International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Development and Analysis of an Item Pool: Assessing Students' Mathematical Conceptual Attainment

Tenzin Pedon[1], Jatinder Grover[2]

[1]Senior Research Fellow, Department of Education, Panjab University, Chandigarh, India
[2]Professor, Department of Education, Panjab University, Chandigarh, India

Abstract
Many students find mathematics difficult, and a lack of mathematical conceptual understanding is one of the key reasons. Assessing students' conceptual knowledge helps in identifying their specific challenges in a particular topic. We develop a conceptual attainment test instrument to measure students' conceptual level in four mathematical topics at the secondary school level. The instrument consists of fifty-five multiple-choice questions (MCQ) and thirty true or false questions framed based on three lower levels of Bloom's taxonomy of learning. The purpose of this study is to develop the instrument, identify item characteristics, and determine its validity and reliability. The data was collected by random sampling from 124 students studying in ninth and tenth grades from two schools. The item analysis was done based on classical test theory, and the reliability of the items was found by calculating Cronbach's alpha coefficient. The overall findings led to the final item bank of fifty MCQ and fifteen true or false with Cronbach's alpha value 0.810, indicating high internal consistency reliability. This item bank can be used for formative assessment to test grade 9th students' conceptual understanding of each topic separately.

Keywords: Mathematical Conceptual Test, Item Analysis, Instrument Development

1. Introduction

Mathematics as a discipline has been very significant since the existence of civilisation, and it has evolved to a very advanced stage to date. The importance of this subject is universally accepted, and it has become a core subject in the school curriculum. However, studies [1] – [3] observed that many students find difficulties in learning mathematics. Students often face difficulties in learning mathematics for a variety of reasons, which can range from cognitive challenges to environmental factors such as the abstract nature of math, complex concepts, cognitive overload, math anxiety, lack of interest or motivation, mathematical language, logic and reasoning, and technical vocabulary, etc. Developing a strong foundation in mathematical concepts and logical structure can greatly enhance a student's ability to navigate and master this subject. Specifically, since the concepts of mathematical topics are interconnected, there are prerequisite concepts that must be mastered to study the next concepts. A conceptual understanding involves the ability to create, connect, and represent information when solving problems. Thus, conceptual understanding is a crucial skill that students must develop in mathematics to build on their prior knowledge and serve as a foundation for solving mathematical or

real-world problems [4]. Studies [5], [6] showed that the better the students' conceptual understanding, the better their mathematics learning outcomes. However, to assess which learning outcomes to measure and how to measure them best, it is important to design and validate assessment tasks that can reflect those measures adequately and accurately.

To achieve increased conceptual understanding, there need to be valid and reliable measures of conceptual understanding [7]. Previous studies on mathematical concept related tests vary in topic, level of schooling, and types of questions. Özyildirim-Gümüs and others (2022) constructed a concept achievement test for elementary students to test memorization and procedures with or without connections on equations, ratios, percentages, lines and angles, and polygons [8]. The items include MCQ, matching, and open-ended. [9] developed a mathematical achievement test (MAT) on number and numeration, algebraic processes, mensuration, plane geometry, trigonometry, and statistics. The type of questions used was problem-solving based open-ended. The targeted group was senior secondary students. Another study [10] developed an achievement test based on Bloom's taxonomy to evaluate students' concept of length measurement in second graders. This study used Test Analysis Program (TAP) and Tetrachoric Factor Analysis for data analysis based on Classical Test Theory.

Our study aims to develop an item pool that would assess students' conceptual understanding of the most common topics in school mathematics. Particularly for this study, we chose topics from ninth grade mathematics. Since at every grade, students learn about numbers, the first topic chosen was the number system. Second, we chose three topics from geometry whose concepts were linked: lines and angles, congruence of triangles, and quadrilaterals. According to the Central Board of Secondary Education's (CBSE) assessment framework for science, mathematics, and English for Classes 6–10 [11], students' outcomes in mathematics are to be assessed with two main objectives. The first objective is to demonstrate knowledge and understanding of mathematical ideas, techniques, and procedures, and the second is to apply this knowledge and understanding to classroom and real-world situations. It was noted by this framework that such assessments would support a move to competency-based education by assessing higher-order thinking skills as well as the underpinning knowledge. This shows that the CBSE aims to assess students based on Bloom's taxonomy on cognitive learning. Bloom's taxonomy consisted of six levels of learning: knowledge, comprehension, application, analysis, synthesis, and evaluation. Following the objectives of CBSE for the assessment of students' mathematics standards and considering the objectives of this study, we delimit our assessment criteria to the three lower levels of Bloom's taxonomy: knowledge, comprehension, and application. In the revised taxonomy, knowledge is at the basis of these six cognitive processes, but knowledge is categorised into four types: factual, conceptual, procedural, and metacognitive knowledge [12]. Our items were framed to assess students' factual knowledge and conceptual knowledge of a particular mathematics idea or topic. While factual knowledge refers to knowledge of terminology, knowledge of specific details and elements, conceptual knowledge refers to knowledge of classifications, knowledge of principles and generalisations, and knowledge of theories, models, and structures [12]. In addition, items were also framed to assess students' understanding of mathematical concepts and the application of this knowledge.

For an effective test development, Downing [13] provided systematic guidance consisting of twelve steps. These steps involve i) an overall plan, ii) content definition, iii) test specifications, iv) item development, v) test design and assembly, vi) test production, vii) test administration, viii) scoring test responses, ix) passing scores, x) reporting test results, xi) item banking, and xii) test technical reports. Particularly, creating and developing effective test items that accurately measure important content at the

right cognitive level is one of the biggest challenges for test developers. Hence, with ninety years of effective use and an extensive research basis on multiple-choice format, Downing [13] suggested the multiple-choice format for the item development. Multiple-choice question (MCQ) is the most common objective-type item. It is generally recommended that one use four or five choices per question, as using fewer alternatives often results in items with inferior characteristics [14]. In our study, we develop items of two types of multiple-choice questions and true or false items. First, MCQ type-1 (MCQ-1) consisted of multiple-choice questions with four choices out of which one is the key. Second, MCQ type-2 (MCQ-2) consisted of multiple-choice questions with five choices that had multiple keys. A key in MCQ is the best or correct answer, while other options are known as distractors [15].

In developing a valid and reliable item bank, it is significant to perform an item analysis. Generally, two common theories support the development of measurement tests: classical test theory and item response theory [16]. Classical Test Theory (CTT) has served as the cornerstone of measurement theory for many years. The core principles, assumptions, and extensions of CTT have enabled the creation of psychometrically robust scales used in educational assessment practices. At the theoretical level, a typical item analysis based on CTT involves key concepts such as item difficulty, item discrimination, reliability, and the standard error of measurement [17]. Since our item bank consists mainly of MCQ, the key to good quality MCQ is based on the availability of good distractors, as it can discriminate between the informed and uninformed students [18]. Another aspect to consider in classical test theory is predictive validity, which is defined as the Pearson correlation between the test score and the validation criteria score [19]. The value of Pearson's product moment correlation when one of the variables is dichotomous, and the other is metric is known as point-biserial correlation [20]. The point-biserial correlation is employed in psychometric item analysis to evaluate the relationship between the total score (excluding the item in question) and the score of a dichotomous item [21]. Validity, from a broad perspective, refers to the evidence we have to support a given use or interpretation of test scores [22]. Test score reliability is a component of validity. Reliability refers to the consistency of a test both within itself and over time, and it can be evaluated through statistical calculations that examine individual items as well as the overall test [23]. If test scores lack reliability, they cannot be considered valid because they will fail to accurately estimate the ability or trait that the test is designed to measure. Reliability is therefore a necessary but not sufficient condition for validity [22].

After several considerations, the researchers are motivated to develop the item pool and analyse the items by considering classical test theory as well as by calculating the point biserial correlation and the reliability of the item bank. The purpose of the study was to develop a valid and reliable item bank to assess students' conceptual attainment of mathematics at the secondary school level. Besides, the study also tried to identify the inappropriate items in the pool for further revision or elimination depending on the item's overall characteristics.

2. Method

2.1 Development of instrument and Procedure

Cohen and Wollack [14] asserted that creating a test blueprint is a prerequisite to developing the test effectively. The test goals, skills to be tested, and the proportional weight of the test assigned to each are all listed in the test blueprint. Thus, firstly, the researchers established the question paper design and the basis of the item bank according to the twelve components of test development that provide a framework for test development and validation [13]. The purpose of the test is to examine ninth graders' concepts

on four mathematical topics as mentioned in the introduction. The evaluation basis was delimited to the three lower levels of Bloom's taxonomy. For each topic, three types of questions were formulated: multiple-choice questions with one correct answer (MCQ-1), multiple-choice questions with multiple correct answers (MCQ-2), and true/false questions (T/F).

The MCQ-1 consisted of questions that assessed the students' conceptual knowledge, understanding of the concept, and application of the concept. Each question has four optional answers, out of which only one is the correct answer. MCQ-2 consisted of multiple-choice questions with five choices, and there was a random number of correct options. Students have to select all the relevant choices that answer the question. The purpose of inclusion of MCQ-2 was to check the in-depth conceptual attainment of a given topic. In the MCQ-1, students might select the correct answer by luck or by guess even though they were not aware of the correct answer. However, MCQ-2 reduces this limitation, as students had to select all the relevant correct options that the question was asking. This type of question helps to examine students' actual knowledge and understanding of a specific concept. The inclusion of a True/False type question is to check the students' assurance of a particular concept. This section not only asks the truth or falsity of a statement, but students are asked to give a short reason or an example to support their answer. The test items were developed following the curriculum specification of ninth grade mathematics set by the CBSE.

The item pool was constructed by referring to the standard mathematical textbooks published by the National Council of Educational Research and Training (NCERT) and multiple sources that produce practice papers on school-based mathematics. The researchers analysed the related topics taught at school level, studied various questions on the selected topics, and finally constructed the items as per the purpose of the present study. Finally, the preliminary question bank was developed by constructing 100 items measuring students' knowledge, understanding, and application of the four selected topics.

2.2 Subject Matter Expert (SME) Panel Review

A panel of subject matter experts was formed to review the initial item pool developed by the researchers. The panel was selected purposefully by considering their specialisation in the subject matter, teaching experience at the secondary school level, and knowledge of the population to be studied. For this study, there were five experts from three different schools affiliated with CBSE. As per the responsibilities of the SME recommended by [24], the panel of SME assessed the construct validity and their alignment with the existing measures, examined the content-related questions, and suggested changes and modifications needed for each item.

The question evaluation grid designed by Caligiore-Gei and Ison [25] was used by the SME to rate each of the questions according to how appropriately they were formulated, taking into account the criteria of clarity, relevance, and sufficiency. These criteria were defined as a) Clarity—degree to which the semantics and syntactic of the question are easily understood; b) Relevance—degree to which the inclusion of the question is relevant to the dimension of category to be evaluated; c) Sufficiency—whether the questions included are enough to obtain the measurement of the particular dimension [25]. The experts rate each question by giving a score based on scoring options: non-compliance with the criterion: 0, compliance at a low level: 1, compliance at a moderate level: 2, or compliance at a high level: 3.

Based on their suggestions, the item bank was finally prepared with eighty-five items: forty questions of MCQ type 1, fifteen questions of MCQ type 2, and thirty questions of T/F type. Each question of MCQ-1 and T/F carries one mark, and each question of MCQ-2 carries two marks, which leads to the total ma-

rk of 100 for eighty-five questions. The detailed blueprint of the question paper is shown in Table 1.

Table 1: Blueprint of the Test Paper

Topics	MCQ-1	MCQ-2	T/F	Total
Number System	9 (9)	4 (8)	7 (7)	20 (24)
Lines and Angles	9 (9)	4 (8)	7 (7)	20 (24)
Triangles	9 (9)	3 (6)	7 (7)	19 (22)
Quadrilaterals	13 (13)	4 (8)	9 (9)	26 (30)
Total	40 (40)	15 (30)	30 (30)	85 (100)

2.3 Sample and Participants

As the study aims to test the concept attainment of particular topics at the grade 9 level, fifty percent of the sample was selected from ninth grade who was at the time of sampling, studying those topics, and fifty percent from tenth grade who had already studied the topics. Thus, a pilot study for this achievement test was conducted with 124 students from two randomly selected schools in Leh, Ladakh, affiliated to the CBSE, India. The test was administered as in the real examination situation in terms of seating arrangement and time allocation. The time permitted to complete the test was one and a half hours.

2.4 Item Analysis

The final analysis of the item was done on 107 responses, as seventeen responses were discarded because of not attempting more than 50% of the test items. To determine which items will be the best to construct the most efficient and reliable test, we conducted an item analysis by calculating the item difficulty index (*p*), item discrimination index (*D*), item distractor index, point-biserial correlation coefficient (*rpbis*), and the reliability of each item. The analysis of the data was done by using MS Excel and IBM Statistical Package for the Social Sciences (SPSS version 25.0).

The item difficulty index (*p*) presents the percentage of students who got the correct answer compared to the total number of students [16]. The range of an acceptable p-value for an item varies according to the type of questions and evaluation criteria. In this study, the difficulty index was determined by using the formula P=R/T, where P is the item difficulty index, R is the number of correct responses, and T is the total number of responses [26]. Based on the difficulty index of each item, we classify items according to the following rules: p: 0.0–0.2, very hard; 0.2–0.4, hard; 0.4–0.6, medium; 0.6–0.8, easy; 0.8–1.0, very easy. The optimal range is (20–80%); a low index may mean that students are attempting the item but are getting it wrong, and a too-high index may mean that regardless of whether poor or good, students are able to get it correct [27].

The item discrimination index (*D*) was calculated to determine how well each item in a test distinguishes between higher-achieving and lower-achieving students. The item discrimination index measures the differences between the percentages of students in the upper group with that of the lower group who obtained the correct responses [26]. In this study, we included the top 27% of the students in the upper group and the bottom 27% in the lower group as suggested by Kelley [1939, as cited in 26] to compute item discrimination. We calculated the value of D by using the formula, where HG is the number of correct responses in the upper group, LG is the number of correct responses in the lower group, and N is the total number of responses. Items were categorised based on guidelines provided by Ebel and Frisbie

[28] as follows: D: negative value, item to be discarded; D: 0.0–0.19, poor item—to be revised; D: 0.2–0.29, acceptable; D: 0.3–0.39, good; D: >0.4, excellent.

Item distractors are the options in the multiple-choice answers that are not the correct answer. Distractor Efficiency (DE) is calculated as a non-functional distractor (NFD) from the distractor that has been selected by less than 5% of the students [18]. Nonfunctional distractors (NFD) are those distractors that give a positive index or zero index. A functional distractor was defined as one that exhibited negative discrimination and was selected by at least 5% of the participants [29]. Distractor efficiency (DE) is expressed as 0%, 33.3%, 66.6%, and 100% depending on the number of NFDs as 3, 2, 1, and 0, respectively [18]. The distractor efficiency is interpreted as low (having 3–4 NFDs), medium (having 1-2 NFDs), and high (having 0 NFD) [30]. This study graded the distractor efficiency of items as poor (3 NFDs), low (2 NFDs), medium (1 NFD), and high (0 NFD). Items with three NFDs are considered not acceptable.

To determine the correlations between the success of students on the item and their success on the whole test, we used the point-biserial correlation coefficient (r_{pbis}). Point-biserial correlation shows how much predictive power an item has and how the item contributes to predictions by estimating the correlation between each test item and the total test score [31]. It reveals how effectively an item measures or discriminates in comparison to the rest of the test. The value of the point-biserial correlation coefficient ranges between -1 and 1. The point-biserial correlation coefficient is considered significant if $rpbis \geq 0.2$ in SPSS in both 0.01 and 0.05 levels of significance. In this study, we interpreted the coefficient as follows: 0-0.10, negligible; 0.10-0.39, weak; 0.40-0.69, moderate; 0.70-0.89, strong; and 0.9-1.00, very strong correlation [32].

To determine the reliability or the items, we measure the internal consistency of the test by using Cronbach's alpha. An extremely common way of evaluating reliability is the internal consistency index, called KR-20 or α (alpha) [22]. Cronbach's alpha reliability is one of the most widely used measures of reliability in the social and organizational sciences [33] that provides a measure of the internal consistency of a test or scale [34]. In this study, the reliability of each item is categorized on the basis of rules provided by Streiner (2003): $\alpha \geq 0.9$ - excellent reliability; $0.7 \leq \alpha < 0.9$ – good; $0.6 \leq \alpha < 0.7$ – acceptable; $0.5 \leq \alpha < 0.6$ – poor; and $\alpha < 0.5$ – unacceptable.

3. Results

3.1 Item analysis

Item analysis refers to a set of descriptive statistics that are useful during the process of developing an item pool for a new test [35]. In this study, dichotomous analysis of test items was used because the types of tests used had multiple-choice questions and true or false questions, both of which had responses either correct or incorrect. The correct answers were given value 1, and the incorrect answers were given value 0. Overall, eighty-five items were analysed based on classical test theory by calculating difficulty index, discrimination index, and distractor efficiency.

3.1.1 Difficulty Indices of the Items

The items were categorised into five difficulty levels, ranging from very easy to very hard. From the analysis, six items) showed a difficulty index of less than 0.2, which implies a very hard difficulty level, and three items were found to be very easy with a difficulty index of more than 0.8. The details of the each item are shown in Table 2. As items with moderate difficulty were preferred compared to those with extreme difficulty levels [36], this study took the optimal range as 0.2–0.8 as suggested by [27].

The items beyond this range were either revised or removed depending on other factors of the analysis. In the acceptable range, there were twenty-eight items of hard level with p-value between 0.2 and 0.4. Twenty-seven items have p-value between 0.4 and 0.6, categorizing them into medium level. With p-value between 0.6 and 0.8, there are twenty-one items of easy level. Questions numbered Q are MCQ-1, those numbered M are MCQ-2, and those with T are of true/false type.

Table 2: Difficulty Levels of the Items

Difficulty Index	Difficulty level	Items	Total
0.0 – 0.2	Very Hard	Q32, Q36, Q38, Q40, M3, M14,	6
0.2 – 0.4	Hard	Q4, Q6, Q8, Q18, Q22, Q23, Q26, Q27, Q29, Q31, Q33, Q35, Q39, M2, M4, M5, M6, M7, M8, M10, M11, T1, T8, T10, T12, T15, T17, T20,	28
0.4 – 0.6	Medium	Q1, Q5, Q10, Q12, Q15, Q21, Q24, Q25, Q30, Q37, M1, M9, M12, M13, M15, T2, T3, T5, T9, T16, T18, T21, T24, T26, T27, T29, T30,	27
0.6 – 0.8	Easy	Q2, Q3, Q7, Q11, Q13, Q16, Q19, Q21, Q28, Q34, T4, T6, T7, T11, T13, T14, T19, T22, T23, T25, T28,	21
0.8 – 1.0	Very Easy	Q9, Q14, Q17	3

3.1.2 Index of Discrimination

Considering the indices of discrimination (D-value) of the items, a total of fourteen items showed a very good discriminant index of value above 0.40. Eighteen items had a reasonably good D-value between 0.3 and 0.39. The analysis revealed twenty-three marginal items with D-value between 0.2 and 0.29. Thirty items have a discrimination index of less than 0.19 categorising them into poor items. The category of each item is shown in table 3. Ebel and Frisbie [28] suggested a need for improvement for the marginal items and rejected or improved the poor items by revision. Hence, the poor items were subjected to removal or improvement depending on other criteria of the analysis.

Table 3: Discrimination Indices of the Test Items

D-Value	Interpretation	Items	Total
0.4 and above	Very Good	Q13, Q18, Q20, Q24, Q29, Q34, T25, T29, M1, M6, M7, M9, M13, M15,	14
0.3 – 0.39	Reasonably Good	Q3, Q5, Q14, Q17, Q19, Q21, Q25, Q26, Q27, Q37, T1, T2, T9, T24, T30, M8, M11, M12,	18
0.2 – 0.29	Marginal item	Q1, Q4, Q6, Q8, Q9, Q10, Q11, Q15, Q31, Q33, Q35, T3, T5, T6, T8, T11, T18, T21, T22, T27, T28, M2, M10,	23
0.19 or less	Poor Item	Q2, Q7, Q12, Q16, Q22, Q23, Q28, Q30, Q32, Q36, Q38, Q39, Q40, T4, T7, T10, T12, T13, T14, T15, T16, T17, T19, T20, A23, A26, M3, M4, M5, M14	30

3.1.3 Item Distractor Analysis

In this study, there were three distractors for each question in MCQ type-1 and different number distractors for each question in MCQ type-2 as these questions have multiple answers, providing 145 distractors. On analysing fifty-five multiple-choice items, we found thirty-two items with 100% distractor efficiency (DE), seventeen items with 66.6% DE, and six items with 33% DE. None of the distractor has 0% DE. This shows that of 145 distractors, 136 were functional. More the functional distractors, the better the quality of the item, as NFDs did not serve any useful purpose [37].

3.2 Point Bi-Serial Correlation of Items

The items were categorised based on the interpretation suggested by Evans [1996, as cited in 38]. The correlation of items was interpreted as follows: *rpbis*: negative value, items to be discarded; 0.0–0.19, very weak; 0.2–0.39, weak; 0.4–0.59, moderate; 0.6–0.79, strong; and 0.8–1.0, very strong. According to the result, there are eight items with moderate correlation, forty-three items with weak correlation, thirty-one with very weak correlation, and three items with negative correlation. This showed that fifty-one items have an acceptable correlation coefficient. Two items (A10, A15) were discarded, while item Q39 was revised as other analysis criteria of this item showed an acceptable result.

3.3 Reliability and Validity

The validity of a research instrument evaluates how accurately it measures what it is intended to measure. In developing this item pool, the content validity, face validity, and construct validity were assured by the evaluation and validation of subject matter experts. The content validity ensured that the test included an adequate set of items that represented the domain of the concept being measured. Face validity refers to the extent to which a test seems to measure what it is intended to measure. Construct validity was concerned with the extent to which a test measures a specific trait or construct. The suggestions and reviews by SMEs were taken into consideration in modifying and validating the items. To produce the trustworthiness of the test, we used Cronbach's alpha coefficient to determine the internal consistency of the item scale. The initial Cronbach's alpha coefficient for the whole item bank is 0.788, with each item having an alpha value greater than 0.7, signifying that the test has good reliability [39]. The Cronbach alpha value if item deleted of each item is shown in Table 4.

Table 4: Overall Item Analysis and Item Selection

Item	P-value	D-Value	No. of NFD	*rpbis*	C-Alpha if Item Deleted	Selection
Q1	0.55 Medium	0.31 Good	1 Medium	.271** Weak	.786 Good	Selected
Q2	0.64 (Easy)	0.24 Marginal	0 High	.106 V Weak	.789 Good	Revised
Q3	0.64 Medium	0.37 Good	0 High	.255** Weak	.786 Good	Selected
Q4	0.38 (Easy)	0.34 Good	0 High	.228* Weak	.787 Good	Selected
Q5	0.44 V Easy	0.34 Good	0 High	.336** Weak	.785 Good	Revised

Q6	0.26 Medium	0.28 Marginal	0 High	.252** Weak	.786 Good	Selected
Q7	0.75 (Easy)	0.17 Poor	1 Medium	.169 V Weak	.788 Good	Revised
Q8	0.36 V Easy	0.31 Good	1 Medium	.208* Weak	.787 Good	Revised
Q9	0.67 Medium	0.48 V Good	1 Medium	.284* Weak	.786 Good	Selected
Q10	0.53 Medium	0.48 V Good	0 High	.316** Weak	.785 Good	Selected
Q11	0.77 (Easy)	0.34 Good	0 High	.249* Weak	.787 Good	Selected
Q12	0.41 Medium	0.10 Poor	1 Medium	.071 V Weak	.790 Good	Revised
Q13	0.74 (Easy)	0.45 V Good	0 High	.380** Weak	.784 Good	Selected
Q14	0.86 V Easy	0.31 Good	0 High	.390** Weak	.785 Good	Revised
Q15	0.55 Medium	0.17 Poor	2 Low	.201* Weak	.788 Good	Revised
Q16	0.64 (Easy)	0.48 V Good	1 Medium	.415** Moderate	.783 Good	Selected
Q17	0.86 V Easy	0.28 Marginal	0 High	.339** Weak	.785 Good	Revised
Q18	0.21 (Hard)	0.17 Poor	0 High	.009 V Weak	.789 Good	Revised
Q19	0.62 (Easy)	0.31 Good	0 High	.232* Weak	.787 Good	Selected
Q20	0.67 (Easy)	0.45 V Good	0 High	.423** Moderate	.783 Good	Selected
Q21	0.48 Medium	0.62 V Good	1 Medium	.325** Weak	.785 Good	Selected
Q22	0.21 (Hard)	0.17 Poor	0 High	.194 V Weak	.788 Good	Revised
Q23	0.31 (Hard)	0.14 (Good)	0 High	.124 V Weak	.789 Good	Revised
Q24	0.51 Medium	0.48 V Good	2 Low	.361** Weak	.784 Good	Revised
Q25	0.55 Medium	0.35 Good	1 Medium	.233* Weak	.787 Good	Selected
Q26	0.35 (Hard)	0.34 Good	0 High	.244* Weak	.787 Good	Selected

Q27	0.31 (Hard)	0.21 Marginal	2 Low	.257** Weak	.786 Good	Selected
Q28	0.76 (Easy)	0.14 Poor	1 Medium	.108 V Weak	.789 Good	Revised
Q29	0.27 (Hard)	0.48 V Good	0 High	.406** Moderate	.783 Good	Selected
Q30	0.57 Medium	0.48 V Good	0 High	.173 V Weak	.788 Good	Revised
Q31	0.37 (Hard)	0.28 Marginal	0 High	.166 V Weak	.788 Good	Revised
Q32	0.21 (Hard)	0.17 Poor	1 Medium	.182 V Weak	.788 Good	Revised
Q33	0.21 (Hard)	0.17 Poor	0 High	0.204* Weak	.787 Good	Revised
Q34	0.68 (Easy)	0.59 V Good	1 Medium	.433** Moderate	.783 Good	Selected
Q35	0.31 (Hard)	0.28 Marginal	0 High	.201* Weak	.787 Good	Selected
Q36	0.17 V Hard	0.17 Poor	1 Medium	.112 V Weak	.789 Good	Discarded
Q37	0.44 (Medium)	0.35 Good	0 High	.271* Weak	.786 Good	Selected
Q38	0.15 (V Hard)	0.10 Poor	0 High	.160 V Weak	.788 Good	Discarded
Q39	0.37 (Hard)	0.03 Poor	1 Medium	-.079 V Weak	.793 Good	Revised
Q40	0.17 (V Hard)	0.31 Good	0 High	.139 V Weak	.788 Good	Revised
M1	0.46 Medium	0.53 V Good	0 High	.487** Moderate	.779 Good	Selected
M2	0.32 Hard	0.22 Marginal	1 Medium	.289** Weak	.786 Good	Selected
M3	0.16 V Hard	0.03 Poor	1 Medium	.085 V Weak	.790 Good	Discarded
M4	0.26 Hard	0.10 Poor	1 Medium	.194 V Weak	.789 Good	Revised
M5	0.24 Hard	0.14 Poor	2 Low	.100 V Weak	.791 Good	Discarded
M6	0.36 Hard	0.45 V Good	0 High	.427** Moderate	.781 Good	Selected

M7	0.38 Hard	0.43 V Good	0 High	.371** Weak	.784 Good	Selected
M8	0.27 Hard	0.31 Good	0 High	.244* Weak	.788 Good	Selected
M9	0.54 Medium	0.62 V Good	0 High	.528** Moderate	.777 Good	Selected
M10	0.29 Hard	0.22 Marginal	0 High	.250** Weak	.788 Good	Selected
M11	0.39 Hard	0.38 Good	2 Low	.363** Weak	.784 Good	Selected
M12	0.32 Hard	0.33 Good	0 High	.330** Weak	.785 Good	Selected
M13	0.48 Medium	0.51 V Good	2 Low	.448** Moderate	.780 Good	Selected
M14	0.17 V Hard	0.22 Marginal	0 High	.305** Weak	.785 Good	Revised
M15	0.49 Medium	0.48 V Good	1 Medium	.365** Weak	.784 Good	Selected
T1	0.34 (Hard)	0.34 Good		.312** Weak	.785 Good	Selected
T2	0.57 Medium	0.31 Good		.242* Weak	.787 Good	Selected
T3	0.46 Medium	0.28 Marginal		.233* Weak	.787 Good	Selected
T4	0.67 (Easy)	- 0.03 Poor		.025 V Weak	.791 Good	Discarded
T5	0.36 (Hard)	0.21 Marginal		.137 V Weak	.789 Good	Revised
T6	0.47 Medium	0.31 Good		.226 Weak	.787 Good	Selected
T7	0.71 (Easy)	0.28 Marginal		.152 V Weak	.789 Good	Revised
T8	0.63 (Easy)	0.24 Marginal		.201* Weak	.787 Good	Revised
T9	0.58 Medium	0.31 Good		.247** Weak	.787 Good	Selected
T10	0.34 (Hard)	0.03 Poor		-.045 V Weak	.792 Good	Discarded
T11	0.75 (Easy)	0.17 Poor		.193 V Weak	.788 Good	Discarded

T12	0.38 (Hard)	0.21 Marginal		.153 V Weak	.789 Good	Revised
T13	0.66 (Easy)	0.14 Poor		.106 V Weak	.789 Good	Discarded
T14	0.71 (Easy)	0.28 Marginal		.171 V Weak	.788 Good	Revised
T15	0.35 (Hard)	0.10 Poor		-.026 V Weak	.792 Good	Discarded
T16	0.50 (Medium)	0.17 Poor		.091 V Weak	.790 Good	Discarded
T17	0.26 (Hard)	0.10 Poor		.083 V Weak	.790 Good	Discarded
T18	0.55 Medium	0.14 Poor		.133 V Weak	.789 Good	Discarded
T19	0.75 (Easy)	0.14 Poor		.070 V Weak	.790 Good	Discarded
T20	0.30 (Hard)	0.10 Poor		.190 V Weak	.790 Good	Discarded
T21	0.45 Medium	0.24 Marginal		.168 V Weak	.788 Good	Revised
T22	0.70 (Easy)	0.10 Poor		.162 V Weak	.788 Good	Discarded
T23	0.64 (Easy)	0.14 Poor		.097 V Weak	.790 Good	Discarded
T24	0.56 Medium	0.52 V Good		.340** Weak	.784 Good	Selected
A25	0.75 (Easy)	0.48 V Good		.345** Weak	.785 Good	Selected
T26	0.52 Medium	0.10 Poor		.035 V Weak	.791 Good	Discarded
T27	0.52 Medium	0.14 Poor		.181 V Weak	.788 Good	Discarded
T28	0.63 (Easy)	0.34 Good		.180 V Weak	.788 Good	Revised
T29	0.55 Medium	0.34 Good		.303** Weak	.785 Good	Selected
T30	0.56 Medium	0.38 Good		.301** Weak	.785 Good	Selected

** Correlation is significant at 0.01 level (2-tailed), * Correlation is significant at 0.05 level (2-tailed)

3.4 Item Selection

The final selection of items was based on all the criteria of the item analysis considered in this study: difficulty level, discrimination level, distraction efficiency, point-biserial correlation coefficient, and internal consistency of the items. Of these five criteria, items having all the criteria lying in the acceptable range were selected, items having at least three criteria with unacceptable values were discarded, and items having one or two criteria with unacceptable values were revised and improved. In particular, items that lie in at least three of the following criteria were discarded: very easy or very hard difficulty level; poor discrimination index; low number of functional distractors; very weak correlation coefficient; and unacceptable alpha value. Items that lie in any one or two of the above ranges were revised.

Thereby, eighteen items (Q36, Q38, M3, M5, T4, T10, T11, T13, T15, T16, T17, T18, T19, T20, T22, T23, T26, T27) were eliminated. Thirty-nine items (Q1, Q3, Q4, Q6, Q9, Q10, Q11, Q13, Q16, Q19, Q20, Q21, Q25, Q26, Q27, Q29, Q34, Q35, Q37, M1, M2, M6, M7, M8, M9, M10, M11, M12, M13, M15, T1, T2, T3, T6, T9, T24, T25, T29, T30) were found to have all the criteria lying within the optimum range, thus were selected. The remaining twenty-eight items (Q2, Q5, Q7, Q8, Q12, Q14, Q15, Q17, Q18, Q22, Q23, Q24, Q28, Q30, Q31, Q32, Q33, Q39, Q40, M4, M14, T5, T7, T8, T12, T14, T21, T28) were further revised for improvement as these items had one or two criteria with unacceptable value. After eliminating the undesirable items, the alpha value of the final item bank that consists of sixty-five items has increased to 0.810.

4. Discussion

In our study, we developed an item pool to test students' mathematical conceptual learning in four topics at the secondary school level. The learning outcome was assessed on three lower levels of the cognitive domain according to Bloom's taxonomy: conceptual knowledge, conceptual understanding, and application of the concepts. Initially, the item bank consisted of one hundred questions but was reduced to eighty-five items after the subject experts' review. The items were of three types: MCQ with one key, MCQ with multiple keys, and true or false items. Multiple-choice questions have now become the most widely applicable, useful, and accepted type of objective assessment as they help assess all-important facets of educational outcomes [40]. The items were analysed on the basis of classical test theory and by calculating Cronbach's alpha to check the internal consistency of the items.

The analysis of items showed that 89% of items have a difficulty index from 0.2 to 0.8, signifying that the majority of test items have an optimum difficulty level. This observation was similar to a study on the development of an instrument of mathematical learning at the high school level that reported 75% of items with a moderate difficulty index [41]. In addition, 65% of items have an acceptable discrimination index of value more than 0.2. Thus, it showed that most of the items were good or satisfactory and would not need any revision as they were able to differentiate between good and weak students. Furthermore, a good item offers equally attractive alternatives to students who do not know the answer. The distractor efficiency of the item also has an impact on its difficulty and discriminatory index. Higher the number of NFDs in an item, the lower will be its discriminatory index and may render the MCQs more difficult [42]. However, for this item pool, more than half of the items have functional distractors, and no item has all distractors non-functional. Only six of them have two NFDs, which will be either discarded or revised.

The relationship between the item difficulty index and discrimination index for each test item was deter-

mined by Pearson correlation analysis using SPSS. Point biserial correlation is a true Pearson product-moment correlation that shows the correlation between the right/wrong scores that students receive on a given item and the total scores that the students receive when summing up their scores across the remaining items [43]. According to our findings, statistically, thirty-three items were significant at the 0.01 level and fourteen items were significant at the 0.05 level. Although studies have shown that "good" items have point-biserials above 0.25, a point-biserial value of at least 0.15 is recommended [43]. Considering this recommendation, only nine of fifty-five MCQs and twelve of thirty true/false questions had point-biserial value less than 0.15. This shows that 75% of items have an acceptable correlation between the students' performance in each item and their overall performance.

In quantitative research, reliability refers to the consistency, stability, and repeatability of results. Better the reliability is performed, more accurate the results are [44]. To examine the internal consistency or reliability of summated rating scales, Cronbach's alpha is used [34]. Initially, the Cronbach's alpha coefficient of eighty-five items in our study is 0.788, calculated by using SPSS version 25. To answer what value of alpha is desirable, Taber [45] noted from many studies that there is no general level (such as 0.70) where alpha becomes acceptable but a high value of alpha was 'desirable' when an instrument was used to assign a score to an individual. To increase alpha, more related items testing the same concept should be added to the test [46], but this leads to an inefficient redundancy [34]. In our study, we did not add any items; rather, we retained, revised, or eliminated items based on cumulative results provided by the item analysis criteria that were used in this study. After discarding undesirable items (2 items of MCQ-1, 2 items of MCQ-2, 14 items of T/F) that do not lie in the acceptable range in the remaining criteria, the alpha value increased to 0.808 for 67 items. From the twenty items that need revision, two items (Q12, T6) were further eliminated because of very weak correlation coefficient. Thus, a final tool of sixty-five items (fifty MCQ and fifteen T/F) were standardized with a high-reliability index of alpha value 0.810. According to Streiner (2003), alpha values of more than 0.8 are considered to have good reliability [39]. Thus, the reliability of the test developed in this study meets the criteria for a reliable test.

5. Conclusion

Our study aimed to develop an item pool to assess secondary school students' mathematical conceptual learning across four topics, focussing on three lower levels of Bloom's taxonomy. The initial item bank of 100 questions was refined to 85 items after expert review, consisting of MCQs (single and multiple keys) and true/false items. The item analysis based on classical test theory and Cronbach's alpha assessed the item characteristics and internal consistency.

Results showed that 89% of items had an optimal difficulty index (0.2-0.8), and 65% had a satisfactory discrimination index (>0.2), indicating good differentiation between strong and weak students. The distractor efficiency analysis found that more than half of the MCQs had functional distractors, with a few requiring revision. Pearson correlation analysis indicated that 75% of items had an acceptable point-biserial correlation (>0.15), aligning individual item performance with overall test performance. Cronbach's alpha initially was 0.788, but after removing 18 underperforming items, it improved to 0.810, suggesting the test had good reliability. The study concludes that the final 65-item pool consisting of 50 MCQ and 15 true or false items is reliable for assessing mathematical conceptual learning in four selected topics at the secondary school level. This item pool can be used by teachers to evaluate students' conceptual attainment in the selected mathematical topics.

REFERENCES

1. O. T. Kaufmann and A. Ryve, "Teachers' framing of students' difficulties in mathematics learning in collegial discussions", *Scandinavian Journal of Educational Research*, vol. 67, no. 7, pp. 1069-1085, Sep. 2022 https://doi.org/10.1080/00313831.2022.2115134
2. N. Guner, "Difficulties Encountered by High School Students in Mathematics", *International Journal of Educational Methodology*, vol. 6, no. 4, pp. 703-713. Nov. 2020. https://eric.ed.gov/?id=EJ1278427
3. L. Salihu, M. Aro, and P. Rasanen, "Children with learning difficulties in mathematics: Relating mathematics skills and reading comprehension", *Issues in Educational Research*, vol. 28, no. 4, pp. 1024–1038, 2018. https://search.informit.org/doi/abs/10.3316/ielapa.022495701125234
4. L. Mason, "High school students' beliefs about maths, mathematical problem solving, and their achievement in maths: A cross-sectional study", *Educ. Psychol. (Lond.)*, vol. 23, no. 1, pp. 73–85, Jan. 2003. https://psycnet.apa.org/doi/10.1080/01443410303216
5. K. Luneta, "Understanding students' misconceptions: An analysis of final Grade 12 examination questions in geometry", *Pythagoras*, vol. 36, no. 1, Jun. 2015.
6. D. P. N. Ningrum, B. Usodo, and S. Subanti, "Students' mathematical conceptual understanding: What happens to proficient students?", in *INTERNATIONAL CONFERENCE OF MATHEMATICS AND MATHEMATICS EDUCATION (I-CMME) 2021*, Surakarta, Indonesia, 2022.
7. M. J. Bisson, C. Gilmore, M. Inglis, and I. Jones, "Measuring conceptual understanding using comparative judgement", *Int. J. Res. Undergrad. Math. Educ.*, vol. 2, no. 2, pp. 141–164, Jul. 2016.
8. Özyildirim-Gümüs, F., Sarpkaya-Aktas, G., & Karaca, H. (2022). Investigation of Achievement Tests Prepared by Elementary Mathematics Teachers and Preservice Teachers. *Acta Didactica Napocensia*, *15*(1), 124-141.
9. Osakuade, Joseph Oluwatayo. "Development and Standardization of Mathematics Achievement Test for Unified Senior Secondary School Class 2 Promotion Examination in Ondo State, Nigeria." *J Adv Educ Philos* 8, no. 3 (2024): 99-106.
10. Rençber, Gizem Nur, and Galip Genç. "Developing an achievement test on length measurement." *Education Mind* 3, no. 3 (2024): 389-405.
11. L. Johnson, *Security controls evaluation, testing, and assessment handbook*, 2nd ed. San Diego, CA: Academic Press, 2019.
12. L. O. Wilson, *Anderson and Krathwohl-Bloom's taxonomy revised. Understanding the new version of Bloom's taxonomy*. 2016.
13. S. M. Downing, *Twelve steps for effective test development. Handbook of test development*. Lawrence Erlbaum Associates, 2006.
14. A. S. Cohen and J. A. Wollack, *Handbook on test development: Helpful tips for creating reliable and valid classroom tests*. Madison, WI, USA, 2015.
15. A. R. Delgado and G. Prieto, "Further evidence favoring three-option items in multiple-choice tests", *Eur. J. Psychol. Assess.*, vol. 14, no. 3, pp. 197–201, Sep. 1998.
16. L. H. Haw, S. B. Sharif, and C. G. K. Han, "Analyzing the science achievement test: Perspective of classical test theory and Rasch analysis", *Int. J. Eval. Res. Educ. (IJERE)*, vol. 11, no. 4, p. 1714, Dec. 2022.

17. G. Janssen, V. Meier, and J. Trace, "Classical test theory and item response theory: Two understandings of one high-stakes performance exam", *Colomb. Appl. Linguist. J.*, vol. 16, no. 2, p. 167, Sep. 2014.
18. I. Burud, K. Nagandla, and P. Agarwal, "Impact of distractors in item analysis of multiple choice questions", *Int. J. Res. Med. Sci.*, vol. 7, no. 4, p. 1136, Mar. 2019.
19. W. Hartono, S. Hadi, R. Rosnawati, and H. Retnawati, "Exploration of Diagnostic Testing Instruments: validity, reliability, and item characteristics", *Pegem Journal of Education and Instruction*, vol. 13, no. 3, pp. 386–394, 2023.
20. D. Kornbrot, *Point biserial correlation*, *Wiley StatsRef: Statistics Reference Online*. 2014.
21. D. G. Bonett, "Point-biserial correlation: Interval estimation, hypothesis testing, meta-analysis, and sample size determination", *Br. J. Math. Stat. Psychol.*, vol. 73 Suppl 1, no. S1, pp. 113–144, Nov. 2020.
22. N. A. Thompson, "Reliability & validity", *Assessment Systems*. 2013
23. J. Day and D. Bonn, "Development of the concise data processing assessment", *Phys. Rev. Spec. Top. - Phys. Educ. Res.*, vol. 7, no. 1, Jun. 2011.
24. E. E. Kisker and K. Boller, "Forming a Team to Ensure High-Quality Measurement in Education Studies, REL 2014-052", *Regional Educational Laboratory*, 2014.
25. C. Gei and M. G. Ison, "Content Validity of a Questionnaire to Assess Parental Involvement in Education", *European Journal of Psychology and educational Research*, vol. 4, no. 2, pp. 83–95, 2021.
26. N. K. Mitra, H. S. Nagaraja, G. Ponnudurai, and J. P. Judson, The levels of difficulty and discrimination indices in type A multiple choice questions of pre-clinical semester 1 multidisciplinary summative tests. *IeJSME*, vol. 3, no. 1, pp. 2-7, 2009.
27. F. Taib and M. S. B. Yusoff, "Difficulty index, discrimination index, sensitivity and specificity of long case and multiple choice questions to predict medical students' examination performance", *J. Taibah Univ. Med. Sci.*, vol. 9, no. 2, pp. 110–114, Jun. 2014.
28. R. L. Ebel and D. A. Frisbie, "Essentials of educational measurement", 1972.
29. S. Testa, A. Toscano, and R. Rosato, "Distractor efficiency in an item pool for a statistics classroom exam: Assessing its relation with item cognitive level classified according to Bloom's taxonomy", *Front. Psychol.*, vol. 9, p. 1585, Aug. 2018.
30. M. Sajjad, S. Iltaf, and R. A. Khan, "Nonfunctional distractor analysis: An indicator for quality of Multiple choice questions", *Pak. J. Med. Sci. Q.*, vol. 36, no. 5, pp. 982–986, Jul. 2020.
31. R. J. Mccowan and S. C. Mccowan, *Item Analysis for Criterion-Referenced Tests. Online Submission*. New York: Center for Development of Human Services, 1999.
32. P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation", *Anesth. Analg.*, vol. 126, no. 5, pp. 1763–1768, May 2018.
33. D. G. Bonett and T. A. Wright, "Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size planning", *J. Organ. Behav.*, vol. 36, no. 1, pp. 3–15, Jan. 2015.
34. L. J. Cronbach, "Coefficient alpha and the internal structure of tests", *psychometrika*, vol. *16, no.* 3, pp. 297-334. 1951
35. B. Forthmann *et al.*, "How much g is in the distractor? Re-thinking item-analysis of multiple-choice items", *J. Intell.*, vol. 8, no. 1, p. 11, Mar. 2020.

36. C. Boopathiraj and K. Chellamani, "Analysis of test items on difficulty level and discrimination index in the test for research in education", *International Journal of Social Science & Interdisciplinary Research*, vol. 2, no. 2, pp. 189–193, 2013.
37. T. Puthiaparampil and M. Rahman, "How important is distractor efficiency for grading Best Answer Questions?", *BMC Med. Educ.*, vol. 21, no. 1, p. 29, Jan. 2021.
38. S. N. Papageorgiou, "On correlation coefficients and their interpretation", *J. Orthod.*, vol. 49, no. 3, pp. 359–361, Sep. 2022.
39. D. L. Streiner, "Starting at the beginning: an introduction to coefficient alpha and internal consistency", *J. Pers. Assess.*, vol. 80, no. 1, pp. 99–103, Feb. 2003.
40. S. Nundy, A. Kakar, and Z. A. Bhutta, "Developing learning objectives and evaluation: Multiple choice questions/objective structured practical examinations", in *How to Practice Academic Medicine and Publish from Developing Countries?*, Singapore: Springer Nature Singapore, 2022, pp. 393–404.
41. B. Harjo, B. Kartowagiran, and A. Mahmudi, "Development of critical thinking skill instruments on mathematical learning high school", *Int. J. Instr.*, vol. 12, no. 4, pp. 149–166, Oct. 2019.
42. M. Ansari, R. Sadaf, A. Akbar, S. Rehman, Z. R. Chaudhry, and S. Shakir, "Assessment of distractor efficiency of MCQS in item analysis", *Prof. Med. J.*, vol. 29, no. 05, pp. 730–734, Apr. 2022.
43. S. Varma, *Preliminary item statistics using point-biserial correlation and p-values. Educational Data Systems Inc*, vol. 16. Morgan Hill CA. Retrieved, 2006, pp. 1–7.
44. H. K. Mohajan, "Two criteria for good measurements in research: Validity and reliability. Annals of Spiru Haret University", *Annals of Spiru Haret University. Economic Series*, vol. 17, no. 4, pp. 59–82, 2017.
45. K. S. Taber, "The use of cronbach's alpha when developing and reporting research instruments in science education", *Res. Sci. Educ.*, vol. 48, no. 6, pp. 1273–1296, Dec. 2018.
46. M. Tavakol and R. Dennick, "Making sense of Cronbach's alpha", *Int. J. Med. Educ.*, vol. 2, pp. 53–55, Jun. 2011.