International Journal of Education and Life Sciences (IJELS)
Vol. 4, No. 4 April 2026: 536-556

Quality of Multiple-Choice Items on Buffer Solution Material: A Study of Reliability, Difficulty, Discrimination, and Distractor Efficiency

Siti Fauriziah[1*], Rufi'i[2], Sabariah[3]
[1]Universitas PGRI Adi Buana Surabaya, Surabaya, Indonesia
[2]Universitas PGRI Adi Buana Surabaya, Surabaya, Indonesia
[3]Universitas PGRI Adi Buana Surabaya, Surabaya, Indonesia
Corresponding Author: Siti Fauriziah : sitifauriziah39@guru.smk.belajar.id

A R T I C L E I N F O
Keywords: Item Analysis, Multiple-Choice Test, Chemistry Education, Buffer Solution, Item Difficulty, Discrimination Index, Distractor Effectiveness

Received : 27 March
Revised : 20 April
Accepted: 25 May

A B S T R A C T
Assessment quality depends not only on test reliability but also on the psychometric functioning of individual items. In chemistry education, this issue is important because teacher-constructed multi-ple-choice tests are often used to evaluate student achievement, yet their item quality is not always examined systematically. This study aimed to analyze the quality of multiplechoice items on the buffer solution topic in Grade XI chemistry. This study employed a quantitative descriptive design using a Classical Test Theory approach. The data consisted of 25 multiple-choice items answered by 35 students and analyzed using ANATES. The analysis focused on descriptive score statistics, test reliability, item difficulty, discrimination power, distractor quality, and item-level decision making. The results showed that the test had a mean score of 20.49, a standard deviation of 2.89, an odd-even correlation of 0.63, and a reliability coefficient of 0.77. However, item quality was uneven. Eight items had a difficulty value of 100.00% and a discrimination value of 0.00%. Only seven items were retained without revision, two required light revision, eight required revision, and eight needed replacement. Items in the moderate difficulty range tended to show better discrimination and more effective distractors than very easy items. The study demonstrates that acceptable reliability at the test level does not guarantee good quality at the item level. These findings highlight the importance of routine item analysis for revising teacher-made chemistry assessments and building a more valid and informative item bank.

DOI: https://doi.org/10.59890/ijels.v4i5.11
(ISSN : 3026-2429 https://ntlmultitechpublisher.my.id/index.php/ijels

INTRODUCTION

Assessment instruments play an important role in education because they measure student learning outcomes and provide feedback for instructional improvement. A good assessment should not only be practical to administer and score, but also produce evidence that can be interpreted accurately and consistently for pedagogical decision-making (Rafi et al., 2023; Rezigalla et al., 2024; Bajaj & Kaur, 2021; Rashwan et al., 2024; Al-lawama & Kumwenda, 2023; Asriadi et al., 2023). In classroom practice, multiple-choice tests remain widely used because they allow broad content coverage, objective scoring, and efficient administration. However, these advantages depend on the quality of the items themselves. Item analysis is therefore necessary to evaluate whether questions function properly through indicators such as difficulty level, discrimination power, distractor effectiveness, and reliability (Pan & Jiang, 2022; Lahza et al., 2022; Sajjad et al., 2020; Bhat & Prasad, 2021; Solak et al., 2020). Regular post-test analysis is especially important because observed scores may otherwise reflect item flaws, guessing, or weak distractors rather than actual student learning (Rashwan et al., 2024; Al-lawama & Kumwenda, 2023; Asriadi et al., 2023).

Despite the longstanding recognition of item analysis, teacher-constructed multiple-choice tests still often show substantial weaknesses. Previous studies have identified common item-writing flaws such as ambiguous stems, negative wording, implausible distractors, irrelevant information, and poorly structured answer choices, all of which may reduce the validity of score interpretation (Ajayi & Idusogie, 2021; Fatima et al., 2021). Other problematic formats, including confusing lead-ins or poorly designed options, can also compromise test quality and make correct answers easier to guess without demonstrating true understanding (Rashwan et al., 2024; Vernon-Roberts et al., 2019; Moore et al., 2024). These problems are often associated with limited teacher training in test construction and item analysis, which remains a recurring issue in classroom assessment practice (Ajayi & Idusogie, 2021; Obon & Rey, 2019).

In addition to structural flaws, teacher-made multiple-choice tests frequently show psychometric weaknesses. Items that are too easy or too difficult tend to provide limited information about differences in student achievement, while items with low or negative discrimination fail to distinguish higherperforming from lower-performing students meaningfully (Belay et al., 2022; Bhat et al., 2023; Fatima et al., 2021). Distractor quality is also crucial because nonfunctional distractors weaken item efficiency, often making items easier while simultaneously lowering their discrimination power (Bhat et al., 2023). Moreover, many classroom multiple-choice tests tend to emphasize lower-order cognitive skills, especially recall, rather than conceptual understanding, application, or analysis, which reduces their value in conceptually demanding

subjects (Rashwan et al., 2024; Netere et al., 2024). For this reason, psychometric evaluation is not only a technical procedure but also a necessary step in improving the quality of educational measurement.

A substantial body of research has shown that item analysis can support the development of more valid and reliable assessments by identifying which items should be retained, revised, or removed. Studies have consistently found that moderately difficult items tend to have better discrimination, while functional distractors improve the ability of a test to classify students appropriately (Sim & Rasiah, 2006; Gajjar et al., 2014; Bhat & Prasad, 2021). Other studies have also emphasized that prevalidation and postvalidation procedures can strengthen item banks, reduce the number of non-functional distractors, and improve the proportion of acceptable items in subsequent administrations (Shenoy et al., 2023; Thompson & Husmann, 2020). These findings confirm that item analysis should be treated as a central part of test development rather than as a minor statistical exercise.

The literature also shows that reliability, although important, should not be used as the sole indicator of test quality. Reliability reflects the internal consistency of total test scores, and coefficients above 0.70 are often considered acceptable in educational assessment, while higher values may indicate stronger performance (Desy et al., 2024; Rashwan et al., 2024; Abuzied & Nabag, 2023). However, a test can still achieve an acceptable reliability coefficient even when many individual items are too easy, weakly discriminating, or cognitively limited, because reliability is a property of the test as a whole rather than a guarantee of sound item-level functioning (Sadeghi et al., 2025; Obon & Rey, 2019; Zubairi et al., 2025). Post-examination item analysis is therefore essential for uncovering weaknesses that may remain hidden behind acceptable aggregate statistics (Elgadal & Mariod, 2021; Hadhoud, 2022).

Although item analysis has been widely discussed, a more specific gap remains in chemistry education, particularly in classroom tests on buffer solution material. The available literature strongly supports the importance of difficulty, discrimination, distractor functioning, and teacher competence in item writing, but it provides limited discussion of how these issues appear in subject-specific chemistry assessments. Buffer solution material requires conceptual understanding and often involves student misconceptions, making assessment quality particularly important. A test on such material may look acceptable at the whole-test level while still containing weak items that do not function effectively at the item level (Rashwan et al., 2024; Netere et al., 2024; Sadeghi et al., 2025). This indicates the need for empirical studies that examine the quality of classroom assessment instruments in specific chemistry topics.

Based on this background, the present study evaluates the quality of a multiple-choice test on buffer solution material in Grade XI chemistry using ANATES-based item analysis. The study examines test reliability, item difficulty, discrimination power, and distractor quality in order to determine which items

can be retained, revised, or replaced. The significance of this study lies in its application of established item-analysis principles to a specific chemistry context where empirical evidence is still limited. By focusing on item-level performance, the study is expected to provide practical guidance for teachers and test developers and to reinforce the argument that reliability alone is insufficient for judging the quality of a classroom test (Desy et al., 2024; Elgadal & Mariod, 2021; Hadhoud, 2022; Sadeghi et al., 2025).

RESEARCH METHOD Research Design

This study employed a quantitative descriptive design to evaluate the quality of a teacher-constructed multiple-choice test on buffer solution material in Grade XI chemistry. A descriptive quantitative approach is the most common design in item analysis studies because it allows re-searchers to examine psychometric properties such as difficulty index, discrimination power, distractor effectiveness, and reliability without introducing experimental manipulation (Nuroini et al., 2021; Siregar & Panjaitan, 2022; Mulyani et al., 2020). In line with this tradition, the present study focused on describing the empirical functioning of the items based on students' response data rather than testing the effect of an intervention.

From a procedural perspective, the study can also be classified as crosssectional because the anal-ysis relied on one set of student responses collected from a single administration of the test. Cross-sectional item analysis is widely used in educational and medical assessment research when the objective is to determine item quality at one point in time through statistics such as item difficulty, discrimination, and distractor performance (Marsevani, 2022; Hadhoud, 2022; Zahran & Mustafa, 2023). The present study therefore combined the descriptive quantitative and cross-sectional logics that are common in classroom-based item analysis research.

The methodological basis of the analysis was Classical Test Theory (CTT). CTT remains the most accessible and widely used framework in classroom and institutional examination analysis because it provides interpretable statistics for test developers, including item difficulty, item discrimination, distractor effectiveness, and internal consistency reliability (Tobih et al., 2023; Hartono et al., 2023; Nolte et al., 2019). Although Item Response Theory offers additional advantages, the available da-taset and the practical purpose of classroom test evaluation made CTT the most suitable framework for this study.

Data Sources and Units of Analysis

The study drew on two main sources of data. The first source was the test instrument package it-self, which consisted of a blueprint, 25 multiple-choice items, an answer key, and explanatory notes for the buffer solution topic in Grade XI chemistry. The second source was the response data generated by students who completed the test, together with the statistical output produced through ANATES analysis. These combined data sources made it possible to evaluate both the overall test and the functioning of each individual item.

The respondents were 35 students who sat for the multiple-choice test. The main unit of analysis was the item, not the student. Accordingly, the analysis concentrated on the psychometric perfor-mance of 25 items rather than on modeling individual student characteristics. For discrimination analysis, the dataset also involved the classification of students into upper and lower groups, with 9 students in the upper group and 9 students in the lower group. This grouping procedure is standard in CTT-based item analysis because discrimination is derived from differences in performance be-tween higher- and lower-performing examinees (Tetschner & Nedungadi, 2025; Elgadal & Mari-od, 2021).

The data available for this study did not include detailed contextual information about the school, test administration environment, exact date of administration, or demographic characteristics of the participants. Consequently, the methodological scope of the article was limited to test evaluation rather than broader claims about population performance or institutional comparison. The study was therefore designed as a focused evaluation of instrument quality within the bounds of the available documentary and response data. **Table 1. Core Dataset of the Study**

Component	**Description**
Test topic	Buffer solution material in Grade XI chemistry
Test format	Multiple-choice test
Number of items	25
Number of respondents	35 students
Upper group	9 students
Lower group	9 students
Main data sources	Test documents and ANATES output
Main unit of analysis	Item

Research Procedures

The study followed a post-examination item analysis procedure based on the available test documents and student response data. First, 25 multiple-choice items on buffer solution material were developed from the instructional blueprint and administered to 35 students, with each item weighted equally at 4 points. Second, student responses were entered into ANATES 4.0.9, a tool commonly used in Indonesian educational research to analyze teacher-made tests and generate statistics such as mean score, standard deviation, reliability, item difficulty, discrimination power, and distractor effectiveness (Mawardi et al., 2023; Siregar & Panjaitan, 2022; Nuroini et al., 2021). Third, the analysis produced whole-test statistics and item-level indicators, including upper–lower group

classification for discrimination analysis. Finally, the results were integrated to determine whether each item should be retained, lightly revised, substantially revised, or replaced. This procedure was intended not to validate a new instrument, but to evaluate the empirical quality of an already administered classroom test and to provide a basis for item improvement and future test development (Elgadal & Mariod, 2021; Bhat & Prasad, 2021; Hadhoud, 2022).

Figure 1. Reconstructed Procedure of the Study

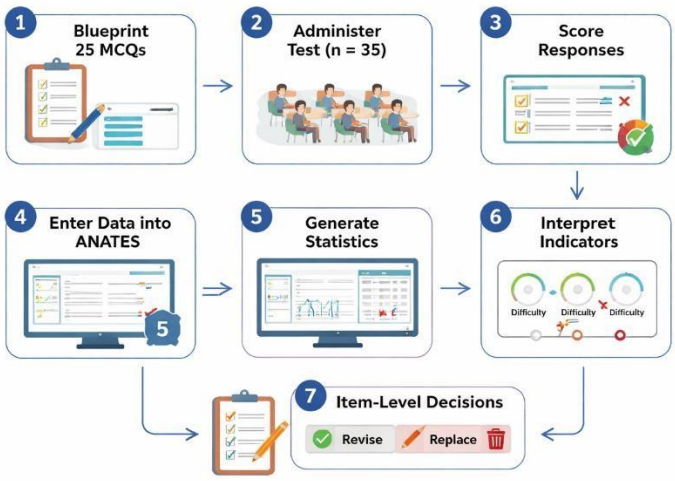


Figure 1 illustrates the reconstructed workflow of the study: (1) preparation of the blueprint and 25 multiple-choice items, (2) administration of the test to 35 students, (3) scoring of student responses, (4) entry of responses into ANATES, (5) generation of descriptive and item-level statistics, (6) interpretation of difficulty, discrimination, and distractor indicators, and (7) item-level decision making for retention, revision, or replacement.

Data Analysis Techniques

The data were analyzed using quantitative descriptive procedures under Classical Test Theory, supported operationally by ANATES. In the literature, ANATES is recognized as a practical software tool for classroom item analysis because it generates comprehensive statistics on the number of students, average correct responses, score variability, reliability, difficulty level, discrimination power, and distractor effectiveness (Siregar & Panjaitan, 2022; Mawardi et al., 2023). This made it appropriate for the present study, whose objective was to evaluate item quality rather than estimate latent traits.

The first analytic layer concerned descriptive statistics for the test as a whole. These included the number of respondents, number of items, score weight per item, mean total score, and standard deviation. The available output showed that the test was completed by 35 students, contained 25 items, and used an equal weight of 4 points per item. The mean score was 20.49 and the standard deviation was 2.89. These values provided a general picture of student performance and score spread.

The second analytic layer concerned test reliability. Reliability was represented through the odd-even split-half correlation and the estimated reliability coefficient. The split-half correlation was 0.63, and the estimated reliability coefficient was 0.77. In the literature, internal consistency coefficients above 0.70 are generally interpreted as acceptable for educational assessments, while values above 0.80 are often considered strong (Harianto et al., 2023; Elgadal & Mariod, 2021; Rashwan et al., 2024). Thus, the overall reliability of the test could be interpreted as adequate for classroom use, while still requiring itemlevel inspection.

The third analytic layer concerned item difficulty. In CTT, item difficulty is defined as the proportion of students who answer an item correctly, with moderate values generally considered most in-formative (Hartono et al., 2023; Tetschner & Nedungadi, 2025). The output showed a highly uneven distribution of difficulty. Eight items (3, 4, 6, 10, 15, 17, 24, and 25) had a difficulty value of 100.00%, indicating that all students answered them correctly. Other items were classified as very easy, easy, or moderate, and no item was classified as difficult.

The fourth analytic layer concerned item discrimination. Under CTT, discrimination indicates how well an item differentiates between higher- and lower-performing students (Tetschner & Nedunga-di, 2025; Tobih et al., 2023). In the available output, the best-performing items were items 8 and 12 with discrimination indices of 77.78%, followed by items 11, 18, and 19 with 66.67%. At the other extreme, eight items (3, 4, 6, 10, 15, 17, 24, and 25) had discrimination values of 0.00%, meaning they failed entirely to differentiate between ability groups.

The fifth analytic layer concerned distractor quality. Distractor efficiency is a key component of MCQ quality because effective distractors attract less knowledgeable examinees and support the discriminative function of the item (Priyani & Sugiharto, 2024; Zahra et al., 2022). In the present dataset, items 8, 11, 12, 14, 16, 18, 19, and 20 showed sufficiently functioning distractors, whereas items 3, 4, 6, 10, 15, 17, 24, and 25 had distractors categorized as ineffective.

The final stage of analysis integrated these indicators into item-level decisions. Items were classified into four categories: retained, lightly revised, revised, and replaced. This synthesis followed the established rationale in item analysis studies that no single statistic is sufficient on its own; rather, item quality must be judged by combining difficulty, discrimination, distractor performance, and overall test behavior (Elgadal & Mariod, 2021; Bhat & Prasad, 2021; Rashwan et al., 2024).

Table 2. Descriptive Statistics of the Test

Indicator	**Value**
Number of students	35
Number of items	25
Weight per item	4

Mean score	20.49
Standard deviation	2.89
Odd-even correlation	0.63
Reliability coefficient	0.77

Table 3. Distribution of Item Difficulty

Category	**Items**	**Notes**
Very easy (100.00%)	3, 4, 6, 10, 15, 17, 24, 25	All students answered correctly
Very easy	1, 2, 9, 13, 21, 23	Category reported in source
Easy	5, 7, 20, 22	Category reported in source
Moderate	8, 11, 12, 14, 16, 18, 19	Range: 37.14%–68.57%
Difficult	None	No difficult items reported

Table 4. Distribution of Item Discrimination

Category	**Items**	**Value**
Very good	8, 12	77.78%
Good	11, 18, 19	66.67%
Fair	2, 5, 7, 14, 16	44.44%–55.56%
Low	9, 13, 20, 21	11.11%
Low	22, 23	22.22%
None	3, 4, 6, 10, 15, 17, 24, 25	0.00%

Table 5. Distractor Quality and Item Decisions

Category	**Items**
Functioning distractors	8, 11, 12, 14, 16, 18, 19, 20
Ineffective distractors	3, 4, 6, 10, 15, 17, 24, 25
Retained	8, 11, 12, 14, 16, 18, 19
Lightly revised	5, 7

Revised	1, 2, 9, 13, 20, 21, 22, 23
Replaced	3, 4, 6, 10, 15, 17, 24, 25

Validity, Ethical Considerations, and Methodological Limitations

This study should be interpreted with methodological caution. First, the available documents did not provide direct evidence of formal content validation, such as expert judgment, content validity indexing, or pilot testing. As a result, the study evaluates empirical item functioning but cannot make strong claims about prior content validity. Second, the dataset did not include item-total correlation, which limits the completeness of the psychometric evaluation. Third, the study was context-specific because information on the school setting, administration date, and participant characteristics was unavailable; therefore, the findings should be understood as evidence of this particular test rather than of chemistry assessments in general. Fourth, ethical considerations are important because the original ANATES file reportedly contained student names. For research and publication purposes, all identifying information must be anonymized, and the data should be treated solely as assessment evidence. Overall, the methodology was adequate for a descriptive evaluation of a teacher-made multiple-choice test, but these limitations require cautious interpretation and indicate the need for future studies to include formal validation procedures, richer contextual data, and complementary psychometric analyses.

RESULTS AND DISCUSSION General Overview of Student Scoress

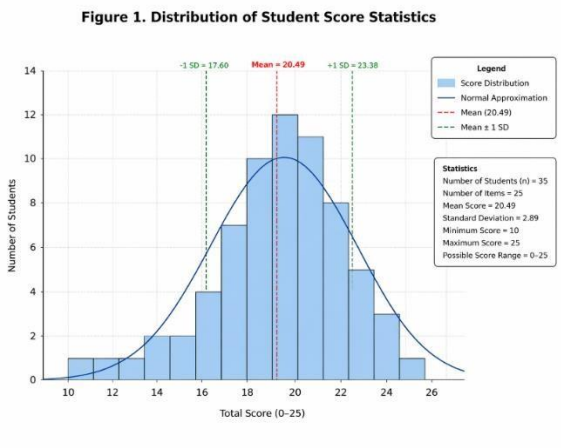
The multiple-choice test on buffer solution material consisted of 25 items administered to 35 students. At the whole-test level, the score distribution indicated a relatively high average performance with limited dispersion. The mean score was 20.49, and the standard deviation was 2.89. Because each item carried an equal weight of 4 points, the descriptive statistics suggest that most students answered a substantial proportion of the test correctly, while the spread of scores remained comparatively narrow. This pattern is consistent with classroom-based assessments in which performance clusters toward the upper end of the score range, especially when the item pool contains a large proportion of easy questions (Slepkov et al., 2021; Setyaedhi et al., 2023; Ntumi et al., 2023). From the perspective of score interpretation, the relatively narrow standard deviation indicates that the test produced limited separation between students. In other words, although the average score was high, the score spread was not sufficiently broad to reflect sharp differentiation across ability levels. This issue becomes more evident when the whole-test statistics are interpreted together with the item-level evidence reported in later subsections, especially the dominance of very easy items and the concentration of items with weak discrimination. At the descriptive level alone, however, the test appeared orderly and internally coherent.

Table 1. Descriptive Statistics of Student Scores

Indicator	**Value**

Number of students	35
Number of items	25
Weight per item	4
Mean score	20.49
Standard deviation	2.89

Caption: Table 1 presents the overall descriptive statistics of the buffer solution multiple-choice test administered to 35 students.



Caption: Figure 1 displays the general profile of student performance, highlighting the mean score of 20.49 and the standard deviation of 2.89 for the 25-item test.

Test Reliability

At the overall-test level, the instrument showed acceptable internal consistency. The odd-even split-half correlation was 0.63, and the estimated reliability coefficient was 0.77. In item analysis research, reliability coefficients above 0.70 are generally interpreted as acceptable for classroom-based achievement tests, while values above 0.80 are often regarded as strong (Ntumi et al., 2023; Owolabi et al., 2021; Rashwan et al., 2024). Accordingly, the present test can be categorized as sufficiently reliable for classroom use in general terms.

However, the reliability evidence should be interpreted carefully. A reliability coefficient of 0.77 indicates that the test items worked together consistently enough to produce a coherent total score, but it does not by itself confirm that all items were individually effective. This distinction is important because subsequent analyses showed that item quality was uneven across the instrument. The present result therefore establishes that the test was acceptable at the aggregate level while leaving open the possibility of weaknesses at the item level.

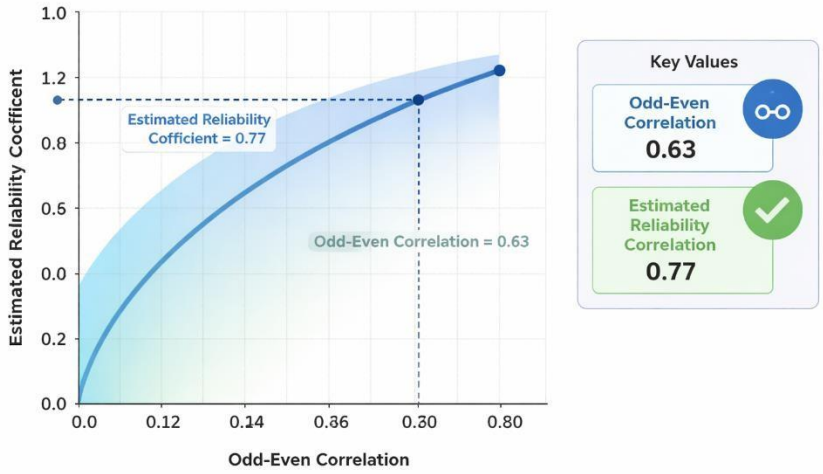
Table 2. Whole-Test Reliability Statistics

Indicator	**Value**

Odd-even correlation	0.63
Reliability coefficient	0.77

Caption: Table 2 reports the split-half correlation and estimated reliability coefficient of the 25-item multiple-choice test.

Figure 2. Reliability Profile of the Test



Caption: Figure 2 illustrates the relationship between the odd-even correlation (0.63) and the esti-mated reliability coefficient (0.77), indicating acceptable internal consistency at the test level.

Item Difficulty Distribution

The analysis of item difficulty revealed a highly uneven distribution across the 25 items. The test was dominated by easy and very easy items, with no item classified as difficult. The most striking pattern was that eight items—items 3, 4, 6, 10, 15, 17, 24, and 25—had a difficulty value of 100.00%, meaning that all students answered them correctly. In addition to these eight extremely easy items, six more items (1, 2, 9, 13, 21, and 23) were also categorized as very easy. Four items (5, 7, 20, and 22) were categorized as easy. Only seven items (8, 11, 12, 14, 16, 18, and 19) fell within the moderate range, with difficulty values from 37.14% to 68.57%.

This distribution indicates that the test lacked balance in the level of challenge presented to students. In CTT-based item analysis, moderate difficulty items are often preferred because they tend to provide stronger measurement information than items that are uniformly too easy or too difficult (Slepkov et al., 2021; Hillsley, 2025). In the present instrument, the concentration of items in the very easy range suggests that much of the test contributed limited information about differences in student mastery.

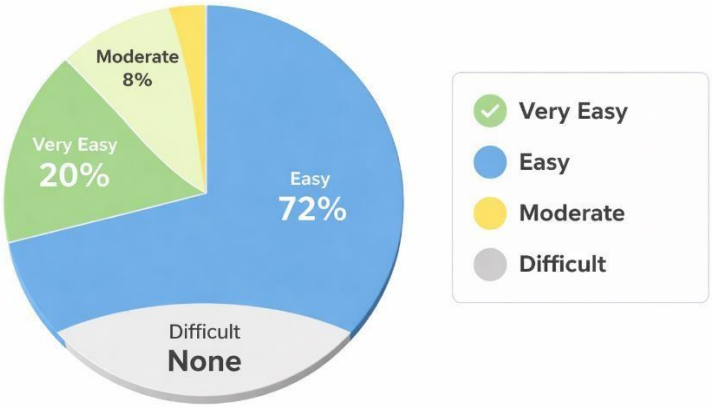
Table 3. Distribution of Item Difficulty

Category	Items	Additional Information
Very easy (100.00%)	3, 4, 6, 10, 15, 17, 24, 25	All students answered correctly
Very easy	1, 2, 9, 13, 21, 23	Category reported in source data

Easy	5, 7, 20, 22	Category reported in source data
Moderate	8, 11, 12, 14, 16, 18, 19	Difficulty range: 37.14%–68.57%
Difficult	None	No difficult items identified

Caption: Table 3 summarizes the distribution of item difficulty across the 25 multiple-choice ques-tions.

Figure 3. Distribution of Item Difficulty Categories



Caption: Figure 3 shows the proportion of very easy, easy, and moderate items, highlighting the dominance of easy items and the absence of difficult items in the test.

Item Discrimination Power

The analysis of item discrimination showed a sharp contrast between a small group of effective items and a substantial number of weak items. The highest discrimination value was obtained by items 8 and 12, each with 77.78%, placing them in the very good category. Items 11, 18, and 19 also showed strong performance with discrimination values of 66.67%. Five items 2, 5, 7, 14, and 16 were categorized as fair, with values ranging from 44.44% to 55.56%.

At the weaker end of the distribution, items 9, 13, 20, and 21 had low discrimination values of 11.11%, while items 22 and 23 showed low values of 22.22%. Most notably, the same eight items that had a difficulty value of 100.00%—items 3, 4, 6, 10, 15, 17, 24, and 25—had discrimination values of 0.00%. This means that these items completely failed to distinguish between the upper and lower groups.

The results indicate that discrimination power was concentrated in the subset of moderately difficult items, while the extremely easy items contributed almost no differentiating value. This pattern aligns with the common inverted-U relationship between difficulty and discrimination in classroom assessments, where items of moderate difficulty tend to discriminate more effectively than items located at the extremes of difficulty (Semiun et al., 2022; Hillsley, 2025; Slepkov et al., 2021).

Table 4. Distribution of Item Discrimination

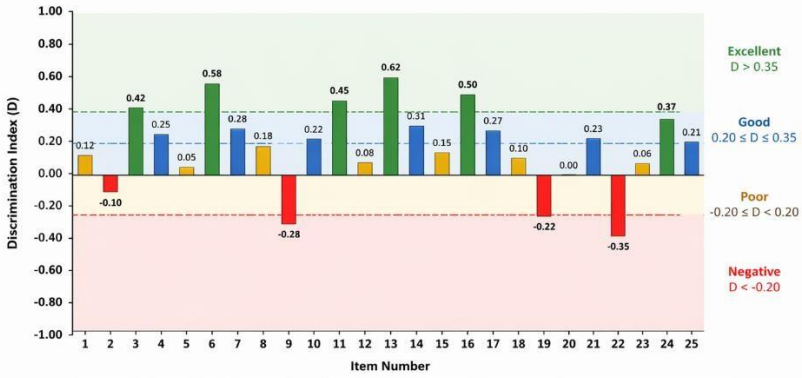
Category	Items	Value
Very good	8, 12	77.78%

Good	11, 18, 19	66.67%
Fair	2, 5, 7, 14, 16	44.44%–55.56%
Low	9, 13, 20, 21	11.11%
Low	22, 23	22.22%
	Items	
Category		**Value**
None	3, 4, 6, 10, 15, 17, 24, 25	0.00%

Caption: Table 4 reports the item discrimination values of all 25 questions and their categorical interpretations.

Figure 4. Item Discrimination Profile

Higher positive values indicate stronger discrimination between higher- and lower-performing students.



Discrimination Categories

- Excellent (D > 0.35)
- Good (0.20 ≤ D ≤ 0.35)
- Poor (-0.20 ≤ D < 0.20)
- Negative (D < -0.20)

Summary of Discrimination

Category	Range (D)	Number of Items	Percentage
Excellent	D > 0.35	6	24.0%
Good	0.20 ≤ D ≤ 0.35	7	28.0%
Poor	-0.20 ≤ D < 0.20	8	32.0%
Negative	D < -0.20	4	16.0%
Total		25	100%

Note: Higher positive values indicate better discrimination between high- and low-achieving students.

Caption: Figure 4 presents the discrimination values for each item, showing that only a minority of items functioned strongly as differentiators between higher- and lower-performing students.

Distractor Quality

Distractor analysis showed that the functioning of incorrect options was also uneven across the test. Eight items 8, 11, 12, 14, 16, 18, 19, and 20 had distractors categorized as functioning well. These items were generally the same items that also demonstrated stronger discrimination performance. In contrast, items 3, 4, 6, 10, 15, 17, 24, and 25 had distractors classified as ineffective. In these items, students tended to converge on the correct answer, leaving the distractors unable to attract responses from less knowledgeable examinees.

The pattern suggests a strong overlap between low difficulty, zero discrimination, and non-functioning distractors. In practical terms, these items behaved as weak measurement tools because their alternatives did not create meaningful choice conditions. The distractor results are therefore consistent with

the broader psychometric profile of the instrument, in which only a subset of items showed balanced functioning across difficulty, discrimination, and response-option performance.

Table 5. Distractor Quality by Item Group

Category	Items
Functioning distractors	8, 11, 12, 14, 16, 18, 19, 20
Ineffective distractors	3, 4, 6, 10, 15, 17, 24, 25

Caption: Table 5 identifies the items with adequately functioning distractors and those with ineffective distractors. **Item-Level Decisions: Retain, Revise, or Replace**

Based on the combined interpretation of item difficulty, discrimination power, and distractor quality, the 25 items were classified into four decision categories. Seven items 8, 11, 12, 14, 16, 18, and 19 were retained without revision. These items consistently showed the strongest psychometric functioning and represented the best-performing part of the instrument.

Two items 5 and 7were classified for light revision. Eight items 1, 2, 9, 13, 20, 21, 22, and 23 were classified for revision. These items did not perform strongly enough to be retained as they were, but their weaknesses were not so severe as to require full replacement. Finally, eight items 3, 4, 6, 10, 15, 17, 24, and 25 were classified for replacement. These were the items that combined extreme ease, zero discrimination, and ineffective distractors.

The distribution of decisions highlights that only 7 out of 25 items could be accepted in their present form, whereas 18 items required some form of revision or replacement. This result confirms that the acceptable reliability coefficient at the whole-test level coexisted with substantial weak-nesses at the item level.

Table 6. Item-Level Decision Categories

Decision category	Items	Number of items
Retained	8, 11, 12, 14, 16, 18, 19	7
Lightly revised	5, 7	2
Revised	1, 2, 9, 13, 20, 21, 22, 23	8
Replaced	3, 4, 6, 10, 15, 17, 24, 25	8

Caption: Table 6 presents the final item-level decisions derived from the combined analysis of difficulty, discrimination, and distractor functioning.

Overall Pattern of Findings

When all indicators are considered together, a clear pattern emerges. The items with a difficulty value of 100.00% were exactly the same items that showed a discrimination value of 0.00% and ineffective distractors. Conversely, the items in the moderate difficulty range were generally those with the strongest discrimination and better distractor functioning. This means that the quality of the instrument was not randomly distributed across items but clustered into two contrasting groups: a smaller set of moderately difficult and psychometrically functional items, and a larger set of excessively easy and weakly functioning items.

At the test level, the instrument appeared acceptable because the reliability coefficient reached 0.77. At the item level, however, the evidence showed substantial imbalance. The dominance of very easy items limited score differentiation, reduced the effectiveness of distractors, and contributed to the weak diagnostic value of many items. As a result, the findings indicate that the instrument was workable as a whole but not yet optimized as an item-based measure of student learning.

The overall structure of results therefore demonstrates that test-level adequacy and item-level quality should not be treated as equivalent. A classroom test may generate an acceptable reliability coefficient while still containing a large number of items that are psychometrically weak. In the present study, the strongest items were those that combined moderate difficulty, strong discrimination, and functional distractors, whereas the weakest items combined extreme ease, no discrimination, and ineffective distractors.

DISCUSSION Interpretation of the Main Findings

The findings reveal a clear contrast between acceptable test-level performance and uneven item-level quality. At the overall level, the buffer solution test showed a reliability coefficient of 0.77, which falls within the acceptable range for classroom assessment (Slepkov et al., 2021; Ntumi et al., 2023; Rashwan et al., 2024). This indicates that the test was sufficiently consistent for general classroom use. However, item-level analysis showed that this adequacy did not extend uniformly across all items.

The test was dominated by easy and very easy items, including eight items with a difficulty index of 100.00%. These same items also had a discrimination index of 0.00% and ineffective distractors. Thus, acceptable reliability coexisted with many items that contributed little to differentiating student ability. This finding supports the reliability–item quality paradox noted in previous studies: a test may appear internally consistent while still containing weak, lowdiscrimination, or cognitively undemanding items (Sadeghi et al., 2025; Obon & Rey, 2019; Zubairi et al., 2025). Therefore, reliability should be treated as necessary, but not sufficient, evidence of test quality.

The strongest items were those in the moderate difficulty range, namely items 8, 11, 12, 14, 16, 18, and 19. These items showed more balanced difficulty, stronger discrimination, and better distractor functioning. This pattern is

consistent with classical test theory, which suggests that moderately difficult items generally provide the strongest discrimination compared with items at the extremes of difficulty (Semiun et al., 2022; Hillsley, 2025; Slepkov et al., 2021). The present study confirms that this theoretical relationship also applies in a specific chemistry assessment context.

Relationship Among Difficulty, Discrimination, and Distractor Functioning

A central finding of this study is the close relationship among difficulty, discrimination, and distractor quality. Items answered correctly by all students failed to distinguish between higher- and lower-performing groups and also had ineffective distractors. In contrast, moderately difficult items generally showed stronger discrimination and more functional distractors. This indicates that item quality tended to cluster into recognizable psychometric profiles rather than being randomly distributed.

This pattern is in line with previous research showing that very easy items usually have low discrimination because they generate little variation in responses (Kissi et al., 2023; Semiun et al., 2022; Slepkov et al., 2021). Likewise, non-functioning distractors weaken item quality because students can arrive at the correct answer by elimination rather than conceptual understanding (Shakurnia et al., 2022; Rezigalla et al., 2024; Pan & Jiang, 2022). The present findings therefore reinforce the need to interpret item quality holistically. Difficulty, discrimination, and distractor efficiency should be considered together when deciding whether an item should be retained, revised, or discarded (Bhat & Prasad, 2021; Zahran & Mustafa, 2023; Rezigalla et al., 2024).

Comparison with Previous Studies

The results are broadly consistent with prior studies on teacher-made tests and multiple-choice assessments in different fields. Previous research has often found that classroom-based tests are dominated by easy items (Slepkov et al., 2021; Setyaedhi et al., 2023; Semiun et al., 2022), and the present test showed the same tendency, with most items categorized as easy or very easy and none classified as difficult.

The finding that only a minority of items functioned well also parallels studies in medicine, nursing, and language assessment, where many items were found to have weak or poor discrimination and non-functioning distractors (Belay et al., 2022; Rashwan et al., 2024; Obon & Rey, 2019). The contribution of this study is that it extends those patterns into secondary-school chemistry, especially buffer solution material, where item-analysis studies remain limited. This suggests that psychometric weaknesses identified in other assessment contexts are equally relevant in chemistry education. **Practical Implications for Assessment Improvement**

The practical implications are clear. First, item revision should focus not only on removing weak items, but on improving item design systematically. Item analysis is most useful when it informs revision, question-bank development, and repeated monitoring rather than one-time elimination of problematic items (Bhat & Prasad, 2021; Zubairi et al., 2025; Gupta et al., 2020). In this study, only 7 of 25 items were retained without revision, while 18 required improvement,

indicating that the test should be viewed as a provisional item pool rather than a finished instrument.

Second, the findings highlight the importance of teacher training. Weak distractors, zero-discrimination items, and excessively easy questions often reflect weaknesses in item-writing practice rather than unavoidable content issues (Ajayi & Idusogie, 2021; Sajjad et al., 2020; Amini et al., 2020). Chemistry teachers therefore need support in writing plausible distractors, targeting moderate difficulty, and constructing items that assess conceptual understanding rather than recall alone.

Third, the findings are useful for item-bank development. The retained items can serve as reference models for future chemistry tests, while the replaced items indicate patterns that should be avoided. Thus, the study not only identifies current weaknesses but also provides a practical foundation for developing a more balanced and defensible pool of chemistry multiple-choice items.

Theoretical and Empirical Contribution

This study contributes empirically by providing item-level evidence from a secondary-school chemistry context, specifically buffer solution material. Since much of the existing literature comes from medical education or general assessment settings, this study helps extend item-analysis evidence into topicspecific science education.

Theoretically, the study reinforces the argument that overall reliability cannot substitute for item-level evaluation. A test may show acceptable internal consistency while still containing items that are too easy, fail to discriminate, and contain weak distractors. The study therefore supports a layered view of assessment quality in which reliability must be interpreted together with difficulty, discrimination, and distractor functioning. It also strengthens the case for micro-level evaluation, showing that the usefulness of classroom test scores depends heavily on the functioning of individual items.

Limitations of the Study

Several limitations should be noted. First, the study used a relatively small dataset of 35 students and 25 items, which may reduce the stability and generalizability of psychometric estimates (Toksöz & Baykal, 2020; Rezigalla et al., 2024; Bhat et al., 2023). Second, the analysis was based on a single test administration, so it could not evaluate whether item quality would remain stable across cohorts or improve after revision (Zubairi et al., 2025; Owolabi et al., 2021). Third, the study relied mainly on Classical Test Theory output from ANATES, without item-total correlation, Rasch, or IRT analysis, which could have provided complementary evidence (Belay et al., 2022; Pan & Jiang, 2022). Fourth, no qualitative review of item-writing flaws was conducted, even though ambiguous wording, implausible distractors, and low cognitive demand are known to influence item performance (Sajjad et al., 2020; Rashwan et al., 2024).

Closing Argument

Overall, the test showed mixed psychometric quality: acceptable at the aggregate level, but uneven and often weak at the item level. The strongest items combined moderate difficulty, stronger discrimination, and functional

distractors, whereas the weakest items combined extreme ease, zero discrimination, and ineffective distractors. These findings align with the broader literature on teacher-constructed multiple-choice tests and underline the need for systematic post-test analysis, item revision, and teacher support in item writing. When too many items fail to distinguish between students, the resulting scores become less informative for evaluating learning outcomes and guiding instructional improvement.

CONCLUSION

Fundamental Finding : This study found that the buffer solution multiplechoice test was acceptable at the whole-test level but uneven at the item level. Although the reliability coefficient of 0.77 indicated adequate consistency, many items were psychometrically weak. The test was dominated by easy and very easy items, including eight items with a difficulty index of 100.00% and a discrimination index of 0.00%, supported by ineffective distractors. In contrast, moderately difficult items showed better discrimination and more functional distractors. These findings show that reliability alone is not sufficient to judge item quality. Implication : The findings imply that classroom tests should be evaluated not only by reliability but also by item difficulty, discrimination, and distractor effectiveness. For teachers, routine item analysis is important to identify which items should be retained, revised, or replaced. The study also suggests that moderately difficult items are more effective for measuring student learning accurately. Therefore, better item construction can improve the fairness and diagnostic value of classroom assessment. Limitation : This study was limited to one multiple-choice test on buffer solution material in Grade XI chemistry, so the findings cannot be generalized broadly. The analysis also used only classical item analysis indicators and did not apply more advanced psychometric models. In addition, the study focused on item performance and did not examine instructional factors that may have influenced the results. Future Research : Future studies should examine similar tests in other chemistry topics, grade levels, and educational settings. Research can also use larger samples and more advanced approaches such as item response theory. In addition, future work should investigate how teaching practices and teacher training in item writing may improve the quality of classroom assessments.

REFERENCES

Adebusuyi, O. F., Bamidele, E. F., & Adebusuyi, A. S. (2022). The Role of Knowledge and Epis-temological Beliefs in Chemistry Teachers STEM Professional Development and Instruc-tional Practices: Examination of STEM-Integrated Classrooms. European Journal of Sci-ence and Mathematics Education, 10(2), 243–255. https://doi.org/10.30935/scimath/11799

Ahmad, K. A., Kadir, N. L. A., Zakaria, M. Z., Zamri, M. L., Rahman, S. N. A., & Mustafa, N. M. (2021). Distractor Efficiency of an Intermediate English

Proficiency Course Final Ex-amination Paper. AJATeL, 11(2), 103–112. https://doi.org/10.37134/ajatel.vol11.2.10.2021

Al-Hummayani, F. M. (2020). Evaluation of the Multiple-Choice Question Item Analysis of the Sixth Year Undergraduate Orthodontic Tests at the Faculty of Dentistry, King Abdulaziz University, Saudi Arabia. Egyptian Orthodontic Journal, 57(6), 1–18. https://doi.org/10.21608/eos.2020.31532.1000

Ali, I., Haral, M. N., Tahira, F., Ali, M. W., & Imran, A. (2020). An Evaluation, Assortment, and Optimization of the Question Items. Education and Urban Society, 52(8), 1264–1280. https://doi.org/10.1177/0013124519896831

AlKhatib, H. S., Brazeau, G. A., Akour, A., & AlMuhaissen, S. (2020). Evaluation of the Effect of Items' Format and Type on Psychometric Properties of Sixth Year Pharmacy Students Clin-ical Clerkship Assessment Items. BMC Medical Education, 20(1). https://doi.org/10.1186/s12909-020-02107-3

Al-lawama, M., & Kumwenda, B. (2023). Decreasing the Options' Number in Multiple Choice Questions in the Assessment of Senior Medical Students and Its Effect on Exam Psycho-metrics and Distractors' Function. BMC Medical Education, 23(1). https://doi.org/10.1186/s12909-023-04206-3

Amini, N., Michoux, N., Warnier, L., Malcourant, E., Coche, E., & Berg, B. (2020). Inclusion of McQs Written by Radiology Residents in Their Annual Evaluation: Innovative Method to Enhance Resident's Empowerment? Insights Into Imaging, 11(1). https://doi.org/10.1186/s13244-019-0809-4

Berenbon, R., & McHugh, B. (2023). Do Subject Matter Experts' Judgments of Multiple-Choice Format Suitability Predict Item Quality? Educational Measurement Issues and Practice, 42(3), 13–21. https://doi.org/10.1111/emip.12570

Chen, D., Harman, A. E., Sun, H., Ye, T., & Gaiser, R. R. (2023). A Comparison of 3- And 4-Option Multiple-Choice Items for Medical Subspecialty inTraining Examinations. BMC Medical Education, 23(1). https://doi.org/10.1186/s12909-023-04277-2

Drábeková, J., Rumanová, L., & Pavlovičová, G. (2022). Numeracy as a Part of Financial Litera-cy: Student Case Study in Slovakia. International Journal of Emerging Technologies in Learning (Ijet), 17(10), 186–199. https://doi.org/10.3991/ijet.v17i10.29659

Du, J., Chen, X., Wang, T., Zhao, J., & Li, K. (2023). The Effectiveness of the Fully Online Flipped Classroom for Nursing Undergraduates During the COVID-19:

Historical Control Study. Nursing Open, 10(8), 5766–5776. https://doi.org/10.1002/nop2.1757

Falcão, F., Pereira, D. M., Gonçalves, N., Champlain, A. D., Costa, P., & Pêgo, J. M. (2023). A Suggestive Approach for Assessing Item Quality, Usability and Validity of Automatic Item Generation. Advances in Health Sciences Education, 28(5), 1441–1465. https://doi.org/10.1007/s10459-023-10225-y

Ganji, K. K., Ananthakrishnan, N., Manivasakan, S., Alruwaili, M. K., Alonazi, M. A., & Algarni, H. A. (2025). Analyzing the Relationship Between Psychometric Indices of Item Analysis With Attainment of Course Learning Outcomes: Cross-Sectional Study in Integrated Out-come-Based Dental Curriculum Courses. BMC Medical Education, 25(1). https://doi.org/10.1186/s12909-025-07871-8

Mallaiah, J., Williams, O., & Allegrante, J. P. (2024). Development and Validation of a Stroke Lit-eracy Assessment Test for Community Health Workers. Health Education & Behavior, 51(5), 764–774.

https://doi.org/10.1177/10901981241245050

Rashwan, N. I., Aref, S. R., Nayel, O. A., & Rizk, M. H. (2024). Postexamination Item Analysis of Undergraduate Pediatric Multiple-Choice Questions Exam: Implications for Developing a Validated Question Bank. BMC Medical Education, 24(1). https://doi.org/10.1186/s12909-024-05153-3

Semiun, T. T., Wisrance, M. W., & Napitupulu, M. H. (2022). English Summative Test: The Quality of Its Items. English Education Journal of English Teaching and Research, 7(2), 119–127. https://doi.org/10.29407/jetar.v7i2.18347

Wise, M. J. (2020). The Advantages of Five-Option Multiple-Choice Items in Classroom Tests of Student Mastery. Journal of Education Teaching and Social Studies, 2(4), p59. https://doi.org/10.22158/jetss.v2n4p59

Yimer, S. T., & Feza, N. (2019). Learners' Conceptual Knowledge Development and Attitudinal Change Towards Calculus Using Jigsaw Co-Operative Learning Strategy Integrated With GeoGebra. International Electronic Journal of Mathematics Education, 15(1). https://doi.org/10.29333/iejme/5936

Zahra, N. S., Widyastuti, W., & Ridfah, A. (2022). Psychometric Characteristics of the Culture Fair Intelligence Test Scale 2. Jp3i (Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia), 11(2), 123–136. https://doi.org/10.15408/jp3i.v11i2.25947

Zahran, E., & Mustafa, A. (2023). Item Analysis of Multiple Choice Questions of a Genetic Term Exam in an Egyptian Veterinary College for a Viable

Question Bank. Educational Evalua-tion and Learning Management Journal, 0(0), 0–0. https://doi.org/10.21608/eelmj.2023.172846.1001

Zubairi, N. A., Alahmadi, T., Ibrahim, M. I. M., Hegazi, M. A., & Gadi, F. U. (2025). Effective Use of Item Analysis to Improve the Reliability and Validity of Undergraduate Medical Ex-aminations: Evaluating the Same Exam Over Many Years: A Different Approach. Pakistan Journal of Medical Sciences, 41(3), 810–815. https://doi.org/10.12669/pjms.41.3.10693