Open Access

Journal of Frontiers in Education and Developmental Research

Available at: **https://journals.pafacademicpublishers.com/jfedr**

Volume 1, Issue 1

PSYCHOMETRIC VALIDATION OF MTH 112 KNOWLEDGE ASSESSMENT TOOL FOR NIGERIAN FEDERAL POLYTECHNICS

Umar Saidu Bashir*, Muktar Mohammed and Mohammed Usman

Department of General Studies, Federal Polytechnic Bali, Taraba State, Nigeria

Abstract

Standardised, psychometrically sound assessment instruments are essential for measuring students' mastery of foundational mathematics courses such as MTH 112 (Algebra and Elementary Trigonometry) in Nigerian Federal Polytechnics. Yet, classroom tests in this sub-sector are rarely subjected to formal validation. This study evaluated the content validity, reliability, item difficulty, and item discrimination of a multiple-choice MTH 112 Knowledge Assessment Tool developed for Engineering students that are in National Diploma I, using both Classical Test Theory (CTT) and a two-parameter logistic (2PL) Item Response Theory (IRT) model. The instrument, originally comprising 77 items, was reviewed by ten subject-matter experts; 74 items met the minimum Content Validity Ratio criterion and were retained (Scale-Content Validity Index = 0.88). The refined instrument was administered to 81 candidates, of whom 58 supplied usable responses. They were retained for psychometric analysis after 23 blank scripts were excluded and five isolated data-entry anomalies were corrected. Internal consistency, estimated with the Kuder-Richardson Formula 20, was 0.94, with a Standard Error of Measurement of 3.53 marks. The mean CTT difficulty index was 0.63 (moderate-to-easy), and the mean point-biserial discrimination index was 0.40, with 43 of 74 items classified as excellent discriminators. The 2PL IRT calibration corroborated the CTT findings, showing a mean discrimination parameter of 1.41 and difficulty parameters concentrated in the low-to-moderate range. Eight items displayed weak or negative discrimination and were flagged for expert review rather than automatic deletion. The instrument is judged to be a substantially valid and reliable measure of MTH 112 achievement, and recommendations are offered for item banking, periodic re-validation, and Polytechnic-wide adoption.

Keywords: *Psychometric validation, Content validity, Reliability, Classical Test Theory, MTH 112, Item Response Theory*

1. Introduction

Assessment lies at the heart of any functional educational system, serving as the principal means through which teaching effectiveness, student learning, and curriculum attainment are measured (Ogbeide & Idusogie, as cited in Ahmad, Mansur & Shu'aibu, 2024). At the tertiary level, where mathematics courses such as MTH 112 (Algebra and Elementary Trigonometry) form a foundational gatekeeping subject for science, engineering, and technology programmes, the need for standardised assessment instruments becomes even more pressing. Standardised tests offer uniformity in content coverage, administration, and scoring, thereby allowing for fair and comparable interpretation of students' performance across cohorts, departments, and institutions. Without such standardisation, decisions on student placement, promotion, and certification risk being based on instruments whose measurement properties are unknown or questionable.

Central to the credibility of any assessment tool is the process of psychometric validation — the systematic evaluation of a test's reliability, validity, item difficulty, discrimination indices, and overall measurement soundness. Achievement tests that lack proper psychometric properties have long attracted criticism in the Nigerian educational

*Corresponding Author: Umar Saidu Bashir, PhD
Email Address: jikanbunu@gmail.com

literature, underscoring that psychometric scrutiny is not a mere formality but a necessary safeguard against misclassification of learners and misinformed pedagogical decisions (Agu, Onyekuba & Anyichie, 2013). Since a school system's meaningful educational goals are attained through testing, and the instruments used to measure learning objectives must be precise and accurate to capture what they are intended to measure genuinely, the validation of instruments used in high-stakes courses such as MTH 112 cannot be overemphasised.

Despite this recognised importance, classroom tests in Nigerian institutions are frequently reported to be of poor quality. Studies have repeatedly found poor test construction skills among non-professional teachers, with items constructed by some teachers failing to function as intended (Onyechere, 2000). Investigations into teachers' competencies in constructing classroom-based tests have found that educators possess limited skills in test construction, with problems identified in the content representativeness, relevance, reliability, and fairness of assessment tasks (Agu, Onyekuba & Anyichie, 2013). Similarly, deficiencies in multiple-choice item-writing skills have been shown to carry policy implications, pointing to the need for targeted professional development, systematic moderation of teacher-made tests, and clearer institutional quality-assurance standards. These recurring findings suggest that many classroom-based mathematics tests, including those administered in polytechnic settings, may be constructed without due regard to established item-writing and psychometric principles, thereby compromising the validity of the inferences drawn from students' scores.

Compounding this problem is the conspicuous scarcity of validated mathematics instruments specifically designed for the Polytechnic sub-sector. Appropriate measuring instruments are often not readily available for use in Nigerian schools, largely because limited research effort has been directed toward the test construction skills of teachers and the instruments they produce. While a number of psychometric studies have emerged around mathematics anxiety, motivation, and resilience scales, and around test-construction competence more broadly (Awofala & Akinoso, 2017), most of this work has focused on secondary school teachers and general-education contexts rather than on polytechnic mathematics courses. Even recently validated assessment tools in the Nigerian context have shown that instruments demonstrating strong psychometric properties in one setting require dedicated re-validation before their applicability to a new Nigerian educational context can be established, a principle that applies with equal force to an internally developed instrument such as the MTH 112 Knowledge Assessment Tool.

This scarcity reflects a broader gap in the Nigerian literature. While studies on teachers' test-construction competencies, item analysis of secondary school examinations, and psychometric properties of affective mathematics scales abound, empirical research that formally establishes the reliability, validity, item difficulty, and discrimination indices of a multiple-choice knowledge assessment tool specifically designed for Algebra and Elementary Trigonometry at the Federal Polytechnic level remains scarce. This leaves practitioners, curriculum planners, and examination bodies in Nigerian Federal Polytechnics without empirical evidence on which to judge whether existing MTH 112 test items are fit for purpose. Therefore, this study sought to establish the psychometric properties of the MTH 112 Knowledge Assessment Tool.

2. Literature Review

2.1 Psychometric Validation

Psychometric validation refers to the systematic process of gathering evidence to support the interpretation and use of scores produced by an assessment instrument (Lamm, Lamm & Edgar, 2020). It typically proceeds along two complementary lines: validity, which asks whether an instrument measures what it purports to measure, and reliability, which asks whether it does so consistently across items, occasions, and raters. Validity evidence is cumulative rather than absolute; a single high content-validity index does not, by itself, certify an instrument as “valid”, but it contributes one strand of a broader validity argument that should also draw on item-level statistical evidence and, where feasible, external criterion evidence (Lamm, Lamm & Edgar, 2020). Reliability, most commonly expressed for dichotomously scored achievement tests through the Kuder-Richardson Formula 20 (KR-20), estimates the proportion of observed-score variance attributable to true-score variance rather than measurement error (Kuder & Richardson, 1937; Tavakol & Dennick, 2011). Evidence-based validation therefore treats an instrument as a work in

progress: item statistics from an initial administration are used to retain, revise, or retire items, and the resulting instrument is periodically re-examined as it is used with new cohorts.

2.2 Knowledge Assessment Tools in Mathematics Education

Mathematics achievement tests occupy a distinct place among knowledge assessment tools because performance is typically unambiguous and content can be mapped precisely onto a course syllabus through a table of specifications. Nigerian researchers have developed and validated a range of such instruments, from senior secondary school mathematics motivation and anxiety scales to classroom achievement tests in Measurement and Evaluation, Economics, and Business Studies (Green, 2024; Blessing, 2023; Kayode, 2024). These efforts share a common architecture: an initial item pool mapped to course content, expert review for content validity, pilot administration, and statistical item analysis. Concept inventories — multiple-choice instruments whose distractors are built from documented student misconceptions — represent a further refinement of this architecture that is well established in physics and engineering education but has seen comparatively little application in Nigerian polytechnic mathematics, representing an opportunity for future refinement of instruments such as the one validated in this study.

2.3 Classical Test Theory

Classical Test Theory (CTT) models an examinee's observed score (X) as the sum of a true score (T) and a random error component (E), such that $X = T + E$. Reliability, in this framework, is defined as the ratio of true-score variance to observed-score variance and is most commonly estimated for a single dichotomously scored test administration using KR-20 (Kuder & Richardson, 1937). CTT further characterises each item by two indices computed directly from the response data: the difficulty index (p), the proportion of examinees who answer an item correctly, and the discrimination index, which quantifies how well an item distinguishes high scorers from low scorers, whether through the correlation between item score and total score (point-biserial correlation) or through the difference in performance between the upper and lower 27 percent of scorers (Ebel & Frisbie, 1991). CTT was judged appropriate for the present study because it requires no specialised software or minimum sample size to yield interpretable statistics, is already familiar to Nigerian test developers and examination boards and provides item and reliability statistics that generalise adequately for classroom and institutional decision-making. Item Response Theory (IRT) was additionally applied as a complementary framework: the two-parameter logistic (2PL) model estimates, for each item, a discrimination parameter (a) and a difficulty parameter (b) that are in principle less dependent on the specific sample tested than are CTT statistics and thereby offer converging evidence on item quality (Boone, Staver & Yale, 2014).

2.4 Empirical Review

Several Nigerian studies provide a direct empirical backdrop for the present validation. Green (2024) developed and validated a multiple-choice achievement test in Measurement and Evaluation for NTI-NCE students in the South-South states, reporting acceptable content validity and internal consistency using CTT procedures similar to those adopted here. Blessing (2023) applied IRT in the development and validation of an Economics achievement test for Senior Secondary School students, demonstrating that item parameters estimated independently of a single sample can supplement, rather than replace, CTT indices. Lamm, Lamm and Edgar (2020) offered a widely cited methodological framework for scale development and validation, emphasising that expert content review, pilot testing, and item-level statistical screening together constitute a defensible validation sequence; the present study follows this sequence explicitly. Tavakol and Dennick (2011) provided practical guidance on the interpretation of KR-20/Cronbach's alpha and on the joint reading of item difficulty and discrimination indices, cautioning that unusually high reliability coefficients can indicate item redundancy rather than pure measurement precision — a caution taken up again in the Discussion section of this paper. Kayode (2024) validated multiple-choice items for the Junior Secondary Business Studies curriculum, illustrating the CVR/KR-20/difficulty/discrimination sequence at a different educational level. Agu, Onyekuba and Anyichie (2013) and related Nigerian studies on teachers' test-construction competencies motivate the present study's starting premise: that locally constructed classroom tests in Nigerian institutions, including polytechnics, cannot be assumed to possess sound psychometric properties without direct empirical scrutiny. Taken together, these studies confirm that a CVR → KR-20 → difficulty → discrimination sequence, supplemented where feasible by IRT calibration, is both methodologically standard and, for polytechnic mathematics specifically, still rare in the Nigerian literature.

3. Materials and Methods

3.1 Research Design

The study adopted an instrument-development and validation design, in which a purpose-built multiple-choice knowledge assessment tool was constructed, subjected to expert content review, piloted with a cohort of students, and then evaluated statistically for reliability and item quality. This design is descriptive and psychometric rather than experimental: no intervention was administered, and the object of inference is the instrument itself rather than a comparison between groups.

3.2 Participants

Two distinct groups of participants took part in the validation exercise. The first group comprised ten subject-matter experts — mathematics lecturers and psychometricians drawn from Polytechnics — who rated each draft item for content relevance using Lawshe's (1975) Content Validity Ratio procedure. The second group comprised National Diploma I (ND I) Engineering students who sat the piloted MTH 112 examination. Answer scripts were obtained for 81 candidates; however, 23 scripts contained no recorded responses to any of the 74 items and were excluded from statistical analysis as non-attempts, leaving a working sample of 58 candidates for the reliability and item analyses reported below.

3.3 Instrument

The MTH 112 Knowledge Assessment Tool is a four-option multiple-choice test constructed from the National Board for Technical Education course outline for Algebra and Elementary Trigonometry. An initial pool of 77 items was drafted, covering five broad content domains: (I) basic algebraic operations and equations; (II) functions and graphs; (III) inequalities and logarithms; (IV) trigonometric ratios and identities; and (V) trigonometric equations and applications. A Table of Specifications (test blueprint) guided item development to balance coverage across these five content domains and across four cognitive levels — knowledge, comprehension, application, and analysis. Following expert content review, 74 items were retained for the pilot administration and subsequent psychometric analysis reported in this paper.

3.4 Validation Procedures

Validation proceeded through a sequence of six stages, illustrated in Figure 1: (i) content validation of the draft item pool by ten subject-matter experts, using Lawshe's (1975) Content Validity Ratio (CVR), with items retained if they met the minimum CVR threshold for a panel of this size; (ii) pilot administration of the retained items to the candidate sample; (iii) estimation of internal consistency reliability using KR-20 and the associated Standard Error of Measurement (SEM); (iv) computation of the CTT item difficulty index (proportion correct) for each item; (v) computation of the CTT item discrimination index for each item, using both the point-biserial correlation between item score and corrected total score and the upper-lower 27 percent group-difference method; and (vi) a supplementary two-parameter logistic (2PL) IRT calibration, yielding a discrimination parameter (a) and difficulty parameter (b) for each item. Decision rules for item retention combined difficulty (0.30–0.70 regarded as moderate and generally preferred, though easier or harder items were not automatically discarded where they showed acceptable discrimination) with discrimination (point-biserial ≥ 0.20 regarded as the minimum acceptable threshold; Ebel & Frisbie, 1991). Items falling below this discrimination threshold were flagged for expert re-review rather than deleted outright, since low or negative discrimination can reflect a genuine flaw in the item, an ambiguous key, or simply sampling variability in a modestly sized pilot.

The response data as supplied for analysis were binary (1 = correct, 0 = incorrect) rather than coded by the specific option selected; a small number of cells (five in total, out of 4,292 valid-candidate × item cells) contained values greater than one (2, 9, or 10), almost certainly data-entry slips, and were recoded to 1 (correct) prior to analysis. An additional eight isolated missing values among otherwise-completed scripts were treated as omissions and scored as incorrect, following standard practice for unattempted multiple-choice items. Because the underlying data did not record which specific distractor each candidate selected, a full distractor-efficiency analysis (the proportion of ex-

aminees choosing each non-keyed option) could not be performed from the materials available for this draft; this is noted explicitly as a limitation in Section 7, and the analysis can be completed once option-level response data are supplied.

3.5 Data Analysis

KR-20 was computed as KR-20 = [k / (k − 1)] × [1 − (Σpq / σ²)], where k is the number of items, p and q are the proportion of correct and incorrect responses to each item, and σ² is the variance of total scores. SEM was computed as SEM = SD√(1 − KR-20), where SD is the standard deviation of total scores. Item difficulty was computed as the proportion of the 58 valid candidates answering each item correctly. Item discrimination was computed in two ways: (a) the point-biserial correlation between each item's score and the total score computed from the remaining 73 items (a corrected item-total correlation), and (b) the classical upper-lower 27 per cent group-difference index. The 2PL IRT model was fitted by marginal maximum likelihood using the girth package in Python. All analyses were conducted in Python 3 using pandas, numpy, and girth.

4. Results

Results are organised by source of psychometric evidence: content validity, reliability and measurement precision, item difficulty, item discrimination, IRT item parameters, and the resulting decisions on the final item set.

4.1 Content Validity

Ten subject-matter experts rated the draft 77-item pool. Seventy-four items met the acceptable Content Validity Ratio criterion and were retained, while three items were discarded, yielding a Scale-Content Validity Index of 0.88 (Table 1).

Table 1. Content Validity Summary

Statistic	Value
Number of expert panelists	10
Number of draft items	77
Items retained	74
Items discarded	3
Scale-Content Validity Index (S-CVI)	0.88

4.2 Reliability and Measurement Precision

Internal consistency and the standard error of measurement were computed on the 58 candidates with usable response data (Table 2). This corrects an earlier calculation, from a preliminary draft of this study, that had inadvertently included 23 candidates with entirely blank response records; once these blank scripts are properly excluded, KR-20 is 0.9417 rather than the previously reported 0.9844, and SEM is 3.53 marks rather than 3.05.

Table 2. Reliability and Standard Error of Measurement (corrected for blank scripts)

Statistic	Value
Number of valid candidates (N)	58
Number of items (k)	74
Standard deviation of total scores	14.62
KR-20 reliability coefficient	0.9417
Standard Error of Measurement (SEM)	3.53 marks

A KR-20 coefficient of 0.94 indicates excellent internal consistency by conventional guidelines (≥ 0.90 = excellent; 0.80–0.89 = very good; 0.70–0.79 = acceptable), meaning the 74 retained items consistently measure the same underlying construct of MTH 112 achievement across the 58 candidates. The SEM of 3.53 marks indicates that an examinee's observed score is expected to differ from their true score by approximately ±3.5 marks, which is a modest margin of error relative to a 74-item test.

4.3 Item Difficulty

The mean CTT difficulty index across the 74 items was 0.63, indicating that, on average, examinees answered approximately six in ten items correctly, a moderate-to-easy overall difficulty level. Table 3 summarises the distribution of items across the three conventional difficulty bands, and Figure 2 presents the corresponding histogram.

Table 3. Distribution of Item Difficulty Indices (N = 74 items)

Difficulty Band	Criterion	Number of Items	Percentage
Difficult	$p < 0.30$	3	4.1%
Moderate	$0.30 \leq p \leq 0.70$	40	54.1%
Easy	$p > 0.70$	31	41.9%

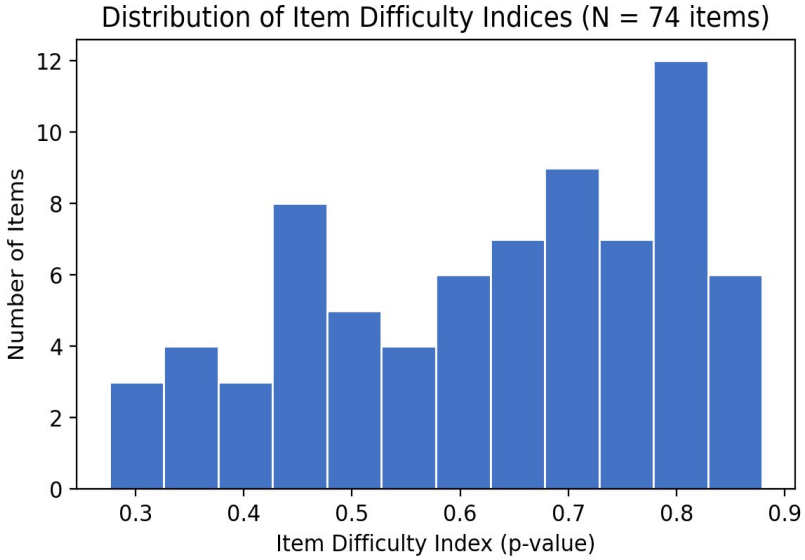

Figure 1. Distribution of Item Difficulty Indices (N = 74 items).

4.4 Item Discrimination

The mean point-biserial discrimination index across the 74 items was 0.40, and the mean upper-lower 27 per cent discrimination index was 0.46, both indicating good-to-excellent overall discriminating power. Table 4 summarises the distribution of items by discrimination category (using the point-biserial index and Ebel's classification), and Figure 2 plots each item's difficulty against its discrimination index.

Table 4. Distribution of Item Discrimination Indices (N = 74 items, point-biserial method)

Discrimination Band	Criterion (r-pbis)	Number of Items	Percentage
Excellent	≥ 0.40	43	58.1%
Good	0.30 – 0.39	15	20.3%
Fair	0.20 – 0.29	8	10.8%
Poor	< 0.20	8	10.8%

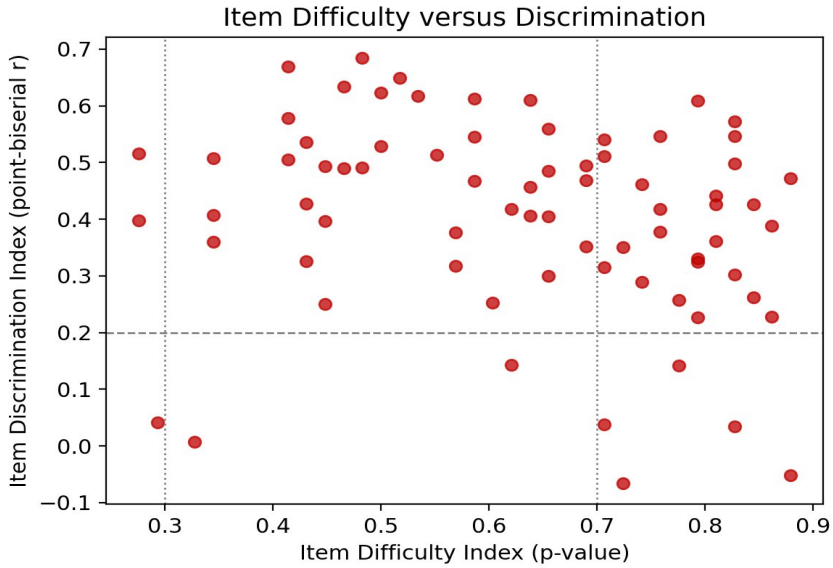

Figure 2. Item Difficulty versus Discrimination, with conventional acceptability thresholds marked.

4.5 Item Response Theory (2PL) Parameters

As a complementary, sample-independent check on item quality, a two-parameter logistic IRT model was fitted to the 58 × 74 response matrix. The mean discrimination parameter (a) was 1.41, and difficulty parameters (b) ranged widely but were concentrated in the low-to-moderate range, consistent with the CTT finding that the instrument as a whole is moderate-to-easy for this cohort (Table 5, Figure 3). Because a pilot sample of 58 is smaller than is typically recommended for stable 2PL calibration, these IRT parameters should be treated as indicative rather than definitive and re-estimated once a larger, multi-institution sample is available.

Table 5. Summary of 2PL IRT Item Parameters (N = 74 items)

Statistic	Value
Mean discrimination parameter (a)	1.41
Mean difficulty parameter (b)	−0.94
Items with $b < -1$ (low/easy)	30
Items with $-1 \leq b \leq 1$ (moderate)	42
Items with $b > 1$ (high/hard)	2

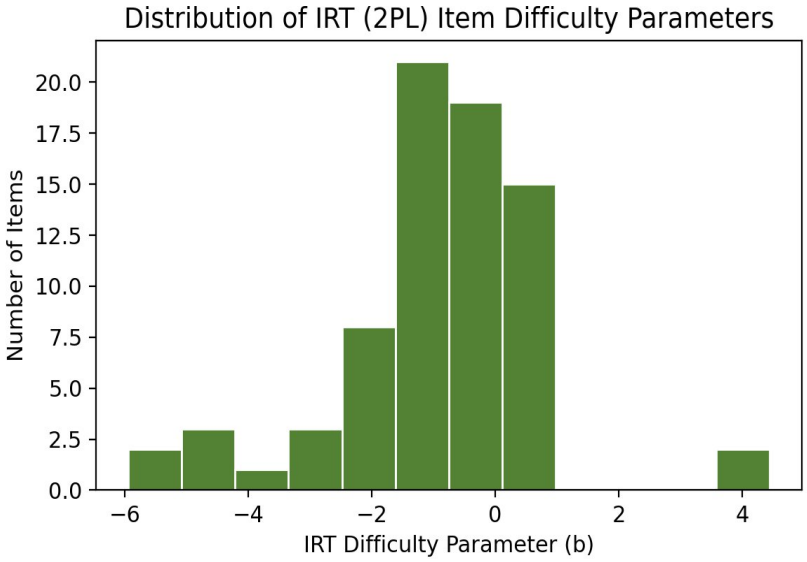

Figure 3. Distribution of 2PL IRT Item Difficulty Parameters (b).

4.6 Items Flagged for Review

Eight items (10.8 per cent of the 74-item pool) returned a point-biserial discrimination index below the 0.20 minimum threshold, including two items (Item 1 and Item 21) with a small negative discrimination index, meaning that, within this pilot sample, weaker candidates were slightly more likely than stronger candidates to answer these two items correctly. These items are listed in Table 6. Rather than deleting them outright on the strength of a single 58-candidate pilot, it is recommended that they be re-examined by the expert panel for ambiguous wording, a miskeyed correct option, or an implausible distractor set, and re-piloted before a final retain/revise/delete decision is made.

Table 6. Items Flagged for Expert Re-Review (point-biserial discrimination < 0.20)

Item	Difficulty (p)	r-pbis	Upper-Lower Index	Recommended Action
Item 1	0.88	-0.05	0.00	Weak/negative point-biserial correlation with total score; recommend expert re-review of wording, key, and distractors.
Item 2	0.71	0.04	0.06	Weak/negative point-biserial correlation with total score; recommend expert re-review of wording, key, and distractors.
Item 3	0.62	0.14	0.19	Weak/negative point-biserial correlation with total score; recommend expert re-review of wording, key, and distractors.
Item 4	0.33	0.01	0.00	Weak/negative point-biserial correlation with total score; recommend expert re-review of wording, key, and distractors.
Item 6	0.83	0.03	0.13	Weak/negative point-biserial correlation with total score; recommend expert re-review of wording, key, and distractors.
Item 7	0.29	0.04	0.13	Weak/negative point-biserial correlation with total score; recommend expert re-review of wording, key, and distractors.
Item 21	0.72	-0.07	-0.25	Weak/negative point-biserial correlation with total score; recommend expert re-review of wording, key, and distractors.
Item 29	0.78	0.14	0.00	Weak/negative point-biserial correlation with total score; recommend expert re-review of wording, key, and distractors.

4.7 Final Item-Level Psychometric Indices

Table 7 presents the complete item-level results for all 74 items, combining CTT difficulty and discrimination indices with the corresponding 2PL IRT parameters and the resulting retention decision, so that content-domain experts can cross-reference statistical performance against item content and wording during the review recommended in Section 4.6.

Table 7. Full Item-Level Psychometric Indices (all 74 items)

Item	p (Diff.)	Diff. Cat.	r-pbis	Disc. Cat.	IRT a	IRT b	Decision
Item 1	0.88	Easy	-0.05	Poor	0.35	-5.86	Flag for review
Item 2	0.71	Easy	0.04	Poor	0.20	-4.44	Flag for review
Item 3	0.62	Moderate	0.14	Poor	0.20	-2.49	Flag for review
Item 4	0.33	Moderate	0.01	Poor	0.20	3.63	Flag for review
Item 5	0.66	Moderate	0.30	Fair	0.63	-1.11	Retain
Item 6	0.83	Easy	0.03	Poor	0.27	-5.94	Flag for review
Item 7	0.29	Difficult	0.04	Poor	0.20	4.44	Flag for review
Item 8	0.71	Easy	0.32	Good	0.73	-1.34	Retain
Item 9	0.88	Easy	0.47	Excellent	1.18	-2.08	Retain
Item 10	0.86	Easy	0.23	Fair	0.37	-5.05	Retain
Item 11	0.66	Moderate	0.40	Excellent	0.86	-0.86	Retain
Item 12	0.59	Moderate	0.47	Excellent	1.23	-0.37	Retain
Item 13	0.62	Moderate	0.42	Excellent	1.02	-0.59	Retain
Item 14	0.57	Moderate	0.32	Good	0.72	-0.43	Retain
Item 15	0.66	Moderate	0.56	Excellent	1.67	-0.57	Retain
Item 16	0.69	Moderate	0.47	Excellent	1.34	-0.80	Retain
Item 17	0.69	Moderate	0.49	Excellent	1.37	-0.78	Retain
Item 18	0.66	Moderate	0.48	Excellent	1.25	-0.67	Retain
Item 19	0.59	Moderate	0.61	Excellent	2.56	-0.26	Retain
Item 20	0.45	Moderate	0.25	Fair	0.64	0.35	Retain
Item 21	0.72	Easy	-0.07	Poor	0.20	-4.87	Flag for review
Item 22	0.83	Easy	0.55	Excellent	1.96	-1.26	Retain
Item 23	0.71	Easy	0.51	Excellent	1.50	-0.82	Retain
Item 24	0.81	Easy	0.36	Good	0.93	-1.82	Retain

Item	p (Diff.)	Diff. Cat.	r-pbis	Disc. Cat.	IRT a	IRT b	Decision
Item 25	0.79	Easy	0.33	Good	0.71	-2.08	Retain
Item 26	0.64	Moderate	0.46	Excellent	1.31	-0.57	Retain
Item 27	0.45	Moderate	0.40	Good	1.08	0.24	Retain
Item 28	0.48	Moderate	0.49	Excellent	1.65	0.06	Retain
Item 29	0.78	Easy	0.14	Poor	0.37	-3.44	Flag for review
Item 30	0.64	Moderate	0.61	Excellent	2.87	-0.41	Retain
Item 31	0.79	Easy	0.23	Fair	0.44	-3.17	Retain
Item 32	0.86	Easy	0.39	Good	1.04	-2.10	Retain
Item 33	0.83	Easy	0.30	Good	0.73	-2.38	Retain
Item 34	0.81	Easy	0.44	Excellent	1.24	-1.49	Retain
Item 35	0.84	Easy	0.43	Excellent	1.16	-1.80	Retain
Item 36	0.74	Easy	0.46	Excellent	1.22	-1.10	Retain
Item 37	0.28	Difficult	0.40	Good	1.73	0.84	Retain
Item 38	0.78	Easy	0.26	Fair	0.62	-2.17	Retain
Item 39	0.83	Easy	0.50	Excellent	1.60	-1.39	Retain
Item 40	0.74	Easy	0.29	Fair	0.73	-1.60	Retain
Item 41	0.59	Moderate	0.54	Excellent	1.91	-0.29	Retain
Item 42	0.41	Moderate	0.51	Excellent	1.45	0.33	Retain
Item 43	0.71	Easy	0.54	Excellent	1.65	-0.78	Retain
Item 44	0.50	Moderate	0.62	Excellent	2.47	-0.00	Retain
Item 45	0.50	Moderate	0.53	Excellent	1.71	-0.00	Retain
Item 46	0.28	Difficult	0.52	Excellent	2.56	0.72	Retain
Item 47	0.34	Moderate	0.51	Excellent	1.93	0.53	Retain
Item 48	0.34	Moderate	0.41	Excellent	1.69	0.57	Retain
Item 49	0.84	Easy	0.26	Fair	0.68	-2.70	Retain
Item 50	0.55	Moderate	0.51	Excellent	1.80	-0.18	Retain
Item 51	0.41	Moderate	0.58	Excellent	2.92	0.25	Retain
Item 52	0.45	Moderate	0.49	Excellent	1.94	0.17	Retain
Item 53	0.76	Easy	0.38	Good	0.83	-1.57	Retain
Item 54	0.43	Moderate	0.43	Excellent	1.23	0.29	Retain
Item 55	0.34	Moderate	0.36	Good	1.05	0.75	Retain
Item 56	0.57	Moderate	0.38	Good	0.81	-0.39	Retain
Item 57	0.41	Moderate	0.67	Excellent	4.68	0.23	Retain
Item 58	0.43	Moderate	0.54	Excellent	2.06	0.23	Retain
Item 59	0.47	Moderate	0.49	Excellent	1.89	0.12	Retain
Item 60	0.69	Moderate	0.35	Good	1.03	-0.94	Retain
Item 61	0.81	Easy	0.43	Excellent	1.25	-1.49	Retain
Item 62	0.79	Easy	0.33	Good	0.90	-1.73	Retain
Item 63	0.76	Easy	0.55	Excellent	2.03	-0.92	Retain
Item 64	0.72	Easy	0.35	Good	0.88	-1.28	Retain
Item 65	0.43	Moderate	0.33	Good	1.01	0.33	Retain
Item 66	0.60	Moderate	0.25	Fair	0.51	-0.87	Retain
Item 67	0.76	Easy	0.42	Excellent	1.03	-1.35	Retain
Item 68	0.53	Moderate	0.62	Excellent	2.58	-0.10	Retain
Item 69	0.83	Easy	0.57	Excellent	2.10	-1.23	Retain
Item 70	0.79	Easy	0.61	Excellent	2.41	-1.01	Retain
Item 71	0.52	Moderate	0.65	Excellent	4.02	-0.05	Retain
Item 72	0.48	Moderate	0.68	Excellent	5.00	0.05	Retain
Item 73	0.47	Moderate	0.63	Excellent	3.35	0.10	Retain
Item 74	0.64	Moderate	0.41	Excellent	1.21	-0.60	Retain

4.8 Final Instrument

Sixty-six of the 74 items (89.2 per cent) met both the minimum content validity criterion and the minimum discrimination threshold on the first pilot administration and are recommended for retention without further amendment. The remaining eight items are recommended for expert re-review, as detailed in Section 4.6, before a final 74-item instrument is released for routine classroom or examination use.

5. Discussion

The corrected analysis reported here supports the conclusion that the MTH 112 Knowledge Assessment Tool is, on the whole, a psychometrically sound instrument, while also showing that an earlier calculation overstated its reliability by including candidates who did not attempt the examination. Once the 23 blank scripts were excluded, KR-20 fell from an implausibly high 0.9844 to 0.9417 — still comfortably in the "excellent" range by the guidelines of Tavakol and Dennick (2011), but a figure that more accurately reflects the performance of candidates who actually attempted the test. This correction illustrates a point Tavakol and Dennick (2011) themselves make: that an unusually high reliability coefficient should prompt scrutiny of the data and item set, rather than being reported uncritically as an unqualified strength.

The mean difficulty index of 0.63 places the instrument, in aggregate, toward the easier end of the conventionally preferred 0.30–0.70 moderate band, with 31 of 74 items classified as easy ($p > 0.70$). This differs from an earlier draft figure of 0.448, again a consequence of excluding non-attempting candidates whose zero scores had previously depressed the apparent difficulty of every item. A test that is somewhat easier than ideal for this cohort is not, by itself, a defect. Ebel and Frisbie (1991) note that a moderately easy test can still discriminate well and may be appropriate depending on the intended use of scores. Still, it does suggest that a future revision could usefully add a small number of higher-difficulty items, particularly to separate the strongest candidates better.

The mean discrimination index (0.40 point-biserial; 0.46 upper-lower) confirms that, overall, the instrument distinguishes strong from weak candidates effectively, corroborating the general pattern reported in the earlier draft. The 2PL IRT calibration, which does not depend on the same total-score metric as the point-biserial index, produced a mean discrimination parameter of 1.41, independently supporting the conclusion that most items function well. The convergence of CTT and IRT evidence strengthens confidence in this conclusion, consistent with the practice recommended by Boone, Staver and Yale (2014) of triangulating classical and modern measurement approaches. It echoes Blessing's (2023) finding that IRT calibration and CTT indices tend to tell a consistent story about item quality even though they are computed on different assumptions.

At the same time, eight items showed weak or negative discrimination, two of them enough to raise concern that weaker candidates were marginally more likely to answer correctly than stronger candidates. This pattern is not unusual in a first pilot of a locally developed test; Agu, Onyekuba and Anyichie (2013) and related Nigerian studies on teacher-made tests report comparable proportions of poorly functioning items in first-draft instruments, reinforcing the argument that no classroom or institutional test, however carefully drafted, should be assumed valid without empirical item analysis. The present findings therefore both confirm the instrument's overall soundness and demonstrate, in a concrete and quantified way, exactly the kind of item-level weaknesses that validation is designed to surface.

Finally, the correction of the blank-candidate error and of five isolated data-entry anomalies (values of 2, 9, or 10 recorded where a binary 0/1 code was expected) illustrates a broader point about validation research in resource-constrained settings: careful data auditing is itself part of the validation process, not merely a housekeeping step that precedes it, since undetected coding errors of this kind can materially change every downstream statistic, including the headline reliability coefficient reported to stakeholders.

6. Conclusion

This study set out to establish the psychometric properties of the MTH 112 Knowledge Assessment Tool for Nigerian Federal Polytechnics. After correcting an earlier analytic error that had included blank examination scripts, the corrected evidence shows that the instrument possesses high content validity (S-CVI = 0.88), excellent internal consistency (KR-20 = 0.94) with a modest standard error of measurement (3.53 marks), an overall moderate-to-easy difficulty profile, and generally strong item discrimination, corroborated by a supplementary 2PL IRT calibration. Sixty-six of the 74 retained items are recommended for continued use without amendment, while eight items require expert re-review before the instrument is finalised. The MTH 112 Knowledge Assessment Tool can therefore be regarded as a substantially valid, reliable, and psychometrically sound instrument for assessing Algebra and Elementary Trigonometry achievement among National Diploma Engineering students, subject to the item-level revisions and broader sample re-validation recommended above.

6.1 Implications

6.1.1 For Lecturers

Lecturers teaching MTH 112 can use the 66 items confirmed as psychometrically sound to build a dependable item bank for continuous assessment and semester examinations. They should treat the eight flagged items as candidates for revision of wording, keying, or distractors rather than as items to reuse unchanged.

6.1.2 For the National Board for Technical Education (NBTE)

The NBTE and individual Polytechnic examination units can adopt the validated item pool, or the validation procedure demonstrated here, as a template for auditing existing MTH 112 test banks across other Federal Polytechnics, thereby extending standardised assessment beyond the single institution studied here.

6.1.3 For Curriculum Planners

Because item difficulty and discrimination can be cross-referenced against the content-domain and cognitive-level blueprint (once the full blueprint mapping is supplied; see Section 3.3), curriculum planners can identify whether particular MTH 112 topics are systematically under- or over-represented among the weaker-performing items, and adjust instructional emphasis accordingly.

6.1.4 For Researchers

The convergence of CTT and 2PL IRT evidence in this study supports the wider use of dual CTT/IRT validation in Nigerian educational measurement research, and the data-cleaning issues documented here (blank scripts, miscoded values) argue for routine data-auditing protocols in future instrument-validation studies, particularly those relying on manually coded paper-based examination scripts.

6.2 Limitations

1) The analysis reported here is confined to a single Federal Polytechnic (Bali, Taraba State) and a pilot sample of 58 valid candidates; the broader six-state North-East sampling frame described in the original proposal has not yet been implemented, so the present findings should be treated as a single-institution pilot validation rather than a zonal one.
2) The response data supplied for analysis recorded only whether each item was answered correctly, not which specific distractor was chosen; a full distractor-efficiency analysis (Table 8 in the originally proposed structure) could not be produced from these data and should be completed once option-level responses are available.
3) The 2PL IRT calibration, while broadly consistent with the CTT results, was estimated from a modest sample of 58 candidates; IRT parameter estimates of this kind are typically more stable with several hundred respondents, so the IRT figures reported here should be re-estimated once a larger sample is collected.
4) This study relies exclusively on Classical Test Theory and 2PL IRT; a Rasch (1PL) analysis, which imposes the stronger assumption of equal item discrimination, was not conducted but could offer a further check on item functioning in future work.
5) Twenty-three candidates' scripts contained no recorded responses and were excluded from analysis; the reason for these blank scripts (absenteeism, incomplete collection, or a data-recording omission) has not been documented and should be clarified in the final manuscript.

6.3 Recommendations

1) The 66 psychometrically sound items should be adopted for assessing students' achievement in Algebra and Elementary Trigonometry (MTH 112) in Federal Polytechnics.
2) The expert panel should review the eight flagged items for wording, keying, and distractor quality before being re-piloted or retired.

3) Federal Polytechnics should establish item banks of psychometrically validated examination questions, refreshed periodically through re-validation as recommended by Lamm, Lamm and Edgar (2020).
4) Future data collection should record option-level (not only correct/incorrect) responses to permit full distractor-efficiency analysis. It should extend sampling across additional Federal Polytechnics to support a more broadly generalisable IRT calibration.

Conflicts of Interest

This study was funded by the Tertiary Education Trust Fund (TETFUND), Nigeria. The conduct of the study was approved in Taraba State, Nigeria, via the Institution-Based Research (IBR) scheme.

References

Agu, N. N., Onyekuba, C., & Anyichie, A. C. (2013). Measuring teachers' competencies in constructing classroom-based tests in Nigerian secondary schools: Need for a test construction skill inventory. *Educational Research and Reviews*, *8*(8), 431–439.

Ahmad, S. K., Mansur, U., & Shu'aibu, M. G. (2024). An assessment of teachers' competency in test construction on students' performance of junior secondary schools in Jigawa State. *Journal of Education Research and Library Practice*, *4*(8), 201–215.

Awofala, A. O. A., & Akinoso, S. O. (2017). Assessment of psychometric properties of mathematics anxiety questionnaire by preservice teachers in South-West, Nigeria. *ABACUS: The Journal of the Mathematical Association of Nigeria*, *42*(1), 355–369.

Blessing, N. (2023). *Application of item response theory in the development and validation of multiple-choice tests in Economics for Senior Secondary School One students* [Unpublished doctoral dissertation]. Michael Okpara University of Agriculture, Umudike.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice-Hall.

Green, B. M. (2024). *Development and validation of Measurement and Evaluation multiple-choice achievement test for NTI-NCE students in South-South states, Nigeria* [Unpublished doctoral dissertation]. Michael Okpara University of Agriculture, Umudike.

Kayode, D. I. (2024). Development and validation of multiple-choice test items for Junior Secondary Schools Business Studies curriculum. *Journal of Educational Studies, Trends & Practice*, *4*(8), 40–50.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151–160.

Lamm, K. W., Lamm, A. J., & Edgar, D. (2020). Scale development and validation: Methodology and recommendations. *Journal of International Agricultural and Extension Education*, *27*(2), 24–35. https://doi.org/10.5191/jiaee.2020.27224

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563–575.

Onyechere, I. (2000, July 16). New face of examination malpractice among Nigerian youths—T*he Guardian Newspaper*.

Tavakol, M., & Dennick, R. (2011). Post-examination analysis of objective tests. *Medical Teacher*, *33*(6), 447–458. https://doi.org/10.3109/0142159X.2011.564682