Volume 11, Issue 3, March – 2026 International Journal of Innovative Science and Research Technology
ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1916

Development and Standardization of an Achievement Test on Mathematics Pedagogy for B.Ed. Pre-Service Teachers

Abdus Safi[1]; Dr. Jaishree Shukla[2]

[1]Research Scholar, Department of Education, Dr. C.V. Raman University Kota, Bilaspur (C.G.)
[2]Associate Professor, Department of Education, Dr. C.V. Raman University Kota, Bilaspur (C.G.)

Publication Date: 2026/04/07

Abstract: This research focuses on developing and standardizing an achievement test in Mathematics Pedagogy specifically designed for B.Ed. pre-service teachers. The preliminary draft of the test consisted of 60 items prepared from ten modules of the mathematics pedagogy curriculum, along with a blueprint and scoring key. The draft was administered to 120 pre-service teachers during the initial try-out. After item analysis, which included calculation of item difficulty and discrimination indices, 50 items were retained for the final test. Test reliability was evaluated using the split-half method and Cronbach's alpha. Content validity was ascertained through an expert panel which thoroughly examined the test items and thus ensured that they were consistent with the curriculum and objectives of mathematics pedagogy.Therefore, the standardized achievement test is a reliable and valid tool for assessing the cognitive skills of B. Ed. pre, service teachers in the area of mathematics pedagogy.

Keywords: *Achievement Test, Discriminating Power, Distractor Efficiency, Item Analysis Mathematics Pedagogy.*

How to Cite: Abdus Safi; Dr. Jaishree Shukla (2026) Development and Standardization of an Achievement Test on Mathematics Pedagogy for B.Ed. Pre-Service Teachers. *International Journal of Innovative Science and Research Technology,* 11(3), 3394-3401. https://doi.org/10.38124/ijisrt/26mar1916

I. INTRODUCTION

According to Gronlund (1982) and Mehrens and Lehmann (1973), achievement tests play a vital role in both educational research and classroom practice. They serve as essential tools for objectively measuring students' knowledge, understanding, and skills. According to 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".

Teachers, by means of these instruments, are capable of quantifying the learning attained and can pinpoint the areas where students are stronger and those where they are weaker. Therefore, achievement tests become a great source of information for educators, which can lead to the refinement of teaching methods and the facilitation of efficient learning. In the case of mathematics education, the role of achievement tests is even more significant as the subject of mathematics is not only one of the logical thinking and problem, solving, but also a base for numerous disciplines as well as the practical life. The importance of "mathematization" of students' thinking, which means helping them to reason, generalize, and solve problems in a meaningful way, has been mentioned in the National Curriculum Framework (NCF, 2005). Nevertheless, to be able to evaluate the achievement of such objectives, it is necessary to have well, crafted achievement tests that are in line with the curriculum.

Brownell (1947) argued that an effective achievement test should assess the student's learning of concepts, not just their memory of facts. This, he said, is essentially the core of a meaningful mathematics learning. Students with conceptual understanding, hence, are more likely to be able to apply their knowledge in different situations, to invent solutions to unfamiliar problems, and to think critically (Anderson & Krathwohl, 2001).

National Education Policy (NEP, 2020), in line with this, places great emphasis on the abandonment of rote memorization and the promotion of comprehension, creativity, and the application of knowledge in education. This change necessitates the transformation of tests in the field of mathematics education to be aligned with these priorities and to be concentrated on higher, order thinking skills, rather than just on the recall of facts.

Putranta and Supahar (2019) note that multiple-choice tests remain among the most commonly used tools for evaluating learning outcomes. They're popular because they're objective, easy to give, and efficient—especially with large groups. Those qualities make multiple-choice tests a

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1916

practical, reliable way to measure students' understanding and performance across many settings. When carefully designed, they can assess more than just surface facts and also tap into reasoning and problem-solving (Wilcox et al., 2015). Two-tier or diagnostic versions of multiple-choice tests have gained recognition for their ability to reveal misconceptions and show how students reason (Treagust, 2006; Chandrasegaran et al., 2007). These tools are particularly useful in teacher education programs, where future teachers need both mathematical knowledge and pedagogical skills.

In the making of achievement tests, standardization is fundamental if the tests are to be reliable and valid. Fraenkel et al. (2012) say reliability means that the results of the test are the same or very similar when the test is given on different occasions under conditions as close to control as possible. It shows how stable and dependable the measuring instrument is. On the other hand, validity as explained by Freeman (1965) is about the extent to which a test measures what it is intended to measure. This means that reliability and validity are not only closely related but also required in the assessment of the same idea. The aspect of content validity in particular stands out as the most important one for educational testing since it literally ensures that the test is aligned with the curriculum and learning objectives. Additionally, item analysis is very essential to the process of polishing test items as it makes it possible to determine the difficulty of a test question, its discrimination power, and the efficiency of its distractors (Quaigrain & Arhin, 2017). Hence, a well, structured and standardized achievement test in mathematics pedagogy has the potential to reveal significant learning outcomes and thereby inform productive teaching strategies.

The present study attempts to construct and standardize an achievement test in Mathematics Pedagogy for B.Ed. pre-service teachers. The test aims to evaluate their knowledge, comprehension, application, and pedagogical skills across ten modules of the curriculum. By ensuring validity and reliability, the study seeks to contribute a useful tool for assessing pre-service teachers' preparedness and supporting teacher education programs.

II. CONCEPTUAL FRAMEWORK OF PREPARATION OF THE TEST

Recent research works on mathematics assessments emphasize the development and standardization of multiple-choice achievement tests using item analysis, reliability measures, and item response theory to measure students' understanding and learning achievement in mathematics effectively (Al Haque & Vandana, 2025; Ibrahim & Sani, 2025; Khoa & Le, 2026; Mata et al., 2026; Ramatu & Kongnyuy, 2025; Wordu & Opusunju, 2025).

The conceptual framework for the standardized achievement test in Mathematics Pedagogy is based on the well-established principles of educational measurement and evaluation. It has been emphasized by Gronlund (1982) that an achievement test, if properly constructed, should measure the desired learning outcomes, which include knowledge and understanding. Mehrens and Lehmann (1973) emphasized the need to ensure the alignment of the test with the objectives of instruction to ensure the content validity of the test. In the present research, the framework has been aligned with the Revised Bloom's Taxonomy of Educational Objectives (Anderson & Krathwohl, 2001), which includes the cognitive domain of remembering, understanding, applying, analyzing, and evaluating. Recent research by Kolagar, Zalkow, and Zarcone (2025) has emphasized the need for the systematic classification of the learning objectives with the levels of Bloom's taxonomy for the precision of the educational assessments, with the results being better with the use of automated classification, assuming the learning objectives are well formulated, which has direct implications for the mathematics pedagogy test.

This test was developed using the systematic approach to test development proposed by Thorndike & Hagen (1977), which entails the formulation of objectives, the preparation of test specifications, the preparation of test items, review of the test items, pilot testing, and the establishment of reliability and validity. This approach to test development is also supported by Nunnally's (1978) psychometric theory, which emphasizes the importance of internal consistency reliability in the measurement of test quality. Additionally, the model of instructional design proposed by Dick & Carey (1996) was used to organize the test items to enhance their alignment to the objectives of instruction. Besser, Hagena, & Kleickmann (2024) empirically support this approach to the integration of the subject matter in the teaching of mathematics. The study, which involved 856 students from 39 mathematics classes, indicated that effective teaching of mathematics offers students an opportunity to attain a deep understanding of mathematics. This finding supports the need for achievement tests that assess both generic pedagogical content knowledge and mathematics-specific instructional skills, a dual requirement that can easily be accommodated by the integrated framework. Recent research by Bhatti and Hashmi (2025) provides specific parameters for item evaluation in achievement testing. Their study, which examined 50 multiple-choice physics test items, indicated that 17 were easy, 31 were moderate, and 2 were difficult, with 48 of them showing adequate discrimination between high and low achievers. All 50 had adequate distractors, which helped to create a high reliability coefficient of 0.92, as measured by Cronbach's alpha. These results provide the basis for the construction of mathematics pedagogy tests, which need to have a range of difficulty levels, adequate discrimination, and distractors that show misconceptions.

Content validity, established through expert review of item-objective alignment, ensures that test items represent the domain of mathematics pedagogy. Kumar and Gangwar (2025) employed ten teacher-educators to determine face and content validity, a practice that should be replicated in mathematics pedagogy test development. Construct validity, demonstrated through relationships between test scores and theoretically related variables, can be established by correlating performance on the pedagogy test with measures of teaching effectiveness or mathematics content knowledge.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1916

It is in this way that a conceptual framework has been developed that combines concepts in educational measurement, instructional design, and cognitive psychology in the creation of a valid and reliable instrument for measuring B.Ed. pre-service teachers' achievement in math pedagogy. Standardization of tests involves item analysis, validity, reliability, and norm development, and it is important in establishing the psychometric qualities of a test. This ensures that the test objectively measures the outcome and is reliable in terms of both formative and summative evaluation.

➢ *Objective of the Study*

- To develop an achievement test in Mathematics Pedagogy for B.Ed. pre-service teachers.
- To standardize the achievement test in Mathematics Pedagogy by determining the difficulty index, discriminating power, distractor efficiency.
- To establish the validity, and reliability of the achievement test.
- To select the best items based on item analysis to design a final test on Mathematics Pedagogy for B.Ed. pre-service teachers.

➢ *Preparation of Achievement Test on Mathematics Pedagogy (ATMP):*

Since a validated test was not available, the investigator thought of developing a tool and validated it for using in the present study.

➢ *Objectives of ATMP*

In the present work, investigator has prepared the achievement test in mathematics pedagogy for B.Ed. pre-service teachers. As the investigator has taken the subject mathematics pedagogy, so major objective were categorized as Remembering, Understanding, Applying, Analyzing and Evaluating, excluding the creating level from revised Bloom's taxonomy of Cognitive domain.

➢ *Content Analysis and Preparation of BluePrint*

A blueprint is an essential plan that guides the systematic construction of an achievement test. It ensures proper representation of subject content, alignment with instructional objectives, and balanced weightage across cognitive domains of Bloom's taxonomy. In the present study, the blueprint was prepared to cover all the ten modules of the course "Mathematics Pedagogy" for B.Ed. pre-service teachers. The e-content for the course is organized into ten weekly modules. Week 1 introduces learners to Mathematics Pedagogy. Next in Week 2 by a focus on the Aims and Objectives of Mathematics Teaching. After that Week 3 explores Strategies and Techniques in Mathematics Pedagogy, while Week 4 emphasizes Inductive and Deductive Approaches as fundamental teaching methods. Next in Week 5, attention shifts to Advanced Methods such as analysis, synthesis, and the project method, and the module on Week 6 highlights Problem-Solving and Heuristic Techniques. Week 7 presents Innovative Methods, including storytelling in mathematics, mathematical induction, and the laboratory method. After this, in the 7th week, "The Use of Learning Resources for Mathematics Instruction" is addressed in Week 8, then "Learning Design in Mathematics Education" in Week 9, and finally in Week 10, "Assessment and Evaluation in Mathematics" is included in order to connect theory with practice in terms of evaluation. The investigator framed different test items. The content areas were distributed under the five cognitive levels: Remembering, Understanding, Applying, Analyzing and Evaluating, excluding the creating level, accoring to revised Bloom's Taxonomy.

Table 1 Blue Print of The Achiemement Test

	Cognitive Domain – Revised Blooms Taxonomy							
Module	**Remembering**	**Understanding**	**Applying**	**Analyzing**	**Evaluating**	**Creating**	**Total Questions**	**%age**
1	1, 2	3, 4	5, 6	7, 8	0	0	8	13%
2	9, 10	11, 12	13, 14	0	15, 16, 17, 18	0	10	16%
3	0	19, 20	21, 22	23, 24	0	0	6	10%
4	0	25, 26	27, 28	0	0	0	4	7%
5	29, 30	0	31, 32	0	33, 34	0	6	10%
6	35, 36	0	0	37, 38	0	0	4	7%
7	0	0	0	39, 40	41, 42	0	4	7%
8	0	43, 44	0	45, 46	0	0	4	7%
9	47, 48	0	49, 50	0	51, 52	0	6	10%
10	53, 54	55, 56	0	57, 58	59, 60	0	8	13%
Total Question	12	12	12	12	12	0	60	100%
%age	20%	20%	20%	20%	20%	0%	100%	

Suitable weightage in terms of the number of questions and percentage for each module has been allocated in the blueprint. The initial blueprint for the test comprises 60 questions, and all five cognitive levels—Remembering, Understanding, Applying, Analyzing, and Evaluating—have been allocated 20% of the questions. No questions are allocated to the "Creating" level, as the questions are in the form of multiple-choice questions.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1916

III. CONSTRUCTION OF TEST ITEMS

The investigator decided to have multiple choice for ATMP. The investigator consulted many standard books of B.Ed. Mathematics Pedagogy written by different authors to prepare good multiple-choice items. The investigator took care to prepare good MCQ items which dealt with significant ideas from developed E-content module. Also unintended clues to the correct answer, repetition of words in the option were avoided. The items were so selected to have equal and uniform representation to all the E-content module. After the investigator himself was satisfied on the item prepared, it was showned to ten Mathematics Teacher and Professors for verifying the suitability of the item for the target pre-service teachers.

➢ *Question Paper*

After carrying out the addition and deletion of items, there were 64 multiple-choice questions (MCQs) were framed, each carrying one mark. Each item consisted of a four alternatives (A, B, C, D), of which one was correct. Care was taken to ensure clarity, objectivity, and simplicity of language to suit B.Ed. pre-service teachers. Necessary instructions were included in the question paper.

➢ *Scoring Key*

A scoring key was prepared to ensure objectivity in scoring. Each correct answer was awarded one mark, while incorrect responses were given zero. This scoring scheme eliminated subjectivity in evaluation.

➢ *Try Out of the Test*

The first draft was administered in three stages:

- *Preliminary Screening:*

At first 64 items were constructed after that Items were reviewed by experts in the field of Mathematics Education to check for ambiguity, language clarity, and relevance to objectives. Based on feedback, some items were revised or reworded.

- *Pre-Try Out:*

The test with 64 items was conducted on a small sample of 30 B.Ed. students to identify difficulties in comprehension and language. Minor modifications were made accordingly. 60 items were retain in the final draft.

- *Final Try Out:*

The refined 60-item test was administered to a larger sample of 120 pre-service teachers for the purpose of item analysis and standardization.

➢ *Item Analysis*

The item analysis was done for selecting the right type of test items for the final draft. For doing item analysis of preliminary draft, it was administered on selected 10 pre-service teachers. Prior permission from the head of the institutions for administering the tool was obtained and the Pre-service teachers were also informed earlier. Sufficient time (1 hour) was given to the pre-service teachers to attend the items. Method teacher was requested to help the investigator for the successful conduct of the tests.

➢ *Item Difficulty Value*

Item difficulty was calculated using the formula:

$$\text{Difficulty Value} = \frac{\text{Number of Students who answer the item correctly}}{\text{total number of students who answer the item}} \quad \ldots\ldots (1)$$

According to Fulcher and Davidson (2007), the difficulty level of test items is commonly defined as the proportion of test-takers who choose the correct answer instead of the distractors on a test question. The following table describes the difficulty level, which helps in analyzing the item difficulty.

Table 2 The Difficulty Level

Index Level	Difficult Category
0.00-0.30	Difficult
0.31-0.70	Moderate
0.71-1.00	Easy

The final data of the difficulty level is shown in the following table.

Table 3 The Final Data of the Difficulty Level

Index Level	Difficult Category	Items	Frequency
0.00-0.30	Difficult	57	1
0.31-0.70	Moderate	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60	50
0.71-1.00	Easy	2, 3, 4, 13, 14, 39, 43, 51	9

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1916

➢ *Test Item Discriminating Power*

Discriminating power indicates the ability of an item to differentiate between high and low achievers. Based on Gronlund (1998), the present study employed the following formula to calculate the discriminating power of the selected test items.

$$\text{Discriminating Power} = \frac{R_U - R_L}{\frac{1}{2}T} \ldots\ldots (2)$$

$R_U =$ Numbers of students in the upper group who answer the question correctly

$R_L =$ Numbers of students in the lower group who answer the question correctly

$\frac{1}{2}T =$ Half of the overall student sample was included in the item analysis

Where *T* is the number of students in either group.

According to Zajda (2006), discriminating power refers to the ability of a test item to distinguish between high-achieving examinees (upper group) and their lower-achieving counterparts (lower group). The levels of discriminating power are generally classified into several categories: excellent, good, satisfactory, and poor or rejected. The table below provides further details.

Table 4 The Level of Discriminating Power

Discrimination Level	Category
0.71-1.00	Excellent
0.41-0.70	Good
0.1-0.40	Satisfied
0.00-0.20	Poor
Negative	Rejected

The final data of the discriminating power is shown in the following table.

Table 5 The Level of Discriminating Power

Discrimination Level	Category	Items	Frequency
0.71-1.00	Excellent	1, 5, 7, 10, 15, 16, 18, 20, 22, 25, 29, 33, 34, 35, 36, 47, 49, 50, 52, 55, 58	21
0.41-0.70	Good	2, 3, 4, 6, 8, 9, 11, 12, 13, 17, 19, 21, 23, 24, 26, 27, 28, 31, 32, 37, 38, 40, 41, 42, 43, 44, 45, 46, 48, 51, 53, 54, 59, 60	34
0.1-0.40	Satisfied	14, 39, 56, 57	4
0.00-0.20	Poor	30	1
Negative	Rejected	Nil	0

➢ *Distractor Analysis*

According to Haladyna & Rodriguez (2013), distractor analysis is an essential step in test validation that examines the effectiveness of each option in a multiple-choice item. A good distractor is one that attracts some of the lower-ability students while being avoided by higher-ability students. Items with non-functional distractors—those rarely chosen—should be reviewed or revised.

Table 6 The Classification of Distractor Efficiency (Malau-Aduli & Zimitat (2012))

Criteria	Category
5% ≥ p and LG > UG	Effective
5% ≥ p and LG < UG	Less effective
p ≤ 5% and LG > UG	Less effective
p ≤ 5% and LG < UG	Ineffective
0	Dysfunctional

Note. 5% ≥ p = The number of students choosing the distractor is 5% or higher; p ≤ 5% = The number of students choosing the distractor is less than 5%; LG = number of students in the lower group choosing the distractor; UG = number of students in the upper group choosing the distractor.

Each distractor was analyzed for effectiveness. Non-functional distractors (selected by less than 5% of students) were revised or replaced. This helped improve the overall quality of the test items.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1916

Table 7 Sample Distractor Analysis Table of Some Items

Item No.	Correct Option	% UG Choosing	% LG Choosing	Distractor				Interpretation
				A (%)	B (%)	C (%)	D (%)	
4	A	85	42	A (85.83)	B (8.33)	C (2.5)	D (3.33)	Distractor C and D is non-functional
14	B	90	35	A (9.17)	B (85.83)	C (0)	D (5)	Distractor C non-functional
26	B	78	30	A (25.83)	B (29.17)	C (23.33)	D (21.67)	Good item, effective distractors

After analysis, 10 items were discarded (due to extreme difficulty, poor discrimination, or non-functional distractors), leaving 50 best-performing items for the final version.

➢ *Designing Of the Final Test*

In order to have a good test items, the investigator eliminated 10 items on the basis of the results of the item analysis. The final achievement test comprised 50 multiple-choice items. The final test ensured proper coverage of all ten modules, balanced weightage across cognitive domains, and optimum psychometric properties. Each item carried one mark, making the test total 50 marks.

➢ *Reliability, Validity and Norms of The Achievement Test*

- *Reliability of the Test*

Reliability was established through two methods:

✓ *Split-Half Reliability*

The test was divided into odd and even items. Correlation between the two halves was calculated and corrected using the Spearman-Brown prophecy formula. The reliability coefficient was found to be 0.928 (60 items), indicating high internal consistency. Correlation between form is 0.865. Guttman Split-Half Coefficient = 0.928.

✓ *Cronbach's Alpha*

Using SPSS (version 23), Cronbach's alpha was calculated as 0.928, confirming strong reliability of the instrument.

➢ *Validity of the Test*

The validity of the standardized achievement test was established through various complementary methods. Content validity was maintained by creating a comprehensive test blueprint that explicitly aligned each question with the course objectives of Mathematics Pedagogy. This alignment ensured that the test effectively measured the desired learning outcomes. Again, expert opinions from professionals in mathematics education and educational measurement were solicited to evaluate the items for their content relevance and representativeness.

Finally, these forms of validity affirmed that the achievement test accurately measured its intended objectives and was suitable for assessing the pedagogical knowledge and skills of pre-service teachers in mathematics.

➢ *Norms of the Test*

In educational assessment, a norm refers to a reference framework that allows interpretation of individual test scores relative to a defined group (Anastasi & Urbina, 1997). Norm-referenced interpretation compares a test taker's performance with that of a representative sample (Crocker & Algina, 2008). According to Anne Anastasi and Susana Urbina (1997), norms are essential for giving meaning to raw scores, as raw scores alone do not indicate the level of performance unless compared to group performance.

To establish interpretive norms for the standardized Achievement Test in Mathematics Pedagogy, a five-point norm-referenced classification was developed based on the combined mean and pooled standard deviation of the total standardization sample (N = 120). The mean score of the sample was 32.29 and the pooled standard deviation was 8.70. Following the principles of norm-referenced interpretation (Crocker & Algina, 2008), performance categories were defined using ±0.5 and ±1 standard deviation units from the mean. Scores at or above +1 SD were categorized as "Very High Achievement," scores between +0.5 SD and +1 SD as "High Achievement," scores within ±0.5 SD as "Average Achievement," scores between −0.5 SD and −1 SD as "Low Achievement," and scores below −1 SD as "Very Low Achievement." This method ensures statistically justified and distribution-based classification of achievement levels.

Table 8 5-Point Norm Table

Raw Score Range	Category	Interpretation
41 – 50	Very High	Excellent Pedagogical Mastery
37 – 40	High	Above Average
28 – 36	Average	Moderate Competency
24 – 27	Low	Below Average
0 – 23	Very Low	Needs Improvement

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1916

IV. LIMITATIONS OF THE STUDY

- No IQ test or semester examination achievement test was used for grouping the students. Only 55% marks obtained in B.Sc. or M.Sc. was taken into account for selecting the pre-service teachers.
- The study has been confined to the pre-service teacher studying B.Ed. Mathematics Method for the convenience of the investigator.
- The investigator prepared the test only some selected content areas of B.Ed. 2nd and 3rd semester Mathematics Method.
- Only MCQ test was administered.

V. CONCLUSION

The present study successfully developed and standardized an achievement test in Mathematics Pedagogy for B.Ed. pre-service teachers. The process of constructing the test included preparing a blueprint, writing items, conducting a pre-tryout, performing a final tryout, and analyzing items (including difficulty index, discrimination power, and distractor analysis). From the initial 60 drafted items, 50 were finalized following the analysis. The investigator found that the test exhibited high reliability (split-half = 0.928; Cronbach's alpha = 0.928) and robust validity (with content, construct, and face validity confirmed).

Consequently, the achievement test developed serves as a reliable, valid, and standardized instrument for evaluating the pedagogical knowledge of B.Ed. pre-service teachers in Mathematics. It can be effectively utilized by teacher educators and researchers to assess competencies in mathematics pedagogy.

REFERENCES

[1]. Al Haque, U. G., & Vandana, D. (2025). Development and standardization a mathematics achievement test for fourth-grade students. *EPRA International Journal of Multidisciplinary Research, 11*(5). https://eprajournals.com/IJMR/article/16159

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

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

[4]. Besser, M., Hagena, M., & Kleickmann, T. (2024). On the added value of considering effects of generic and subject-specific instructional quality on students' achievements: An exploratory study on the example of implementing formative assessment in mathematics education. ZDM–Mathematics Education, 56(5), 815–830. https://doi.org/10.1007/s11858-024-01562-2

[5]. Bhatti, S. A., & Hashmi, M. A. (2025). Semi standardization of an achievement test in the subject of physics at secondary level. Annals of Human and Social Sciences, 6(3), 394–407. https://ojs.ahss.org.pk/journal/article/view/1050

[6]. Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals.* Longman.

[7]. Brownell, W. A. (1947). *The place of meaning in the teaching of arithmetic*. The Elementary School Journal, 47(5), 256–265.

[8]. Chandrasegaran, A. L., Treagust, D. F., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions. *Chemistry Education Research and Practice, 8*(3), 293–307.

[9]. Crocker, L., & Algina, J. (2008). Introduction to classical and modern test theory. Cengage Learning.

[10]. Dick, W., & Carey, L. (1996). *The systematic design of instruction* (4th ed.). HarperCollins College Publishers.

[11]. Downie, N.M. (1961) Fundamentals of Measurement. Oxford University Press, New York

[12]. Fraenkel, J. R., & Wallen, N. E. (1990). How to design and evaluate research in education. Order Department, McGraw Hill Publishing Co., Princeton Rd., Hightstown, NJ 08520.

[13]. Freeman, F. S. (1965). *Theory and practice of psychological testing* (3rd ed.). New Delhi: Oxford & IBH.

[14]. Fulcher, G., & Davidson, F. (2007). Language testing and assessment: An advanced resource book. Routledge.

[15]. Gronlund, N. E. (1976). Measurement and evaluation in teaching (3rd ed.). New York: Macmillan.

[16]. Gronlund, N. E. (1982). *Constructing achievement tests.* Prentice Hall.

[17]. Gronlund, N. E. (1998). Assessment of student achievement. Allyn & Bacon Publishing, Longwood Division, 160 Gould Street, Needham Heights, MA 02194-2310; tele.

[18]. Gronlund, N. E., & Linn, R. L. (1990). Measurement and evaluation in teaching (6th ed.). Macmillan.

[19]. Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items* (3rd ed.). Routledge.

[20]. Ibrahim, B. M., & Sani, M. I. (2025). Development of standardized mathematics achievement test instruments for measuring senior secondary school students' learning outcomes. *Eureka: Journal of Educational Research, 3*(2). https://doi.org/10.56773/ejer.v3i2.56

[21]. Khoa, N. V., & Le, P. (2026). Developing a multiple-choice question bank to assess learning outcomes for the course "Foundations of Mathematics in Elementary Education 1" at Ho Chi Minh City University of Education. *Vietnam Journal of Education.* https://www.vjol.info.vn/index.php/sphcm/article/view/131698

[22]. Kolagar, Z., Zalkow, F., & Zarcone, A. (2025). *Investigating methods for mapping learning objectives to Bloom's revised taxonomy in course descriptions for higher education.* In Proceedings of the 20th Workshop on Innovative Use of NLP for Building

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1916

Educational Applications (BEA 2025) (pp. 415–445). Association for Computational Linguistics.

[23]. Kumar, G., & Gangwar, S. (2025). Construction and standardization of achievement test: Achievement of B.Ed. pupil teachers in educational psychology. Far Western Journal of Education, 2(1), 42–56. https://nepjol.info/index.php/fwje/article/view/83293

[24]. Malau-Aduli, B. S., & Zimitat, C. (2012). Peer review improves the quality of MCQ examinations. *In Assessment and Evaluation in Higher Education* (Vol. 37, Issue 8, pp. 919–931). https://doi.org/10.1080/02602938.2011.586991

[25]. Mata, P. V., Abao, Q. R. H., Bihasa, R. K. S., Kiamco, G. N., & Neiz, A. M. C. (2026). Development and validation of curriculum-based problem-solving MCQ classroom tests in mathematics. *International Journal of Evaluation and Research in Education, 15*(1). https://doi.org/10.11591/ijere.v15i1.35698

[26]. Mehrens, W. A., & Lehmann, I. J. (1973). *Measurement and evaluation in education and psychology* (2nd ed.). Holt, Rinehart & Winston.

[27]. Mehrens, W. A., & Lehmann, I. J. (1978). Measurement and evaluation in education and psychology.

[28]. National Curriculum Framework (NCF). (2005). *National curriculum framework 2005*. New Delhi: NCERT.

[29]. National Education Policy (NEP). (2020). *National education policy 2020*. New Delhi: Ministry of Human Resource Development, Government of India.

[30]. Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

[31]. Putranta, H., & Supahar, S. (2019). Development of physics-tier tests (PysTT) to measure students' conceptual understanding and creative thinking skills: a qualitative synthesis. *Journal for the Education of Gifted Young Scientists,* 7(3), 747-775.

[32]. Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education, 4(1), 1301013.*

[33]. Ramatu, N., & Kongnyuy, P. (2025). An analysis of item difficulty and item discrimination of mathematics multiple choice items (2023 and 2024) in Bamboutos Division of Cameroon. *East African Journal of Education Studies, 8*(3), 505–511. https://doi.org/10.37284/eajes.8.3.3600

[34]. Thorndike, R. L., & Hagen, E. (1977). *Measurement and evaluation in psychology and education.* Wiley.

[35]. Treagust, D. F. (2006). Diagnostic assessment in science as a means to improving teaching, learning and retention. *In Proceedings of The Australian Conference on Science and Mathematics Education.*

[36]. Wilcox, B. R., Lauffer, H., & Gouvea, J. S. (2015). Designing multiple-choice assessments to probe student reasoning. *The Physics Teacher*, 53(9), 564–568.

[37]. Wordu, H., & Opusunju, M. O. (2025). Development, standardization and benchmark of mathematics proficiency test for senior secondary school students using item response theory. *UNIJERPS, 7*(5). https://mail.unijerps.org/index.php/unijerps/article/view/944

[38]. Zajda, J. (2006). Learning and teaching. James Nicholas Publisher Pty Ltd.