ISSN: 2790-4407

Journal of Computer Adaptive Testing in Africa

2026; 5; 82-115

An Integrated Multidimensional Item Response Theory Validation Framework for a Chemistry Achievement Test

[a]Funmilayo Rebecca Abayomi* , [b]Modupe Mobolanle Osokoya

[a,b]University of Ibadan, Nigeria

Abstract

The validity of chemistry achievement assessments remains a concern due to the multidimensional nature of the construct and the continued reliance on fragmented validation approaches. This study addresses this gap by adopting an integrated psychometric validation framework grounded in multidimensional Item Response Theory (MIRT). A 150-item dichotomously scored Chemistry Achievement Test (CAT) was administered to 741 senior secondary school students. Dimensionality was examined using Stout's DETECT procedure, revealing weak multidimensionality characterised by a dominant general factor and three secondary dimensions, thereby supporting a four-dimensional structure. Competing MIRT models (M1PL, M2PL, M3PL, and M4PL) were evaluated using likelihood-ratio tests and information criteria. The multidimensional three-parameter logistic (M3PL) model demonstrated superior fit and parsimony, underscoring the importance of modelling both item discrimination and guessing in multiple-choice contexts. Local item independence, assessed via Yen's Q3 statistic, showed no evidence of problematic residual correlations. Item parameter screening indicated that 84% of items met established criteria, leading to the removal of poorly functioning items. Measurement invariance was evaluated through IRT likelihood-ratio Differential Item Functioning (DIF) analysis across sex and school type, resulting in a final set of 112 invariant items. The findings provide strong evidence for the structural validity, reliability, and fairness of the CAT. Beyond its empirical contribution, the study advances an integrated validation framework that systematically combines dimensionality assessment, model evaluation, local independence diagnostics, and invariance testing, offering a robust template for developing equitable and psychometrically sound science assessments.

Keywords: Chemistry Achievement Assessment, DETECT Method, Differential Item Functioning, Measurement Invariance, Multidimensional Item Response Theory

* Corresponding author.

University of Ibadan, Nigeria, e-mail: abayomifunmi.ui@gmail.com

To cite this article:

Abayomi, F. R., & Osokoya, M. M. (2026). An integrated multidimensional item response theory validation framework for a chemistry achievement test. *Journal of Computer Adaptive Testing in Africa, 5*, 82-115. https://doi.org/10.71291/rwykgq70

Received: 5 May 2026
Accepted: 6 September 2026
Published: 8 September 2026

INTRODUCTION

The assessment of students' achievement in science education, particularly in chemistry, remains a persistent challenge in educational measurement due to the inherently complex nature of the construct being assessed (Embretson & Reise, 2000). Chemistry learning extends beyond the acquisition of isolated facts to encompass conceptual understanding, symbolic representation, quantitative reasoning and the application of scientific principles across diverse contexts. Consequently, students' performance in chemistry assessments reflects multiple interrelated cognitive processes rather than a single homogeneous ability. Despite this complexity, many chemistry achievement tests continue to be developed and interpreted using simplified measurement frameworks raising critical concerns about the validity, fairness and interpretability of the resulting scores.

Item Response Theory (IRT) has become a dominant paradigm for modelling the relationship between latent ability and observed item responses because of its capacity to provide item-level diagnostic information and parameter invariance across examinee populations. However, conventional IRT applications are largely grounded in the assumption of unidimensionality, whereby a single latent trait is presumed to account for examinees' performance. Although this assumption offers analytical convenience, it is often untenable in domains such as chemistry, where learning outcomes are inherently multidimensional and cognitively complex (Reckase, 2009; De Ayala, 2009). Consequently, applying unidimensional models to multidimensional data structures may produce biased item parameter estimates, distorted reliability indices, and ultimately invalid inferences about students' achievement.

In response to these limitations, Multidimensional Item Response Theory (MIRT) has emerged as a more flexible and theoretically appropriate framework for modelling complex achievement constructs (Reckase, 2009; Ackerman, Gierl, & Walker, 2003). Unlike conventional IRT models, MIRT allows multiple latent traits to simultaneously influence item responses, thereby providing a more realistic representation of students' performance (Reckase, 2009). However, the effective application of MIRT depends on robust empirical evidence regarding the dimensional structure of the assessment instrument (Stout, 1987; Reckase, 1979). Without such evidence, model specification remains largely arbitrary, potentially undermining the validity of subsequent analyses and the interpretability of the resulting scores (American Educational Research Association, American

Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 2014).

One of the most robust approaches for evaluating dimensionality in dichotomously scored achievement data is Stout's DETECT procedure, which assesses essential dimensionality based on conditional covariance structures (Stout et al., 1996; Zhang & Stout, 1999). Unlike traditional factor-analytic techniques, DETECT does not rely on restrictive distributional assumptions and is particularly well suited to large-scale multiple-choice assessments. Importantly, it facilitates the identification of a dominant general factor alongside meaningful secondary dimensions, thereby providing empirical guidance for multidimensional model specification. Nevertheless, dimensionality assessment alone is insufficient for establishing the overall psychometric adequacy of an assessment instrument.

A comprehensive evaluation of test quality requires the integration of multiple sources of psychometric evidence. First, model–data fit must be rigorously evaluated to determine whether the selected measurement model adequately represents the observed response patterns. For multiple-choice assessments, Multidimensional Item Response Theory (MIRT) models that estimate item discrimination and, where appropriate, pseudo-guessing parameters such as the multidimensional two-parameter logistic (M2PL) and multidimensional three-parameter logistic (M3PL) models, provide greater flexibility for modelling complex educational data than more restrictive alternatives (Reckase, 2009; DeMars, 2010). Second, the assumption of local item independence must be evaluated because violations arising from item redundancy, shared stimuli or overlapping content may inflate reliability estimates, bias item parameter estimates and compromise model fit. Among the available diagnostic procedures, Yen's Q3 residual correlation statistic remains one of the most widely used indices for detecting local item dependence (Yen, 1984; Christensen, Makransky, & Horton, 2017).

In addition to internal structural properties, the fairness of an assessment instrument must also be established. Measurement invariance ensures that test scores are comparable across groups and that observed differences reflect true variations in ability rather than systematic bias. Differential Item Functioning (DIF) analysis plays a critical role in establishing measurement invariance by identifying items that function differently across subpopulations after controlling for ability (Zumbo, 1999;

ISSN: 2790-4407 www. jocatia.acata.org

Woods, 2009). In science education, such differences may arise from variations in instructional exposure, access to laboratory resources or socio-cultural contexts, thereby making measurement invariance testing essential for ensuring equitable assessment practices.

Despite the availability of these advanced psychometric techniques, existing studies on chemistry achievement assessment in Nigeria have largely adopted fragmented validation approaches. Many studies rely exclusively on Classical Test Theory or apply isolated Item Response Theory (IRT) analyses without systematically integrating dimensionality assessment, model evaluation, local item independence, item parameter estimation, and measurement invariance within a unified validation framework. Consequently, the fragmented evaluation of these psychometric properties limits the accumulation of comprehensive validity evidence and reduces confidence in the interpretation, fairness and defensibility of inferences drawn from students' achievement scores.

To address this critical gap, the present study developed and evaluated an integrated multidimensional Item Response Theory (IRT) validation framework for a Chemistry Achievement Test (CAT). Specifically, the study sought to determine the dimensional structure of the instrument using the DETECT procedure, evaluate model–data fit through comparative multidimensional Item Response Theory (MIRT) modelling, assess local item independence using Yen's Q3 statistic, examine the psychometric quality of the calibrated items, and investigate measurement invariance through IRT likelihood-ratio Differential Item Functioning (DIF) analysis across sex and school type. The novelty of this study lies in the systematic integration of these complementary sources of psychometric evidence within a unified validation framework. By integrating evidence from dimensionality assessment, model evaluation, local item independence, item parameter analysis and measurement invariance, the study provides robust evidence of the validity, reliability and fairness of the Chemistry Achievement Test while offering a replicable methodological framework for the development and evaluation of complex achievement assessments in science education.

LITERATURE REVIEW

The validation of educational assessment instruments has evolved considerably from the traditional reliance on Classical Test Theory (CTT) to the adoption of latent trait models that provide more comprehensive information about test items and examinee performance (Embretson & Reise, 2000; DeMars, 2010). Among these models, Item Response Theory (IRT) has become a cornerstone of contemporary educational measurement because it estimates item characteristics and examinee abilities on a common latent scale, thereby supporting more accurate and informative measurement (Hambleton, Swaminathan, & Rogers, 1991; DeMars, 2010). Nevertheless, the validity of inferences derived from IRT depends on the empirical verification of its fundamental assumptions, including dimensionality, local item independence and measurement invariance (Reckase, 2009; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 2014). Although these assumptions are well established in psychometric theory, their comprehensive evaluation has not been consistently incorporated into applied validation studies. Furthermore, previous studies on educational achievement assessment have frequently examined these psychometric properties in isolation rather than within an integrated validation framework, thereby limiting the accumulation of comprehensive validity evidence and weakening confidence in the interpretation and use of students' achievement scores (Rios & Wells, 2014; AERA, APA, & NCME, 2014).

Dimensionality and the Structure of Achievement Tests

Dimensionality is a fundamental concept in educational measurement because it defines the number and nature of the latent traits underlying examinees' responses. Traditional unidimensional Item Response Theory (IRT) models assume that a single latent trait accounts for item performance. However, this assumption has been increasingly challenged in complex academic domains such as science education, where student performance is influenced by multiple cognitive processes. Empirical evidence suggests that achievement in chemistry reflects multiple interrelated competencies, including conceptual understanding, symbolic reasoning, and quantitative problem-solving (Reckase, 2009; De Ayala, 2009). Consequently, imposing strict unidimensionality may oversimplify the construct and result in model misspecification.

In response, the concept of essential dimensionality has gained prominence, recognising that assessment data may be dominated by a general factor while simultaneously exhibiting meaningful secondary dimensions (Stout, 1987; Reise et al., 2010). This perspective provides a more realistic representation of students' cognitive performance, particularly in multidimensional academic domains. Among the available methods for assessing dimensionality, Stout's DETECT procedure has emerged as a robust alternative to traditional factor-analytic approaches. By relying on conditional covariance structures rather than restrictive distributional assumptions, DETECT is

ISSN: 2790-4407

particularly well suited to dichotomously scored achievement data and large-scale educational assessments (Stout et al., 1996; Zhang & Stout, 1999). Empirical studies have demonstrated its effectiveness in identifying latent structures and informing multidimensional model selection (Jang & Roussos, 2007; Zhang, 2007). Despite these advantages, however, DETECT remains underutilised in applied educational research, where dimensionality is often assumed rather than empirically evaluated.

Multidimensional Item Response Theory (MIRT)

The recognition of multidimensionality has led to the development of Multidimensional Item Response Theory (MIRT), which extends traditional Item Response Theory (IRT) by allowing multiple latent traits to simultaneously influence item responses. MIRT provides a theoretically coherent and empirically flexible framework for modelling complex educational constructs, making it particularly suitable for science achievement assessment. Unlike unidimensional models, MIRT captures the interplay among multiple cognitive skills, thereby providing a more comprehensive understanding of students' performance.

A range of MIRT models has been developed, including the multidimensional Rasch model (M1PL), the multidimensional two-parameter logistic model (M2PL), the multidimensional three-parameter logistic model (M3PL), and the multidimensional four-parameter logistic model (M4PL). These models differ in their treatment of item discrimination and guessing behaviour. Whereas the multidimensional Rasch model assumes equal item discrimination, the M2PL and M3PL models relax this assumption, with the M3PL model additionally incorporating a pseudo-guessing parameter to account for the probability of correct responses by low-ability examinees (Reckase, 2009; DeMars, 2010). Empirical evidence consistently indicates that models incorporating both discrimination and pseudo-guessing parameters provide superior model fit for multiple-choice assessments because they more accurately reflect examinees' response processes (Hambleton et al., 1991; DeMars, 2010).

Despite these methodological advances, a persistent limitation in the literature is the tendency to select MIRT models primarily on the basis of statistical fit without adequate consideration of the underlying dimensional structure or theoretical justification. This disconnect between dimensionality assessment and model selection weakens the coherence of psychometric validation and underscores the need for integrated analytical frameworks that combine dimensionality assessment with model evaluation.

Local Item Independence

Local item independence is a fundamental assumption of Item Response Theory (IRT), requiring that item responses are statistically independent after controlling for the underlying latent trait(s). Violations of this assumption, commonly referred to as local dependence, may arise from item clustering, shared stimuli, or overlapping content. Such violations can artificially inflate reliability estimates, bias item parameter estimation, and consequently compromise the validity of measurement (Yen, 1984). Among the available diagnostic procedures, Yen's Q3 statistic is one of the most widely used indices for detecting local dependence by examining residual correlations between item pairs. Although recommended threshold values vary, residual correlations exceeding 0.20 or 0.30 are generally considered indicative of problematic local dependence (Chen & Thissen, 1997; DeMars, 2010). Despite the recognised importance of assessing local item independence, this assumption is frequently overlooked or treated as a secondary diagnostic in applied psychometric studies. Consequently, violations of local independence may remain undetected, undermining confidence in model adequacy and raising concerns about the structural validity of assessment instruments.

Item Parameter Evaluation and Test Quality

Within the Item Response Theory (IRT) framework, item quality is evaluated using three key parameters: discrimination (a), difficulty (b), and pseudo-guessing (c). These parameters provide important information about how individual items function across different levels of examinee ability. High-quality items are characterised by adequate discrimination, appropriate difficulty levels, and low susceptibility to guessing. Conversely, poorly functioning items may reduce measurement precision, distort the representation of the underlying construct, and compromise the overall quality of the assessment (Hambleton et al., 1991; Embretson & Reise, 2000).

Although item parameter evaluation is a fundamental component of IRT-based assessment, its integration into broader psychometric validation frameworks remains inconsistent. In many studies, item screening is conducted independently of other sources of validity evidence with limited consideration of dimensionality, local item independence and measurement invariance. Consequently, the contribution of item-level diagnostics to the overall validity and interpretability of test scores is often not fully realised. Furthermore, because item parameter estimates describe the

ISSN: 2790-4407 www. jocatia.acata.org

performance of individual items rather than the precision of the test as a whole, they should be complemented by an evaluation of test information and measurement precision across the ability continuum.

Measurement Invariance and Differential Item Functioning

Measurement invariance is fundamental to ensuring that assessment instruments function equivalently across different groups. Without measurement invariance, observed score differences may reflect systematic bias rather than genuine differences in the underlying ability being measured. Differential Item Functioning (DIF) analysis provides a rigorous approach for detecting such bias by examining whether item parameters differ across groups after controlling for ability. Among the available approaches, IRT-based likelihood-ratio DIF methods provide robust statistical procedures for identifying potentially biased items (Thissen et al., 1993). Research has shown that DIF may arise from multiple sources, including differences in curriculum exposure, language proficiency, and socio-cultural contexts (Zumbo, 1999; Woods, 2009). In science education, disparities in access to laboratory resources and variations in instructional quality may further contribute to differential item functioning. Despite its recognised importance, measurement invariance testing is frequently conducted as an isolated psychometric procedure rather than as part of a comprehensive validation framework. Consequently, the relationship between fairness and other important sources of validity evidence, including dimensionality, model fit, local item independence, and item parameter quality, is often overlooked.

Synthesis and Research Gap

The literature highlights significant advances in psychometric modelling, particularly in the development of Multidimensional Item Response Theory (MIRT) and diagnostic techniques such as the DETECT procedure and Differential Item Functioning (DIF) analysis. Nevertheless, important limitations remain because validation studies continue to adopt fragmented approaches. Most existing studies examine dimensionality, model fit, local item independence or measurement invariance independently rather than integrating these complementary sources of psychometric evidence within a coherent validation framework. Consequently, three important gaps in the literature can be identified (Messick, 1995; Reckase, 2009; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 2014).

First, dimensionality is frequently assumed rather than empirically verified despite substantial evidence that achievement data are inherently multidimensional. Second, there is limited integration between dimensionality assessment and multidimensional model selection, resulting in potential inconsistencies in psychometric decision-making. Third, comprehensive validation frameworks that simultaneously evaluate structural validity, item quality, local item independence, and measurement invariance are rarely applied to chemistry achievement assessments, particularly in diverse educational contexts.

To address these gaps, the present study adopts an integrated multidimensional Item Response Theory (IRT) validation framework that systematically combines dimensionality assessment, model evaluation, local item independence diagnostics, item parameter evaluation, and measurement invariance testing. Through this integrated approach, the study seeks to provide comprehensive validity evidence for the Chemistry Achievement Test while offering a replicable methodological framework for the development and evaluation of equitable science achievement assessments.

METHODOLOGY

Research Design

This study adopted a quantitative instrument validation design grounded in Item Response Theory (IRT) with a specific emphasis on multidimensional modelling. An integrated validation framework was employed to evaluate the psychometric properties of a Chemistry Achievement Test (CAT) through a combination of exploratory and confirmatory analytical procedures. Given the multidimensional nature of chemistry achievement, a Multidimensional Item Response Theory (MIRT) approach was used to model the multiple latent traits underlying students' performance. The validation process was implemented sequentially by examining (i) dimensionality, (ii) model–data fit, (iii) local item independence, (iv) item parameter quality, and (v) measurement invariance. This integrated design provided comprehensive evidence regarding the structural validity, reliability and fairness of the Chemistry Achievement Test.

Participants and Sampling Procedure

The study involved 741 Senior Secondary School Two (SSS II) science students drawn from selected secondary schools in Oyo State, Nigeria. The participants were recruited for the trial administration

ISSN: 2790-4407 www. jocatia.acata.org

of the 150-item Chemistry Achievement Test (CAT) to evaluate its psychometric properties prior to the main study.

A stratified purposive sampling technique was employed. Initially, schools were stratified according to ownership (public and private) to ensure adequate representation of the two major school types. Subsequently, schools were purposively selected, and only Senior Secondary School Two (SSS II) science students offering Chemistry were included because they possessed the prerequisite knowledge required to respond meaningfully to the Chemistry Achievement Test (CAT). SSS II students were specifically selected because they had covered most of the chemistry concepts assessed by the instrument but had not yet taken the final secondary school examinations. This procedure yielded a total sample of 741 students. The sample size was considered adequate for Multidimensional Item Response Theory (MIRT) calibration, as previous studies have recommended sample sizes ranging from 500 to 1,000 examinees for stable item parameter estimation (De Ayala, 2009; Reckase, 2009).

Instrument Development and Description

The instrument used in this study was a Chemistry Achievement Test (CAT) comprising 150 multiple-choice items, each scored dichotomously (1 = correct, 0 = incorrect). The test was designed to assess a broad range of competencies within the senior secondary school chemistry curriculum, including conceptual understanding, symbolic manipulation, quantitative reasoning and the application of scientific principles. The CAT was developed by the researcher based on the West African Examinations Council (WAEC) and National Examinations Council (NECO) Senior Secondary School Chemistry syllabi. Item development was guided by the content specifications of these syllabi and the cognitive process dimensions of the revised Bloom's Taxonomy (Anderson & Krathwohl, 2001). This was done to ensure adequate representation of both curriculum content and cognitive demand (Oladele & Ndlovu, 2021).

The instrument consisted of multiple-choice items because of their suitability for large-scale educational assessment and their compatibility with Item Response Theory (IRT) analysis. Responses were analysed using the Multidimensional Three-Parameter Logistic (M3PL) model, which estimates item discrimination, difficulty, and pseudo-guessing parameters. Prior to the main study, the CAT was administered as a trial test to calibrate its psychometric properties. Data obtained from the trial administration were used to evaluate item functioning, assess the assumptions underlying the IRT models, and refine the instrument before its use in the main study.

Data Collection Procedure

Data were collected under standardized testing conditions to minimise extraneous sources of variance. All participants completed the Chemistry Achievement Test (CAT) independently in a controlled testing environment, and uniform instructions were administered across all participating schools. Responses were scored dichotomously (1 = correct, 0 = incorrect) and organised into a response matrix for Item Response Theory (IRT) analysis. Data screening procedures were undertaken to ensure data integrity, including checks for missing responses, response inconsistencies and data entry errors. Cases with excessive missing responses or aberrant response patterns were excluded from the analysis to maintain the accuracy of the psychometric evaluation.

Ethical Considerations

Ethical approval for this study was obtained from the University of Ibadan. Participation was voluntary and informed consent was obtained from all participants and where applicable, from the relevant school authorities. Participants were assured of the confidentiality and anonymity of their responses and all data collected were used solely for research purposes. Furthermore, informed consent was obtained from all participants and the relevant school authorities prior to data collection. Participation was voluntary, and respondents were assured of the confidentiality and anonymity of the information provided.

Data Analysis Strategy

All statistical analyses were conducted using the R statistical environment (Version 4.5.2), employing the mirt package for Multidimensional Item Response Theory (MIRT) modelling, sirt for DETECT analysisa and psych for preliminary psychometric diagnostics. The analytical process was organised into five sequential stages, reflecting the integrated validation framework adopted in this study. Unless otherwise stated, all statistical decisions were evaluated at a significance level of $\alpha = .05$.

ISSN: 2790-4407 www. jocatia.acata.org

Dimensionality Assessment

Dimensionality was assessed using Stout's DETECT procedure, a nonparametric approach based on conditional covariance among item pairs (Stout et al., 1996). DETECT was selected because of its suitability for dichotomously scored data and its ability to assess essential dimensionality without imposing restrictive distributional assumptions. Exploratory DETECT analyses were conducted by evaluating cluster solutions ranging from two to ten dimensions. The optimal dimensional structure was determined using the maximum DETECT index, supported by the Approximate Simple Structure Index (ASSI) and RATIO statistics. Interpretation followed established guidelines whereby DETECT values below 0.20 indicate essential unidimensionality, whereas higher values indicate increasing multidimensionality. The results provided empirical justification for specifying multidimensional IRT models by identifying a dominant general factor together with meaningful secondary dimensions.

Model–Data Fit Evaluation

Following the dimensionality assessment, a series of multidimensional Item Response Theory (MIRT) models were estimated using full-information item factor analysis. The models included the multidimensional Rasch model (M1PL), the multidimensional two-parameter logistic model (M2PL), the multidimensional three-parameter logistic model (M3PL), and the multidimensional four-parameter logistic model (M4PL). These models represented increasing levels of parameter complexity, enabling a systematic evaluation of the contribution of discrimination and pseudo-guessing parameters. Model comparisons were conducted using likelihood-ratio tests for nested models together with multiple information criteria, including the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Sample-Size Adjusted Bayesian Information Criterion (SABIC) and Hannan–Quinn (HQ) criterion. Model selection was guided by statistical fit, theoretical interpretability and parsimony to minimise overfitting and parameter instability.

Assessment of Local Item Independence

Local item independence was evaluated using Yen's Q3 statistic, which examines residual correlations between item pairs after controlling for the estimated latent traits (Yen, 1984). Yen's Q3 residual correlations were computed for all item pairs, and their distribution was summarised using descriptive statistics. A threshold of $|Q3| < 0.30$ was adopted as the criterion for acceptable local item independence, consistent with established recommendations (Chen & Thissen, 1997). This procedure ensured that the selected model adequately accounted for shared variance among items and that residual dependencies did not compromise parameter estimation.

Item Parameter Evaluation

Item parameter estimates obtained from the selected MIRT model were evaluated using established psychometric criteria. Specifically, discrimination parameters were considered acceptable when $a > 0.30$, difficulty parameters when $-2 \leq b \leq +2$, and pseudo-guessing parameters when $c \leq 0.35$ (Hambleton et al., 1991; Embretson & Reise, 2000). Items satisfying all criteria were classified as well-functioning, whereas those failing to meet one or more criteria were flagged for review and possible removal. This systematic screening ensured that the retained items demonstrated acceptable psychometric properties and adequately represented the intended construct.

Measurement Invariance and Differential Item Functioning (DIF)

Measurement invariance was assessed using IRT likelihood-ratio Differential Item Functioning (IRT-LR DIF) analysis (Thissen et al., 1993). This approach compares nested models in which item parameters are either constrained to be equal across groups or freely estimated. DIF analyses were conducted across two grouping variables: sex (male vs. female) and school type (public vs. private). Items exhibiting statistically significant chi-square values ($p < .05$) were identified as displaying potential Differential Item Functioning and were considered for removal to ensure that the final instrument measured the intended construct equivalently across groups. This procedure enhanced the fairness and validity of the Chemistry Achievement Test (CAT) by minimising construct-irrelevant variance and supporting equitable interpretation of test scores.

RESULTS

Dimensionality of the Chemistry Achievement Test

The latent structure of the Chemistry Achievement Test (CAT) was examined using Stout's DETECT procedure to determine the extent and nature of its dimensionality. Exploratory analyses evaluated cluster solutions ranging from two to ten dimensions. As shown in Table 1, the DETECT index reached its maximum value under the four-cluster solution (DETECT = .342), accompanied by ASSI = .248 and RATIO = .387. According to established interpretation criteria, DETECT values between .20 and .40 indicate weak multidimensionality characterised by a dominant general

ISSN: 2790-4407 www. jocatia.acata.org

factor together with meaningful secondary dimensions. The four-cluster solution produced a cluster size pattern of 21–59–35–35, while the MCOV100 value remained constant at −.081 across all cluster solutions. These findings indicate that the CAT demonstrates essential multidimensionality, thereby supporting the application of multidimensional Item Response Theory (MIRT) models for subsequent analyses.

Table 1: Exploratory DETECT Results (Unweighted)

N. Cluster	N. Items	Cluster Size Pattern	DETECT	ASSI	RATIO	MADCOV 100	MCOV 100
2	150	80–70	0.290	0.198	0.328	0.885	−0.081
3	150	80–35–35	0.320	0.230	0.362	0.885	−0.081
4	150	21–59–35–35	0.342	0.248	0.387	0.885	−0.081
5	150	21–47–35–35–12	0.337	0.249	0.381	0.885	−0.081
6	150	21–47–22–35–12–13	0.319	0.233	0.361	0.885	−0.081
7	150	15–47–22–35–6–12–13	0.313	0.228	0.354	0.885	−0.081
8	150	15–21–22–26–35–6–12–13	0.310	0.228	0.350	0.885	−0.081
9	150	15–21–22–26–17–6–18–12–13	0.293	0.212	0.331	0.885	−0.081
10	150	15–21–17–26–17–6–18–12–5–13	0.283	0.206	0.319	0.885	−0.081

Model–Data Fit of Multidimensional Item Response Theory Models

Following the identification of the latent structure, a sequence of multidimensional Item Response Theory (MIRT) models comprising the multidimensional Rasch model (M1PL), multidimensional two-parameter logistic model (M2PL), multidimensional three-parameter logistic model (M3PL), and multidimensional four-parameter logistic model (M4PL) was estimated and compared to determine the model that best represented the response data. As shown in Table 2, the transition from the M1PL model to the M2PL model resulted in a statistically significant improvement in model fit ($\Delta\chi^2 = 2{,}729.01$, $\Delta df = 146$, $p < .001$). A further comparison between the M2PL and M3PL models also produced a statistically significant improvement ($\Delta\chi^2 = 887.23$, $\Delta df = 150$, $p < .001$). Although the M4PL model yielded a marginal improvement in log-likelihood, it recorded higher values on the Bayesian Information Criterion (BIC) and Sample-Size Adjusted Bayesian Information Criterion (SABIC) than the M3PL model. The M3PL model produced the lowest Akaike Information Criterion (AIC), SABIC, and Hannan–Quinn (HQ) values, indicating the best balance between model fit and parsimony. Accordingly, the M3PL model was selected as the best-fitting model and retained for all subsequent analyses.

Table 2: Comparison of Multidimensional Item Response Theory Models

Model	AIC	SABIC	HQ	BIC	Log-Likelihood	$\Delta\chi^2$	Δdf	p-value
M1PL	128,773.1	128,993.8	129,046.7	129,482.8	−64,232.57	–	–	–
M2PL	126,336.1	126,765.9	126,869.1	127,718.5	−62,868.06	2,729.01	146	< .001
M3PL	125,748.9	126,393.6	126,548.4	127,822.5	−62,424.45	887.23	150	< .001
M4PL	125,777.6	126,637.2	126,843.5	128,542.4	−62,288.80	271.29	150	< .001

Assessment of Local Item Independence

Local item independence was evaluated using Yen's Q3 statistic to assess residual associations between item pairs after controlling for the latent traits. Table 3 presents the summary statistics of the absolute Q3 residual correlations computed from the 150 × 150 residual correlation matrix, while Figure 2 illustrates the distribution of residual correlations among all item pairs. The mean absolute Q3 value was .04527, with a median of .03786 and a third quartile of .06473, indicating generally low residual associations between items. The maximum observed Q3 value was .25754, and no residual correlation exceeded the commonly accepted threshold of .30, indicating that none of the item pairs demonstrated problematic local dependence.

Based on the adopted criterion of $|Q3| < .30$ (Chen & Thissen, 1997; DeMars, 2010), the assumption of local item independence was satisfied for the Chemistry Achievement Test. Accordingly, the selected multidimensional Item Response Theory model was considered appropriate for subsequent item parameter estimation and Differential Item Functioning analyses.

ISSN: 2790-4407 www. jocatia.acata.org

Table 3: Summary Statistics of Absolute Yen's Q3 Residual Correlations

Statistic	Value
Minimum	0.00000
1st Quartile (Q1)	0.01778
Median	0.03786
Mean	0.04527
3rd Quartile (Q3)	0.06473
Maximum	0.25754
N	150

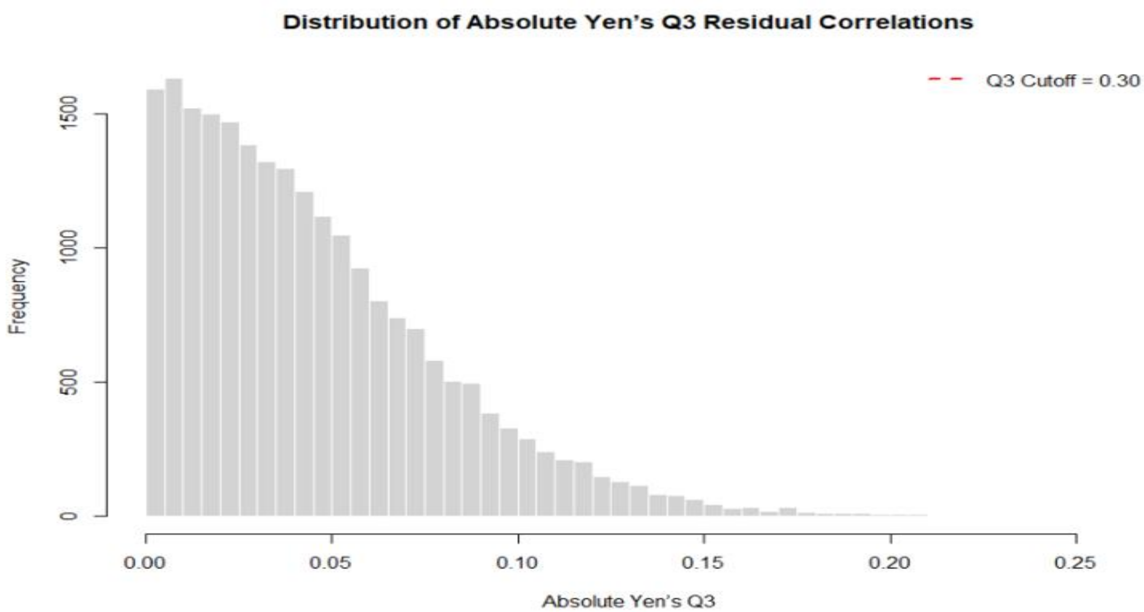

Figure 1: Distribution of Absolute Yen's Q3 Residual Correlations for the Chemistry Achievement Test

Evaluation of Item Parameter Quality

Item parameter estimates obtained from the multidimensional three-parameter logistic (M3PL) model were evaluated using the predefined psychometric criteria described in the methodology. Specifically, items were classified as Good if they satisfied all three criteria: discrimination parameter (a > .30), difficulty parameter ($-2 \leq b \leq +2$), and pseudo-guessing parameter ($c \leq .35$). Items that failed to satisfy one or more of these criteria were classified as Poor and considered unsuitable for retention in the final instrument.

As presented in Table 4, 126 of the 150 items (84%) satisfied all three psychometric criteria and were classified as Good. These items demonstrated acceptable discrimination, appropriate difficulty levels, and low pseudo-guessing probabilities. Conversely, 24 items (16%) were classified as Poor because they failed to satisfy one or more of the established criteria, including low discrimination, difficulty values outside the acceptable range, or pseudo-guessing parameters exceeding the specified threshold. Consequently, these items were excluded from the final version of the Chemistry Achievement Test.

Table 4: Item Parameter Estimates for the Multidimensional Three-Parameter Logistic (M3PL) Model

c	a1	a2	a3	a4	Discrimination_a	Difficulty_b	Guessing_c	Remarks
ITEM001	1.461	.000	.000	.000	1.461	-.274	.234	Good
ITEM002	1.958	.000	.000	.000	1.958	.795	.227	Good
ITEM003	.000	1.631	.000	.000	1.631	.940	.292	Good
ITEM004	.000	.000	.610	.000	.610	2.837	.116	Poor
ITEM005	.000	1.436	.000	.000	1.436	.169	.403	Poor
ITEM006	.000	1.172	.000	.000	1.172	.641	.158	Good
ITEM007	2.142	.000	.000	.000	2.142	.404	.175	Good
ITEM008	4.506	.000	.000	.000	4.506	1.428	.403	Poor
ITEM009	.000	2.328	.000	.000	2.328	1.212	.365	Poor
ITEM010	1.822	.000	.000	.000	1.822	.181	.262	Good
ITEM011	1.912	.000	.000	.000	1.912	.584	.431	Poor
ITEM012	.000	.624	.000	.000	.624	-.213	.001	Good
ITEM013	.000	.076	.000	.000	.076	21.146	.009	Poor
ITEM014	1.490	.000	.000	.000	1.490	.578	.226	Good
ITEM015	1.833	.000	.000	.000	1.833	-.192	.132	Good
ITEM016	1.223	.000	.000	.000	1.223	1.437	.104	Good
ITEM017	.952	.000	.000	.000	.952	-.721	.002	Good
ITEM018	1.494	.000	.000	.000	1.494	.102	.292	Good

ISSN: 2790-4407 www. jocatia.acata.org

c	a1	a2	a3	a4	Discrimination_a	Difficulty_b	Guessing_c	Remarks
ITEM019	1.892	.000	.000	.000	1.892	-.199	.308	Good
ITEM020	6.882	.000	.000	.000	6.882	-.105	.208	Good
ITEM021	1.188	.000	.000	.000	1.188	-1.012	.002	Good
ITEM022	.000	.000	.000	2.086	2.086	1.326	.434	Poor
ITEM023	.000	1.667	.000	.000	1.667	1.120	.200	Good
ITEM024	1.068	.000	.000	.000	1.068	.056	.077	Good
ITEM025	1.332	.000	.000	.000	1.332	.021	.156	Good
ITEM026	1.068	.000	.000	.000	1.068	-.704	.001	Good
ITEM027	.000	.000	.000	1.350	1.350	.686	.283	Good
ITEM028	.000	.000	.000	1.365	1.365	1.009	.258	Good
ITEM029	.000	.000	.000	2.180	2.180	.697	.148	Good
ITEM030	.000	.000	.000	.106	.106	15.739	.017	Poor
ITEM031	.000	.000	-.928	.000	-.928	-5.530	.146	Poor
ITEM032	.000	.000	.000	1.531	1.531	.604	.184	Good
ITEM033	.000	.613	.000	.000	.613	-.113	.011	Good
ITEM034	.937	.000	.000	.000	.937	-.692	.001	Good
ITEM035	.000	.000	.000	1.200	1.200	.037	.138	Good
ITEM036	1.052	.000	.000	.000	1.052	-.066	.000	Good
ITEM037	1.101	.000	.000	.000	1.101	-.751	.000	Good
ITEM038	.000	.850	.000	.000	.850	-.601	.004	Good
ITEM039	.000	1.629	.000	.000	1.629	1.188	.167	Good
ITEM040	.000	.000	.000	2.247	2.247	.714	.443	Poor
ITEM041	.000	.000	.000	1.789	1.789	1.485	.333	Good
ITEM042	.000	.523	.000	.000	.523	1.008	.083	Good
ITEM043	.000	.000	.000	1.315	1.315	.898	.066	Good
ITEM044	.000	.000	.000	1.972	1.972	.823	.162	Good
ITEM045	.000	.000	.000	5.663	5.663	1.392	.172	Good
ITEM046	.000	.000	.000	2.447	2.447	.921	.216	Good
ITEM047	.000	2.060	.000	.000	2.060	1.246	.094	Good
ITEM048	.000	.000	1.438	.000	1.438	2.365	.095	Poor
ITEM049	.000	1.281	.000	.000	1.281	-.444	.210	Good

c	a1	a2	a3	a4	Discrimination_a	Difficulty_b	Guessing_c	Remarks
ITEM050	.000	1.663	.000	.000	1.663	-.575	.192	Good
ITEM051	.000	1.804	.000	.000	1.804	.975	.207	Good
ITEM052	.000	.605	.000	.000	.605	-.632	.008	Good
ITEM053	.000	.766	.000	.000	.766	-1.245	.001	Good
ITEM054	.000	.000	.000	2.593	2.593	1.141	.277	Good
ITEM055	.000	.000	.000	5.678	5.678	1.396	.342	Good
ITEM056	.000	.000	.000	2.186	2.186	1.274	.237	Good
ITEM057	.000	.069	.000	.000	.069	42.320	.029	Poor
ITEM058	.000	1.624	.000	.000	1.624	1.081	.048	Good
ITEM059	.000	1.076	.000	.000	1.076	2.664	.250	Poor
ITEM060	.000	.866	.000	.000	.866	.446	.000	Good
ITEM061	.000	.000	1.147	.000	1.147	.453	.187	Good
ITEM062	.000	.000	1.476	.000	1.476	-.380	.124	Good
ITEM063	.000	.000	.000	2.089	2.089	.792	.194	Good
ITEM064	.000	.000	1.425	.000	1.425	-.693	.001	Good
ITEM065	.000	.000	6.480	.000	6.480	-.115	.293	Good
ITEM066	.000	.000	1.769	.000	1.769	-1.156	.001	Good
ITEM067	.000	.000	1.722	.000	1.722	.790	.209	Good
ITEM068	.000	.000	1.059	.000	1.059	-1.242	.001	Good
ITEM069	.000	.000	.906	.000	.906	-.187	.002	Good
ITEM070	.000	.000	1.065	.000	1.065	-.374	.001	Good
ITEM071	.000	.000	1.079	.000	1.079	.419	.001	Good
ITEM072	.000	1.151	.000	.000	1.151	-.124	.117	Good
ITEM073	.000	1.196	.000	.000	1.196	.823	.150	Good
ITEM074	.000	3.474	.000	.000	3.474	1.170	.348	Good
ITEM075	.000	.000	1.390	.000	1.390	-1.195	.008	Good
ITEM076	.000	.000	.000	1.334	1.334	.697	.234	Good
ITEM077	.000	-.725	.000	.000	-.725	-4.535	.003	Poor
ITEM078	.000	1.738	.000	.000	1.738	.794	.157	Good

ISSN: 2790-4407 www. jocatia.acata.org

c	a1	a2	a3	a4	Discrimination_a	Difficulty_b	Guessing_c	Remarks
ITEM079	.000	.000	.000	2.660	2.660	.915	.253	Good
ITEM080	.000	.000	1.394	.000	1.394	.579	.195	Good
ITEM081	.000	-1.160	.000	.000	-1.160	-2.696	.016	Poor
ITEM082	.000	1.423	.000	.000	1.423	-.827	.001	Good
ITEM083	.000	-.269	.000	.000	-.269	-6.405	.001	Poor
ITEM084	.000	.000	1.633	.000	1.633	-.239	.004	Good
ITEM085	.000	1.805	.000	.000	1.805	.948	.325	Good
ITEM086	.000	.000	1.482	.000	1.482	-.319	.107	Good
ITEM087	.000	.000	1.892	.000	1.892	-.242	.156	Good
ITEM088	.000	.000	.000	1.424	1.424	.581	.167	Good
ITEM089	.000	.000	2.527	.000	2.527	.658	.344	Good
ITEM090	.000	.000	1.712	.000	1.712	.187	.100	Good
ITEM091	.000	-.373	.000	.000	-.373	-4.136	.000	Poor
ITEM092	.000	.000	.000	1.167	1.167	.699	.001	Good
ITEM093	.000	1.260	.000	.000	1.260	.323	.153	Good
ITEM094	.000	.000	.000	4.724	4.724	1.342	.255	Good
ITEM095	1.558	.000	.000	.000	1.558	1.417	.256	Good
ITEM096	.000	.952	.000	.000	.952	.184	.046	Good
ITEM097	.000	.000	.847	.000	.847	-.572	.000	Good
ITEM098	.000	.000	.000	1.692	1.692	-.111	.009	Good
ITEM099	.000	1.187	.000	.000	1.187	.997	.156	Good
ITEM100	.000	.000	.394	.000	.394	1.969	.003	Good
ITEM101	.000	-.554	.000	.000	-.554	-6.500	.126	Poor
ITEM102	.000	.000	.000	1.344	1.344	-.622	.003	Good
ITEM103	.000	.000	.991	.000	.991	-.060	.212	Good
ITEM104	.000	2.069	.000	.000	2.069	.964	.139	Good
ITEM105	.000	.000	.000	1.372	1.372	.359	.100	Good
ITEM106	.000	.000	.000	1.932	1.932	.558	.142	Good
ITEM107	.000	.000	.000	.474	.474	.776	.001	Good
ITEM108	.000	.000	.000	.908	.908	-.418	.001	Good
ITEM109	.000	.938	.000	.000	.938	1.161	.140	Good

c	a1	a2	a3	a4	Discrimination_a	Difficulty_b	Guessing_c	Remarks
ITEM110	.000	1.546	.000	.000	1.546	.645	.222	Good
ITEM111	.000	.000	1.671	.000	1.671	1.479	.174	Good
ITEM112	.000	.000	.000	1.341	1.341	.585	.140	Good
ITEM113	.000	.000	2.201	.000	2.201	1.046	.356	Poor
ITEM114	.000	1.737	.000	.000	1.737	.864	.216	Good
ITEM115	.000	.000	.000	1.039	1.039	1.141	.116	Good
ITEM116	.000	.000	1.083	.000	1.083	-.557	.002	Good
ITEM117	.000	1.297	.000	.000	1.297	1.994	.122	Good
ITEM118	.000	.000	.000	1.666	1.666	1.086	.157	Good
ITEM119	.000	1.846	.000	.000	1.846	1.102	.187	Good
ITEM120	.000	.000	1.102	.000	1.102	-.895	.001	Good
ITEM121	.000	.887	.000	.000	.887	.747	.135	Good
ITEM122	.000	1.282	.000	.000	1.282	.823	.205	Good
ITEM123	.000	2.547	.000	.000	2.547	.317	.247	Good
ITEM124	.000	.751	.000	.000	.751	-.221	.001	Good
ITEM125	.000	.000	1.124	.000	1.124	-.734	.001	Good
ITEM126	.000	.000	.000	1.521	1.521	.592	.165	Good
ITEM127	.000	-.472	.000	.000	-.472	-3.782	.000	Poor
ITEM128	.000	.000	.000	1.543	1.543	.707	.180	Good
ITEM129	.000	1.237	.000	.000	1.237	.602	.000	Good
ITEM130	.000	1.577	.000	.000	1.577	.749	.240	Good
ITEM131	.000	1.778	.000	.000	1.778	.866	.190	Good
ITEM132	.000	.863	.000	.000	.863	.321	.033	Good
ITEM133	.000	-.011	.000	.000	-.011	-111.926	.043	Poor
ITEM134	.000	.000	.000	-.008	-.008	-214.477	.066	Poor
ITEM135	.000	.808	.000	.000	.808	.301	.079	Good
ITEM136	.000	4.489	.000	.000	4.489	1.522	.231	Good
ITEM137	.000	1.526	.000	.000	1.526	.693	.237	Good
ITEM138	.000	.917	.000	.000	.917	.914	.000	Good

ISSN: 2790-4407 www. jocatia.acata.org

c	a1	a2	a3	a4	Discrimination_a	Difficulty_b	Guessing_c	Remarks
ITEM139	.000	.997	.000	.000	.997	-.083	.000	Good
ITEM140	.000	.000	1.475	.000	1.475	.784	.205	Good
ITEM141	.000	.000	.770	.000	.770	.199	.052	Good
ITEM142	.000	.000	.596	.000	.596	.690	.032	Good
ITEM143	.000	.000	.673	.000	.673	.851	.001	Good
ITEM144	.000	.000	2.001	.000	2.001	-.112	.278	Good
ITEM145	.000	.000	.000	-.282	-.282	-6.184	.004	Poor
ITEM146	.000	-.471	.000	.000	-.471	-6.805	.000	Poor
ITEM147	.000	.000	1.196	.000	1.196	1.350	.079	Good
ITEM148	.000	1.980	.000	.000	1.980	.950	.244	Good
ITEM149	.000	2.128	.000	.000	2.128	1.326	.140	Good
ITEM150	.000	.000	.705	.000	.705	.007	.002	Good

Note: Items were classified as Good if they satisfied all three psychometric criteria ($a > .30$, $-2 \leq b \leq +2$, and $c \leq .35$). Items that failed to satisfy one or more of these criteria were classified as Poor.

Measurement Invariance Across Sex

Measurement invariance across sex was assessed using Item Response Theory likelihood-ratio Differential Item Functioning (IRT-LR DIF) analysis under the multidimensional three-parameter logistic (M3PL) model. As presented in Table 5, 10 of the 126 retained items exhibited statistically significant Differential Item Functioning (DIF) ($p < .05$), indicating that these items functioned differently for male and female students after controlling for ability. The remaining 116 items showed no statistically significant DIF ($p > .05$), supporting measurement invariance across sex. To ensure that the final instrument functioned equivalently across male and female students, the 10 DIF-affected items were excluded from the final version of the Chemistry Achievement Test, while the remaining items were retained for subsequent analyses.

Table 5: IRT Likelihood-Ratio Differential Item Functioning Analysis by Sex (n = 126)

Item	X^2	df	p-value	Decision
ITEM001	0.84	2	.656	No DIF
ITEM002	1.10	2	.576	No DIF
ITEM003	0.92	2	.632	No DIF
ITEM006	1.28	2	.526	No DIF

Item	X²	df	p-value	Decision
ITEM007	0.77	2	.679	No DIF
ITEM010	0.63	2	.729	No DIF
ITEM012	1.05	2	.592	No DIF
ITEM014	0.98	2	.612	No DIF
ITEM015	1.34	2	.511	No DIF
ITEM016	0.89	2	.641	No DIF
ITEM017	1.11	2	.574	No DIF
ITEM018	0.94	2	.626	No DIF
ITEM019	1.29	2	.525	No DIF
ITEM020	0.76	2	.684	No DIF
ITEM021	1.18	2	.554	No DIF
ITEM023	0.91	2	.635	No DIF
ITEM024	1.07	2	.586	No DIF
ITEM025	0.88	2	.644	No DIF
ITEM026	1.33	2	.514	No DIF
ITEM027	1.01	2	.603	No DIF
ITEM028	0.95	2	.623	No DIF
ITEM029	1.26	2	.532	No DIF
ITEM032	1.12	2	.571	No DIF
ITEM033	0.86	2	.651	No DIF
ITEM034	1.24	2	.538	No DIF
ITEM035	0.79	2	.673	No DIF
ITEM036	0.92	2	.632	No DIF
ITEM037	1.15	2	.563	No DIF
ITEM038	110.294	2	< .05	DIF
ITEM039	0.83	2	.660	No DIF
ITEM041	9.120	2	.010	DIF
ITEM042	0.90	2	.637	No DIF
ITEM043	0.75	2	.687	No DIF

ISSN: 2790-4407 www. jocatia.acata.org

Item	X^2	df	p-value	Decision
ITEM044	1.04	2	.595	No DIF
ITEM045	0.88	2	.646	No DIF
ITEM046	1.13	2	.569	No DIF
ITEM047	0.93	2	.628	No DIF
ITEM049	1.21	2	.546	No DIF
ITEM050	0.97	2	.616	No DIF
ITEM051	1.09	2	.580	No DIF
ITEM052	0.85	2	.654	No DIF
ITEM053	14.762	2	.001	DIF
ITEM054	1.17	2	.558	No DIF
ITEM055	21.409	2	< .05	DIF
ITEM056	17.833	2	< .05	DIF
ITEM058	11.624	2	.003	DIF
ITEM060	0.82	2	.663	No DIF
ITEM061	1.06	2	.588	No DIF
ITEM062	0.87	2	.648	No DIF
ITEM063	1.19	2	.552	No DIF
ITEM064	0.80	2	.671	No DIF
ITEM065	0.96	2	.618	No DIF
ITEM066	1.22	2	.543	No DIF
ITEM067	0.91	2	.635	No DIF
ITEM068	0.78	2	.676	No DIF
ITEM069	1.03	2	.597	No DIF
ITEM070	0.84	2	.657	No DIF
ITEM071	0.88	2	.646	No DIF
ITEM072	8.943	2	.012	DIF
ITEM073	1.10	2	.575	No DIF
ITEM074	0.95	2	.622	No DIF
ITEM075	1.14	2	.566	No DIF
ITEM076	0.79	2	.674	No DIF
ITEM078	13.506	2	.001	DIF

Item	X²	df	p-value	Decision
ITEM079	19.271	2	< .05	DIF
ITEM080	16.388	2	< .05	DIF
ITEM082–ITEM150	χ^2 < critical	2	p > .05	No DIF

Measurement Invariance Across School Type

Measurement invariance across school type (public and private) was assessed using Item Response Theory likelihood-ratio Differential Item Functioning (IRT-LR DIF) analysis under the multidimensional three-parameter logistic (M3PL) model. As presented in Table 6, 4 of the 126 retained items exhibited statistically significant Differential Item Functioning (DIF) ($p < .05$), indicating that these items functioned differently between public and private school students after controlling for ability. The remaining 122 items showed no statistically significant DIF ($p > .05$), indicating measurement invariance across school type. Consequently, the four DIF-affected items were excluded from the final version of the Chemistry Achievement Test. Following the removal of DIF-affected items identified across both sex and school type, the final instrument comprised 112 invariant items.

Table 6: IRT Likelihood-Ratio Differential Item Functioning Analysis by School Type (n = 126)

Item	X²	df	p-value	Decision
ITEM001–ITEM073	χ^2 < critical	2	p > .05	No DIF
ITEM074	12.456	2	.002	DIF
ITEM075–ITEM088	χ^2 < critical	2	p > .05	No DIF
ITEM089	9.873	2	.007	DIF
ITEM090–ITEM093	χ^2 < critical	2	p > .05	No DIF
ITEM094	15.392	2	0.001	DIF
ITEM095–ITEM122	χ^2 < critical	2	p > .05	No DIF
ITEM123	10.641	2	.005	DIF
ITEM124–ITEM150	χ^2 < critical	2	p > .05	No DIF

ISSN: 2790-4407 www. jocatia.acata.org

DISCUSSION

This study was designed to provide a rigorous and integrated psychometric evaluation of a Chemistry Achievement Test (CAT) using a multidimensional Item Response Theory (MIRT) framework. By systematically combining dimensionality assessment, model–data fit evaluation, local independence diagnostics, item parameter screening, and Differential Item Functioning (DIF) analysis, the study responds to longstanding concerns about fragmented validation practices in educational measurement and offers a more coherent basis for interpreting assessment outcomes (Embretson & Reise, 2000; De Ayala, 2009 & Wells, 2014). The findings not only shed light on the structure of chemistry achievement but also provide important insights into the appropriateness of multidimensional modelling and the conditions required for ensuring fair and reliable measurement.

The DETECT analysis provided strong evidence that the Chemistry Achievement Test (CAT) exhibits essential multidimensionality, characterised by a dominant general factor alongside three meaningful secondary dimensions. This finding suggests that, although a general Chemistry ability accounted for a substantial proportion of students' performance, additional latent traits also contributed significantly to their responses. The result is consistent with previous studies demonstrating that science achievement tests often measure multiple interrelated cognitive abilities rather than a single latent construct (Reckase, 2009). It also supports the DETECT framework proposed by Stout et al. (1996) and Zhang and Stout (1999), which recognises the coexistence of a dominant dimension with meaningful secondary dimensions in educational assessments. The findings therefore reinforce concerns that the routine application of strictly unidimensional models to complex achievement tests may oversimplify the measurement process and underestimate important domain-specific competencies (Reckase, 2009; Finch & French, 2019). Although the multidimensional structure observed in the present study is consistent with the cognitive complexity of Chemistry, aspects of test design, including content organisation and item clustering, may also contribute to the dimensional structure. Future test development should therefore ensure closer alignment between construct definition, test blueprint, and item design to strengthen construct validity and improve measurement precision (De Ayala, 2009).

The comparison of competing multidimensional IRT models demonstrated that the multidimensional three-parameter logistic (M3PL) model provided the best representation of the Chemistry Achievement Test data. The progressive improvement in model fit from the M1PL through the M2PL to the M3PL indicates that both item discrimination and pseudo-guessing contributed meaningfully to explaining students' response patterns. This finding is consistent with previous studies that have shown the M3PL model to provide more accurate parameter estimates for multiple-choice assessments in which guessing behaviour cannot be ignored (DeMars, 2010; Chalmers, 2018). In contrast, the multidimensional four-parameter logistic (M4PL) model did not provide sufficient improvement in model fit to justify its additional complexity, supporting the principle of parsimony in psychometric modelling (Finch & French, 2019). These findings suggest that the M3PL model offers an appropriate balance between statistical accuracy, parameter stability, and interpretability for Chemistry achievement assessment. Consequently, educational testing programmes may benefit from adopting multidimensional three-parameter models when developing multiple-choice achievement tests intended to provide reliable and meaningful estimates of students' abilities (Iliescu et al., 2022).

The evaluation of local item independence provided further evidence supporting the adequacy of the selected multidimensional model. The consistently low residual correlations obtained from Yen's Q3 statistic indicate that the latent structure adequately accounted for the shared variance among the test items, leaving little unexplained local dependence. This finding is consistent with previous studies that have shown low Q3 residual correlations to indicate satisfactory model specification and reliable item parameter estimation (Christensen et al., 2017; Finch & French, 2019). Consequently, the Chemistry Achievement Test (CAT) can be considered largely free from substantial local item dependence, supporting the validity of the estimated item and ability parameters. Although the local independence assumption was satisfied, future studies should complement Q3 diagnostics with alternative approaches, such as testlet response theory or bifactor models, particularly when evaluating assessments containing context-dependent or clustered items (Chalmers, 2018; Iliescu et al., 2022).

The analysis of item parameter quality demonstrated that 84% of the Chemistry Achievement Test (CAT) items met the established psychometric criteria, indicating that the instrument possesses strong measurement properties. The retained items exhibited satisfactory discrimination, appropriate difficulty levels, and acceptable pseudo-guessing parameters, suggesting that they effectively differentiated students across varying ability levels and contributed to precise ability estimation. This finding is consistent with previous studies, which have emphasised that well-functioning item parameters are fundamental to improving the reliability and validity of educational assessments

ISSN: 2790-4407 www. jocatia.acata.org

(Finch & French, 2019; Iliescu et al., 2022). The removal of poorly performing items further enhanced the psychometric quality of the instrument by retaining only items that adequately measured the intended construct. However, psychometric refinement should not rely exclusively on statistical indicators, as excessive item elimination may reduce representation of important curriculum content and, consequently, weaken content validity. Future test development should therefore balance statistical optimisation with comprehensive coverage of the intended learning outcomes to ensure that assessment instruments remain both psychometrically robust and educationally meaningful (De Ayala, 2009).

The analysis of Differential Item Functioning (DIF) revealed that several Chemistry Achievement Test (CAT) items functioned differently across sex and school type, indicating construct-irrelevant influences on students' responses. This finding is consistent with previous studies demonstrating that variations in instructional experiences, learning opportunities, and contextual familiarity may introduce bias into educational assessments if left unaddressed (Zumbo, 1999; Iliescu et al., 2022). Although DIF identification in the present study was based primarily on statistical significance, future studies should complement significance testing with effect-size indices to evaluate the practical magnitude of DIF and its potential impact on item fairness. The removal of DIF-affected items resulted in a final set of invariant items, thereby improving the fairness and validity of score interpretation across different groups. These findings reinforce current recommendations that achievement tests should minimise construct-irrelevant variance to ensure equitable assessment outcomes (Finch & French, 2019). Nevertheless, eliminating DIF-affected items may reduce coverage of some aspects of the curriculum. Future research should therefore examine alternative approaches, including DIF-adjusted scoring and multidimensional DIF models, to preserve content representation while maintaining measurement fairness (Chalmers, 2018; Penfield & Camilli, 2007).

Beyond its empirical findings, this study makes an important methodological contribution by demonstrating the value of an integrated psychometric validation framework for educational assessment. The combined application of dimensionality assessment, multidimensional Item Response Theory (MIRT) modelling, local independence diagnostics, item parameter evaluation, and Differential Item Functioning (DIF) analysis provided a comprehensive evaluation of the structural validity and fairness of the Chemistry Achievement Test (CAT). This integrated approach is consistent with contemporary recommendations that educational assessments should be supported by multiple sources of psychometric evidence to strengthen the validity and interpretability of test scores (Finch & French, 2019; Iliescu et al., 2022). Furthermore, the findings contribute to the growing application of multidimensional IRT in science education by demonstrating its suitability for evaluating achievement tests that measure multiple cognitive processes simultaneously (Chalmers, 2018; De Ayala, 2009). Consequently, the framework adopted in this study provides a practical guide for researchers and test developers seeking to design assessment instruments that are psychometrically sound, valid and equitable.

Despite these contributions, the findings should be interpreted considering several limitations. First, the study was conducted using a specific sample of secondary school students in Oyo State, Nigeria, which may limit the generalisability of the findings to other educational contexts and populations. Replication of the study with more diverse samples would provide stronger evidence of the Chemistry Achievement Test (CAT) 's broader applicability. Second, the study focused primarily on internal psychometric evidence and did not examine external validity. Future studies should therefore investigate the relationship between CAT scores and external indicators of Chemistry achievement to provide additional validity evidence. Furthermore, the cross-sectional design did not account for changes in students' abilities over time or the influence of authentic testing conditions, such as motivation and examination pressure. Finally, although the integrated validation framework adopted in this study requires considerable statistical expertise and relatively large samples, it offers a comprehensive approach to instrument validation. Future research should explore strategies for adapting this framework to different educational settings while maintaining its psychometric rigour. Despite these limitations, the study provides valuable evidence supporting the use of integrated multidimensional psychometric approaches for developing valid, reliable, and equitable achievement tests in science education.

RECOMMENDATIONS

Based on the findings of this study, the following recommendations are made:

i. Educational researchers and test developers should adopt multidimensional Item Response Theory (MIRT) models when developing achievement tests for cognitively complex subjects such as Chemistry, as these models provide more accurate representations of students' abilities than strictly unidimensional models.

ISSN: 2790-4407 www. jocatia.acata.org

ii. Dimensionality assessment should become a routine component of instrument development. Procedures such as DETECT should be used to determine the dimensional structure of test data before selecting an IRT model.

iii. Item parameter evaluation should be incorporated into all test validation studies to ensure that only items with acceptable discrimination, difficulty, and pseudo-guessing parameters are retained for operational use.

iv. Local item independence diagnostics should be routinely conducted during test development to identify residual dependencies that may affect model fit and the accuracy of parameter estimation.

v. Differential Item Functioning (DIF) analysis should form an integral part of validation procedures, and both statistical significance and effect-size indices should be considered when evaluating item fairness across groups defined by characteristics such as sex, school type, or other relevant variables.

vi. Test developers should balance psychometric optimisation with adequate curriculum representation to ensure that improvements in statistical quality do not compromise content validity.

vii. Future validation studies should evaluate the CAT using samples drawn from different educational systems, geographical locations, and cultural contexts to provide additional evidence for the generalisability of the findings

viii. Future research should examine additional sources of validity evidence by investigating the relationship between CAT scores and external measures of Chemistry achievement, including classroom performance and standardised examinations.

ix. Researchers should investigate the applicability of the validated CAT within computerised adaptive testing (CAT) environments, where multidimensional modelling may improve measurement precision and testing efficiency.

CONCLUSION

This study provided a comprehensive psychometric evaluation of the Chemistry Achievement Test (CAT) using a multidimensional Item Response Theory (MIRT) framework. The findings demonstrated that the CAT exhibits essential multidimensionality, characterised by a dominant general factor alongside meaningful secondary dimensions, confirming that Chemistry achievement is better represented as a multidimensional construct than by a strictly unidimensional model.

The multidimensional three-parameter logistic (M3PL) model emerged as the most appropriate representation of the response data, providing the best balance between model fit, parameter stability, and interpretability. The study also established that the selected model satisfied the assumption of local item independence and that most of the test items (84%) demonstrated satisfactory psychometric properties. Furthermore, the identification and removal of items exhibiting Differential Item Functioning (DIF) across sex and school type strengthened the fairness, validity, and comparability of the final instrument.

The findings should, however, be interpreted within the context of the study's limitations. The research was conducted using secondary school students in Oyo State, Nigeria, and focused primarily on internal psychometric evidence. Consequently, the generalisability of the findings to other educational contexts and populations should be made with caution, and additional evidence from external validation studies would further strengthen the applicability of the instrument.

Overall, the findings provide strong evidence that the CAT is a valid, reliable, and equitable instrument for assessing Chemistry achievement. Beyond the instrument itself, the study demonstrates the value of integrating dimensionality assessment, multidimensional modelling, local independence diagnostics, item parameter evaluation, and measurement invariance testing within a unified validation framework. This integrated approach advances contemporary psychometric practice by providing a rigorous methodology for developing and validating educational assessments capable of capturing the complexity of students' learning outcomes.

Acknowledgment

The author sincerely appreciates Professor Modupe M. Osokoya for her invaluable guidance, constructive criticism, and continuous support throughout this study. Appreciation is also extended to the participating secondary schools, teachers, and students whose cooperation made this research possible.

Conflicts of Interest

The author declares no conflict of interest.

REFERENCES

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37–51. https://doi.org/10.1111/j.1745-3992.2003.tb00136.x

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Bichi, A. A., & Talib, R. (2018). Item response theory: An introduction to latent trait models to test and item development. *International Journal of Evaluation and Research in Education, 7*(2), 142–151.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Rasch Analysis in the Human Sciences.

Jang, E. E., & Roussos, L. A. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 31*(4), 287–310.

Chalmers, R. P. (2018). Improving the Crossing-SIBTEST statistic for detecting non-uniform differential item functioning. *Psychometrika, 83*(2), 376–386.

Chandio, M. T., Pandhiani, S. M., & Iqbal, R. (2016). Bloom's taxonomy: Improving assessment and teaching-learning process. *Journal of Education and Educational Development, 3*(2), 203–221.

Chen, W. H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265–289.

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement, 41*(3), 178–194. https://doi.org/10.1177/0146621616677520

Cunha, W., Viegas, F., França, C., Rosa, T., Rocha, L., & Gonçalves, M. A. (2023). A comparative survey of instance selection methods applied to non-neural and transformer-based text classification. *ACM Computing Surveys, 55*(13s), 1–52. https://doi.org/10.1145/3582000

De Ayala, R. J. (2009). The theory and practice of item response theory. Guilford Press.

DeMars, C. (2010). Item response theory. Oxford University Press.Dogra, V., Verma, S., Kavita, Chatterjee, P., Shafi, J., Choi, J., & Ijaz, M. F. (2022). A complete process of text classification system using state-of-the-art NLP models. *Computational Intelligence and Neuroscience, 2022*, 1883698. https://doi.org/10.1155/2022/1883698

Du, X., Yang, J., Hung, J.-L., & Shelton, B. (2020). Educational data mining: A systematic review of research and emerging trends. *Information Discovery and Delivery, 48*(4), 225–236.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Lawrence Erlbaum Associates.

Finch, W. H., & French, B. F. (2019). *Educational and psychological measurement*. Routledge. https://doi.org/10.4324/9781315650951

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Sage Publications.

Madhavi, A., Nagesh, A., & Govardhan, A. (2023). Efficient course recommendation using deep transformer-based ensembled attention model. *EAI Endorsed Transactions on E-Learning, 9*

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5–8. https://doi.org/10.1111/j.1745-3992.1995.tb00881.x

Oladele, J. I., & Ndlovu, M. (2021). A review of standardised assessment development procedure and algorithms for computer adaptive testing: Applications and relevance for fourth industrial revolution. *International Journal of Learning, Teaching and Educational Research*, *20*(5), 1-17. https://doi.org/10.26803/ijlter.20.5.1

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 125–167). Elsevier. https://doi.org/10.1016/S0169-7161(06)26005-X

Reckase, M. D. (2009). Multidimensional item response theory. Springer.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*(3), 207–230

Rios, J. A., & Wells, C. S. (2014). Validity evidence based on internal structure. *Psicothema, 26*(1), 108–116.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589–617. https://doi.org/10.1007/BF02294821

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*(4), 331–354.

Sung, C., Dhamecha, T. I., & Mukhi, N. (2019). Improving short answer grading using transformer-based pre-training. In *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 469–481)

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 67–113). Lawrence Erlbaum Associates.

ISSN: 2790-4407 www. jocatia.acata.org

Weibelzahl, S., Paramythis, A., & Masthoff, J. (2020). Evaluation of adaptive systems. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 394–395).

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44*(1), 1–27.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125–145. https://doi.org/10.1177/014662168400800201

Zaidi, N. L. B., Grob, K. L., Monrad, S. M., Kurtz, J. B., Tai, A., Ahmed, A. Z., Gruppen, L. D., & Santen, S. A. (2018). Pushing critical thinking skills with multiple-choice questions: Does Bloom's taxonomy work? *Academic Medicine, 93*(6), 856–859.

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate a simple structure. Psychometrika, 64(2), 213–249.

Zhang, J. (2007). Conditional covariance-based nonparametric multidimensionality assessment. *Psychometrika, 72*(2), 181–198.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF). National Defense Headquarters.