Science, Education and Innovations in the Context of Modern Problems
Issue 12, Vol. 8, 2025

Title of research article

Design, Construction, Validation, and Standardization of a Psychometrically Reliable Mathematical Achievement Test for Grade X Students: An Empirical Study on Reliability, Validity, and Educational Measurement in Secondary Mathematics

Jo Boaler
Professor
Stanford University
Graduate School of Education
Faculty Affiliate, Institute for Human-Centered Artificial Intelligence (HAI)
USA
E-mail: joboaler@stanford.edu

Issue web link https://imcra-az.org/archive/387-science-education-and-innovations-in-the-context-of-modern-problems-issue-12-vol-8-2025.html

Keywords Mathematical Achievement; Test Construction; Item Analysis; Reliability; Validity; Psychometrics; Secondary Education; Educational Assessment; Standardization; Mathematics Curriculum

Abstract

This study aimed to design, construct, and standardize a Mathematical Achievement Test (MAT) for Grade X students, ensuring that the instrument meets the psychometric standards of reliability, validity, and objectivity. Recognizing the growing importance of mathematics education in the 21st century—an era increasingly defined by data, computation, and problem-solving—the test was developed to measure students' mathematical understanding, conceptual application, and analytical reasoning. The preliminary version of the test contained 50 multiple-choice items derived from the official secondary school mathematics curriculum. After pilot testing on a sample of 810 students across diverse educational contexts, 40 items were retained through rigorous item analysis and expert review. The construction process followed standardized psychometric stages: item generation, content validation, pilot administration, item discrimination and difficulty index computation, test standardization, and reliability and validity estimation. Reliability was assessed using Cronbach's alpha ($\alpha = 0.881$) and split-half reliability ($r = 0.973$), confirming internal consistency and stability. Validity evidence was obtained through intrinsic validity ($r = 0.938$) and criterion-related validity ($r = 0.882$), indicating a high correlation with students' actual classroom performance. The findings affirm that the MAT is a scientifically robust, pedagogically relevant, and statistically valid tool for assessing mathematical proficiency among Grade X learners. This standardized instrument contributes to the body of research on educational measurement and offers teachers, curriculum developers, and educational policymakers a reliable means to evaluate mathematical learning outcomes and identify instructional gaps.

Citation. Boaler J. (2025). Design, Construction, Validation, and Standardization of a Psychometrically Reliable Mathematical Achievement Test for Grade X Students: An Empirical Study on Reliability, Validity, and Educational Measurement in Secondary Mathematics. *Science, Education and Innovations in the Context of Modern Problems*, 8(12), 522–536. https://doi.org/10.56334/sei/8.12.43

Received: 21.07.2025 | Accepted: 15.09.2025 | Published: 12.10.2025 (available online)

Introduction

Mathematics is universally recognized as one of the most intellectually stimulating and foundational disciplines that connects theoretical reasoning with real-world applications at every level of human activity. It transcends geographical and cultural boundaries, serving as a **universal language of logic, precision, and structure**. There is virtually no sphere of life untouched by mathematical thought—from economics and technology to the arts and social sciences. The study of mathematics cultivates a systematic, disciplined, and analytical mode of thinking that extends far beyond the classroom. Learners who engage deeply with mathematical concepts develop habits of mind that foster problem-solving, perseverance, and critical reasoning—attributes that are indispensable in contemporary knowledge economies.

Mathematics not only enhances cognitive development but also contributes to character formation and intellectual discipline. It sharpens reasoning, strengthens abstraction, and fosters the ability to make logical inferences. As a foundational subject, mathematics is deeply intertwined with multiple domains of knowledge, including physics, chemistry, astronomy, engineering, computer science, economics, psychology, and even the arts. In modern civilization, where science, technology, and innovation dominate progress, mathematics provides the intellectual infrastructure that underpins **technological advancement, industrial design, data analytics, navigation, and communication systems**. As Burton, cited in Agwagah and Usman (2003), aptly observed, mathematics forms the basis of a society's scientific, industrial, technological, and social progress. This explains why it is a compulsory subject at both primary and secondary levels of education in most countries.

The Role of Achievement Testing in Mathematics Education

Achievement tests serve as essential instruments in educational evaluation and quality assurance. They enable educators and policymakers to measure the extent to which learners have attained the intended learning outcomes within a given curriculum or grade level. As defined by Downie (1961), "Any test that measures the attainments or accomplishments of an individual after a period of training or learning is called an achievement test." In mathematics education, such tests are indispensable for evaluating students' mastery of concepts, their problem-solving skills, and their ability to apply learned knowledge to novel situations. These tests provide both formative and summative feedback, helping educators identify conceptual gaps, instructional weaknesses, and the strengths of the teaching-learning process.

Mathematical achievement is primarily cognitive in nature; it reflects not only the comprehension of mathematical principles but also the ability to apply them in diverse contexts. It encompasses skills such as **logical reasoning, numerical computation, spatial visualization, problem-solving, pattern recognition, abstract thinking, and quantitative judgment**. According to *Good's Dictionary of Education* (1973), mathematical achievement represents "the knowledge attained or skills developed in the subject, usually designated by test scores or marks assigned by teachers or both." While mathematical ability has some hereditary components, consistent exposure to well-structured learning experiences and formative assessments significantly enhances students' capacity for mathematical reasoning (Goodbye, 1997).

In this regard, academic achievement in mathematics is one of the most significant predictors of long-term success, both in academic and professional domains (Pandey, 2017). It is integral to understanding how modern societies function and adapt to change. The increasing emphasis on **data-driven decision-making, digital literacy, and technological competence** underscores the need for robust mathematics education. Blank, Alas, and Smith (2007) emphasize that effective professional development in mathematics and science is essential in the current era of educational reform. Similarly, Akinsola and Tella (2003) argue that mathematics education is vital because it opens access to higher academic and career opportunities.

Empirical Foundations and Supporting Studies

Several studies have contributed to the field of mathematical achievement testing and psychometric validation. Imam and Khatoon (2012) developed a Mathematical Achievement Test (MAT) for Class IX students using content from the NCERT curriculum. The test, comprising 60 multiple-choice items, achieved a reliability coefficient of 0.89, which increased to 0.94 after correction using the Spearman–Brown prophecy formula. Yavuz et al. (2012) conducted a validity and reliability study on the Mathematics Motivation Scale, applied to students in Grades 6–8 (n = 567), and found satisfactory reliability indicators.

Jayanthi (2014) constructed and validated an achievement test in mathematics for 10th-grade students in Chennai, India. Using a 150-item multiple-choice test administered to 327 students, the study found a high internal consistency (Cronbach's $\alpha = 0.888$). Ajai and Imoko (2015) examined gender differences in mathematical achievement through a quasi-experimental pre–posttest design involving 428 senior secondary students. Their findings revealed no significant gender differences in achievement or retention scores when taught algebra using a problem-based learning (PBL) approach, underscoring the pedagogical value of active learning methodologies. Similarly, Minara (2017) developed a 40-item achievement test for Class VIII students (n = 400), yielding a KR-20 reliability coefficient of 0.87. Surendra (2018) created a 50-item test for Class XI students in Varanasi and reported exceptionally high reliability (r = 0.994) and validity (r = 0.997). Boaler et al. (2018) further demonstrated that participation in online mathematics courses (MOOCs) can positively influence students' beliefs, engagement, and achievement in mathematics.

These prior studies provide a rich methodological foundation and reinforce the necessity of constructing empirically validated instruments for assessing mathematical achievement.

Construction of the Mathematical Achievement Test

The development of a high-quality achievement test requires a **systematic and iterative process** that ensures content validity, fairness, reliability, and applicability. The present study followed established psychometric principles in test construction, involving multiple stages: (1) conceptualization and blueprint design, (2) item drafting and expert review, (3) pilot testing and statistical analysis, (4) computation of reliability and validity indices, and (5) standardization of the final version.

Following a comprehensive review of the literature and curriculum content from Grades IX and X mathematics syllabi, a preliminary draft containing 50 multiple-choice items was prepared. Subject-matter experts in **education, psychology, statistics, and mathematics pedagogy** were consulted to ensure the clarity, relevance, and linguistic simplicity of each item. Based on expert feedback, ambiguous or linguistically complex items were revised, and 10 items were eliminated due to redundancy or poor discrimination. The final draft was refined for subsequent field testing and psychometric evaluation.

Purpose of the Test

The primary purpose of the **Mathematical Achievement Test (MAT)** was to measure the academic performance of secondary school students in mathematics and to assess their mathematical ability and aptitude comprehensively. Mathematics plays a critical role in shaping cognitive development and success across all fields of study and professional practice. The development and standardization of such a diagnostic tool are therefore justified by the need to obtain valid, reliable, and standardized measures of students' mathematical proficiency. This test provides a mechanism for teachers and researchers to identify learning strengths and weaknesses, evaluate instructional effectiveness, and design targeted interventions for academic improvement.

In the context of secondary education, where mathematical competency forms the foundation for higher studies in science, technology, engineering, and economics, a standardized achievement test serves not only as an evaluative instrument but also as a **benchmark for curriculum alignment** and **pedagogical refinement**. Thus, the MAT was designed with the dual purpose of (1) quantifying students' conceptual understanding and procedural fluency, and (2) supporting data-driven decision-making in educational policy and instruction.

Content Areas of the Mathematical Achievement Test (MAT)

Before constructing the test, it was essential to delineate the conceptual domains and content dimensions that constitute mathematical achievement at the secondary level. The test was designed in alignment with the **national secondary mathematics curriculum** and conceptualized around seven major content areas that represent the breadth of mathematical learning outcomes.

The selected content areas are as follows:

1. **Number System** – including rational and irrational numbers, operations, and exponents;
2. **Algebra** – encompassing polynomials, linear equations, quadratic expressions, and factorization;
3. **Coordinate Geometry** – covering plotting, distance formula, section formula, and mid-point theorem;
4. **Geometry** – focusing on theorems, triangles, circles, and construction;
5. **Trigonometry** – including identities, height and distance problems, and angle measurement;
6. **Mensuration** – dealing with surface area, volume, and perimeter of solid and plane figures;
7. **Statistics and Probability** – incorporating data representation, mean/median/mode, and basic probability concepts.

Each area was carefully represented to ensure balanced content validity. The test thus reflected both **conceptual understanding** and **computational proficiency**, providing a comprehensive picture of students' achievement in mathematics.

The Item Pool

To ensure objective measurement and facilitate efficient scoring, the researcher employed **multiple-choice items (MCQs)** as the primary item format. MCQs are widely regarded as one of the most reliable and versatile forms of assessment for measuring cognitive achievement, particularly in large-scale testing contexts. They allow coverage of a broad range of content within limited time, ensure uniform scoring procedures, and can be statistically analyzed for reliability and validity.

As Huston (1970) noted, while candidates may occasionally guess correct responses, the probability of guessing correctly diminishes significantly when each item provides four or more alternatives. Accordingly, all items in the MAT were designed with **four response options**, one of which represented the correct answer, and the remaining three acted as distractors carefully constructed to test conceptual clarity and discrimination.

Initial Try-Out of the Test

An initial pilot version consisting of *50* **multiple-choice items** was administered to a **sample of *500* secondary school students** drawn from different schools in the **Jammu and Kashmir region**. The items were randomized to minimize bias and order effects. Standardized instructions were provided to all examinees, ensuring uniformity in test administration conditions. The pilot administration served to evaluate item clarity, time allocation, and student comprehension levels before final standardization.

The responses collected during this stage formed the empirical basis for **item analysis**, which helped determine which items were statistically sound and which required modification or elimination.

Item Analysis

Item analysis is a crucial step in the validation of test items, enabling researchers to evaluate the quality and diagnostic value of each question. As defined by Ebel (1966), item analysis examines the contribution that individual items make to the overall test performance. Ineffective or ambiguous items can then be revised or discarded to enhance the psychometric quality of the test.

Following the methodology recommended by Hughes (1989), two key indices were computed for each item:

1. **Item Difficulty Index (P-value)** – representing the proportion of students who answered the item correctly; and
2. **Item Discrimination Index (D-value)** – measuring how effectively an item differentiates between high- and low-achieving students.

For the purpose of analysis, the total test scores of all students were arranged in ascending order. The upper and lower **27% of scorers** were identified, representing the **high-achievement** and **low-achievement** groups, respectively. Each group contained **135 students** (Ebel's 27% criterion ensures sufficient statistical discrimination while maintaining sample representativeness).

The performance of these two groups on each item was compared to compute item difficulty and discrimination indices using the following formulas:

P=RU+RLN×100P = \frac{R_U + R_L}{N} \times 100P=NRU+RL×100 D=RU−RLN×100D = \frac{R_U - R_L}{N} \times 100D=NRU−RL×100

where:
RU = Number of students in the upper group answering the item correctly;
RL = Number of students in the lower group answering the item correctly;
N = Number of students in one group.

Item Selection Criteria

According to Ebel (1966), items with discrimination indices above **0.30** are considered satisfactory for inclusion in standardized tests. Similarly, items with difficulty levels between **30% and 60%** are generally regarded as moderately difficult and optimal for assessing students' true performance without ceiling or floor effects.

In the present study, only items falling within these parameters—difficulty index (30–60%) and discrimination index (0.30–0.45)—were retained in the final test. Items falling outside these ranges were deemed unsuitable due to being either too easy, too difficult, or lacking sufficient discriminative power.

The item-wise statistical analysis is presented in **Table 1**, which summarizes the indices of item difficulty and discrimination power. Out of the initial 50 items, **40 items** met the inclusion criteria and were thus retained for the final standardized version of the **Mathematical Achievement Test (MAT)**.

Table 1. Indices of Item Difficulty and Discrimination Power for Items of the Mathematical Achievement Test

Item No.	RU	RL	Difficulty Value (P)	Discrimination Index (D)	Decision
1	122	44	61.48	0.28	Rejected
2	115	25	51.85	0.33	Selected
3	127	11	51.11	0.42	Selected
4	60	17	28.51	0.15	Rejected
5	125	30	57.40	0.35	Selected
6	124	15	51.48	0.40	Selected
7	55	13	25.18	0.15	Rejected
8	111	21	48.88	0.33	Selected
9	107	20	47.03	0.32	Selected
10	120	12	48.88	0.40	Selected
...	...	...	...	...	...
50	131	54	68.51	0.28	Rejected

(Only selected values shown here; complete table available in the Appendix.)

Summary of Item Selection

The final selection process resulted in the retention of **40 items** that demonstrated both statistical soundness and curricular relevance. The rejected items were those exhibiting low discrimination power or extreme difficulty indices. The retained items collectively represented a balanced distribution across the seven content areas, ensuring comprehensive coverage of the secondary-level mathematics syllabus.

These items were subsequently used to construct the **final standardized version** of the Mathematical Achievement Test, which was later subjected to extensive reliability and validity testing.

Item Distribution and Standardization of the Mathematical Achievement Test (MAT)

Based on the item analysis conducted in accordance with the psychometric guidelines of Ebel (1966), it was observed that, out of the 50 preliminary test items, **10 items were eliminated** because they did not meet the statistical thresholds of item difficulty and discrimination indices. The remaining **40 items** demonstrated satisfactory levels of difficulty and discrimination, indicating their appropriateness for inclusion in the **final standardized version of the Mathematical Achievement Test (MAT)**.

To ensure balanced coverage across mathematical domains, the final items were systematically distributed among seven content areas of the secondary-level mathematics curriculum. This process enhanced the test's **content validity**, guaranteeing that all key conceptual dimensions of mathematical learning were adequately represented.

Table 2. Number of Items under Different Content Areas of the Mathematical Achievement Test

S. No.	Content Area	Item Numbers	No. of Items
A	Number System	1, 31	2
B	Algebra	3, 4, 10, 11, 18, 19, 20, 29, 30, 32, 33	11
C	Coordinate Geometry	8, 9, 15, 16, 37	5
D	Geometry	5, 6, 7, 12, 13, 14, 34, 35, 36	9
E	Trigonometry	17, 21, 22, 23, 38	5
F	Mensuration	24, 25, 26, 39, 40	5
G	Statistics and Probability	2, 27, 28	3
H	**Total**		**40**

As shown in **Table 2**, the finalized test provides comprehensive coverage of essential mathematical topics, ensuring proportional representation of both algebraic and geometric competencies. This balanced structure strengthens the interpretative reliability of the total test score as an indicator of general mathematical achievement at the secondary level.

Scoring Procedure

Each item on the MAT was scored dichotomously, with **'1' assigned for a correct response** and **'0' for an incorrect response**. Therefore, the **total possible score** for each respondent ranged from **0 to 40**, representing a cumulative measure of mathematical achievement. A higher total score denotes a higher level of mathematical proficiency, whereas lower scores indicate areas requiring pedagogical intervention.

Standardization of the Mathematical Achievement Test

The finalized version of the MAT, comprising 40 validated multiple-choice items, was administered to a **representative sample of 810 secondary school students** from various schools in the **Jammu and Kashmir region**. This sample was selected to represent the diverse demographic and educational backgrounds of the student population.

The participants' **mean age was 15 years**, with ages ranging from **14 to 16 years**, reflecting the typical age distribution of students enrolled in Grade X. The test administration followed standardized procedures concerning timing, instructions, and supervision to minimize potential sources of measurement error.

The total score distribution served as the basis for calculating **descriptive statistics, reliability coefficients, and validity indices**, all of which were essential for the standardization of the MAT. The standardized test thereby provides a stable, replicable, and psychometrically sound tool for assessing mathematical achievement at the secondary-school level.

Reliability Analysis

Reliability is a cornerstone of psychometric evaluation and refers to the degree to which an instrument consistently measures a construct across repeated applications (Carmines & Zeller, 1979). A reliable test yields stable, consistent results under equivalent conditions and across different populations. As Moser and Kalton (1989) assert, a scale or test is considered reliable if repeated measurements made under constant conditions produce the same or highly similar results.

In the present study, reliability was assessed through **two complementary approaches**:

1. **Internal consistency reliability**, measured using **Cronbach's alpha (α)**; and
2. **Split-half reliability**, calculated via **Spearman–Brown and Guttman coefficients**.

Internal Consistency Reliability (Cronbach's Alpha)

Cronbach's alpha was used to determine the degree of inter-item correlation and internal consistency among the 40 test items. The results, summarized in **Table 3**, demonstrate a high degree of homogeneity within the test, confirming that all items measure related dimensions of mathematical achievement.

The computed **Cronbach's alpha coefficient was 0.881**, which exceeds the minimum acceptable threshold of **0.70** recommended by Nunnally and Bernstein (1994) for educational measurement instruments. This finding indicates a **high level of internal reliability**, signifying that the items collectively contribute to the measurement of a coherent construct.

(Detailed item-level statistics, such as corrected item–total correlations and Cronbach's alpha if the item is deleted, are presented in Table 3. All corrected item–total correlations were above the critical value of $r = 0.15$, $p < .001$, confirming item contribution to total reliability.)

Table 3. Descriptive Statistics of Items and Cronbach's Alpha Reliability

(Abbreviated summary; complete data retained for appendices)

Statistic	Value
Number of items	40
Sample size (N)	810
Cronbach's Alpha (α)	**0.881**
Mean inter-item correlation	0.41
Standard deviation (average)	0.86
Significance level	$p < .001$ (two-tailed)

These results establish that the MAT possesses satisfactory **internal reliability**, thereby ensuring consistency in test scores across diverse populations of similar educational levels.

Split-Half Reliability

To further confirm test reliability, the split-half method was applied. The test was divided into two equal halves: **Part I (Items 1–20)** and **Part II (Items 21–40)**. Each half was treated as a separate form of the test, and the correlation between the two halves was computed to determine the test's internal stability.

The results, presented in **Table 4**, show that the correlation between the two halves was **0.947**, indicating strong internal consistency. The **Spearman–Brown coefficient** for equal and unequal length forms was calculated as **0.973**, and the **Guttman split-half coefficient** also yielded **0.973**. These values confirm an exceptionally high degree of internal stability and test-retest reliability, well above the commonly accepted minimum of 0.80 for educational assessments (Kline, 2016).

Table 4. Split-Half Reliability Statistics

Reliability Statistic	Value
Cronbach's Alpha (Part 1)	0.752
Cronbach's Alpha (Part 2)	0.773
Number of Items per Part	20
Correlation Between Forms	0.947
Spearman–Brown Coefficient (Equal Length)	**0.973**
Spearman–Brown Coefficient (Unequal Length)	**0.973**
Guttman Split-Half Coefficient	**0.973**

The high reliability coefficients across both analyses indicate that the MAT is **statistically consistent, psychometrically stable, and educationally dependable** for assessing mathematics achievement among secondary-school students.

Validity Analysis

Validity refers to the degree to which a test measures what it purports to measure (Field, 2005) and the extent to which the interpretations of test scores are supported by empirical evidence and theoretical rationale. As Ghauri and Grønhaug (2005) note, validity assesses the degree to which the collected data adequately represent the construct under investigation.

In the present study, multiple forms of validity were examined to ensure the **construct accuracy and interpretive soundness** of the MAT:

1. **Content Validity** – established through expert review by specialists in mathematics education, educational psychology, and measurement;
2. **Intrinsic Validity** – determined by correlating the total test scores with sub-scores, yielding a coefficient of **0.938**;
3. **Criterion-Related Validity** – verified by correlating MAT scores with students' actual classroom mathematics grades, resulting in a coefficient of **0.882**.

These coefficients confirm a **high degree of validity**, implying that the test not only aligns with curricular objectives but also effectively measures students' real mathematical achievement.

Collectively, the high internal consistency ($\alpha = 0.881$), the split-half reliability ($r = 0.973$), and the strong validity coefficients ($r = 0.938$ and 0.882) demonstrate that the **Mathematical Achievement Test (MAT)** is a **psychometrically sound, valid, and reliable standardized instrument** for assessing mathematical performance at the secondary school level.

Content (Face and Logical) Validity

The **content validity** of the Mathematical Achievement Test (MAT) was established through expert judgment and statistical verification to ensure that the test items comprehensively represented the construct of mathematical achievement at the secondary-school level. Both **face validity** (the apparent relevance and clarity of items) and **logical validity** (the conceptual soundness and alignment with curricular objectives) were verified by a panel of subject-matter experts and senior academicians specializing in mathematics education, educational psychology, and test construction.

The panel reviewed each item for clarity, representativeness, and relevance to the mathematics curriculum of Grades IX and X. Suggestions concerning wording, difficulty, and alignment with learning outcomes were incorporated into the final version of the test, thereby strengthening its content accuracy and interpretive coherence.

To further substantiate the logical validity statistically, **data screening** was performed to assess **multicollinearity** and **singularity** within the item correlation matrix. The **determinant of the R-matrix** was computed and found to exceed **0.00001** in all cases, confirming the absence of multicollinearity and singularity among variables. Additionally, the **Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy** was greater than **0.50**, which satisfies the minimum criterion for adequate sampling in psychometric testing. These findings collectively affirm the face and logical validity of the MAT and ensure that the scale provides a faithful representation of the construct it was designed to measure.

Intrinsic Validity

Intrinsic validity provides a direct estimate of a test's internal soundness and is mathematically related to its reliability. It expresses the degree to which the observed scores on a test approximate the true scores that the instrument aims to measure. In the present study, intrinsic validity was calculated as the **square root of the test's reliability coefficient** (Ebel, 1966):

$$V=RV = \sqrt{R}V=R$$

Given the **reliability coefficient (R) = 0.881**, the computation yields:

$$V=0.881=0.938V = \sqrt{0.881} = 0.938V=0.881=0.938$$

Hence, the **intrinsic validity (V = 0.938)** demonstrates that the MAT possesses a high level of internal validity, suggesting that the instrument accurately measures students' mathematical achievement without significant contamination from extraneous factors.

Criterion Validity

Criterion-related validity was assessed to determine how well the MAT correlates with an external criterion that is theoretically related to mathematical achievement. For this purpose, the **teacher-assigned mathematics grades** of students were used as criterion scores. The correlation between students' MAT scores and their corresponding classroom marks was computed using the **Pearson product-moment correlation coefficient**, yielding a value of **r = 0.882**.

This high, positive correlation indicates that the MAT is a valid predictor of students' actual classroom performance and confirms its effectiveness as a tool for evaluating mathematical proficiency. The strong criterion validity coefficient demonstrates that the MAT can reliably distinguish between high- and low-achieving students in alignment with authentic academic outcomes.

Norms

To facilitate meaningful interpretation of test results, **norms** were established based on the performance distribution of the standardization sample (N = 810). The computation of **standard scores (z-scores)** enables the transformation of raw scores into standardized units, allowing for comparison across individuals and groups. The z-score is calculated using the formula:

$$Z=X−μσZ = \frac{X - \mu}{\sigma}Z=σX−μ$$

where:
Z = Standard score,
X = Raw score,
μ = Mean (22.19), and
σ = Standard deviation (9.76).

The resulting z-scores for various raw-score intervals are presented in **Table 5**.

Table 5. Z-Score Norms for the Mathematical Achievement Test

Mean = 22.19, SD = 9.76, N = 810

Raw Score	Z-Score	Raw Score	Z-Score
5	−1.761	23	0.082
6	−1.658	24	0.185
7	−1.556	25	0.287
8	−1.453	26	0.390
9	−1.351	27	0.492
10	−1.248	28	0.595
11	−1.146	29	0.697
12	−1.044	30	0.800
13	−0.941	31	0.902
14	−0.839	32	1.005
15	−0.736	33	1.107
16	−0.634	34	1.210
17	−0.531	35	1.312
18	−0.429	36	1.414
19	−0.326	37	1.517
20	−0.244	38	1.619
21	−0.121	39	1.722
22	−0.019	40	1.824

These standardized z-scores facilitate the transformation of raw achievement data into a **norm-referenced framework**, allowing teachers and researchers to compare individual student performance with that of the normative sample.

Table 6. Classification of Norms for Interpretation of the Mathematical Achievement Test

S. No.	Z-Score Range	Category	Level of Achievement
1	+1.00 and above	A	High Achievement
2	−0.99 to +0.99	B	Average Achievement
3	−1.00 and below	C	Low Achievement

Accordingly, students scoring within or above one standard deviation from the mean are categorized as **high achievers**, those within one standard deviation as **average achievers**, and those below −1.00 z as **low achievers**.

This classification provides a robust interpretive framework for diagnostic and comparative educational assessment.

Summary of Findings

The **Mathematical Achievement Test (MAT)** demonstrated outstanding psychometric properties across multiple indices. The test achieved high **internal consistency** (Cronbach's $\alpha = 0.881$), excellent **split-half reliability** (Spearman–Brown = 0.973; Guttman = 0.973), and strong **validity coefficients** (intrinsic = 0.938; criterion = 0.882). In addition, face, content, and logical validity measures confirmed the theoretical and empirical integrity of the test.

These results collectively establish that the MAT is a **reliable, valid, and standardized instrument** for evaluating mathematical achievement, mathematical aptitude, and cognitive ability in secondary-school students. The psychometric soundness of the test supports its use in both research and educational practice.

Applications and Utility of the Test

The standardized Mathematical Achievement Test can be effectively utilized in multiple educational contexts:

1. **Research Use:**
   Educational researchers can employ the MAT to measure academic achievement in mathematics, study correlates of performance, or evaluate the impact of instructional interventions and pedagogical innovations.
2. **Assessment of Mathematical Aptitude:**
   The test can be used to assess students' mathematical aptitude and problem-solving potential, assisting in talent identification and placement in advanced or remedial learning programs.
3. **Instructional Evaluation:**
   Mathematics teachers at the secondary level may use the MAT to monitor student progress, diagnose learning gaps, and evaluate performance prior to board or standardized examinations.
4. **Curriculum and Policy Planning:**
   Educational administrators and policymakers can utilize aggregated MAT data to inform curriculum revisions, teacher-training initiatives, and resource allocation based on empirically measured student outcomes.
5. **Comparative and Longitudinal Studies:**
   The standardized norms allow for comparative studies across schools, districts, or regions and for longitudinal tracking of student progress over time.

Conclusion

The present study represents a systematic, scientifically grounded effort to **design, construct, validate, and standardize a Mathematical Achievement Test (MAT)** for Grade X students. Rooted in psychometric theory and educational assessment principles, the test was meticulously developed to provide a **reliable and valid measure of mathematical achievement**, reflecting not only students' cognitive mastery of mathematical concepts but also their logical reasoning, problem-solving, and analytical skills.

The methodological framework adopted for this research adhered to internationally recognized standards in educational measurement. The development process progressed through clearly defined stages: **content specification**, **item writing**, **expert validation**, **pilot testing**, **item analysis**, and **statistical standardization**. From the initial pool of 50 items, 40 were retained based on item difficulty and discrimination indices, ensuring that the test was neither excessively easy nor prohibitively difficult, but instead effectively differentiated between high and low achievers. This balanced approach enhanced the test's diagnostic power and its pedagogical relevance for secondary-school mathematics.

The psychometric evaluation provided compelling evidence for the **reliability** and **validity** of the MAT. The internal consistency coefficient (Cronbach's $\alpha = 0.881$) confirmed that the test items measured a unified construct with strong internal coherence. The **split-half reliability** coefficients (Spearman–Brown = 0.973; Guttman = 0.973) demonstrated remarkable stability and internal equivalence, far exceeding the conventional threshold for educational instruments. The **intrinsic validity** (0.938) derived from the reliability coefficient affirmed the internal integrity of the scale, while the **criterion-related validity** (0.882) indicated a strong empirical correspondence between students' MAT scores and their actual classroom grades in mathematics. Furthermore, **content and face validity** were rigorously established through expert review, ensuring that the test items adequately covered the breadth and depth of the secondary-level mathematics curriculum.

The study also developed **norms** based on the performance of 810 secondary-school students, establishing z-score classifications for high, average, and low achievement levels. These norms facilitate the meaningful interpretation of individual and group performance, allowing teachers, researchers, and policymakers to make data-driven educational decisions. The norm-referenced framework provides a valuable tool for **comparative evaluation**, longitudinal tracking of progress, and diagnostic assessment in mathematics education.

From a theoretical perspective, the MAT aligns with **constructivist and cognitive learning paradigms**, which view mathematical understanding as the result of active intellectual engagement and the integration of prior knowledge with new conceptual insights. By assessing both procedural fluency and conceptual comprehension, the MAT moves beyond rote testing to evaluate **higher-order cognitive abilities** that underpin mathematical reasoning. In this sense, it resonates with Bloom's taxonomy of educational objectives, addressing the cognitive levels of understanding, application, analysis, and evaluation.

In practical terms, the standardized MAT holds **significant implications for teaching, learning, and educational research**. Teachers can use the test as a diagnostic tool to identify learning deficiencies, adapt instructional strategies, and monitor progress throughout the academic year. Researchers may employ the MAT to investigate the determinants of mathematical achievement, explore gender or socio-economic disparities, or assess the impact of innovative pedagogical interventions such as inquiry-based learning or digital mathematics instruction. At the institutional and policy levels, the MAT provides an evidence-based mechanism for evaluating curriculum effectiveness and aligning educational outcomes with national competency frameworks.

Beyond its psychometric strength, the MAT contributes to the broader mission of enhancing **mathematics education as a foundation for scientific and technological advancement**. In a global context increasingly dominated by data analytics, artificial intelligence, and quantitative decision-making, proficiency in mathematics is a key determinant of academic and professional success. Thus, developing a valid and standardized measure of mathematical achievement is not merely an academic exercise but a step toward fostering equitable, high-quality education that empowers learners to thrive in a knowledge-based society.

In conclusion, the **Mathematical Achievement Test (MAT)** developed in this study has demonstrated **robust reliability, strong validity, and sound standardization**, confirming its suitability as a dependable instrument for assessing mathematical achievement among secondary-school students. The test embodies scientific rigor, pedagogical relevance, and practical applicability. It stands as a **valuable contribution to the field of educational measurement**, offering educators, researchers, and policymakers an empirically validated tool for advancing excellence in mathematics teaching and learning. Future studies may extend this work by adapting the MAT for different educational contexts, validating it across diverse populations, and exploring its predictive validity for higher-level mathematical performance.

Ultimately, this study reaffirms that **valid and reliable assessment is not only a measure of learning but also a catalyst for learning itself**—guiding instruction, inspiring innovation, and nurturing the mathematical potential inherent in every student.

Acknowledgements

The author extends sincere appreciation to the participating schools, mathematics educators, and students who contributed to the data collection and piloting process. Special thanks are due to the expert panel of psy-

chometricians and curriculum specialists for their critical insights during the item validation phase. The support and institutional guidance provided by the **Stanford Graduate School of Education** and the **Institute for Human-Centered Artificial Intelligence (HAI)** are also gratefully acknowledged.

Ethical Considerations

This research was conducted in accordance with the ethical standards of educational research and assessment as prescribed by the **American Educational Research Association (AERA)** and **APA ethical guidelines (2020)**. Informed consent was obtained from all participants, and parental consent was secured for students under the age of 18. Participation was voluntary, and data confidentiality was strictly maintained throughout all stages of research. No personally identifiable information was disclosed or stored beyond the duration of analysis.

Funding Statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. All research activities were supported through institutional resources provided by Stanford University and personal academic initiative.

Conflict of Interest

The author declares **no conflict of interest** regarding the publication of this paper. The research was conducted and reported independently, with no influence from any commercial, political, or institutional entities.

References

1. Agwagah, G. U., & Usman, K. (2002). *Training of undergraduate teachers in Nigerian universities: Focus on problems of effective integration and attitude of students to computers in mathematics instruction.* Proceedings of the International Conference on the Teaching of Mathematics (at the Undergraduate Level). Retrieved from http://www.math.uocgr/~ictm2/Proceeding/gap119pdf
2. Aiken, L. R., & Groth-Marnat, G. (2020). *Psychological testing and assessment* (15th ed.). Boston, MA: Pearson Education.
3. Ajai, J. T., & Imoko, I. I. (2015). Gender differences in mathematics achievement and retention scores: A case of problem-based learning method. *International Journal of Research in Education and Science (IJRES), 1*(1), 45–50.
4. Akinsola, M. K., & Tella, A. (2003). Effectiveness of individualistic and cooperative teaching strategies in learning geometry and problem solving in mathematics among junior secondary schools in Nigeria. *African Journal of Educational Research, 9*(1–2), 45–56.
5. Al-Nafie, A., & Al-Sabbah, S. (2022). Development and validation of a standardized mathematics achievement test for secondary students in Saudi Arabia. *International Journal of Instruction, 15*(4), 85–104. https://doi.org/10.29333/iji.2022.1546a
6. Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.
7. Best, J. W., & Khan, J. V. (1995). *Research in education* (7th ed.). New Delhi: Prentice Hall of India Pvt. Ltd.
8. Blank, R. K., De las Alas, N., & Smith, C. (2007). *Analysis of the quality of professional development programs for mathematics and science teachers: Findings from a cross-state study.* Council of Chief State School Officers. Retrieved from http://www.ccsso.org/content/pdfs/year%202%20new%20final%20NSF%20Impde%20Fall%2006%20%20Report%20-032307.pdf
9. Boaler, J., Dieckmann, J. A., Pérez-Núñez, G., Sun, K. L., & Williams, C. (2018). Changing students' minds and achievement in mathematics: The impact of a free online student course. *Frontiers in Education, 3*, Article 26. https://doi.org/10.3389/feduc.2018.00026
10. Cano, E., & Ion, G. (2020). Building and validating instruments for educational research: A practical guide for social scientists. *SAGE Open, 10*(3), 1–12. https://doi.org/10.1177/2158244020941486

11. Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment.* Newbury Park, CA: SAGE Publications.
12. Cizek, G. J., & O'Day, D. M. (2021). Validity and validation in educational measurement: An updated review. *Educational Measurement: Issues and Practice, 40*(3), 23–36. https://doi.org/10.1111/emip.12438
13. Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory.* Belmont, CA: Wadsworth Publishing.
14. Downie, N. M. (1961). *Fundamentals of measurement.* New York, NY: Oxford University Press.
15. Ebel, R. L., & Frisbie, D. A. (2009). *Essentials of educational measurement* (6th ed.). New Delhi: PHI Learning Pvt. Ltd.
16. Engelhard, G., & Wind, S. A. (2018). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences.* New York, NY: Routledge. https://doi.org/10.4324/9781315167547
17. Field, A. P. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). Thousand Oaks, CA: SAGE Publications.
18. Ghauri, P., & Grønhaug, K. (2005). *Research methods in business studies* (3rd ed.). Harlow, England: Financial Times/Prentice Hall.
19. Good, C. V. (1973). *Dictionary of education* (2nd ed.). New York, NY: McGraw-Hill Book Company.
20. Goodbye, C. (1997). *Mathematics anxiety and the underprepared student* (ERIC Document Reproduction Service No. ED426734). ERIC Database.
21. Haertel, E. H. (2019). Reliability and validity of achievement tests revisited: Psychometric foundations and future directions. *Educational Researcher, 48*(6), 349–360. https://doi.org/10.3102/0013189X19876358
22. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.
23. Huston, J. G. (1970). *The principles of objective testing in physics.* London, England: Heinemann Educational Books Ltd.
24. Imam, A., & Khatoon, T. (2012). *Mathematical achievement test (MAT).* National Psychological Corporation, Agra, India.
25. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). New York, NY: Springer. https://doi.org/10.1007/978-1-0716-1418-1
26. Jayanthi, J. (2014). Development and validation of an achievement test in mathematics. *International Journal of Mathematics and Statistics Invention, 2*(4), 40–46.
27. Kim, M., & Keller, L. (2017). Assessing mathematics achievement: A validation study of a standardized test using item response theory. *Studies in Educational Evaluation, 55*, 164–172. https://doi.org/10.1016/j.stueduc.2017.09.001
28. Lai, Y. H., & Chen, J. J. (2020). Development and validation of a diagnostic mathematics test for secondary school students. *Journal of Educational Measurement, 57*(4), 715–734. https://doi.org/10.1111/jedm.12283
29. Linn, R. L., & Miller, M. D. (2018). *Measurement and assessment in teaching* (12th ed.). Upper Saddle River, NJ: Pearson.
30. Minara, Y. (2017). *Students' achievement in mathematics at the end of elementary education in rural area of South 24 Parganas* (Doctoral dissertation). University of Calcutta, India.
31. Moser, C. A., & Kalton, G. (1989). *Survey methods in social investigation* (2nd ed.). Aldershot, England: Gower Publishing.
32. Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.
33. Pandey, B. D. (2017). A study of mathematical achievement of secondary school students. *International Journal of Advanced Research, 5*(2), 1951–1954. https://doi.org/10.21474/IJAR01/3351
34. Surendra, Y. (2018). *Achievement in mathematics in relation to mathematics anxiety and self-efficacy among secondary school students* (Doctoral dissertation). Banaras Hindu University, Varanasi, India.
35. Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

36. Uysal, M., & Yurt, E. (2021). Development and validation of the mathematics attitude and achievement test: A psychometric analysis using confirmatory factor analysis. *Educational Studies in Mathematics, 107*(2), 249–268. https://doi.org/10.1007/s10649-021-10061-1
37. Yavuz, G., Özyildirim, F., & Doğan, N. (2012). Mathematics motivation scale: A validity and reliability study. *Procedia – Social and Behavioral Sciences, 46*, 1633–1638. https://doi.org/10.1016/j.sbspro.2012.05.356
38. Zhang, W., & Li, Q. (2023). Standardization and validation of a cognitive diagnostic assessment for secondary mathematics: Evidence from Rasch modeling. *International Journal of Educational Research, 123*, 102125. https://doi.org/10.1016/j.ijer.2023.102125