



## ITEM DIFFICULTY AND DISCRIMINATION INDICES IN OSUN STATE, NIGERIA

BY

ALADE, Muideen Olatunji

Department of Science Education,

Faculty of Education, Adeyemi Federal University of Education, Ondo, Nigeria

alademo@afued.edu.ng; alademdn@gmail.com | +2348039551057

### Abstract

*This study evaluated the psychometric quality of 40 secondary school chemistry multiple-choice items using item difficulty and discrimination indices. The population comprised 5,831 senior secondary students in selected schools in Ilesa East and Egbedore Local Government Areas of Osun State, Nigeria. A sample of 374 students was obtained through stratified random sampling, with sample size determined by a finite-population formula at 5% precision. Data were collected with the Chemistry Multiple-Choice Achievement Test covering organic chemistry, chemical combination, electrolysis, redox reactions, thermodynamics and the periodic table. Classical Test Theory guided the analysis. The instrument produced a Kuder-Richardson Formula 20 coefficient of .70, indicating minimum acceptable but borderline internal consistency. Approximately 72% of the items fell within the acceptable difficulty range of .30-.70, while approximately 68% attained an acceptable discrimination index of at least .30. The documented item examples yielded difficulty values from .24 to .83 and discrimination values from .04 to .60. The findings show that most items functioned adequately, although a substantial minority require revision. Routine item analysis, expert review and a continuously updated chemistry item bank were recommended.*

**Keywords:** Chemistry assessment, Classical test theory, Difficulty index, Discrimination index, Item analysis

### Introduction

Assessment is central to chemistry teaching and learning because it provides evidence of students' conceptual understanding, supports instructional decisions and informs promotion, placement and certification. Multiple-choice questions are widely used because they can sample broad curriculum content, support objective scoring and permit efficient comparison of performance. Their efficiency, however, does not guarantee measurement quality. Ambiguous stems, implausible distractors, cueing, content underrepresentation and inappropriate cognitive demand can introduce error and weaken the meaning of test scores. Recent psychometric research therefore emphasises the need to evaluate how individual items function, rather than relying only on total scores (Rashwan et al., 2024; Rezigalla et al., 2024).

Under Classical Test Theory (CTT), the difficulty index is the proportion of examinees who answer an item correctly, while the discrimination index indicates the extent to which the item differentiates stronger from weaker examinees. These indices are

complementary: an item may be easy but still discriminate well, or possess moderate difficulty but discriminate poorly because of ambiguity, a defective key or non-functioning distractors. International studies show that systematic post-examination analysis can identify defective items, strengthen question banks and improve assessment governance (Eleragi et al., 2025; Gebremichael et al., 2025; Hillsley, 2025).

Four gaps justify the present study. Empirically, much recent work on multiple-choice item quality has been conducted in higher education and medical education, with comparatively limited evidence from secondary school chemistry in Nigeria. Methodologically, studies do not always combine difficulty and discrimination evidence into explicit retain-revise-remove decisions based on stated criteria. Theoretically, CTT is often applied procedurally without explaining observed, true and error scores or contrasting CTT with Item Response Theory (IRT). Contextually, evidence on locally developed chemistry tests in Osun State, particularly across schools in different senatorial districts, remains limited.

These gaps constrain evidence-based improvement of classroom chemistry assessment (Ayanwale et al., 2022; Butakor, 2022; Logayah et al., 2024).

This study addresses the gaps by evaluating 40 senior secondary chemistry multiple-choice items with explicit difficulty and discrimination criteria, interpreting the evidence within CTT and identifying items that should be retained, revised or removed. The study is intended to strengthen the validity, reliability and fairness of locally developed chemistry assessments and support teachers in building higher-quality item banks.

### Statement of the Problem

Teachers and school administrators commonly interpret chemistry test scores as evidence of achievement, yet the quality of the individual questions producing those scores may remain unknown. Where items are excessively easy, excessively difficult, ambiguously worded or unable to distinguish high- from low-achieving students, total scores may provide misleading feedback and unfairly influence academic decisions. In Osun State, there is limited documented evidence showing whether locally administered secondary school chemistry multiple-choice items meet explicit psychometric standards. A systematic evaluation is therefore required to combine item difficulty and discrimination indices, ground interpretation in a clear measurement framework and identify questions that should be retained, revised or removed.

### Purpose of the Study

The study evaluated the psychometric quality of secondary school chemistry multiple-choice tests in selected schools in Osun State, Nigeria. Specifically, it sought to:

1. determine the difficulty levels of the chemistry multiple-choice items;
2. determine the extent to which the items discriminate between high- and low-performing students; and
3. classify the items according to whether they should be retained, revised or removed.

### Research Questions

1. What are the difficulty levels of the secondary school chemistry multiple-choice items?
2. To what extent do the items discriminate between high- and low-performing students?

3. Which items meet the stated psychometric criteria, and which require revision or removal?

### Literature Review

The psychometric quality of a test item concerns the extent to which it contributes useful and interpretable information about intended learning outcomes. Difficulty is conventionally represented by  $p$ , the proportion answering correctly. For achievement testing, a range of difficulties is desirable, although moderately difficult items commonly provide greater differentiation because they generate more response variation. Discrimination represents the relationship between item performance and overall achievement; positive values show that higher-achieving examinees are more likely to answer correctly. Negative or near-zero values are warning signs requiring investigation of the key, wording, curriculum alignment, distractors and scoring (Hillsley, 2025; Rezigalla et al., 2024).

Recent evidence demonstrates the value of systematic item analysis. Rashwan et al. (2024) showed that quantitative indices combined with qualitative review can identify defective items and support a validated question bank. Rezigalla et al. (2024) found that distractor efficiency is related to both difficulty and discrimination, showing that item quality cannot be inferred from one statistic. Eleragi et al. (2025), in a three-year evaluation of 210 items, reported that continuous analysis supports quality improvement, while Gebremichael et al. (2025) similarly emphasised removal or revision of excessively easy, difficult or non-discriminating items. In general education, Logayah et al. (2024) also demonstrated the usefulness of difficulty and discrimination statistics for judging the quality of olympiad multiple-choice items.

Validity concerns the strength of evidence supporting the intended interpretation and use of test scores. Content evidence requires systematic alignment between items, curriculum domains and intended cognitive processes. Expert review may be quantified through item-level and scale-level content validity indices when validators provide structured relevance ratings (Zapata-Ospina & Garcha-Valencia, 2022). Reliability concerns score consistency. For dichotomously scored

items, KR-20 estimates internal consistency; a coefficient of .70 is generally a minimum exploratory threshold, but it is borderline rather than strong and should not be treated as sufficient for consequential individual decisions without further evidence (Reynolds et al., 2021).

Classical Test Theory provides the framework for the study. It assumes that an examinee's observed score (X) consists of a true-score component (T) and an error component (E):

$$r_{xx} = \text{Var}(T) / \text{Var}(X) = \text{Var}(T) / [\text{Var}(T) + \text{Var}(E)]$$

Within CTT, the difficulty index for a dichotomously scored item is:

$$P = R / N$$

where R is the number of correct responses and N is the number of examinees. The upper-lower discrimination index is:

$$D = (RU - RL) / n_g$$

where RU and RL are the numbers answering correctly in the upper and lower groups, respectively, and  $n_g$  is the number of students in either group. Positive values approaching 1.00 indicate stronger discrimination; values near zero indicate little differentiation, while negative values suggest a potentially defective item.

IRT models the probability of a correct response as a function of latent ability and item parameters such as difficulty and discrimination. Compared with CTT, IRT can produce item and ability estimates that are less sample-dependent when assumptions, calibration size and model fit are adequate. It nevertheless requires specialised estimation and more demanding diagnostics. CTT was selected because the study sought direct classroom-level screening of a 40-item dichotomous test using an available sample of 374 students and decision rules that teachers can readily interpret. The choice is therefore proportionate to the study purpose, while future studies may compare CTT findings with IRT estimates (Ayanwale et al., 2022; Butakor, 2022).

## Methodology

The study adopted a descriptive psychometric evaluation design. The design was appropriate because students' responses to an existing chemistry achievement test were analysed without manipulation of variables. The focus was item functioning under CTT rather than estimation of a causal treatment effect.

The population comprised 5,831 senior secondary school students in selected schools

$$X = T + E$$

The true score is the expected score across an indefinitely large number of equivalent administrations, whereas error represents unsystematic influences such as guessing, temporary inattention, unclear wording, scoring inconsistencies and occasion-specific conditions. Reliability is the proportion of observed-score variance attributable to true-score variance:

in Ilesa East Local Government Area, Osun East Senatorial District, and Egbedore Local Government Area, Osun West Senatorial District, Osun State, Nigeria. The target sample was determined at 5% precision using the finite-population formula  $n = N/[1 + N(e^2)]$ . Thus,  $n = 5,831/[1 + 5,831(0.05^2)] = 374.32$ , which was rounded to 374 students. Stratified random sampling was used to obtain proportional representation by local government area, school and class level, after which students were selected randomly within the relevant strata.

Data were collected with the Chemistry Multiple-Choice Achievement Test (CMCA), a 40-item dichotomously scored instrument covering organic chemistry, chemical combination, electrolysis, redox reactions, thermodynamics and the periodic table. Each item had one keyed answer and distractors. A correct response attracted one mark, while an incorrect or omitted response attracted zero; there was no penalty for guessing. The maximum obtainable score was 40.

Face validity was addressed through expert review of clarity, language, format, keying and distractor plausibility. Content validity was addressed through a table of specifications mapping the 40 items to the approved senior secondary chemistry curriculum and intended cognitive demands. Experienced chemistry teachers and a specialist in test construction reviewed the items for relevance, accuracy and representativeness. For quantified content validity, the item-level content validity index is  $I-CVI = A/N_e$ , where A is the number of

experts rating an item as relevant and  $N_e$  is the total number of experts. The scale-level average is  $S-CVI/Ave = \sum(I-CVI)/k$ , where  $k$  is the number of items (Zapata-Ospina & Garcha-Valencia, 2022).

The instrument was pilot-tested with 40 students from schools outside the main study areas to reduce contamination. The pilot responses were analysed using Kuder-Richardson Formula 20: The coefficient was .70. This meets the minimum threshold for exploratory classroom use but is borderline and should be interpreted cautiously. The reported pilot review led to minor revisions of items with unclear distractors or extreme difficulty.

The final CMCA was administered during a 60-minute session under standardised classroom conditions. Uniform instructions and supervision were provided. Response sheets were scored dichotomously, and each student's total score was obtained by summing correct responses.

For each item, difficulty was calculated as  $P = R/N$ . Values below .30 were classified

as difficult, values from .30 to .70 as acceptable or moderate, and values above .70 as easy. Students were ranked by total score for upper-lower discrimination analysis, and  $D$  was computed as  $(RU - RL)/ng$ . Values of .30 or greater were treated as acceptable; values below .30 required qualitative review, revision or removal. Decisions were based jointly on difficulty, discrimination, curriculum relevance and item-writing quality rather than on a single statistic.

## Results and Discussion

**Research Question 1:** What are the difficulty levels of the secondary school chemistry multiple-choice items?

The reviewer-supplied summary indicates that approximately 72% of the 40 items were within the acceptable difficulty interval of .30-.70. Expressed as whole items, this corresponds to 29 items (72.5%), while 11 items (27.5%) were outside the interval and required review for being too easy or too difficult.

**Table 1: Summary of psychometric classifications for the 40-item CMCA**

| Indicator            | Criterion             | Acceptable items | Acceptable (%) | Items requiring review |
|----------------------|-----------------------|------------------|----------------|------------------------|
| Difficulty index     | $.30 \leq P \leq .70$ | 29               | 72.5           | 11 (27.5%)             |
| Discrimination index | $D \geq .30$          | 27               | 67.5           | 13 (32.5%)             |

**Research Question 2:** To what extent do the items discriminate between high- and low-performing students?

Approximately 68% of the items attained an acceptable discrimination index of at least .30. This corresponds to 27 items (67.5%). The remaining 13 items (32.5%) had discrimination below .30 and therefore required inspection of their keys, stems, distractors, content alignment and cognitive demand.

**Research Question 3:** Which items meet the stated psychometric criteria, and which require revision or removal?

Table 2 combines the difficulty and discrimination evidence available for the five item-level examples contained in the submitted manuscript. Item 5 has been reclassified as moderate rather than difficult because  $P = .40$  falls within the stated .30-.70 interval. Items 1 and 5 were retained; Item 2 may be retained after review of its distractors because it is easy but discriminates well. Items 3 and 4 require revision or removal because they combine extreme difficulty with poor discrimination

**Table 2: Combined item statistics and decisions for the five documented items**

| Item | Correct | P   | Level     | Upper | Lower | D   | Decision                   |
|------|---------|-----|-----------|-------|-------|-----|----------------------------|
| 1    | 220     | .59 | Moderate  | 48    | 18    | .60 | Retain                     |
| 2    | 280     | .75 | Easy      | 50    | 30    | .40 | Retain; review distractors |
| 3    | 90      | .24 | Difficult | 30    | 28    | .04 | Revise/remove              |
| 4    | 310     | .83 | Very easy | 40    | 32    | .10 | Revise/remove              |
| 5    | 150     | .40 | Moderate  | 38    | 20    | .36 | Retain                     |

### Discussion of Findings

The first major finding was that 29 of the 40 items (72.5%) fell within the acceptable difficulty range. This indicates that most questions provided a reasonable level of challenge. The pattern is broadly consistent with contemporary item-analysis studies in which a substantial proportion of items fell within usable difficulty ranges, although exact proportions differ across disciplines and populations (Eleragi et al., 2025; Logayah et al., 2024; Rezigalla et al., 2024). The 11 items outside the interval may reflect low-level recall, insufficient curriculum coverage, ambiguous wording or weak distractors. Educationally, acceptable items may be retained, but extreme items should undergo targeted qualitative review before reuse. Future studies should report cognitive level and distractor efficiency for every item so that genuine content difficulty can be separated from item-writing defects.

The second major finding was that 27 items (67.5%) attained discrimination of at least .30, whereas 13 items did not. The acceptable proportion shows that most questions differentiated stronger from weaker students, but the number of weak items remains educationally significant. Recent studies report that discrimination is affected by key accuracy, distractor plausibility, item format, curriculum alignment and the match between item demand and examinee preparation (Gebremichael et al., 2025; Hillsley, 2025; Rashwan et al., 2024). The very low values for Items 3 and 4 suggest that their difficulty did not produce useful information about achievement. They may have confused both groups or encouraged guessing. Such items should be checked for miskeying, rewritten and re-piloted. Future research should supplement upper-lower discrimination with point-biserial correlations and differential item functioning.

The third finding concerned joint retain-revise-remove decisions. Items 1 and 5 combined moderate difficulty with acceptable discrimination, while Item 2, although easy, still discriminated well and may be retained after checking its curricular purpose and distractors. Items 3 and 4 should not be reused unchanged. This decision-oriented approach agrees with the view that quantitative indices should trigger, rather than replace, expert review (Eleragi et al., 2025; Rashwan et al., 2024). The practical implication is that school item banks should preserve each item's content domain, cognitive level, difficulty, discrimination, distractor performance and revision history. Subsequent administrations can then show whether revised items function more effectively.

The KR-20 coefficient of .70 provides minimum acceptable but borderline evidence of internal consistency. It is lower than preferred for consequential decisions about individual students. Under CTT, weakly discriminating items and heterogeneous content can increase error variance and depress reliability (Reynolds et al., 2021). Revising or removing defective items may improve score consistency, but reliability must be recalculated after revision rather than assumed to improve. Future research should report confidence intervals for reliability, examine dimensionality and, where feasible, compare CTT results with an IRT model.

### Limitations of the Study

The study was limited to selected schools in two local government areas, so the findings should not be generalised to all schools in Osun State without caution. The analysis relied on CTT indices, which are partly sample-dependent, and did not report distractor efficiency, point-biserial correlations, differential item functioning or IRT information functions. In addition, the

submitted files contained complete item-level statistics for only five of the 40 items, so the appendix requires completion from the original analysis output before resubmission.

### Conclusion

The 40-item chemistry multiple-choice test demonstrated generally useful but uneven psychometric quality. Approximately 72% of the items had acceptable difficulty and approximately 68% had acceptable discrimination. The evidence supports retention of many items, but a substantial minority require revision, re-keying checks, distractor improvement or removal. The KR-20 coefficient of .70 is adequate only at a minimum exploratory level. Systematic item analysis, expert review and re-piloting are therefore necessary before the test is reused for consequential assessment.

### Recommendations

1. Chemistry teachers should conduct item analysis after every major test and preserve the results in a school-based item bank.
2. Items with P outside .30-.70 or D below .30 should undergo expert review of the stem, key, distractors, curriculum alignment and cognitive demand before reuse.
3. Teacher professional development should include blueprint construction, distractor writing, content-validity procedures and interpretation of CTT indices.
4. Revised items should be re-piloted, and KR-20 should be recalculated after weak items are corrected or removed.
5. Future studies should use larger and more diverse samples and extend the analysis to point-biserial correlation, distractor efficiency, differential item functioning and IRT.

### References

- Ayanwale, M. A., Chere-Masopha, J., & Morena, M. C. (2022). The classical test or item response measurement theory: The status of the framework at the Examination Council of Lesotho. *International Journal of Learning, Teaching and Educational Research*, 21(8), 384-406.  
<https://doi.org/10.26803/ijlter.21.8.22>
- Butakor, P. K. (2022). Using classical test and item response theories to evaluate psychometric quality of teacher-made test in Ghana. *European Scientific Journal, ESJ*, 18(1), 139-165.
- <https://doi.org/10.19044/esj.2022.v18n1p139>
- Eleragi, A. M. S., Miskeen, E., Hussein, K., Rezigalla, A. A., Adam, M. I. E., Al-Faifi, J. A., Alhalafi, A., Al Ameer, A. Y., & Mohammed, O. A. (2025). Evaluating the multiple-choice questions quality at the College of Medicine, University of Bisha, Saudi Arabia: A three-year experience. *BMC Medical Education*, 25, Article 233.  
<https://doi.org/10.1186/s12909-025-06700-2>
- Gebremichael, M. W., Baraki, B., Mehari, M.-A., & Assalfew, B. (2025). Item analysis of multiple choice questions from assessment of health sciences students, Tigray, Ethiopia. *BMC Medical Education*, 25, Article 441.  
<https://doi.org/10.1186/s12909-025-06904-6>
- Hillsley, K. (2025). Question format is the best predictor of item discrimination: A multivariable analysis. *Journal of Microbiology & Biology Education*, 26(3), Article e00205-25. <https://doi.org/10.1128/jmbe.00205-25>
- Logayah, D. S., Ruhimat, M., Arrasyid, R., & Islamy, M. R. F. (2024). Item analysis of National Geography Olympiad multiple-choice questions (MCQs) in Indonesia. *Cogent Social Sciences*, 10(1), Article 2354971.  
<https://doi.org/10.1080/23311886.2024.2354971>
- 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, Article 168.  
<https://doi.org/10.1186/s12909-024-05153-3>
- Reynolds, C. R., Altmann, R. A., & Allen, D. N. (2021). *Mastering modern psychological testing: Theory and methods* (2nd ed.). Springer. <https://doi.org/10.1007/978-3-030-59455-8>
- Rezigalla, A. A., Eleragi, A. M. E. S. A., Elhussein, A. B., Alfaifi, J., AlGhamdi, M. A., Al Ameer, A. Y., Yahia, A. I. O., Mohammed, O. A., & Adam, M. I. E. (2024). Item analysis: The impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education*, 24, Article 445.  
<https://doi.org/10.1186/s12909-024-05433-y>
- Zapata-Ospina, J. P., & Garcha-Valencia, J. (2022). Validity based on content: A challenge in health measurement scales. *Journal of Health Psychology*, 27(2), 481-493.  
<https://doi.org/10.1177/1359105320953477>