematical and Statistical Psychology**, v. n/a, n. n/a, 2025. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1111/bmsp.12396

BEN-ZVI, Dani. Statistical reasoning learning environment. **Em Teia-Revista de Educacao Matemática e Tecnológica Iberoamericana**, v. 2, n. 2, p. 1–13, 2011. Disponível em: https://periodicos.ufpe.br/revistas/emteia/article/view/2152/1721

BOUD, David; DAWSON, Phillip. What feedback literate teachers do: an empirically-derived competency framework. **Assessment & Evaluation in Higher Education**, v. 48, n. 2, p. 158–171, 17 fev. 2023. Disponível em: https://doi.org/10.1080/02602938.2021.1910928

BRAZIL. **Relatório síntese de área - Medicina [Area summary report - Medicine]**. Brasília/DF: Inep, 2024. Disponível em: https://download.inep.gov.br/educacao_superior/enade/relatorio_sintese/2023/medicina.pdf. Acesso em: 21 maio. 2024.

BURRILL, Gail; PFANNKUCH, Maxine. Emerging trends in statistics education. **ZDM – Mathematics Education**, v. 56, n. 1, p. 19–29, 1 fev. 2024. Disponível em: https://doi.org/10.1007/s11858-023-01501-7

DUVAL, Raymond. Registros de representação semiótica e funcionamento cognitivo do pensamento: Registres de représentation sémiotique et fonctionnement cognitif de la pensée (Trad. Méricles Thadeu Moretti). **Revista Eletrônica de Educação Matemática**, Tradução revisada (2023). v. 7, n. 2, p. 266–297, 13 dez. 2012. Disponível em: https://periodicos.ufsc.br/index.php/revemat/article/view/1981-1322.2012v7n2p266

ELGADAL, Amani H.; MARIOD, Abdalbasit A. Item Analysis of Multiple-choice Questions (MCQs): Assessment Tool For Quality Assurance Measures. **Sudan Journal**

of Medical Sciences, v. 16, n. 3, p. 334–346, 3 nov. 2021. Disponível em: https://www.ajol.info/index.php/sjms/article/view/216888

FERRETTI, Federica; GAMBINI, Alessandro; SPAGNOLO, Camilla. Management of semiotic representations in mathematics: Quantifications and new characterizations. **European Journal of Science and Mathematics Education**, v. 12, n. 1, p. 11–20, 1 jan. 2024. Disponível em: https://www.scimath.net/article/management-of-semiotic-representations-in-mathematics-quantifications-and-new-characterizations-13827

GAL, Iddo. Statistical Literacy. *In*: BEN-ZVI, Dani; GARFIELD, Joan (Orgs.). **The Challenge of Developing Statistical Literacy, Reasoning and Thinking**. Dordrecht: Springer Netherlands, 2004. p. 47–78. Disponível em: https://doi.org/10.1007/1-4020-2278-6_3

GROTH, Randall E. The Relevance of Statistical Knowledge for Teaching to Health Care Professionals: Reflections on a COVID-19 Press Briefing. **Journal of Statistics and Data Science Education**, v. 29, n. 1, p. 84–94, 23 jan. 2021. Disponível em: https://doi.org/10.1080/10691898.2020.1851160

HAMAMOTO FILHO, Pedro Tadao; BICUDO, Angélica Maria. Improvement of Faculty's Skills on the Creation of Items for Progress Testing Through Feedback to Item Writers: a Successful Experience. **Revista Brasileira de Educação Médica**, v. 44, n. 1, p. e018, 2020. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-55022020000100401&tlng=en

HICKS, Jenna *et al.* Paired Multiple-Choice Questions Reveal Students' Incomplete Statistical Thinking about Variation during Data Analysis. **Journal of Microbiology & Biology Education**, v. 22, n. 2, p. 10.1128/jmbe.00112-21, 31 maio 2021. Disponível em: https://journals.asm.org/doi/10.1128/jmbe.00112-21

LOCUS PROJECT. **Levels of Conceptual Understanding in Statistics - LOCUS**. Disponível em: https://locus.statisticseducation.org. Acesso em: 10 set. 2025.

MCLEAN, Susan F. Case-Based Learning and its Application in Medical and Health-Care Fields: A Review of Worldwide Literature. **Journal of Medical Education and Curricular Development**, v. 3, p. JMECD.S20377, 2016. Disponível em: https://pubmed.ncbi.nlm.nih.gov/29349306/

MEE, Janet *et al.* An experimental comparison of multiple-choice and short-answer questions on a high-stakes test for medical students. **Advances in Health Sciences Education**, v. 29, n. 3, p. 783–801, jul. 2024. Disponível em: https://link.springer.com/10.1007/s10459-023-10266-3

NAIR, Geethu G.; FEROZE, M. Effectiveness of multiple-choice questions (MCQS) discussion as a learning enhancer in conventional lecture class of undergraduate medical students. **Medical Journal of Dr. D.Y. Patil Vidyapeeth**, v. 16, n. 8, p. 183–188, 1 jan. 2023. Disponível em: https://journals.lww.com/mjdy/fulltext/2023/16002/effectiveness_of_multiple_choice_questions__mcqs_.3.aspx

PAREKH, Priya *et al.* The Utility of Multiple-Choice Assessment in Current Medical Education: A Critical Review. **Cureus**, v. 16, 7 maio 2024. Disponível em: https://www.cureus.com/articles/248393-the-utility-of-multiple-choice-assessment-in-current-medical-education-a-critical-review

PELLEGRINO, James W.; DIBELLO, Louis V.; GOLDMAN, Susan R. A Framework for Conceptualizing and Evaluating the Validity of Instructionally Relevant Assessments. **Educational Psychologist**, v. 51, n. 1, p. 59–81, 2 jan. 2016. Disponível em: https://doi.org/10.1080/00461520.2016.1145550

PEREIRA, Rodrigo Fioravanti; DUFRANC, Ileana Maria Greca; VILLAGRA, Jesus Angel Meneses. Ways of teaching statistics for the health area. **Revista Latinoamericana de Investigación en Matemática Educativa**, v. 22, n. 1, mar. 2019. Disponível em: https://www.relime.org/index.php/relime/article/view/24

PERRETT, Jamis. Revising the Guidelines for Assessment and Instruction of Statistics Education (GAISE) College Report. **Scatterplot**, v. 1, n. 1, p. 2401637, 31 dez. 2024. Disponível em: https://doi.org/10.1080/29932955.2024.2401637

PHADKE, Sayali; BECKMAN, Matthew; MORGAN, Kari Lock. Examining the role of context in statistical literacy assessment. **Statistics Education Research Journal**, v. 23, n. 1, p. 4–4, 28 jul. 2024. Disponível em: https://iase-pub.org/ojs/SERJ/article/view/529

REZIGALLA, Assad Ali *et al.* Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. **BMC Medical Education**, v. 24, n. 1, p. 445, 24 abr. 2024. Disponível em: https://doi.org/10.1186/s12909-024-05433-y

RIZOPOULOS, Dimitris. **ltm: An R Package for Latent Variable Modeling and Item Response Analysis**. , 2007. Disponível em: https://www.jstatsoft.org/index.php/jss/article/view/v017i05

SISTERMANS, Ilse Johanna. Integrating competency-based education with a case-based or problem-based learning approach in online health sciences. **Asia Pacific Education Review**, v. 21, n. 4, p. 683–696, 1 dez. 2020. Disponível em: https://link.springer.com/article/10.1007/s12564-020-09658-6

TAVAKOL, Mohsen; DENNICK, Reg. Making sense of Cronbach's alpha. **International Journal of Medical Education**, v. 2, p. 53–55, 27 jun. 2011. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4205511/

ZAIDI, Nikki L. Bibler *et al.* Pushing Critical Thinking Skills With Multiple-Choice Questions: Does Bloom's Taxonomy Work? **Academic Medicine**, v. 93, n. 6, p. 856, jun. 2018. Disponível em: https://pubmed.ncbi.nlm.nih.gov/29215375/

ZIKMUND-FISHER, Brian J.; THORPE, Alistair; FAGERLIN, Angela. How to Communicate Medical Numbers. **Clinical Review & Education**, 2025. Disponível em: https://jamanetwork.com/journals/jama/article-abstract/2839303#google_vignette

Submetido em 22/10/2025.

Aprovado em 11/02/2026.