analytics

MDPI

Article

Operationalising CTT and IRT in Spreadsheets: A Methodological Demonstration for Classroom Assessment

António Faria [1,*] and Guilhermina Lobato Miranda [2]

[1] UIDEF, Instituto de Educação, Universidade de Lisboa, 1649-013 Lisboa, Portugal
[2] Instituto de Educação, Universidade de Lisboa, 1649-013 Lisboa, Portugal; gmiranda@edu.ulisboa.pt
* Correspondence: antonio.faria@edu.ulisboa.pt

Abstract

The evaluation of student performance often relies on basic spreadsheet outputs that provide limited insight into item functioning. This study presents a methodological demonstration showing how widely available spreadsheet software can be transformed into a practical environment for psychometric analysis. Using a simulated dataset of 40 students responding to 20 dichotomous items, spreadsheet formulas were developed to compute descriptive statistics and Classical Test Theory (CTT) indices, including item difficulty, discrimination, and corrected item–total correlations. The demonstration was extended to Item Response Theory (IRT) through the implementation of 1PL, 2PL, and 3PL logistic models using forward-calculated item parameters. A smaller dataset of 10 students and 10 items was used to illustrate the interpretability of the indices and the generation of Item Characteristic Curves (ICCs). Results show that spreadsheets can support teachers in in-terpreting test data beyond total scores, enabling the identification of weak items, refinement of distractors, and construction of small-scale item banks aligned with competence-based curricula. The approach contributes to Sustainable Development Goal 4 (SDG 4) by promoting accessible, equitable, and high-quality assessment practices. Limitations include the instability of IRT parameter estimation in small samples and the need for teacher training. Future research should apply the approach to real classroom data, explore automation within spreadsheet environments, and examine the integration of artificial intelligence for adaptive assessment.

Keywords: classical test theory; competence-based education; item analysis; item response theory; psychometrics; spreadsheets; SDG 4

1. Introduction

The fair, reliable, and credible assessment of student learning remains a critical and sensitive issue across educational systems. Although learning primarily occurs during the instructional process [1,2], assessment practices traditionally focus on learning outcomes, whether through formative, ongoing evaluation or summative final testing. Throughout the learning process, corrective and cognitive feedback play a crucial role [3], as does self-assessment, which enables learners to monitor and evaluate their own understanding [4,5]. In formative assessment, teachers employ a wide range of techniques, including portfolios, case studies, group work, quizzes, and multiple-choice tests. Summative assessment typically relies on written examinations and closed-response formats.

This article focuses exclusively on closed-response assessment instruments, including multiple-choice items and other fixed-response formats (e.g., true/false, matching, and short-answer with predefined responses). For such instruments to be considered valid

Academic Editor: Binshan Lin

Received: 8 January 2026
Revised: 9 February 2026
Accepted: 14 February 2026
Published: 24 February 2026

and credible, careful item design is essential, and items must undergo scientific content validation [6]. Content validation seeks to ensure that test items adequately represent the domain of knowledge being assessed and is commonly achieved through expert review supported by quantitative indicators, such as the Content Validity Index (CVI) or Content Validity Ratio (CVR). This process typically involves defining the content domain, selecting subject-matter experts, calculating validity indices, and refining items based on relevance and representativeness.

Despite the importance of this rigour, closed-response tests are not always constructed following systematic validation procedures, even though they are widely used throughout compulsory education for both formative and summative purposes. In many contexts, teachers continue to rely on paper-based scoring grids, often consisting of manually completed tables. Others use digital scoring templates provided by educational publishers alongside their assessment instruments.

The widespread availability of computers and digital technologies has facilitated the use of software tools capable of replacing handwritten assessment records. Numerous free and commercial spreadsheet applications are now available, and many educational publishers offer digital scoring grids integrated with their tests. However, the analytical potential of these tools is frequently underused. In most cases, they provide only basic outputs, such as: (i) individual student total scores; (ii) score conversion to qualitative grading scales; (iii) class averages; and (iv) graphical distributions of class results.

Integrating more robust forms of test analysis into everyday assessment practices can improve the alignment of teaching strategies with students' learning needs and, consequently, enhance student performance. To achieve this, teachers require not only basic digital competencies but also foundational knowledge of statistical analysis that can be readily applied within commonly used tools such as spreadsheets. When appropriately designed, these tools can support more detailed analyses of student performance at both individual and class levels, as well as provide valuable information about the quality of the assessment instrument itself (e.g., item difficulty, discrimination, and reliability). Classical Test Theory and Item Response Theory are theories that can support teachers in making student assessment more tailored to their performance and making the assessment more reliable [7–9]. In addition to its methodological contribution, this study aligns with Sustainable Development Goal 4 (SDG 4), which emphasises inclusive and equitable quality education [10]. Although the analysis of dichotomous multiple-choice items represents only one component of the broader assessment landscape, strengthening teachers' capacity to interpret item functioning contributes to more transparent, fair, and pedagogically informed decision-making. By enabling teachers to diagnose item quality and detect sources of difficulty or discrimination using widely available spreadsheet tools, the proposed approach supports more equitable assessment practices within classroom contexts. In this sense, the operationalisation of psychometric analysis in accessible digital environments reinforces the principles of SDG 4 by promoting fairer evaluation processes, enhancing data-informed pedagogical decisions, and expanding teachers' ability to implement high-quality assessment practices without the need for specialised software.

2. Theoretical Framework

2.1. Classical Test Theory and Item Response Theory

Classical Test Theory (CTT) has long been used in the analysis of objective tests, particularly in post-exam evaluations at higher education levels, due to its simplicity and accessibility [11]. Its core concepts form the foundation of modern assessment and rely on elementary algebraic and statistical operations that can be implemented in common spreadsheet software [12–14]. The classical model, $X = T + E$, conceptualises the observed

score (X) as the sum of the true score (T) and measurement error (E) [14–17]. CTT's weak assumptions make it suitable for practical educational contexts, and it provides reliable estimates even with relatively small samples [11–13,18]. It also tolerates suboptimal item formulation, as deficiencies can be compensated by increasing the number of items [12,19]. Despite these advantages, CTT presents significant limitations that motivated the development of Item Response Theory (IRT) [13,20]. Item parameters such as difficulty and discrimination are sample-dependent [12,19–23], meaning that a high-ability group may artificially inflate difficult items [19,24,25]. CTT is also test-dependent: easier tests yield higher scores even when actual performance is unchanged [13,19,23]. The assumption of uniform measurement error is unrealistic, as error varies across performance levels [13,19]. Furthermore, CTT focuses on the total score [19,21,23], offering limited insight into item-level functioning [11]. It assumes equal item contribution to the total score [11,12], which may distort results in heterogeneous populations [18], and it struggles in contexts where all examinees answer items identically, sometimes producing technical errors such as divisions by zero [18].

IRT was developed to address these limitations by modelling the probability of a correct response as a function of latent ability [11,13,17,19,20]. A key advantage is parameter invariance: item parameters are independent of the sample, and ability estimates are independent of the specific items administered [12–14,19]. IRT provides detailed item-level analysis, enabling the detection of low or negative discrimination and supporting more accurate estimation of competence [11,13,19,23]. The Rasch Model places item difficulty and person ability on the same logistic scale, clarifying their relationship [18]. IRT also offers more sophisticated approaches to modelling measurement error [14], detecting bias [18], and accounting for guessing through the three-parameter logistic model (3PL) [23]. It further supports the construction of calibrated item banks, enabling consistent difficulty across test administration [21]. However, IRT requires strong assumptions, such as unidimensionality and local independence, and is more complex to implement [11–13,15,24,25].

In summary, IRT provides a more rigorous and precise measurement than CTT [12,14,18], but CTT remains widely used due to its simplicity and minimal requirements [11–13,19]. Many authors recommend combining both approaches: CTT for basic statistical analysis and IRT for item-level modelling and ability estimation [14,19].

2.2. Logistic Models in Item Response Theory (1PL, 2PL, 3PL)

IRT logistic models describe the probability that an examinee with latent ability θ answers an item correctly [9,26,27]. Parameter invariance is central to these models: item parameters do not depend on the examinee sample, and ability estimates do not depend on the specific items administered [9,28,29]. The One-Parameter Logistic Model (1PL), or Rasch model, assumes that item difficulty (b) is the only parameter influencing the probability of success, with discrimination fixed at $a = 1$ and no guessing ($c = 0$) [26,27,29]. The Two-Parameter Logistic Model (2PL) adds the discrimination parameter (a), which determines how sharply the item differentiates between examinees of different abilities [26–28]. The Three-Parameter Logistic Model (3PL) incorporates difficulty (b), discrimination (a), and a guessing parameter (c), making it particularly suitable for multiple-choice items where low-ability examinees may answer correctly by chance [26–28,30]. The c-parameter represents the lower asymptote of the item characteristic curve. However, parameter recovery becomes less precise as model complexity increases, often resulting in higher Root Mean Square Error (RMSE) values [9,27].

In summary, the 1PL, 2PL, and 3PL models differ in the number of parameters considered, and model selection should reflect the characteristics of the data and the intended use of the assessment [26].

2.3. Competence-Based Assessment Using CTT and IRT

Competence-Based Education (CBE) has become a central paradigm in international educational policy, emphasising the integrated mobilisation of knowledge, skills, attitudes, and values [26,31,32]. Assessment plays a strategic role in CBE, as students must demonstrate mastery of clearly defined learning goals [33–35]. Assessing transversal competences—such as critical thinking, creativity, or digital literacy—poses particular challenges, often requiring more sophisticated measurement approaches [36,37]. CTT and IRT contribute to competence-based assessment by supporting the development of reliable and valid instruments [38]. Item-level analysis enables the refinement of items, the identification of weaknesses, and the construction of high-quality item banks with known psychometric properties [38]. IRT, in particular, supports mastery-based progression by estimating competence levels independently of instructional time [35] and by providing diagnostic information that guides pedagogical decisions [36]. Competence-based formative assessment relies on timely, actionable feedback that helps students monitor their learning and supports teachers in adapting instruction [35,39]. Reliable data are essential for informing educational policy and evaluating the impact of curricular reforms [36,37]. Psychometrically sound assessments also support the use of AI-driven personalised learning pathways [40]. Assessment fulfils multiple functions within competence-based education, including diagnostic, formative, and summative purposes. Diagnostic assessment identifies students' initial levels of mastery and informs instructional planning; formative assessment supports ongoing learning through timely feedback; and summative assessment certifies achievement at the end of an instructional period [32,37]. The quality of the items used in each of these functions directly affects the validity of the inferences drawn [15,23,26]. CTT and IRT contribute to these functions by providing statistical evidence that helps teachers select, refine, and validate items, ensuring that assessment decisions are grounded in reliable and interpretable data [12,13,17,18,20].

In summary, CTT and IRT provide complementary tools for ensuring that competence-based assessment is conducted with scientific rigour, enabling valid inferences about student mastery and supporting continuous improvement of assessment instruments [35,36,38].

2.4. Multiple-Choice Items and Psychometric Analysis

Multiple-choice questions (MCQs) are widely used due to their efficiency, objectivity, and capacity to cover broad curricular content [41–47]. Well-constructed MCQs can assess higher-order cognitive processes, including analysis, evaluation, and problem-solving, in line with Bloom's taxonomy [42,43,48–50]. However, constructing high-quality MCQs is demanding and prone to error [41,49]. IRT is particularly relevant for MCQ analysis because it evaluates item difficulty and discrimination with greater precision than CTT [41]. The discrimination parameter is essential for determining an item's ability to differentiate between competence levels [41,51]. Item analysis—both pre-validation and post-validation—is indispensable for building item banks [42,50,52,53]. Research also shows that three-option MCQs (one correct, two distractors) are often as reliable as four- or five-option items, as plausible distractors are difficult to construct [46–49,54]. Distractor quality is critical: functional distractors attract at least 5% of examinees, while non-functional distractors should be revised or removed [41,49,52,55]. Effective distractors must be plausible, homogeneous, grammatically consistent, and free of clues [50,51,55].

In summary, MCQs are versatile and reliable when constructed and analysed rigorously. CTT and IRT provide essential tools for evaluating item quality and ensuring alignment with cognitive and curricular goals.

2.5. The Literature Gap: Democratising Psychometric Analysis Through Accessible Tools

Despite the shift toward competence-based assessment and the recognised need for psychometric rigour, a significant gap remains in the practical application of CTT and IRT by classroom teachers. Most existing research focuses on large-scale assessments using specialised software, creating technical and financial barriers for everyday practitioners. Teachers typically rely on basic spreadsheet outputs—such as total scores and class averages—that provide limited insight into item functioning.

Recent studies highlight inconsistencies in item quality across years in state-level examinations [56] and point to emerging psychometric models, such as the four-parameter logistic model (4PLM), which account for guessing and carelessness [57]. However, these advanced models remain inaccessible to most teachers due to the absence of simplified implementation guides. There is also a lack of pedagogical literature translating psychometric formulas into spreadsheet-based functions, despite the potential of tools such as Excel and Google Sheets to support detailed item analysis [58].

3. Research Questions and Objectives

This study seeks to address three main research questions, together with the objectives to be achieved:

RQ1—Which statistical analyses, feasible through the use of spreadsheets, can improve the robustness of test analysis?
O1—To enhance spreadsheet-based scoring tools with accessible formulas that support improved statistical analysis of assessment instruments.
RQ2—How can statistical results generated in spreadsheets be meaningfully interpreted by teachers?
O2—To present the foundational principles necessary for effective statistical analysis of applied assessment instruments.
RQ3—How can statistical analysis results be used to improve item construction or support the development of an assessment item database?
O3—To use statistical evidence to refine test items, inform assessment strategies, and structure a reusable item database.

4. Methodology

The study adopts an applied methodological demonstration design, illustrating how Classical Test Theory (CTT) and Item Response Theory (IRT) analyses can be operationalised using accessible spreadsheet tools [8,50]. The intention is not to conduct an empirical validation based on real student data, but rather to show the feasibility and pedagogical relevance of spreadsheet-based psychometric analysis. To this end, all analytical procedures are demonstrated using a simulated dataset constructed to reflect typical response patterns found in closed-response assessment instruments. This approach allows a transparent exposition of the computational steps involved while avoiding the ethical and contextual constraints associated with real assessment data. Real classroom data were not included due to ethical and administrative constraints, particularly the requirement for institutional authorisation to collect, store, and process student responses, even in anonymised form. Although the use of anonymised data might appear feasible, institutional guidelines stipulate that formal approval is required for any handling of student assessment information. To ensure full compliance with these requirements, simulated data were used. This approach also provides complete control over item characteristics and response patterns, ensuring that the examples presented in the manuscript are pedagogically transparent, replicable, and suitable for teacher training and professional development contexts. Although the dataset is simulated, it reflects typical response patterns observed in lower- and upper-secondary

education, particularly in subjects such as science and mathematics where closed-response formats are frequently used. The study addresses the identified gap by demonstrating how widely available spreadsheet software can be transformed into a robust analytical environment for psychometric analysis. By enabling teachers to compute corrected item–total correlations and apply Rasch or logistic models, spreadsheet-based procedures can support item refinement, distractor evaluation, and the development of calibrated item banks without requiring specialised software or programming skills. The methodological demonstration responds directly to the research questions by identifying the statistical indices that can be feasibly computed in spreadsheets (RQ1), clarifying how these indices may be interpreted by teachers (RQ2), and illustrating how such evidence can inform item refinement and the development of assessment item banks (RQ3). The overarching aim is to translate core psychometric principles into accessible analytical processes that can be implemented by educators with limited statistical training.

A simulated dataset was therefore created to represent plausible student responses to a closed-response assessment. The dataset comprises 40 hypothetical students responding to 20 dichotomously scored items (1 = correct; 0 = incorrect), each with four response options. The spreadsheet structure mirrors typical teacher-created scoring grids: students are represented in rows, items in columns, and additional columns compute total scores, corrected totals, and the upper and lower groups required for discrimination analysis. For each item, the correct answer is stored in a dedicated cell (e.g., E7), allowing the spreadsheet to automatically compare each student's response with the key and compute item-level statistics without manual recoding. The sample size of 40 students is adequate for illustrating CTT procedures, which are known to produce stable estimates even with relatively small groups. In contrast, IRT models generally require larger samples for stable parameter estimation; however, because this study does not estimate parameters but instead demonstrates forward calculation of item response probabilities, the sample size is appropriate for the methodological purpose of the analysis. CTT procedures implemented in the spreadsheet include item difficulty, Rasch-based difficulty, item discrimination, corrected item–total correlations, and the probability of random guessing, following established approaches [20,25]. The IRT component extends the demonstration by computing response probabilities under the 1PL, 2PL, and 3PL logistic models using assigned item parameters. The two datasets served distinct analytical purposes. The larger 40×20 matrix was used for the full CTT and IRT demonstrations, including item difficulty, discrimination, corrected item–total correlations, and Rasch-based difficulty. The smaller 10×10 dataset was used exclusively for illustrative purposes, allowing readers to follow the computational steps more easily and to interpret the resulting indices in a compact and visually accessible format. This distinction ensures both methodological rigour and pedagogical clarity. All mathematical expressions and spreadsheet formulas used to operationalise these procedures are provided in Appendices A and B.

5. Results

The results illustrate how spreadsheet-based CTT and IRT analyses can be applied to a simulated dataset of 10 students responding to 10 dichotomous items. The purpose is to demonstrate interpretability, replicability, and pedagogical relevance rather than to generalise findings. The spreadsheet was organised to mirror typical teacher-created scoring grids: students were listed in rows, items in columns, and additional columns computed total scores, corrected totals, and the upper and lower groups required for discrimination analysis. The structure of the spreadsheet and all formulas used are fully documented in Appendices A and B, ensuring transparency and replicability.

5.1. Descriptive Statistics of Total Scores

Total simulated scores ranged from 3 to 9 points, with a mean of 6.2, a median of 6, and a standard deviation of 1.93. The distribution was approximately symmetric, indicating that the dataset was suitable for illustrating item-level analysis and the behaviour of the statistical indices.

5.2. Classical Test Theory Results

5.2.1. Item Difficulty

Item difficulty values (p-values) ranged from 0.20 to 0.90. Seven items fell within the recommended range of 0.30 to 0.80, indicating balanced difficulty. Item 4 was considered very easy (p = 0.90), and Item 8 was very difficult (p = 0.20). Because their p-values fall outside the recommended range, Items 4 and 8 would require revision in a real assessment context.

5.2.2. Rasch-Based Difficulty

Rasch difficulty parameters (b-values) ranged from −2.20 (very easy) to +1.39 (very difficult). Items with negative b-values (e.g., Items 1, 4, 9) were easier, and items with positive b-values (e.g., Items 5 and 8) were more challenging. The Rasch transformation aligned closely with the CTT difficulty results, confirming consistency between the two approaches.

5.2.3. Item Discrimination

Using the 27% method (top 3 vs. bottom 3 students), discrimination values ranged from 0.05 to 0.52. Items 1, 5, and 9 showed strong discrimination, and Items 4 and 8 showed weak discrimination. This indicates that the stronger items effectively differentiated between higher- and lower-performing students, whereas the weaker items did not.

5.2.4. Corrected Item–Total Correlations

Corrected item–total correlations ranged from 0.10 to 0.48. Items 1, 5, and 9 showed good alignment with the overall construct ($r_{it} > 0.30$), and Items 4 and 8 showed weak alignment ($r_{it} < 0.20$). The convergence between discrimination and corrected correlations reinforces the identification of strong and weak items.

5.2.5. Summary of CTT Findings

A compact summary of the CTT indices is presented in Appendix A, Table A2. Items 1, 5, and 9 consistently demonstrated desirable psychometric properties, while Items 4 and 8 underperformed across all indices and would require revision or replacement in a real assessment context.

5.3. Item Response Theory Results

5.3.1. Parameter Behaviour (Demonstration)

Using assigned parameters for demonstration, the 1PL, 2PL, and 3PL models produced probability curves consistent with expected psychometric behaviour. Items 1, 5, and 9 had strong discrimination values (a > 1.20), and Items 4 and 8 had weak discrimination (a < 0.70). Difficulty parameters (b) were consistent with the CTT and Rasch results. Guessing parameters (c) were appropriate for four-option multiple-choice items.

5.3.2. Item Characteristic Curves Interpretation

Item Characteristic Curves (ICCs) were generated in Excel using the θ-P(θ) values computed for three representative items (Items 1, 5, and 8). The θ scale ranged from −4 to +4 in increments of 0.5, and probabilities were calculated using the 3PL model with item-specific parameters. The resulting curves are shown in Figure 1, and the underlying

data are presented in Table 1. The ICCs display the expected psychometric behaviour for items with differing difficulty and discrimination levels. Item 1 shows a steep curve with an early rise, indicating an easy item with strong discrimination. Item 5 presents a similar steep slope but shifts slightly to the right, reflecting moderate difficulty and high discrimination. In contrast, Item 8 exhibits a shallow slope and a higher lower asymptote, suggesting weak discrimination and potential distractor issues. These patterns align with the CTT results, reinforcing the identification of strong and weak items.

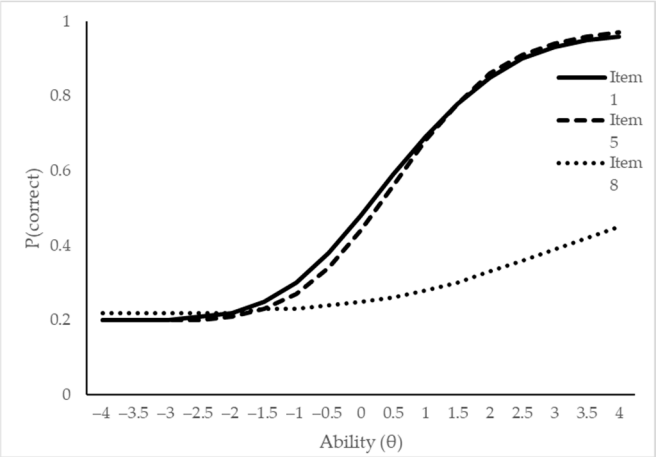


Figure 1. Item Characteristic Curves (selected items).

Table 1. θ-P(θ) Values Used to Generate Item Characteristic Curves.

θ	P Item 1	P Item 5	P Item 8
−4	0.20	0.20	0.22
−3.5	0.20	0.20	0.22
−3	0.20	0.20	0.22
−2.5	0.21	0.20	0.22
−2	0.22	0.21	0.22
−1.5	0.25	0.23	0.23
−1	0.30	0.27	0.23
−0.5	0.38	0.34	0.24
0	0.48	0.44	0.25
0.5	0.59	0.56	0.26
1	0.69	0.68	0.28
1.5	0.78	0.78	0.30
2	0.85	0.86	0.33
2.5	0.90	0.91	0.36
3	0.93	0.94	0.39
3.5	0.95	0.96	0.42
4	0.96	0.97	0.45

5.3.3. Test Information Function (Conceptual)

The simulated test provided the highest measurement precision for ability levels between θ = −0.5 and +0.8, indicating that the test is most informative for students of average to slightly above-average proficiency. Items 5 and 9 contributed most to test information, while Items 4 and 8 contributed minimally.

5.4. CTT–IRT Comparison

The comparison between CTT and IRT results showed strong convergence: Items 1, 5, and 9 consistently emerged as strong performers across all indices, while Items 4

and 8 consistently underperformed. Rasch difficulty values aligned with CTT *p*-values, and 2PL discrimination parameters mirrored CTT discrimination. This coherence demonstrates that spreadsheet-based CTT and IRT analyses can jointly support evidence-based item refinement.

5.5. Spreadsheet Structure

The spreadsheet layouts used for the analyses are present in Appendices A and B. The CTT layout includes columns for item responses, total scores, difficulty, discrimination, and corrected totals. The IRT layout includes ability values and probability outputs for the 1PL, 2PL, and 3PL models, with item parameters stored in fixed cells. These descriptions allow teachers to replicate the analyses.

5.6. Pedagogical Interpretation

The combined CTT and IRT analyses demonstrate that spreadsheet-based psychometric tools can support teachers in identifying weak items, refining distractors and stems, building calibrated item banks, strengthening formative assessment practices, and aligning evaluation procedures with competence-based curricula. Within the simulated dataset, Items 4 and 8 clearly illustrate how statistical evidence can signal the need for item revision, whereas Items 1, 5, and 9 exemplify desirable psychometric behaviour and show how well-functioning items contribute to reliable and interpretable assessment results.

6. Discussion

This study aimed to demonstrate how spreadsheet software can be used to operationalise Classical Test Theory (CTT) and Item Response Theory (IRT) analyses in a manner accessible to teachers. The results obtained from the simulated dataset illustrate that spreadsheets can support a level of psychometric analysis typically associated with specialised software, thereby addressing the gap identified in the literature regarding the practical application of item analysis in everyday educational contexts [56–59]. All mathematical expressions and spreadsheet formulas referenced in this section are provided in Appendices A–C.

6.1. Interpretation of CTT Findings

The CTT indices generated in the spreadsheet—item difficulty, discrimination, and corrected item–total correlation—produced patterns consistent with established psychometric standards. Items with difficulty values within the recommended range and discrimination values above 0.30 were identified as strong performers, aligning with the literature that emphasises balanced difficulty and adequate discrimination for valid inferences [11,19,23]. The convergence between discrimination and corrected item–total correlations reinforces the reliability of the spreadsheet-based approach, as both indices consistently identified Items 1, 5, and 9 as high-quality items and Items 4 and 8 as problematic.

These findings support the argument that CTT remains a valuable framework for classroom assessment due to its conceptual simplicity and suitability for small samples [11–13,18]. By enabling teachers to compute these indices directly in spreadsheets, the approach enhances their capacity to evaluate item functioning without requiring advanced statistical tools.

6.2. Interpretation of IRT Findings

The IRT analyses provided complementary insights that extend beyond CTT. The discrimination parameter in the 2PL model mirrored the CTT discrimination index, while the difficulty parameter aligned with both CTT difficulty values and Rasch logits. The 3PL model further highlighted the influence of guessing, particularly for items with weak

distractors, consistent with research emphasising the importance of accounting for lower asymptotes in multiple-choice items [26–30].

It is important to acknowledge, however, that IRT parameter estimation typically requires large samples to ensure stable and interpretable estimates [20]. In this study, parameters were assigned for demonstration purposes, avoiding the instability that would arise from estimating them with small samples. This limitation does not compromise the pedagogical value of the demonstration but should be considered when applying IRT in real educational contexts.

6.3. Pedagogical Implications for Competence-Based Assessment

The results demonstrate that spreadsheet-based psychometric analysis can meaningfully support competence-based assessment. Teachers can use these tools to:

- identify weak items and revise distractors, stems, or cognitive alignment;
- ensure that assessments measure intended competences with appropriate difficulty;
- build small-scale item banks with known psychometric properties;
- interpret student performance beyond total scores;
- provide targeted feedback based on item-level evidence.

These implications align with the literature emphasising the importance of rigorous assessment practices in competence-based education, where valid and reliable evidence is essential for diagnosing learning needs, supporting formative assessment, and guiding instructional decisions [33–39].

6.4. Feasibility, Practical Constraints and Ethical Considerations

While the spreadsheet-based approach is accessible and transparent, several practical constraints must be acknowledged: (i) although CTT indices can be computed reliably with small samples, IRT parameter estimation typically requires larger datasets, particularly for 2PL and 3PL models [20]; (ii) teachers require basic training in interpreting psychometric indices; without such training, there is a risk of misinterpretation or over-reliance on numerical outputs; (iii) spreadsheets are limited in their capacity to automate complex analyses or handle large datasets. Nevertheless, they remain a powerful entry point for teachers who lack access to specialised software.

Ethical issues arise when assessment data are used to inform pedagogical decisions. Teachers must ensure confidentiality, avoid over-interpretation of statistical indices, and maintain transparency regarding how assessment results are used. The spreadsheet-based approach supports ethical practice by making analytical procedures visible and replicable, reducing the opacity often associated with proprietary software.

6.5. Revisiting the Research Questions and Achievement of Objectives

The study's research questions and objectives were fully addressed through the analyses conducted. The first research question (RQ1) examined whether spreadsheet software could operationalise CTT and IRT procedures in a teacher-friendly manner. The results confirmed that spreadsheets can compute key psychometric indices and implement 1PL, 2PL, and 3PL models, demonstrating that such tools offer an accessible alternative to specialised software. This directly fulfilled the objective of developing a transparent and replicable analytical framework.

The second research question (RQ2) explored the interpretability and pedagogical value of spreadsheet-generated indices. The findings showed consistent patterns across CTT and IRT, with strong items identified by multiple indicators and weaker items highlighted for revision. This confirmed that the spreadsheet-based approach yields meaningful insights aligned with established psychometric literature, thereby achiev-

ing the objective of illustrating how these analyses can support item refinement and competence-based assessment.

The third research question (RQ3) focused on the practical contribution of these analyses to educational practice. The study demonstrated that teachers can use item-level evidence to diagnose learning needs, improve test quality, and construct small-scale item banks. These outcomes fulfilled the objective of demonstrating the pedagogical relevance and practical feasibility of integrating psychometric analysis into everyday assessment.

Together, these results confirm that the study successfully met its aims: to provide an accessible methodological approach, to demonstrate the interpretability of spreadsheet-based psychometrics, and to show how such analyses can meaningfully support competence-based assessment.

6.6. Limitations and Future Research

A key methodological limitation concerns the estimation of parameters in the 2PL and 3PL models. These models generally require large samples to produce stable and interpretable discrimination and guessing parameters, and classroom-sized datasets may lead to substantial instability. Although this study avoided these issues by not estimating parameters, teachers applying IRT in real contexts should be aware that 2PL and 3PL estimation typically require specialised software and larger datasets to ensure psychometric validity.

The use of a simulated dataset also limits the ecological validity of the findings. Future research should apply the spreadsheet-based approach to real classroom data to examine its practical utility and robustness. Additional work could explore the automation of item analysis within spreadsheet environments, as well as the integration of artificial intelligence to support adaptive item selection and personalised learning pathways. Further investigation is also needed into the feasibility of using spreadsheets for larger datasets and more complex IRT models, including the 4PL model highlighted in the recent literature [57].

This study focuses specifically on dichotomous multiple-choice items and therefore does not address the assessment of complex performances typically associated with competency-based education. The methodology is most appropriate for formative purposes, supporting teachers in diagnosing item functioning and improving the quality of their test items. It does not replace assessment approaches that capture higher-order reasoning, argumentation, or the integration of knowledge, skills, and attitudes in authentic contexts. These limitations should be considered when interpreting the potential contribution of the proposed approach to broader educational goals.

7. Conclusions

This study demonstrates that spreadsheet software can serve as an accessible and transparent environment for conducting Classical Test Theory (CTT) and Item Response Theory (IRT) analyses in educational contexts. By operationalising key psychometric procedures within a familiar tool, the study addressed the need for practical, teacher-friendly approaches to item analysis and test quality monitoring. The spreadsheet-generated indices behaved consistently with established psychometric expectations, confirming that this approach can provide meaningful insights into item functioning, test reliability, and student performance. The findings showed that teachers can use CTT and IRT indicators to identify weak items, refine distractors, adjust difficulty levels, and construct small-scale item banks aligned with competence-based assessment. These outcomes reinforce the pedagogical value of integrating psychometric reasoning into everyday assessment practice, supporting more informed instructional decisions and more accurate interpretations of student learning. In doing so, the study contributes to the broader goals of Sustainable Development Goal 4 (SDG 4), which emphasises inclusive, equitable, and high-quality education for all,

as well as the need to strengthen assessment systems that support learning and fairness [60]. The study also highlighted important methodological considerations. While spreadsheets can effectively compute CTT indices and illustrate IRT models, the estimation of 2PL and 3PL parameters requires larger datasets and specialised software to ensure stable and interpretable results. The use of simulated data further limits the ecological validity of the findings, underscoring the need for future research using real classroom datasets. Despite these limitations, the study achieved its objectives: it provided a replicable analytical framework, demonstrated the interpretability of spreadsheet-based psychometrics, and illustrated how such analyses can support competence-based assessment. Future work should explore automation within spreadsheet environments, the integration of artificial intelligence for adaptive assessment, and the feasibility of applying spreadsheets to larger datasets and more complex IRT models. Overall, the study contributes to bridging the gap between psychometric theory and classroom practice, offering teachers a practical and transparent method for improving the quality of their assessment instruments while supporting the global agenda for high-quality education embodied in SDG 4.

Author Contributions: Conceptualisation, A.F. and G.L.M.; methodology, A.F.; software, A.F.; validation, A.F.; formal analysis, A.F.; investigation, A.F.; resources, A.F.; data curation, A.F.; writing—original draft preparation, A.F.; writing—review and editing, A.F.; visualisation, A.F.; supervision, A.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by National Funds through FCT-Portuguese Foundation for Science and Technology, I.P., under the scope of UIDEF—Unidade de Investigação e Desenvolvimento em Educação e Formação, UID/04107/2025, https://doi.org/10.54499/UID/04107/2025.

Data Availability Statement: All data are presented in the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

This appendix provides the simulated dataset and all spreadsheet-generated indices referenced in Section 5 (Results). Tables A1–A5 correspond directly to the analyses described in Sections 5.1–5.5.

Table A1. Simulated Dataset, referenced in Section 5.1 (10 Students × 10 Items).

Student	1	2	3	4	5	6	7	8	9	10	Total
S1	1	1	0	1	0	1	1	0	1	0	6
S2	0	1	1	1	0	1	0	0	1	1	6
S3	1	0	1	1	1	0	1	0	1	1	7
S4	1	1	1	1	0	1	0	0	1	0	6
S5	0	0	0	1	1	1	1	0	0	0	4
S6	1	1	0	1	1	0	1	1	1	1	8
S7	0	1	1	1	0	1	0	0	1	0	5
S8	1	0	1	1	1	1	1	0	1	1	8
S9	1	1	1	1	1	1	1	1	1	0	9
S10	0	0	0	0	0	1	0	0	0	1	2

Table A2. CTT indices, referenced in Sections 5.2.1–5.2.5 (difficulty p, Rasch difficulty b and corrected item–total correlation r_{it}).

Item	p-Value	b-Value	Discrimination	r_{it}
I1	0.8	−1.39	0.45	0.42
I2	0.6	−0.41	0.28	0.29
I3	0.6	−0.41	0.31	0.27
I4	0.9	−2.2	0.1	0.15
I5	0.4	0.41	0.52	0.48
I6	0.7	−0.85	0.33	0.31
I7	0.6	−0.41	0.29	0.26
I8	0.2	1.39	0.05	0.1
I9	0.8	−1.39	0.44	0.39
I10	0.5	0	0.22	0.24

Table A3. IRT parameters, referenced in Section 5.3.1 (simulated discrimination a, difficulty b, and guessing c parameters).

Item	a	b	c
I1	1.2	−1.1	0.2
I4	0.55	−2	0.25
I5	1.45	0.6	0.18
I8	0.6	1.3	0.3
I9	1.5	−0.9	0.22

Table A4. Spreadsheet structure for CTT (layout used to compute p, b and r_{it}).

Column	Description
A	Student
B–K	Item responses (1/0)
L	Total score (=SUM(B2:K2))
M	Corrected total (=L2 − B2)
N	Difficulty p (=COUNTIF(B2:B11,B1)/10)
O	Rasch b (=LN((1 − N)/N))
P	Discrimination (=AVG(upper) − AVG(lower))
Q	r_{it} (=CORREL(B2:B11,M2:M11))

Table A5. Spreadsheet structure for IRT (layout used to compute 1PL, 2PL, and 3PL probabilities for a range of θ values).

Column	Description
A	Ability values θ (from −4 to +4)
B	1PL probability
C	2PL probability
D	3PL probability
F2	Difficulty parameter b
G2	Discrimination parameter a
H2	Guessing parameter c

Appendix B

This appendix consolidates all mathematical expressions and spreadsheet formulas used in the study, including descriptive statistics, Classical Test Theory (CTT) indices, Item Response Theory (IRT) models, and Excel implementations. The construction of assessment instruments aims to evaluate latent traits in students, and Item Response

Theory (IRT) is widely applied for this purpose [1]. Although item parameters can be estimated even in non-representative samples, the literature emphasises that large samples are generally required—particularly for one-parameter (1PL) and two-parameter (2PL) models—to ensure stable and interpretable estimates [20].

Descriptive statistics provide essential information about the distribution of student performance. The sample mean is a measure of central tendency that summarises the average performance of a group of students, while the population mean is used when the complete dataset of all students is available for a given assessment element. These measures are meaningful only for quantitative variables and support the interpretation of test results within both CTT and IRT frameworks [59].

The following sections present the mathematical definitions and corresponding Excel formulas used throughout the study.

Table A6. Descriptive statistics.

Function	Mathematical Formula	Excel
Sample mean	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	=AVERAGE(range)
Population mean	$\mu = \frac{1}{n}\sum_{i=1}^{N} X_i$	=AVERAGE(range)
Corrected sample variation	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2$	=VAR.S(range)
Population variance	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (X_i - \mu)^2$	=VER.P(range)
Population standard deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}}$	=STDEV.P(range)
Minimum		=MIN(range)
Maximum		=MAX(range)
Range		=MAX(range) − MIN(range)

Table A7. CTT formulas.

Function	Mathematical Formula	Excel
Item difficulty (p)	$p = \frac{numberofcorrectresponses}{N}$	=COUNTIF(E8:E47,E7)/COUNT(E8:E47)
Rasch difficulty (b)	$b = \ln\left(\frac{1-p}{p}\right)$	=LN((1 − (COUNTIF(E8:E47,E7) /COUNT(E8:E47)))/(COUNTIF(E8:E47,E7)/COUNT(E8:E47)))
Item discrimination (27% method)	$D = X_{upper} - X_{lower}$	=AVERAGE(E8:E18) − AVERAGE(E37:E47)
Corrected item–total correlation	r_{it} = correlation between item and corrected total	=CORREL(E8:E39,AW8:AW39)
Random guess probability	$g = \frac{1}{m}$	=1/(number of MCQ options)

Table A8. IRT models.

Model	Mathematical Formula	Excel
One-parameter logistic model (1PL/Rasch)	$P(\theta) = \frac{1}{1+e^{-(\theta-b)}}$	=1/(1 + EXP(−(A8 − $B$2)))
Two-parameter logistic model (2PL)	$P(\theta) = \frac{1}{1+e^{-\alpha(\theta-b)}}$	=1/(1 + EXP(−$C$2 × (A8 − $B$2)))
Three-parameter logistic model (3PL)	$P(\theta) = c + (1-c)\frac{1}{1+e^{-\alpha(\theta-b)}}$	=$D$2 + (1 − $D$2)/(1 + EXP(−$C$2 × (A8 − $B$2)))

Table A9. IRT models and Excel formulations.

Model	Mathematical Formula	Excel Formula	Parameters	Pedagogical Interpretation
1PL (Rasch)	$P(\theta) = \frac{1}{1+e^{-(\theta-b)}}$	=1/(1 + EXP(−(θ − b)))	b = difficulty	Fixed discrimination; suitable for standardised tests
2PL	$P(\theta) = \frac{1}{1+e^{-\alpha(\theta-b)}}$	=1/(1 + EXP(−a × (θ − b)))	α = discrimination b = difficulty	Variable discrimination; higher α indicates better item sensitivity
3PL	$P(\theta) = c + (1-c)\frac{1}{1+e^{-\alpha(\theta-b)}}$	=c + (1 − c) × (1/(1 + EXP(−a × (θ − b))))	α = discrimination b = difficulty c = guessing	Accounts for random guessing; ideal for multiple-choice items

Table A10. Supporting Excel indicators.

Indicator	Excel Formula	Pedagogical Interpretation
Proportion correct (p)	=COUNTIF(responses;1)/COUNT(responses)	Classical difficulty index
Logit transformation (b)	=LN((1 − p)/p)	Converts p to logit scale
Item–ability correlation	=CORREL(item;ability)	Proxy for discrimination (approximates α)
Corrected item–total correlation	=CORREL(item;total score)	Used in test revision and quality monitoring

Appendix C

Table A11. Item difficulty classification.

Proportion of Correct Responses	Interpretation
0.00–0.30	Difficult item
0.31–0.70	Moderate difficult item
0.71–1.00	Easy item

Table A12. Item correlation interpretation.

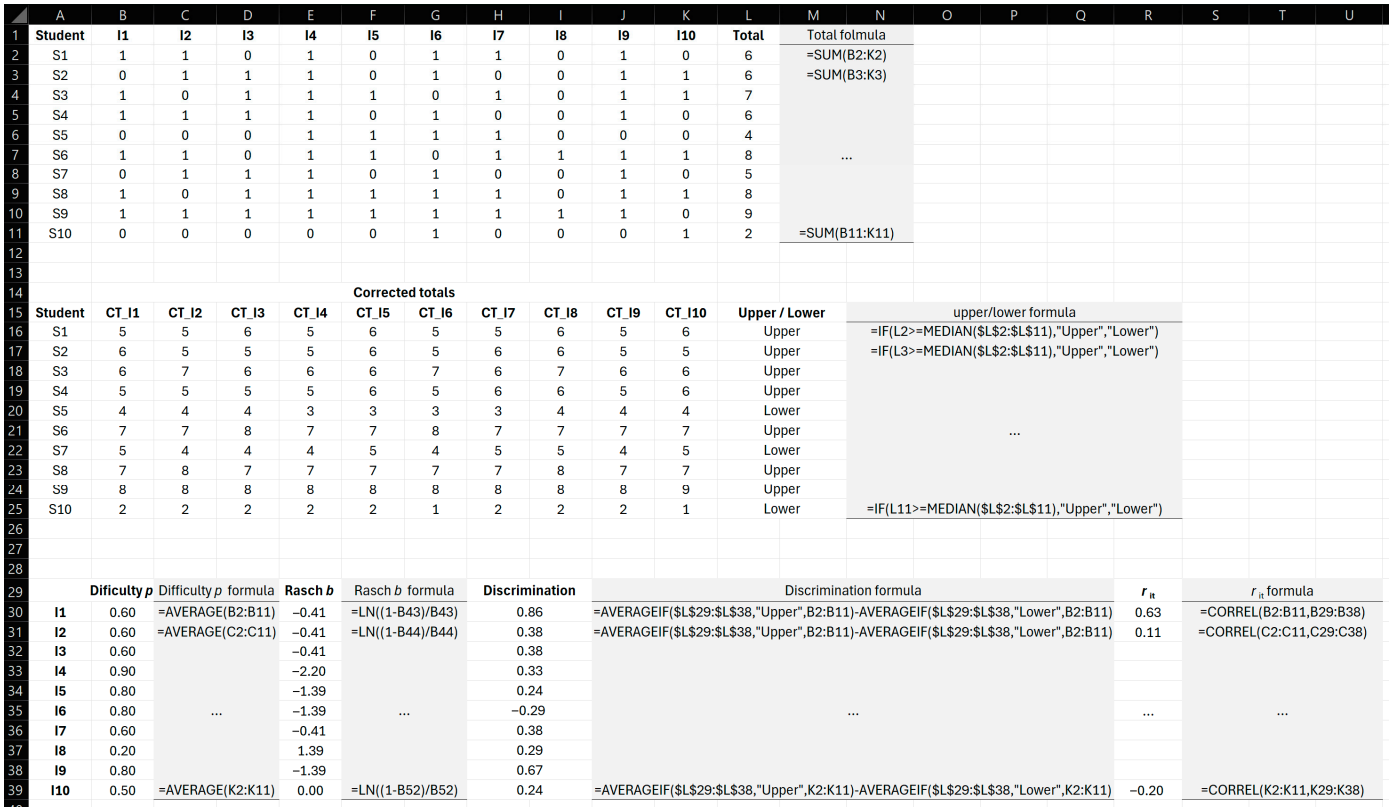
Item–Total Correlation	Interpretation
≥0.40	Excellent discrimination
0.30–0.39	Good discrimination
0.20–0.29	Acceptable; could be improved
<0.20	Weak discrimination—item may be poorly formulated
Negative	Problematic item—inverse to expected performance trend
Blank cell	Non-discriminating item—should be reviewed

Table A13. Additional interpretation guidelines.

Indicator	Interpretation
Item difficulty	If p is close to the guessing value, the item may be poorly constructed or too difficult.
Item discrimination	High guessing probability and low discrimination suggests the item does not differentiate ability levels.
Distractor analysis	If many students select the correct option by chance, distractors may not be functioning effectively.

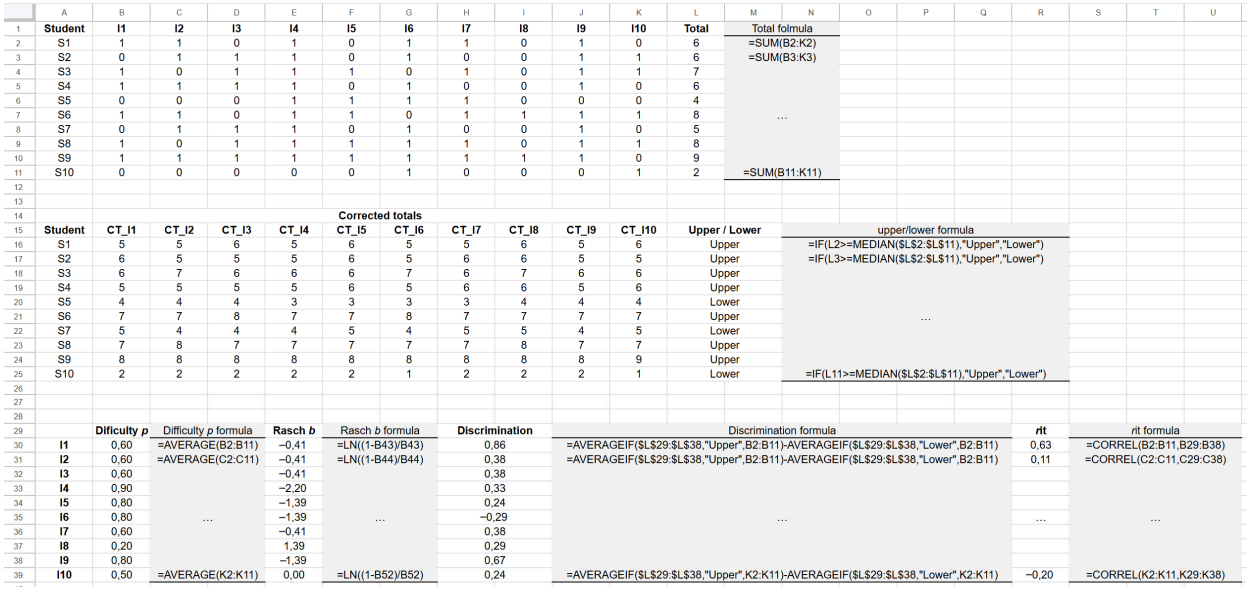
Appendix D

The first screenshot (Figure A1) illustrates the organisation of the student x item response matrix, total scores, corrected totals, and item-level indices used in the Classical Test Theory procedures.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	**Student**	**I1**	**I2**	**I3**	**I4**	**I5**	**I6**	**I7**	**I8**	**I9**	**I10**	**Total**	Total folmula								
2	S1	1	1	0	1	0	1	1	0	1	0	6	=SUM(B2:K2)								
3	S2	0	1	1	1	0	1	0	0	1	1	6	=SUM(B3:K3)								
4	S3	1	0	1	1	1	0	1	0	1	1	7									
5	S4	1	1	1	1	0	1	0	0	1	0	6									
6	S5	0	0	0	1	1	1	1	0	0	0	4									
7	S6	1	1	0	1	1	0	1	1	1	1	8	...								
8	S7	0	1	1	1	0	1	0	0	1	0	5									
9	S8	1	0	1	1	1	1	1	0	1	1	8									
10	S9	1	1	1	1	1	1	1	1	1	0	9									
11	S10	0	0	0	0	0	1	0	0	0	1	2	=SUM(B11:K11)								
12																					
13																					
14							**Corrected totals**														
15	**Student**	**CT_I1**	**CT_I2**	**CT_I3**	**CT_I4**	**CT_I5**	**CT_I6**	**CT_I7**	**CT_I8**	**CT_I9**	**CT_I10**	**Upper / Lower**	upper/lower formula								
16	S1	5	5	6	5	6	5	5	6	5	6	Upper	=IF(L2>=MEDIAN($L$2:$L$11),"Upper","Lower")								
17	S2	6	5	5	5	6	5	6	6	5	5	Upper	=IF(L3>=MEDIAN($L$2:$L$11),"Upper","Lower")								
18	S3	6	7	6	6	6	7	6	7	6	6	Upper									
19	S4	5	5	5	5	6	5	6	6	5	6	Upper									
20	S5	4	4	4	3	3	3	3	4	4	4	Lower									
21	S6	7	7	8	7	7	8	7	7	7	7	Upper	...								
22	S7	5	4	4	4	5	4	5	5	4	5	Lower									
23	S8	7	8	7	7	7	7	7	8	7	7	Upper									
24	S9	8	8	8	8	8	8	8	8	8	9	Upper									
25	S10	2	2	2	2	2	1	2	2	2	1	Lower	=IF(L11>=MEDIAN($L$2:$L$11),"Upper","Lower")								
26																					
27																					
28																					
29		**Dificulty *p***	Difficulty *p* formula	**Rasch *b***	Rasch *b* formula	**Discrimination**	Discrimination formula											r_{it}	r_{it} formula		
30	**I1**	0.60	=AVERAGE(B2:B11)	−0.41	=LN((1-B43)/B43)	0.86	=AVERAGEIF($L$29:$L$38,"Upper",B2:B11)-AVERAGEIF($L$29:$L$38,"Lower",B2:B11)											0.63	=CORREL(B2:B11,B29:B38)		
31	**I2**	0.60	=AVERAGE(C2:C11)	−0.41	=LN((1-B44)/B44)	0.38	=AVERAGEIF($L$29:$L$38,"Upper",B2:B11)-AVERAGEIF($L$29:$L$38,"Lower",B2:B11)											0.11	=CORREL(C2:C11,C29:C38)		
32	**I3**	0.60		−0.41		0.38															
33	**I4**	0.90		−2.20		0.33															
34	**I5**	0.80		−1.39		0.24															
35	**I6**	0.80	...	−1.39	...	−0.29	...											...	...		
36	**I7**	0.60		−0.41		0.38															
37	**I8**	0.20		1.39		0.29															
38	**I9**	0.80		−1.39		0.67															
39	**I10**	0.50	=AVERAGE(K2:K11)	0.00	=LN((1-B52)/B52)	0.24	=AVERAGEIF($L$29:$L$38,"Upper",K2:K11)-AVERAGEIF($L$29:$L$38,"Lower",K2:K11)											−0.20	=CORREL(K2:K11,K29:K38)		

Figure A1. Classical Test Theory Analysis Spreadsheet (Excel extract).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	**Student**	**I1**	**I2**	**I3**	**I4**	**I5**	**I6**	**I7**	**I8**	**I9**	**I10**	**Total**	Total folmula								
2	S1	1	1	0	1	0	1	1	0	1	0	6	=SUM(B2:K2)								
3	S2	0	1	1	1	0	1	0	0	1	1	6	=SUM(B3:K3)								
4	S3	1	0	1	1	1	0	1	0	1	1	7									
5	S4	1	1	1	1	0	1	0	0	1	0	6									
6	S5	0	0	0	1	1	1	1	0	0	0	4									
7	S6	1	1	0	1	1	0	1	1	1	1	8	...								
8	S7	0	1	1	1	0	1	0	0	1	0	5									
9	S8	1	0	1	1	1	1	1	0	1	1	8									
10	S9	1	1	1	1	1	1	1	1	1	0	9									
11	S10	0	0	0	0	0	1	0	0	0	1	2	=SUM(B11:K11)								
12																					
13																					
14							**Corrected totals**														
15	**Student**	**CT_I1**	**CT_I2**	**CT_I3**	**CT_I4**	**CT_I5**	**CT_I6**	**CT_I7**	**CT_I8**	**CT_I9**	**CT_I10**	**Upper / Lower**	upper/lower formula								
16	S1	5	5	6	5	6	5	5	6	5	6	Upper	=IF(L2>=MEDIAN($L$2:$L$11),"Upper","Lower")								
17	S2	6	5	5	5	6	5	6	6	5	5	Upper	=IF(L3>=MEDIAN($L$2:$L$11),"Upper","Lower")								
18	S3	6	7	6	6	6	7	6	7	6	6	Upper									
19	S4	5	5	5	5	6	5	6	6	5	6	Upper									
20	S5	4	4	4	3	3	3	3	4	4	4	Lower									
21	S6	7	7	8	7	7	8	7	7	7	7	Upper	...								
22	S7	5	4	4	4	5	4	5	5	4	5	Lower									
23	S8	7	8	7	7	7	7	7	8	7	7	Upper									
24	S9	8	8	8	8	8	8	8	8	8	9	Upper									
25	S10	2	2	2	2	2	1	2	2	2	1	Lower	=IF(L11>=MEDIAN($L$2:$L$11),"Upper","Lower")								
26																					
27																					
28																					
29		**Dificulty *p***	Difficulty *p* formula	**Rasch *b***	Rasch *b* formula	**Discrimination**	Discrimination formula											**rit**	*rit* formula		
30	**I1**	0,60	=AVERAGE(B2:B11)	−0,41	=LN((1-B43)/B43)	0,86	=AVERAGEIF($L$29:$L$38,"Upper",B2:B11)-AVERAGEIF($L$29:$L$38,"Lower",B2:B11)											0,63	=CORREL(B2:B11,B29:B38)		
31	**I2**	0,60	=AVERAGE(C2:C11)	−0,41	=LN((1-B44)/B44)	0,38	=AVERAGEIF($L$29:$L$38,"Upper",B2:B11)-AVERAGEIF($L$29:$L$38,"Lower",B2:B11)											0,11	=CORREL(C2:C11,C29:C38)		
32	**I3**	0,60		−0,41		0,38															
33	**I4**	0,90		−2,20		0,33															
34	**I5**	0,80		−1,39		0,24															
35	**I6**	0,80	...	−1,39	...	−0,29	...											...	...		
36	**I7**	0,60		−0,41		0,38															
37	**I8**	0,20		1,39		0,29															
38	**I9**	0,80		−1,39		0,67															
39	**I10**	0,50	=AVERAGE(K2:K11)	0,00	=LN((1-B52)/B52)	0,24	=AVERAGEIF($L$29:$L$38,"Upper",K2:K11)-AVERAGEIF($L$29:$L$38,"Lower",K2:K11)											−0,20	=CORREL(K2:K11,K29:K38)		

Figure A2. Classical Test Theory Analysis Spreadsheet (GoogleSheet extract).

This third screenshot (Figure A3) shows the θ grid, the 1PL, 2PL, and 3PL probability formulas, and the item parameter table used to generate the Item Characteristic Curves.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	**Student**	**I1**	**I2**	**I3**	**I4**	**I5**	**I6**	**I7**	**I8**	**I9**	**I10**	**Total**					
2	S1	1	1	0	1	0	1	1	0	1	0	6					
3	S2	0	1	1	1	0	1	0	0	1	1	6					
4	S3	1	0	1	1	1	0	1	0	1	1	7					
5	S4	1	1	1	1	0	1	0	0	1	0	6					
6	S5	0	0	0	1	1	1	1	0	0	0	4		Item	a	b	c
7	S6	1	1	0	1	1	0	1	1	1	1	8		I1	1.20	−1.1	0.20
8	S7	0	1	1	1	0	1	0	0	1	0	5		I4	0.55	−2.00	0.25
9	S8	1	0	1	1	1	1	1	0	1	1	8		I5	1.45	0.60	0.18
10	S9	1	1	1	1	1	1	1	1	1	0	9		I8	0.60	1.30	0.30
11	S10	0	0	0	0	0	1	0	0	0	1	2		I9	1.50	−0.9	0.22
12																	
13																	
14	θ	1PL	1PL formula			2PL	2PL formula				3PL	3PL formula					
15	−4	0.05	=1/(1+EXP(−(A23−$C$15)))			0.03	=1/(1+EXP(-−$B$15×(A23−$C$15)))				0.22	=$D$15+(1−$D$15)/(1+EXP(−$B$15×(A23−$C$15)))					
16	−3.5	0.08	=1/(1+EXP(−(A24−$C$15)))			0.05	=1/(1+EXP(−$B$15×(A24−$C$15)))				0.24	=$D$15+(1−$D$15)/(1+EXP(−$B$15×(A24−$C$15)))					
17	−3	0.13				0.09					0.27						
18	−2.5	0.20				0.16					0.33						
19	−2	0.29				0.25					0.40						
20	−1.5	0.40				0.38					0.51						
21	−1	0.52				0.53					0.62						
22	−0.5	0.65				0.67					0.74						
23	0	0.75		...		0.79		...			0.83			...			
24	0.5	0.83				0.87					0.90						
25	1	0.89				0.93					0.94						
26	1.5	0.93				0.96					0.97						
27	2	0.96				0.98					0.98						
28	2.5	0.97				0.99					0.99						
29	3	0.98				0.99					0.99						
30	3.5	0.99				1.00					1.00						
31	4	0.99	=1/(1+EXP(−(A39−$F$23)))			1.00	=1/(1+EXP(−$B$15×(A39−$C$15)))				1.00	=$D$15+(1−$D$15)/(1+EXP(−$B$15×(A39−$C$15)))					
32																	
33																	
34	θ	P_Item1	P_item formula			θ	P Item 1	P Item5	P Item8	Note: The θ-P(θ) values presented in Table 1 are pedagogical							
35	−4	0.22	=K21			−4	0.20	0.20	0.22	approximations chosen to illustrate the qualitative form of the							
36	−3.5	0.24	=K22			−3.5	0.20	0.20	0.22	item characteristic curves. They are not intended to represent the							
37	−3	0.27				−3	0.20	0.20	0.22	exact probabilities produced by the 3PL model using the							
38	−2.5	0.33				−2.5	0.21	0.20	0.22	parameters calculated here.							
39	−2	0.40				−2	0.22	0.21	0.22								
40	−1.5	0.51				−1.5	0.25	0.23	0.23								
41	−1	0.62				−1	0.30	0.27	0.23								
42	−0.5	0.74				−0.5	0.38	0.34	0.24								
43	0	0.83	...			0	0.48	0.44	0.25								
44	0.5	0.90				0.5	0.59	0.56	0.26								
45	1	0.94				1	0.69	0.68	0.28								
46	1.5	0.97				1.5	0.78	0.78	0.30								
47	2	0.98				2	0.85	0.86	0.33								
48	2.5	0.99				2.5	0.90	0.91	0.36								
49	3	0.99				3	0.93	0.94	0.39								
50	3.5	1.00				3.5	0.95	0.96	0.42								
51	4	1.00	=K37			4	0.96	0.97	0.45								
52																	

Figure A3. Item Response Theory Analysis Spreadsheet (Excel extract).

In Item Response Theory (IRT), the parameters a, b, and c describe how each item behaves across different levels of ability. The discrimination parameter (a) indicates how effectively an item differentiates between respondents whose ability levels (θ) are close to one another; higher values produce a steeper item characteristic curve, meaning the item is more sensitive to small differences in ability. The difficulty parameter (b) represents the point on the ability scale at which the probability of answering the item correctly (excluding guessing) is approximately 0.50, with higher values corresponding to more difficult items. The guessing or pseudo-guessing parameter (c) reflects the lower asymptote of the curve, representing the probability that a respondent with very low ability answers the item correctly by chance. In multiple-choice items, this value often approximates the reciprocal of the number of response options (e.g., around 0.25 for a four-option item).

In operational testing, these parameters are typically estimated from response data using specialised statistical software and likelihood-based methods. However, when the purpose is pedagogical rather than inferential, it is common to assign plausible values to these parameters in order to illustrate how discrimination (a), difficulty (b), and guessing (c) influence the shape of the item characteristic curve. In this study, the values of a, b, and c were therefore specified theoretically to support the demonstration of the 1PL, 2PL, and 3PL models.

References

1. Bruner, J. *The Process of Education*; Harvard University Press: Cambridge, MA, USA, 1960.
2. Bruner, J. *Toward a Theory of Instruction*; Harvard University Press: Cambridge, MA, USA, 1966.
3. van Merriënboer, J.J.G.; Clark, R.E.; de Croock, M.B.M. Blueprints for complex learning: The 4C/ID-model. *ETR&D* **2002**, *50*, 39–61. [CrossRef]
4. Uner, O.; Tekin, E.; Roediger, H.L. True-false tests enhance retention relative to rereading. *J. Exp. Psychol. Appl.* **2022**, *28*, 114–129. [CrossRef] [PubMed]
5. Roediger, H.L.; Brown, P.C. The importance of testing as a learning strategy. *Sch. Adm.* **2019**, *76*, 35–37. Available online: http://psychnet.wustl.edu/memory/wp-content/uploads/2021/01/Roediger_Brown_School-Administrator.pdf (accessed on 15 November 2025).
6. Pinto, A.C. Factores relevantes na avaliação escolar por perguntas de escolha múltipla. *Psicol. Educ. Cult.* **2001**, *5*, 23–44. Available online: https://www.fpce.up.pt/docentes/acpinto/artigos/15_pergunt_escolha_multipla.pdf (accessed on 2 November 2025).
7. Gasigwa, T.; Bimenyimana, S.; Nteziryimana, J. Teachers' attitude towards the use of excel software, to teach and learn statistics on learner's performance in rwanda's kicukiro public upper secondary schools. *J. Res. Innov. Implic. Educ.* **2024**, *8*, 246–252. [CrossRef]
8. Muhammad, G.A.; Adamu, A.; Zubair, S.I.; Usman, H.A. Excel As an ICT Tool For Increasing Teacher Proficiency Towards Quality Education: A Panacea for Addressing Challenges Confronting Nigeria Education System. *Int. J. Educ. Eval.* **2024**, *10*, 40–56. Available online: https://ijee.io/Abstract/3857/excel-as-an-ict-tool-for-increasing-teacher-proficiency-towards-quality-education-a-panacea-for-addressing-challenges-confronting-nigeria-education-system (accessed on 16 November 2025).
9. Hambleton, R.; Swaminathan, H.; Rogers, H. *Fundamentals of Item Response Theory*; SAGE: Hemet, CA, USA, 1991.
10. Haleem, A.; Javaid, M.; Qadri, M.A.; Suman, R. Understanding the role of digital technologies in education: A review. *Sustain. Oper. Comput.* **2022**, *3*, 275–285. [CrossRef]
11. Akbari, A. The rasch analysis of item response theory: An untouched area in evaluating student academic translations. *SKASE J. Transl. Interpret.* **2025**, *18*, 50–77. [CrossRef]
12. Hu, Z.F.; Lin, L.; Wang, Y.H.; Li, J.W. The integration of classical testing theory and item response theory. *Psychology* **2021**, *12*, 1397–1409. Available online: https://www.scirp.org/journal/paperinformation?paperid=111936 (accessed on 15 November 2025).
13. Ayanwale, M.A.; Chere-Masopha, J.; Morena, M.C. The classical test or item response measurement theory: The status of the framework at the examination council of Lasotho. *Int. J. Learn. Teach. Educ. Res.* **2022**, *21*, 384–406. [CrossRef]
14. Eleje, L.I.; Onah, F.E.; Abanobi, C.C. Comparative study of classical test theory and item response theory using diagnostic quantitative economics skill test item analysis results. *Eur. J. Educ. Soc. Sci.* **2018**, *3*, 57–75. Available online: https://dergipark.org.tr/tr/pub/ejees/issue/40156/477675 (accessed on 2 November 2025).
15. Allen, M.; Yen, W. *Introduction to Measurement Theory*; Brooks/Cole Publishing Company: Pacific Grove, CA, USA, 1979; p. 57.
16. Kurniawan, D.D.; Syifa, A.; Huda, N.; Kusuma, M. Item analysis of teacher made test in Biology subject. In *5th International Conference on Current Issues in Education (ICCIE 2021)*; Atlantis Press: Dordrecht, The Netherlands, 2022; pp. 312–317. [CrossRef]
17. Butakor, P.K. Using classical test and item response theories to evaluate psychometric quality of teacher-made test in Ghana. *ESJ* **2022**, *18*, 139. Available online: https://eujournal.org/index.php/esj/article/view/15098 (accessed on 15 November 2025).
18. Priyani, T.; Sugiharto, B. Analysis of biology midterm exam items using a comparison of the classical theory test and the Rasch model. *JPBI* **2024**, *10*, 939–958. [CrossRef]
19. Nasir, M. Application of Classical Test Theory and Item Response Theory to Analyze Multiple Choice Questions. Doctoral Thesis, University of Calgary, Calgary, AB, Canada, 2014. Available online: http://hdl.handle.net/11023/1917 (accessed on 2 November 2025).
20. Sartes, L.; de Souza-Formigoni, M. Avanços na psicometria: Da teoria clássica dos testes à teoria de resposta ao item. *Psicol. Reflexão Crítica* **2013**, *26*, 241–250. [CrossRef]
21. Bhakta, B.; Tennant, A.; Horton, M.; Lawton, G.; Andrich, D. Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education. *BMC Med. Educ.* **2005**, *5*, 9. [CrossRef]
22. Hamidah, N. The quality of test on national examination of natural science in the level of elementary school. *Int. J. Eval. Res. Educ.* **2022**, *11*, 604–616. [CrossRef]
23. Brown, G.; Abdulnabi, H. Evaluating the quality of higher education instructor-constructed multiple-choice tests: Impact on student grades. *Front. Educ.* **2017**, *2*, 24. [CrossRef]
24. Janssen, G.; Meier, V.; Trace, J. Classical test theory and item response theory: Two understandings of one high-stakes performance exam. *Colomb. Appl. Linguist. J.* **2014**, *16*, 167–184. [CrossRef]
25. Lahza, H.; Smith, T.G.; Khosravi, H. Beyond item analysis: Connecting student behaviour and performance using e-assessment logs. *Br. J. Educ. Technol.* **2022**, *54*, 335–354. [CrossRef]
26. Furr, R.; Bacharach, V. *Psychometrics: An Introduction*; SAGE: Hemet, CA, USA, 2018; pp. 314–334.

27. Saatçioğlu, F.; Atar, H. Investigation of the effect of parameter estimation and classification accuracy in mixture IRT models under different conditions. *Int. J. Assess. Tools Educ.* **2022**, *9*, 1013–1029. [CrossRef]
28. Jumini, J.; Retnawati, H. Estimating item parameters and student abilities: An IRT 2PL analysis of mathematics examination. *Al-Ishlah J. Pendidik.* **2022**, *14*, 385–398. [CrossRef]
29. Liu, D.; Mueller, C.; Sedaghat, A. A scoping review of Rasch analysis and item response theory in otolaryngology: Implications and future possibilities. *Laryngoscope Investig. Otolaryngol.* **2024**, *9*, e1208. [CrossRef] [PubMed]
30. Barbetta, P.; Trevisan, L.; Tavares, H.; Azevedo, T. Aplicação da Teoria da Resposta ao Item uni e multidimensional. *Estud. Em Avaliação Educ.* **2014**, *25*, 280–302. [CrossRef]
31. Howells, K. *The Future of Education and Skills: Education 2030: The Future We Want*; OECD: Paris, France, 2018. Available online: https://repository.canterbury.ac.uk/download/96f6c3f39ae6dcffa26e72cefe47684172da0c93db0a63d78668406e4f478ae8/3102592/E2030%20Position%20Paper%20%2805.04.2018%29.pdf (accessed on 16 November 2025).
32. Dillon, S. *OECD Future of Education and Skills 2030: OECD Learning Compass 2030*; OECD: Paris, France, 2019. Available online: https://www.oecd.org/content/dam/oecd/en/about/projects/edu/education-2040/1-1-learning-compass/OECD_Learning_Compass_2030_Concept_Note_Series.pdf (accessed on 16 November 2025).
33. Le, C.; Wolfe, R.; Steinberg, A. *The Past and the Promise: Today's Competency Education Movement*; Students at the Center: Competency Education Research Series; Jobs for the Future: Boston, MA, USA, 2014. Available online: https://files.eric.ed.gov/fulltext/ED561253.pdf (accessed on 11 November 2025).
34. Education Commission of the States. Available online: https://www.ecs.org/wp-content/uploads/CBE-Toolkit-2017.pdf (accessed on 15 November 2025).
35. Evans, C.M.; Landl, E.; Thompson, J. Making sense of K-12 competency-based education: A systematic literature review of implementation and outcomes research from 2000 to 2019. *J. Competency-Based Educ.* **2020**, *5*, e01228. [CrossRef]
36. Looney, J.; Kelly, G. *Assessing Learners' Competences—Policies and Practices to Support Successful and Inclusive Education—Thematic Report*; Publications Office of the European Union: Luxembourg, 2023. Available online: https://curated-library.iiep.unesco.org/library-record/assessing-learners-competences-policies-and-practices-support-successful-and (accessed on 11 November 2025).
37. European Commission/EACEA/Eurydice. *Developing Key Competences at School in Europe: Challenges and Opportunities for Policy*; Eurydice Report; Publications Office of the European Union: Luxembourg, 2012. Available online: https://eurydice.eacea.ec.europa.eu/publications/developing-key-competences-school-europe-challenges-and-opportunities-policy (accessed on 11 November 2025).
38. Pacheco, L.; Degering, L.; Mioto, F.; Gresse von Wangenheim, C.; Borgato, A.; Petri, G. Improvements in bASES21: 21st-Century Skills Assessment Model to K12. In *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU 2020)*; SciTePress: Setúbal, Portugal, 2020; Volume 1, pp. 297–307. [CrossRef]
39. European Commission, Directorate-General for Education and Culture. *Key Competences for Lifelong Learning*; Publications Office of the European Union: Luxembourg, 2019. Available online: https://www.fi.uu.nl/publicaties/literatuur/2018_eu_key_competences.pdf (accessed on 11 November 2025).
40. European Education Area. Available online: https://education.ec.europa.eu/document/action-plan-on-basic-skills-graphic-version (accessed on 16 November 2025).
41. Kumar, A.P.; Nayak, A.; Shenoy, M.; Goyal, S. A novel approach to generate distractors for multiple choice questions. *Expert Syst. Appl.* **2023**, *225*, 120022. [CrossRef]
42. Kar, S.S.; Lakshminarayanan, S.; Mahalakshmy, T. Basic principles of constructing multiple choice questions. *Indian J. Community Fam. Med.* **2015**, *1*, 65–69. [CrossRef]
43. Shin, J.; Guo, Q.; Gierl, M.J. Multiple-choice item distractor development using topic modeling approaches. *Front. Psychol.* **2019**, *10*, 825. [CrossRef]
44. Kiat, J.; Ong, A.R.; Ganesan, A. The influence of distractor strength and response order on MCQ responding. *Educ. Psychol.* **2018**, *38*, 368–380. [CrossRef]
45. Toksöz, S.; Ertunc, A. Item Analysis of a Multiple-Choice Exam. *Adv. Lang. Lit. Stud.* **2017**, *8*, 141–146. [CrossRef]
46. Vegada, B.; Shukla, A.; Khilnani, A.; Charan, J.; Desai, C. Comparison between three option, four option and five option multiple choice question tests for quality parameters: A randomized study. *Indian J. Pharmacol.* **2016**, *48*, 571–575. [CrossRef]
47. Nwadinigwe, P.I.; Naibi, L. The number of options in a multiple-choice test item and the psychometric characteristics. *J. Educ. Pract.* **2013**, *4*, 189–196. Available online: https://www.iiste.org/Journals/index.php/JEP/article/view/9944/10148 (accessed on 2 November 2025).
48. Dehnad, A.; Nasser, H.; Hosseini, A. comparison between three-and four-option multiple choice questions. *Procedia-Soc. Behav. Sci.* **2014**, *98*, 398–403. [CrossRef]
49. Tarrant, M.; Ware, J.; Mohammed, A.M. An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. *BMC Med. Educ.* **2009**, *9*, 40. [CrossRef]

50. Romão, G.; Sá, M. Como elaborar questões de múltipla escolha de boa qualidade. *Femina* **2019**, *47*, 561–564. Available online: https://docs.bvsalud.org/biblioref/2019/12/1046547/femina-2019-479-561-564.pdf (accessed on 17 November 2025).
51. Lai, H.; Gierl, M.J.; Touchie, C.; Pugh, D.; Boulais, A.P.; De Champlain, A. Using automatic item generation to improve the quality of MCQ distractors. *Teach. Learn. Med.* **2016**, *28*, 166–173. [CrossRef]
52. Ansari, M.; Sadaf, R.; Akbar, A.; Rehman, S.; Chaudhry, Z.; Shakir, S. Assessment of distractor efficiency of MCQS in item analysis. *Prof. Med. J.* **2022**, *29*, 730–734. [CrossRef]
53. Royal, K.; Stockdale, M.R. The impact of 3-option responses to multiple-choice questions on guessing strategies and cut score determinations. *J. Adv. Med. Educ. Prof.* **2017**, *5*, 84. Available online: https://pmc.ncbi.nlm.nih.gov/articles/PMC5346173/ (accessed on 17 November 2025).
54. Loudon, C.; Macias-Muñoz, A. Item statistics derived from three-option versions of multiple-choice questions are usually as robust as four-or five-option versions: Implications for exam design. *Adv. Physiol. Educ.* **2018**, *42*, 565–575. [CrossRef]
55. Rezigalla, A.; Eleragi, A.; Elhussein, A.; Alfaifi, J.; ALGhamdi, M.; Al Ameer, A.; Yahia, A.; Mohammed, O.; Adam, M. Item analysis: The impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Med. Educ.* **2024**, *24*, 445. [CrossRef]
56. Opesemowo, O.A.G.; Opatunji, K.O.; Babatimehin, T.; Opesemowo, T.R. Analysis of 2022 and 2023 Osun State basic education certificate examination mathematics items using item response theory: Implications for large scale assessment. *Soc. Sci. Humanit. Open* **2026**, *13*, 102381. [CrossRef]
57. Kasali, J.; Opesemowo, O.A.; Faremi, Y.A. Psychometric analysis of senior secondary school certificate examination (SSCE) 2017 NECO English language multiple choice test item in KWARA state using item response theory. *J. Appl. Res. Multidiscip. Stud.* **2022**, *3*, 83–102. [CrossRef]
58. Kumar, M. A study on importance of microsoft excel data analysis statistical tools in research works. *J. Manag. Educ. Res. Innov.* **2023**, *1*, 25–33. [CrossRef]
59. Marôco, J. *Analise Estatística Com o SPSS Statistics*, 8th ed.; ReportNumber: Lisboa, Portugal, 2021.
60. Goals 4 Ensure Inclusive and Equitable Quality Education and Promote Lifelong Learning Opportunities for All. Available online: https://sdgs.un.org/goals/goal4 (accessed on 24 November 2025).