E-ISSN: 1115 - 8425 P-ISSN: 3027-012X
MPRI-JAEE | INTERNATIONAL JOURNAL OF ASSESSMENT & EVALUATION IN EDU.
VOL. 11 NO. 8 – JANUARY. 2026

PSYCHOMETRIC ASSESSMENT OF A CHEMISTRY ACHIEVEMENT TEST AS A PREDICTOR OF ACADEMIC PERFORMANCE AMONG SENIOR SCHOOLS STUDENTS IN ONDO WEST, NIGERIA

AKINGBADE, O.; & OGUNMOLA, GRACE T.

Department of Educational Psychology and Counselling, Adeyemi Federal University of Education, Ondo, Nigeria.
Corresponding Author: akingbadeolumide@gmail.com
DOI: https://doi.org/10.70382/mejaee.v11i8.094

Abstract

The study assessed the psychometric of a Chemistry Achievement Test (CAT) as a predictor of academic performance among students in Ondo West secondary schools. Three research questions were raised and one hypothesis was formulated and tested at 0.05 level of significant to guide the study. The study was anchored on Item Response Theory (IRT) as the underpinning theoretical framework for modern, sophisticated test analysis and validation. The study used survey design type of descriptive research and the population comprised of chemistry students in Ondo West LGAs. One hundred and eighty students were drawn using purposive and proportionate sampling techniques. Two instruments including Chemistry Achievement Test (CAT), a 40- multiple-choice items and a secondary score data were used. The CAT's consistency was determined using KR_{20} and a reliability coefficient of .74 was obtained while its accuracy was found with face and content validity methods by two psychometricians and two chemistry teachers. The psychometrics of the CAT was determined using 2-PL Model. The hypothesis formulated was tested using independent t-test. Findings revealed that 30 out of 40 items were considered suitable; Kuder–Richardson Formula 21 (KR-21) coefficient of 0.7351 suggests good internal consistency, Fisher z- test value of 2.214 is significant at the 0.05 level of significant ($p<0.05$); significant

Keywords:
Achievement Test, Predictive Validity, Multiple-Choice, Item Response Theory, Chemistry Education, Psychometric Evaluation.

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

difference in the chemistry performance of male and female students; t (178) = 3.448, p<0.05, eta square (η^2) = 0.063. It was concluded that the Chemistry Achievement Test as a reliable test to is good for state mock-examination after the non-effective ones have been removed or adjusted for. The study recommended that the Ministry of Education should consider using Chemistry Achievement Test items for subsequent state mock chemistry examination among SS2 students and also, item bank of the Chemistry Achievement Test should be rigorously reviewed.

Introduction

Education is the cornerstone of national development, and its quality is often measured by the performance of students in both internal and external examinations. In recent years, technological advancements and new educational methodologies have contributed to the continuous improvement of evaluation tools. Modern assessments incorporate diverse formats, including project-based tasks, hands-on activities, and digital platforms, allowing for a more holistic measurement of student performance. As education continues to evolve, there remains a need to refine assessment strategies to ensure they accurately reflect students' knowledge, skills, and readiness for future challenges as educators and psychologists sought more objective ways to measure student achievement. In Nigeria's educational landscape, particularly in science education, these concepts have evolved significantly over the decades.

Historically, the development of structured evaluation methods gained prominence in the late 19th century when researchers sought to refine the way knowledge and intellectual abilities were measured. Early efforts in educational assessment primarily focused on testing memory and factual recall, often neglecting critical thinking and problem-solving skills (Thorndike, 1918). However, as education systems advanced, scholars began to emphasize the importance of designing assessments that align with broader learning objectives.

Throughout the 20th century, innovations in educational research introduced new ways of structuring examinations to better capture students' understanding and application of concepts. These improvements aimed to bridge the gap between theoretical knowledge and practical problem-solving, ensuring that assessments truly reflected learners' competencies (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956). The introduction of structured frameworks for evaluating learning outcomes further enhanced the effectiveness of academic assessments.

Chemistry education plays a foundational role in secondary schools, both as a pillar of science learning and as a tool for national development. In Nigeria and other parts of the

world, it serves as a bridge between theory and practice, underpinning advancements in health, industry, agriculture, and environmental management.

Objectives of Chemistry Education in Secondary Schools (*as outlined in Nigerian curriculum*)

i. Stimulate and sustain students' interest in chemistry.
ii. Enable students to acquire basic knowledge and practical skills in chemistry.
iii. Develop in students an appreciation of chemistry's role in industry, economic development, and everyday life.
iv. Equip students with scientific attitudes such as curiosity, honesty, objectivity, and critical thinking.
v. Prepare students for further studies in chemistry-related fields.

A chemistry achievement test, according to various authors, is an assessment tool designed to measure a student's acquired knowledge and understanding of chemistry concepts and skills, typically covering material taught in a specific chemistry course, allowing educators to evaluate how well students have mastered the subject matter through questions that test their ability to apply chemical principles to different scenarios; essentially, it's a test that gauges a student's current level of chemistry proficiency after completing a course of study. A chemistry achievement test primarily focuses on evaluating a student's understanding of core chemistry concepts, including chemical equations, periodic table trends, chemical bonding, reaction types, and stoichiometry. These tests can include multiple-choice questions, short answer questions, essay questions, and even practical application problems that require students to demonstrate their understanding through calculations or analysis of experimental data. However, an effective chemistry test requires to measure consistently and accurately, a true mastery level of students.

Concept of Reliability in Testing

Reliability is defined as "the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and repeatable" (AERA et al., 2014). A reliable test minimizes measurement error the random and systematic inconsistencies that obscure true scores (Thorndike & Thorndike-Christ, 2010).

Reliability is crucial in educational testing because unreliable results can lead to unfair academic decisions such as misplacement, incorrect grading, or unjust promotion and demotion of students. In classroom settings, reliability ensures that tests are free from random errors and reflect true student performance.

Reliability is important because:

i. It ensures fairness and consistency in assessment.
ii. It supports the accuracy of decision-making in education and psychology.
iii. It is a necessary condition for validity without reliable measurement, validity is compromised.

Kuder-Richardson Reliability

The Kuder-Richardson method refers to a family of formulas used to estimate the internal consistency reliability of tests composed of dichotomous items, that is, items with two possible responses (e.g., true/false, yes/no, right/wrong). Developed by G. Frederic Kuder and M. W. Richardson in 1937, these formulas are foundational in classical test theory and remain widely cited in psychometrics and educational assessment literature. Among the several formulas proposed by Kuder and Richardson, the most well-known and widely used are KR-20 and KR-21. Both estimate how consistently a test measures a single construct, and they are closely related to Cronbach's Alpha, which generalizes these formulas for items with more than two response options.

KR-20: The Kuder-Richardson Formula 20

The KR-20 formula is applicable when the following conditions are met:

a. The test items are dichotomously scored (e.g., 0 = incorrect, 1 = correct).
b. The items are assumed to be measuring a single construct (i.e., unidimensionality).
c. The test takers are from a homogeneous population.
d. The items are not all equally difficult, which distinguishes it from KR-21.

The KR-20 formula is as follows:

$$\text{KR-20} = \frac{K}{K-1} \times \left(1 - \frac{\boldsymbol{\Sigma Pq}}{\boldsymbol{\sigma^2}}\right)$$

Where:

K = Number of test items

P = Proportion of students answering an item correctly

q = Proportion of students answering an item incorrectly (1 – p)

σ^2 = Variance of total test scores

The product represents the item variance for dichotomous items. The KR-20 therefore estimates the proportion of test score variance that is attributable to true score variance, rather than random measurement error. The higher the value of KR-20 (closer to 1), the more reliable the test is considered to be.

The Unitary Concept of Validity

Messick's (1989) groundbreaking work consolidated earlier frameworks into a unitary concept of validity. He defined validity as "an overall evaluative judgment of the degree

to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores".

Messick emphasized six aspects of validity:

i. Content: Adequacy of the test content in representing the domain.
ii. Substantive: Theoretical rationale behind responses to test items.
iii. Structural: Match between test structure and construct structure.
iv. Generalizability: Extent to which score interpretations can be applied across contexts.
v. External: Relations to other variables, including convergent and discriminant evidence.
vi. Consequential: Consideration of social consequences and value implications.

Messick's model unified earlier perspectives and introduced the ethical dimension of validity—especially the consequences of test use, including fairness and bias.

Validity in the Nigerian Context

In Nigeria, the application of modern validity principles in educational testing faces challenges due to a lack of technical expertise, insufficient test documentation, and irregularities in assessment practices (Onuka & Amoo, 2011). While national examinations such as WAEC and NECO aim for content and construct validity, issues such as examination malpractice and political pressures often undermine the validity of test score interpretations (Amoo & Onuka, 2015).

To improve the validity of educational assessments in Nigeria, stakeholders must invest in psychometric training, enforce test development standards, and promote a culture of evidence-based evaluation (Yusuf & Bello, 2020). Academic success is a key indicator of students' readiness for higher education and their potential for future careers. Chemistry, as a core science subject, is often used to assess students' academic capabilities, especially in STEM (Science, Technology, Engineering, and Mathematics) fields. The effectiveness of Chemistry achievement tests in predicting overall academic success remains a critical area of study. This finding will provide valuable insights for educators, school administrators, and curriculum developers in improving assessment strategies and enhancing the overall quality of science education in the region.

Statement of the Problem

Despite the importance of chemistry education in Nigeria's development, a critical gap exists in our understanding of chemistry achievement tests' effectiveness. This gap manifests in the troubling disconnect between students' test performance and their subsequent academic outcomes, where high scorers often struggle in higher education while average performers sometimes excel. This inconsistency raises fundamental

questions about what these assessments actually measure. The high-stakes nature of these assessments often determining access to competitive university programs and scholarships means that validity and reliability deficiencies may perpetuate systemic inequities, particularly affecting disadvantaged students. Additionally, as chemistry education globally evolves toward emphasizing scientific practices and real-world applications, it remains uncertain whether local assessment methods have adapted accordingly.

A comprehensive investigation into these tests' validity and reliability is therefore essential not only to improve assessment practices but also to ensure equitable educational opportunities and informed decision-making for students pursuing chemistry-related fields in Ondo West Local Government Area. The technical quality of locally developed chemistry tests remains largely unexamined, unlike national standardized examinations. These local tests, which significantly influence students' educational advancement, have not been adequately scrutinized for content validity, construct validity, and reliability, casting doubt on their legitimacy as assessment tools.

Objectives of the Study

The study seeks to:

1. determine the extent to which experts and students view the Chemistry Achievement Test as a relevant measure of chemistry knowledge and skills.
2. examine the internal consistency reliability of the Chemistry Achievement Test using the Kuder-Richardson Formula 20 (KR-20).
3. investigate the strength of the correlation between students' performance in the Chemistry Achievement Test and their subsequent academic success

Research Questions

The following research questions were generated to guide the study:

1. To what extent do experts and students view the chemistry achievement test as a relevant measure of chemistry knowledge and skills?
2. What is the internal consistency of the chemistry achievement test as measured by the Kuder-Richardson Formula 20 (KR-20)?
3. What is the strength of correlation between students' performance on chemistry achievement tests and their subsequent academic success?

Research Hypotheses

For the purpose of this research, one null hypothesis was formulated and tested at 0.05 level of significance.

1. There is no significant difference between the academic performance of female and male students in measuring the validity and reliability of chemistry achievement test.

Methodology

This study employed a survey design of descriptive research. The population comprised of all the science students in senior secondary school 1 & 2 and chemistry teachers in Ondo West Local Government Area, Ondo state. One hundred and eighty (180) students and five (5) chemistry teachers were selected from the population. A purposive and proportionate sampling techniques were used to draw adequate representations across public and private schools. The study made use of Chemistry Achievement Test (CAT), a 40- multiple-choice items with alternative A-D, designed to evaluate mastery of key concept in Chemistry and a secondary data describing the students' initial performance in chemistry subject. The reliability of the instrument was tested through Kuder-Richardson Formula 20 (KR-20). A reliability co- efficient of 0.74 was obtained which indicated a reliable instrument. The instrument was validated through face and content validity by two psychometricians and two chemistry teachers. The data collected was analyzed with the aid of SPSS vs.27.0 and BiLog-MG. Data collected to answer the research questions raised were analyzed using 2-PL Model and Fisher z-test while the hypothesis formulated was tested using independent t-test.

Result and Discussion of Findings

To answer the research question one, an item analysis; Two-Parameter Logistic Model (2PL) was done using BiLog-MG to determine the difficulty and discrimination indices of each items scored as shown in Table 1 below.

Chemistry Achievement Test (CAT)

1. These are the basic particles from which matter could be made except (a) salt (b) atom (c) ion (d) molecule
2. From $XNH_3 + YO_2 = ZNO + QH_2O$, the value of Z is (a) 4 (b) 7 (c) 6 (d) 5
3. Atomicity of ozone is (a) 1 (b) 2 (c) 3 (d) 4
4. A chemical reaction is always associated with (a) an increase in the composition of one of the substances (b) a change in the volume of the reactants (c) a change in the nature of the reactants (d) the formation of new substances
5. Which of the following is an example of a chemical change? (a) Rusting of iron (b) the formation of a mixture (c) dissolution of salt in water (d) an endothermic change

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

6. An atom with 17 protons, 17 electrons, and 18 neutrons has a mass number of (a) 17 (b) 18 (c) 35 (d) 34
7. If an atom of an element is presented as 40, 20 Y, this shows that it has (a) 40 neutrons (b) mass number of 20 (c) 20 protons (d) atomic number of 40
8. Which of the following elements has no unpaired electrons? (a) Li 3 (b) Ne 10 (c) P 15 (d) K 19
9. Which of the following atoms contain the highest number of electrons in the outermost shell? (a) O 18 (b) Ne 10 (c) P 15 (d) K 19
10. Which of the following electronic configuration represent that of a noble gas? (a) 2,8,8,2 (b) 2,8,2 (c) 2,8 (d) 2,6
11. Electrons will occupy equivalent orbitals singly as far as possible with the same spin is a statement of (a) Hund's rule (b) Pauli exclusion principle (c) periodic law (d) Aufbau principle
12. What is the value of -14°C on the kelvin temperature scale? (a) 259k (b) 235k (c) 287K (d) 290k
13. The combining power of an element is known as (a) equation (b) valency (c) compound (d) atom
14. Elements are arranged on the periodic table according to their (a) atomic number (b) mass number (c) atomic mass (d) valence electrons
15. The component of an atom that contribute least to it mass is (a) nucleus (b) neutron (c) electron (d) proton
16. A particle that contains 9 protons, 10 neutrons and 10 electrons is a (a) positive ion (b) neutral atom of a metal (c) neutral atom of a non-metal (d) negative ion
17. Which of these is the same in isotopes of an element? (a) Mass number (b) number of neutrons (c) atomic number (d) relative atomic mass
18. The major reason why chemical reaction occurs among elements is that they have the tendency to (a) attain the nearest noble gas structure (b) become a metal (c) become a non-metal (d) become any noble element
19. Which of the three states of matter has no fixed shaped, no fixed volume and least dense? (a) Gas (b) Liquid (c) Solid (d) Crystals
20. The phenomenon whereby the atmospheric pressure equals the saturated vapour pressure is called (a) freezing (b) latent heat (c) boiling (d) normal pressure
21. $P_1V_1 = P_2V_2$ supports (a) Charle's law (b) Boyle's law (c) Graham's law (d) Avogadro's law
22. Which of the following is a noble gas? (a) Oxygen (b) Nitrogen (c) Argon (d) Hydrogen
23. Which of these is a chemical change? (a) Boiling water (b)Melting ice (c)Dissolving sugar (d) Burning wood

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

24. What is the atomic number of carbon? (a) 12 (b) 6 (c) 8 (d) 4
25. Which of these is a mixture? (a) Water (b) Carbon dioxide (c) Air (d)Sodium chloride
26. The process of changing a solid directly to gas is called: (a) Condensation (b) Sublimation (c) Evaporation (d) Melting
27. Which instrument is used to measure temperature (a) Barometer (b) Thermometer (c)Hydrometer (d) Calorimeter
28. The simplest particle of an element that can take part in a chemical reaction is: (a) Ion (b) Molecule (c) Atom (d) Proton
29. Which of the following is a compound? (a) Air (b) Water (c) Oxygen (d) Copper
30. A molecule of water contains how many atoms? (a) 1 (b) 2 (c) 3 (d) 4
31. The number of protons in an atom is called: (a) Mass number (b) Atomic number (c)Isotope (d) Ion number
32. Which of these is not a physical property of matter? (a) Mass (b) Volume (c) Shape (d) Reactivity
33. What is the chemical symbol for potassium? (a) K (b) P (c) Po (d) Pt
34. Which of the following is an acid? (a) NaOH (b) KOH (c) HCl (d) $Ca(OH)_2$
35. What is the valency of hydrogen? (a) 2 (b) 3 (c) 1 (d) 0
36. Which of the following is a physical change? (a) Rusting of iron (b) Boiling of water (c)Burning of paper(d) Digestion of food
37. The relative atomic mass of an element is: (a) The total number of protons and neutrons (b) The number of electrons (c) The mass of the atom (d)The number of protons only
38. What is the chemical formula for ammonia? (a) H_2O(b) NH_3 (c) CO_2 (d) HCl
39. An example of a triatomic molecule is: (a) O_2 (b) H_2 (c) CO_2 (d) NaCl
40. The process used to separate a mixture of salt and water is: (a) Filtration (b) Decantation (c) Distillation (d) Evaporation

Table 1: Item Difficulty and discrimination Analysis of the Chemistry achievement test

Item no	Difficulty	Discrimination	Item No	Difficulty	Discrimination
1	53.20	.65	**21**	65.70	.57
2	62.50	.75	**22**	39.80	.37
3	60.60	.63	**23**	47.70	.75
4	53.70	.37	**24**	46.30	.54
5	47.70	.42	**25**	62.50	.58
6	47.20	.52	**26**	48.80	.60
7	61.60	.51	**27**	60.60	.49
8	63.90	.64	**28**	56.00	.69

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

Item no	Difficulty	Discrimination	Item No	Difficulty	Discrimination
9	57.90	.73	**29**	60.60	.62
10	51.40	.44	**30**	42.60	.50
11	46.30	.56	**31**	78.70	.31
12	53.20	.58	**32**	69.40	.30
13	52.80	.46	**33**	60.20	.48
14	51.40	.74	**34**	58.80	.38
15	54.60	.30	**35**	55.10	.79
16	60.20	.20	**36**	45.00	.36
17	50.90	.56	**37**	49.00	.78
18	48.10	.78	**38**	48.00	.42
19	48.60	.66	**39**	38.00	.41
20	56.00	.55	**40**	83.00	.50

Key: p= Difficulty index, N_T = 98, $HR_{27\%}$= 49, $LR_{27\%}$= 49

Source: (Field, 2025)

N = Number of items (40)

Rule of thumb used for Difficulty Index

Range	Difficulty level
0-20	Very difficult
21-39	Difficult
40-60	Average/Moderate
61-80	Easy
81-100	Very easy

Interpretation of Discrimination power

Range	Description
0.40 and above	Very good Item
0.30-0.39	Good Item
0.20-0.29	Fair Item
0.10-0.19	Poor Item
0.01-0.09	Very poor items
≤0.00	Mis keyed

Table 1 reveals the result of items facility based on examinees' performances to each item. Items with a difficulty index ranging from 40-60 are only retained. Therefore, items 2, 7, 8, 21, 25, 31 and 32 are easy for the examinees (61≤p =80), only item 40 is very easy. Also, two items (22, and 39) appeared difficult for examinees. This implies that only 30 out of 40-item were considered suitable for inclusion in the subsequent examinations for being neither (too) difficult nor easy but moderately difficult. More so, items with discrimination power ranging from 0.40 and above are considered very good and to be only retained; eight items (4, 15, 16, 22, 31, 32, 34, and 36) are considered good for discrimination but still not suitable for inclusion in the subsequent exam. Invariably, only 32 items are able to discriminate between high achievers and low achievers.

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

Table 2: Reliability coefficient summary showing variances in internal consistencies of CAT

Reliability Methods	Indices
Guttman's L2	0.6409
KR21	0.7351
Feldt-Gilmer	0.6018
Feldt-Brennan	0.7482
Raju's Beta	0.7006

The result in Table 2 reveals the reliability indices of the chemistry achievement test. Kuder–Richardson Formula 21 (KR-21) coefficient of 0.7351 suggests good internal consistency, meaning that the test items are reasonably homogeneous and measure the same knowledge. Feldt–Gilmer Reliability of 0.602 indicates moderate reliability, this suggest that while the test items are reasonably consistent, there is room for improvement in item homogeneity, Feldt–Brennan Reliability coefficient of 0.748 shows acceptable reliability, implying the test halves correlate well, and internal consistency is satisfactory. Lastly, Raju's Beta of 0.701 which indicate correction for item intercorrelations and test length, shows strong internal consistency.

To answer this question the r-value obtained from predictive validity report on the two chemistry tests (CAT and future performance) were compared using z-test. The result is presented in Table 3.

Table 3: Comparison of the validity of performance on chemistry achievement tests (CAT) and their subsequent academic success.

Academic performance	n	r	Z	z^2	P
CAT_1	180	0.422	0.454	2.214	<0.05
CAT_2	180	0.542	0.611		

From table 3, the numbers of items are 40, number of respondents are 180 and r-values are 0.422 and 0.542 for CAT_1 and CAT_2, respectively. Using the z-table, the r-values when transformed become z= 0.454 and 0.611, respectively. Comparing these z-values using Fisher z- test gave a z-value of 2.214 which is significant at the 0.05 level of significant ($p<0.05$). This reveals that there is a significant difference in the validity of the tests.

Hypothesis Testing:

Table 4: Result of independent t-test analysis of the influence of gender on academic performance in secondary school.

Variable	Gender	N		St.D	df	t	Sig	P	η^2
Academic performance	Male	79	13.64	2.075	178	3.448	.000	<.05	0.0626
	Female	101	10.38	2.517					

The result presented on the Table 4 shows that there is a significant difference in the chemistry performance of male and female students; t (178) = 3.448, p<0.05, eta square (η^2) = 0.0626. Thus, the null hypothesis is rejected. The table further reveals that mean rating of male students ($\overline{X}$ = 13.64) is slightly higher than their female counterparts ($\overline{X}$ = 10.38). Size of effect (η^2 = 0.0626) reveals that gender had low effect (according to Cohen 1988 and Field 2000 rule of thumb for size of effect) on the academic performance of students; that is, gender accounted for 6.3% change in the academic performance of chemistry students.

Discussion of Findings

The item analysis in Table 1 indicated that 30 out of 40 items had moderate difficulty indices (40% ≤ p ≤ 60%), showing that most items were appropriately challenging for the target population. This finding corroborates the work of Crocker and Algina (2008), who maintained that moderately difficult items produce optimal score variability and improve discrimination. This study also agrees with Oludipe and Awokoya (2018), who found that discrimination indices above 0.40 are ideal for cognitive tests assessing science reasoning and conceptual mastery. However, about ten items fell outside the optimal range—some being too easy (p > 60) and others too difficult (p < 40). This result marks a line of demarcation from Akpan (2017) and Nworgu (2019), whose studies reported a smaller number of misfitting items in their chemistry test validations. The discrepancy may be due to differences in curriculum coverage, instructional time, or students' exposure to specific chemistry topics.

Additionally, eight items exhibited low discrimination despite appropriate difficulty levels. This outcome corroborates Downing and Haladyna (2006), who argued that item ambiguity, poor distractor construction, or content misalignment can weaken item discrimination even when difficulty is acceptable. A further insight from recent research on scoring methods in multiple-choice testing provides another perspective on these findings. Studies such as Fehintola and Akingbade (2019) and Kasali and Adeyemi (2022) compared the psychometric properties of multiple-choice agricultural science tests using number-right (NR) and confidence scoring methods (CSM) among Nigerian secondary school students. They found that the scoring method significantly influenced both item

difficulty and discrimination indices. For instance, Fehintola and Akingbade reported mean difficulty indices of 55.42% for NR and 44.01% for CSM, and mean discrimination indices of 0.57 and 0.52 respectively, indicating that number-right scoring tends to produce higher difficulty and discrimination levels.

The internal consistency indices as shown in Table 2, KR-21 = 0.7351 and Feldt–Brennan = 0.7482—fall within the acceptable reliability threshold (≥0.70). These findings are supported by Emaikwu (2012) and Okonkwo (2015), who obtained reliability coefficients between 0.70 and 0.80 for science achievement tests among Nigerian secondary students.

However, the relatively lower Feldt–Gilmer coefficient (0.602) indicates moderate reliability, representing a line of demarcation from the higher reliabilities (>0.80) reported by Akpan (2017) and Ojerinde and Popoola (2020). This may reflect variations in item homogeneity or the multidimensional nature of chemistry content.

Nonetheless, Raju's Beta (0.701) supports the test's internal stability, consistent with Miller, Linn, and Gronlund (2009), who emphasized that coefficients around 0.70 denote satisfactory reliability for classroom-based assessments. Thus, the CAT demonstrates acceptable consistency, though slight item revisions could further enhance its psychometric soundness.

The Chemistry Achievement Test's predictive validity showed positive, significant correlations between students' test scores and their subsequent academic achievement ($r = 0.422$ for CAT1; $r = 0.542$ for CAT2, $p < 0.05$). This corroborates earlier findings by Anastasi and Urbina (1997) and Bennett et al. (2005), who established that well-constructed achievement tests moderately predict future academic performance. The stronger correlation for CAT2 suggests improved predictive capacity under this scoring method. This result agrees with Kpolovie (2010), who observed that weighting cognitive and application-level items more heavily increases test validity.

The findings of Fehintola and Akingbade (2019) and Kasali and Adeyemi (2022) further support the notion that scoring method influences validity. Confidence scoring has been shown to reduce random guessing and better reflect true knowledge, thereby enhancing construct and predictive validity. Although the present study applied the number-right method, which yielded moderate predictive validity ($r = 0.542$), incorporating confidence scoring in future versions of the CAT may potentially increase its predictive accuracy by integrating both correctness and confidence levels in scoring.

However, the moderate strength of the correlation contrasts with the stronger predictive coefficients ($r \approx 0.7$) reported by Ojerinde and Popoola (2020) in national standardized tests. This line of demarcation may stem from contextual factors such as differences in instructional quality, school resources, or assessment environment.

The t-test results showed that male students ($\bar{x}$ = 13.64) performed significantly better than female students ($\bar{x}$ = 10.38), t(178) = 3.448, p < 0.05, η^2 = 0.0626. This corroborates study by Ugwu and Opara (2019), who reported male advantage in science performance, often linked to differences in confidence and task persistence.

Nevertheless, the small effect size (η^2 = 0.06) aligns with Stoet and Geary (2018), who found that gender gaps in science achievement are statistically significant but practically small, influenced more by sociocultural and instructional factors than by innate ability.

This finding also demarcates itself from Ezeudu and Ezinwa (2018), who observed no significant gender differences when cooperative and inquiry-based learning approaches were used. Thus, gender disparities may emerge more prominently in traditional teaching settings, emphasizing the need for inclusive instructional strategies.

Conclusion

The findings of this study conclusively affirm the Chemistry Achievement Test (CAT) as a reliable test to be used for state mock-examination. The test demonstrates robustness across key indicators of measurement quality, lending confidence to its adoption within educational assessment frameworks. The observed gender difference, while statistically verifiable, possesses negligible practical significance, thus cautioning against the implementation of broad, gender-specific interventions. Instead, the evidence strongly suggests that the focus of academic support and pedagogical enhancement should be directed toward optimizing instructional quality, customizing teaching methods, and fostering motivational engagement for all students.

Recommendations

Based on the findings and conclusion, the following recommendations were made

1. The state government through the Ministry of Education should consider using Chemistry Achievement Test (CAT) items considered very good for subsequent state mock chemistry examination among SS2 students. This instrument has been validated and proven to be highly reliable and valid with the exception of those minor items
2. The item bank of the Chemistry Achievement Test (CAT) should be rigorously reviewed. Test developers are advised to revise or replace the specific items identified as possessing low discrimination indices and homogeneity values to enhance the instrument's overall diagnostic precision.
3. Interventions designed to improve chemistry performance should primarily target non-gender-specific variables, such as student self-efficacy, interest in the subject, study skills training, and addressing academic anxiety, as these are likely to yield greater practical gains than gender-focused programs.

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

Counselling Implications

The results of this psychometric evaluation carry distinct implications for school counsellors and educational psychologists involved in academic and career guidance:

1. Counsellors must utilize the CAT results diagnostically rather than categorically. Given the small practical effect size of gender, counsellors should actively dismantle gender stereotypes regarding aptitude in chemistry and other STEM fields. Guidance should be based on individual student scores, specific item weaknesses, and observed motivational levels.
2. Counselling sessions should leverage the CAT data to pinpoint precise areas of content deficiency (e.g., stoichiometry, thermodynamics). Interventions should focus on connecting students with targeted academic resources and developing tailored study plans, ensuring the support is needs-based, not gender-based.
3. A core focus of counselling must be on bolstering student confidence and self-efficacy in science. Counsellors should work to reframe perceived failures, encourage participation in high-level science coursework, and actively promote awareness of career pathways that utilize chemistry for both male and female students.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Amoo, S. A., & Onuka, A. O. (2015). Examination malpractice: A threat to educational assessment in Nigeria. *Journal of Educational Assessment in Africa, 3*(1), 102–118.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Prentice Hall.

Bennett, R. E., Jenkins, F., & Persky, H. R. (2005). The impact of test design on predictive validity in educational measurement. *Educational Measurement Quarterly, 12*(3), 18–29.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Cengage Learning.

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development*. Lawrence Erlbaum.

Emaikwu, S. O. (2012). Reliability of science achievement tests in Nigerian secondary schools. *Nigerian Journal of Educational Research, 8*(2), 55–68.

Ezeudu, F. O., & Ezinwa, E. (2018). Gender and inquiry-based learning in chemistry achievement. *Journal of Science Education Research, 5*(3), 41–53.

Fehintola, J.O. & Akingbade, O. (2019). Comparison of the psychometric properties of multiple- choice test using confidence and number right scoring methods on Agricultural Science test among senior secondary school students in Ibadan metropolis. *Journal of Education, AL-HALKIM University, Ilorin (AHJE)* 2(1), 12-25

Kasali, T. A., & Adeyemi, O. A. (2022). Comparative analysis of confidence and number-right scoring methods. *Nigerian Journal of Psychometric Research, 11*(1), 23–38.

Kpolovie, P. J. (2010). *Advanced research methods*. Springfield Publishers. Messick, S. (1989). *Validity*. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education.

Okonkwo, C. E. (2015). Psychometric analysis of achievement tests among secondary school students. *Journal of Educational Assessment, 8*(1), 73–85.

Onuka, A. O. U., & Amoo, S. A. (2011). Teacher assessment literacy and test development in Nigeria. *African Journal of Educational Research, 7*(2), 102–115.

Stoet, G., & Geary, D. C. (2018). The gender-equality paradox in STEM education. *Psychological Science, 29*(4), 581–593.

Thorndike, E. L. (1918). *An introduction to the theory of mental and social measurements*. Science Press.

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education* (8th ed.). Pearson Education.

Ugwu, E. C., & Opara, I. (2019). Gender differences in students' academic achievement in science subjects. *Nigerian Journal of Educational Studies, 15*(2), 23–33.

Yusuf, M. A., & Bello, M. B. (2020). Enhancing psychometric quality of tests through technology integration in Nigerian schools. *Journal of Educational Technology and Evaluation, 12*(1), 15–28.