Research Journal of Social Sciences & Economics Review
Vol. 5, Issue 3, 2024 (July – September)
ISSN 2707-9023 (online), ISSN 2707-9015 (Print)
ISSN 2707-9015 (ISSN-L)

RJSSER
Research Journal of Social Sciences & Economics Review

Development of Quality Mathematics Test items through the Application of Classical Test Theory at the Elementary Level

* Muhammad Touseef Khizar, MPhil Scholar

** Dr. Fareeha Sami, Assistant Professor

*** Syeda Hoor Fatima, Lecturer

Abstract

The study examined development of the quality mathematics test items at the elementary level by using classical test theory. This quantitative study employed development and descriptive methods of research. The population comprised all the students of Mathematics Grade 7th of tehsil Sahiwal district Sargodha. A sample of 100 students was selected for multi-Stage sampling technique. The researcher developed a test of multiple-choice items on an elementary-level course of mathematics. This test was used as the instrument to gather data from the respondents. Microsoft Excel was a suitable tool for item analyses for assessment used to examine the information and account the study findings. Reliability of the test noted (0.78) the test measured under the application of CTT. The validity was also ensured by discussing with 5 experts having Ph.D. qualifications. 33 items were found to be of moderate difficulty level (having item difficulty of 0.25 to 0.75), these were accepted whereas, 16 items were rejected and 6 items were rewritten for the enhancement of excellence of mathematics assessment matters at the elementary level. On the other hand, 16 items with a discrimination index below 0.20 were discarded because they failed to differentiate between high and low achievers. Whereas, 6 items with a discrimination index below 0.30 were revised because they did not adequately discriminate between the groups. 28 items with a discrimination index between 0.4 and 0.1 were accepted because they were effectively discriminating between the high and low achievers. Keeping in view the given results it is recommended that QAED academy may arrange the training for elementary school teachers to develop their tests in a standardized way. It would be beneficial for the development of an item bank.

Keywords: Classical Test Theory, Item Difficulty, Item Discrimination, Reliability, Validity

Introduction

Assessment is a process of obtaining information for decision-making for all subjects at the elementary level. Assessment is a key aspect that provides valuable feedback regarding the understanding of individuals and pertinent groups (Koçdar et al. 2016) mention it as Assessment is without any doubt, a key aspect of education. It is used to provide valuable feedback on understanding an individual and the group. It is also useful in the recognition of syllabi and certain teaching methods, which can help in reviewing. Therefore, it is to be defined assessment as a "systematic process of gathering information that is educationally relevant to make decisions about the provision of instruction and legal matters. (Gyamfi, 2022) Mathematics takes an essential place in education and learning. Moreover, the subject of Mathematics is significant, and human being must be able to achieve some basic calculation in order to contribute efficiently to his or her society. Its knowledge improves the competencies of human attention, which in turn, enables the expansion of scientific and technological advancement. Mathematics is well thought-out as an essential subject and is predictable to develop the ability to reason and analyze in solving daily life problems (Gravemeijer, Stephan, Julie, Lin, & Ohtani, 2017). (Schoenfeld, 2022) contends that mathematics is not a material of incoming facts into formulations and execution committal to memory procedures. However, mathematics is a method of rational and interrogative somewhat indefinite to students. (Arseven. 2015) added that Problem-solving skills can be developed by performing mathematics activities. Mathematical doings must wisely be done in the class according to practical things.

* Email: touseef.dude123@gmail.com
** University of Lahore Sargodha Campus Email: fareeha.sami@ed.uol.edu.pk
*** University of Lahore Sargodha Campus Email: hoor.fatima@ed.uol.edu.pk

Mathematics is used as an instrument to comprehend the patterns that occur in the world everywhere, as well as the patterns that happen in our minds. In this view, mathematics can be understood as a comprehensive science of pattern incisive (Zippert et.al. 2021). In the 20th century, the typical assessment in mathematics consisted of giving students a piece of paper with tasks, to which they had to find the 'correct' answer (Radmehr, 2020). Different test formats; objective and essay, exist for use in the classroom. The objective test comprises multiple choices, true or false, matching, fill and short answers. Generally, objectives test, especially the multiple-choice format is mostly used because it is easy to score, have high content validity, suitable for a large population, and are susceptible to statistical analysis. In this regard Classical Test Theory (CTT) provides a framework for evaluating the quality of test items grounded on their algebraic possessions and presentation. In the context of elementary-level mathematics assessments, CTT can be utilized to ensure that test items effectively measure students' understanding and mastery of mathematical concepts (Ayanwale et.al, .2023).

By taking into account the importance of decision making; the scope of this research was to evaluate the application of Classical Test Theory to optimize the quality of classroom mathematics test items at the elementary level. Moreover, it intends to enhance the validity, reliability, fairness, and instructional utility of assessments, ultimately benefiting both students and educators.

Statement of problem

Classroom assessment plays a crucial role in gauging students' understanding and mastery of mathematical concepts at the elementary level. However, the effectiveness of these assessments relies heavily on the quality of test items used. Keeping in view this fact the study was conducted to develop the Quality Mathematics Test items through the Application of Classical Test Theory at the Elementary Level.

Significance of Study

1. Optimizing the quality-based mathematics test items at the elementary level would help teachers to improve the teaching-learning process.
2. Similarly, this study would be helpful for teachers can ensure the validity, reliability, and fairness of classroom assessments, leading to more accurate evaluations of students' mathematical proficiency.
3. Students may able to improve their abilities, as critical thinking and problem-solving skills, ensuring equity in assessment, preparing students for standardized testing and fostering confidence and motivation in mathematical learning.
4. The findings from this research may be helpful for institutes to develop quality classroom mathematics test items using CTT at elementary level institutions by enhancing assessment validity and reliability, supporting data-driven instructional decisions, ensuring curriculum alignment and improvement, evaluating teaching effectiveness, monitoring student progress, maintaining quality assurance and accountability, preparing students for standardized testing.
5. At the end, the findings from this study may make it easier for educational policies makers to make informed decisions about educational policies, resource allocation, and professional development initiatives aimed at improving mathematics education.

Research Objective

The purposes of the study were:

1. To develop quality-based test items at the elementary level in the subject of mathematics.
2. To assess the quality-based test items by using item difficulty
3. To assess the quality-based test items based on the discrimination index.

Research Questions

1. To what extent does the item difficulty affect the quality of test items?
2. To what extent does the discrimination index affect the quality of test items at the elementary level?
3. What is the reliability of the test items of mathematics at the elementary level?

Methodology

A sample was selected for multi-stage sampling technique. This sampling was carried out according to the following steps:

1. First of all, from all seven Tehsils of District Sargodha, Tehsil Sahiwal was selected conveniently.

2. On the basis of 50% selection 13 public Elementary schools of tehsil Sahiwal were selected among 26 public Elementary schools.
3. Two groups were made on gender basis, i.e., male and female.
4. There were 7 male and 6 female public Elementary schools were included in each group.
5. Total 100 students were selected from 13 public Elementary schools.

Data were collected from respondents i.e., grade 7 students. Detail of tests was as follows:
The correct answer indicated value of 1, and wrong answer stated a value of 0. Total number of instrument items (tests) was 50.

Instrument

Self-developed test was used and developed by using the bloom taxonomy as an instrument which was comprises of 50 items.

Validity of research instrument

The test was validated on the following standards:

1. ***Concurrent validity*** was ensured by five experts by taking their suggestions for improvement.
2. ***Face validity*** was ensured by taking suggestion from experts and after that incorporating these suggestions.
3. ***Content validity*** was ensured by developing the table of specification under the blooms 'taxonomy cognitive domain.

Table of Specification

Topics / Domains	Knowledge	Comprehension	Application and above	Total	Percentage
Algebra/Number Sequence	3	6	4	13	22%
Real Life Problems	0	4	5	9	15%
Algebraic Expressions	2	1	2	5	15%
Like Terms&Unlike Terms	2	1	2	5	9%
Equation and Inequality	1	2	2	5	9%
Polynomials	1	1	2	4	10%
Linear Equations	1	1	1	3	5%
Real life problems involving linear equations	1	2	3	6	15%
Total	11	18	21	50	100%
Percentage	28%	35%	37%		

Table 1 shows the two-way chart of cognitive objectives and selected topics. Most of the questions were developed on application level which were having 37% ratio. As it is mathematics test that's why it was mandatory to have more questions in application level.

Reliability of the instrument of study

The reliability of the test was calculated by using the Kuder and Richardson Formula

$$Reliability = \frac{k}{k-1}\left(1 - \frac{\sum PQ}{\sigma^2}\right)$$

Where k = number of items

&σ²=variance

$$Reliability = \frac{50}{50-1}\left(1 - \frac{10.65}{43.77}\right)$$

$$=(1.020)\,(0.76)$$

$$=0.78$$

Overall reliability of the test was 0.78 which is fallen under the satisfactory range (Khairul Zahreen et.,al 2018).

Data Analysis

The data were investigated on the basis of CTT by using Microsoft Excel. The results obtained from this application were tabulated according to their sequences.

Item Difficulty Index

We denoted item difficulty index by p.it is calculated as $\boldsymbol{P = \frac{R}{T}}$

Where P = item difficulty index

R = Number of students who got the item correct

T = Total number of students

There are three levels of item difficulty index. Detail is given in the Table 2.

Range (P)	Decision
0.00-0.25	Rejected
0.25-0.75	Accepted
0.75-1.00	Revised

Gul, N., Shagufta, S., & Parveen, S. (2022). Item Difficulty in Item Analysis of Intelligence Test Items. *Pakistan Social Sciences Review*, *6*(2), 97–108. Retrieved from https://ojs.pssr.org.pk/journal/article/view/117

The item difficulty index is often referred as" P-value" in classical test theory, measures how tough a test item is for an assumed group of exam takers. It is calculated as the number of students who replied the item correctly.

Table 3: Item Difficulty Index (*MS Excel* – CTT)

P	Decision	Item No.	Total
0.00 – 0.25	Rejected	0	0
0.25 – 0.75	Accepted	1, 2, 3, 4, 5, 6, 10, 11, 12, 13, 16, 17, 18, 19, 20, 21, 22,23, 24, 25, 26, 27,28,29,30,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50	44
0.75 – 1.00	Revised	07, 08, 09, 14, 15, 31	06
		Total	**50**

According to the table 3 it is shown that none of the item were below 0.25 range. 6 items including item number 07,08,09,14,15,31, had P value above than the acceptable range, were most easy items so these were tended to be revised, whereas, rest of items which were found to be in moderate difficulty level (having item difficulty of 0.25 to 0.75), were accepted (Smith, et.al. 2023).

Item Discrimination Index (D) with *MS Excel* – CTT

The item discrimination index is represented by d. it is calculated using the following formula.

$$D = \frac{R_h - R_l}{Either\ Group}$$

R□ = the number of respondents in the high achiever's group (27 %)

R□ = the number of respondents in the low achiever's group (27 %)

"Either group" refers to the number of respondents in R□ or R□

The item discrimination index was calculated for each test item by using 27% ratio for high and low achievers.

There are four levels of item discrimination index. Detail is given in the Table 4

Item Discrimination Index Range

Range (D)	Status
≤ 0.19	Rejected
0.20-0.29	Revised
0.30-0.39	Marginally Accepted
≥ 0.4	Accepted

Mahjabeen, W., Alam, S., Hassan, U., Zafar, T., Butt, R., Konain, S., & Rizvi, M. (2017). Difficulty index, discrimination index and distractor efficiency in multiple choice questions. *Annals of PIMS-Shaheed Zulfiqar Ali Bhutto Medical University*, *13*(4), 310-315

Table 5: Item Discrimination index (*MS Excel* – CTT)

D	Status	Item No.	Total
≤ 0.19	Rejected	01,02,03,04,05,07,08,09,12,13,14,15,17,23,49,50	16
0.20 – 0.29	Revised	06, 10, 16, 22, 39,44	06
0.30 – 0.39	Marginally Accepted	19,24,31,37	04
≥ 0.4	Accepted	11,18,20,21,25,26,27,28,29,30,32,33,34,35,36,38,40,41,42,43,45,46,47,48	24
		Total	**50**

According to the table 5 it is shown that 16 items were below 0.19 range so these were rejected.5 items including item number 06, 10, 16, 22, 39, 44 had D value less than the 0.30 range, these items not adequately discriminating between the high and low achievers so these were tended to be revised, whereas, four items which were found to be in marginally accepted discrimination index range (having item discrimination index of 0.30 to 0.39), were marginally accepted, rest of the 24 items having item discrimination index above than 0.4 were accepted.

Conclusion

Based on item difficulty and discrimination index 16 items were rejected, 12 items tended to be revised and according to discrimination index four items were marginally accepted while only 18 items were accepted. The reliability of the test was 0.78 by using the Kuder & Richardson formula.

Discussion

According to the result it was shown that, for measuring the content validity, the table of specification was developed under the blooms 'taxonomy cognitive domain. The same results were supported by (Gyamfi, 2022). It was shown in his study that table of specification may be a good source to ensure the content validity because table of specification is a two-way chart and has ability to compare the cognitive objectives with selected topics. In view of the study findings, it was concluded that Kuder & Richardson was the suitable technique to find out the consistency of the objective test in all schools where the test was administered. These findings were supported by (Ajayi, 2017) in which he concluded that among all the reliability analysis methods, Kuder -Richardson was the most suitable technique for objective test, since in all the schools in which the tests were directed, Kuder-Richardson may have the highest scores.

To determine the item degree of difficulty its difficulty was measured. This method for identifying the difficulty was supported by (Gyamfi, 2022) in which he concluded that item difficulty indices is a sign of proportion of the examinees who replied to the item correctly.

It was concluded that for finding the discrimination power the discrimination index was used. This finding was supported by (Panjaitan, et. al.,2017) according to his point of view A researcher cannot judge the quality of an item from calculation only, without taking Discrimination index (D) into consideration.

In short, Classical test theory (CTT) also recognized as true score theory was a deterministic method and needed many familiar concepts such as the reliability and the validity of instruments. At the item level, the CTT model is comparatively simple. CTT does not appeal to a composite theoretic model to narrate an examinee's capability to succeed on a specific item. Instead, CTT collectively considers a pool of examinees and empirically examines their success rate on an item (assuming it is dichotomously scored).

Recommendations

On the basis of the results following recommendations were made:

1. It was concluded that measuring the reliability coefficient and validity of the test improves the quality of the test. So, it is suggested that Quaid e Azam Academy of Educational Development (QAED) may organize the trainings for the school teachers so that they may expand the excellence of tests.
2. According to the conclusions, it was shown that the Item difficulty and Discrimination index are the best methods to improve the items' quality. So, it is recommended that QAED academy may arrange training for the school teachers so that they may improve the quality of test items.

References

Ajayi, B. K. (2017). A comparative analysis of reliability methods. *Journal of Education & Practice, 8*(25), 160-163 Retrieved from https://core.ac.uk/download/pdf/234640929.pdf

Arseven, A. (2015). Mathematical Modelling Approach in Mathematics Education. *Universal Journal of Educational Research, 3(12), 973-980* Retrieved from https://files.eric.ed.gov/fulltext/EJ1083314.pdf.

Ayanwale, M. A. (2023). Test score equating of multiple-choice mathematics items: techniques from characteristic curve of modern psychometric theory. *Discover Education, 2(1), 30* Retrieved from https://www.researchgate.net/publication/373546054_Test_score_equating_of_multiple-choice_mathematics_items_techniques_from_characteristic_curve_of_modern_psychometric_theory

Brod, M., Waldman, L. T., Shu, A. D., & Smith, A. (2023). Content validation of the SF-36v2® Health Survey Acute for use in hypoparathyroidism. *Quality of Life Research*, 32(6), 1795-1806 retrieved from https://www.researchgate.net/publication/368391514_Content_validation_of_the_SF-36v2R_Health_Survey_Acute_for_use_in_hypoparathyroidism.

Brookhart, S. M., & Nitko, A. J. (2011). Strategies for constructing assessments of higher-order thinking skills. *Assessment of Higher Order Thinking Skills*, 1, 327-59 Retrieved from https://jurnal.usk.ac.id/JPSI/article/view/23968.

Gravemeijer, K. (2024). Mathematics and STEM, Preparing Students for Their Future. In Disciplinary and Interdisciplinary Education in STEM: *Changes and Innovations (pp. 13-31). Cham: Springer Nature Switzerland* Retrieved from https://www.researchgate.net/publication/379090985_Mathematics_and_STEM_Preparing_Students_for_Their_Future.

Gul, N., Shagufta, S., & Parveen, S. (2022). Item Difficulty in Item Analysis of Intelligence Test Items. *Pakistan Social Sciences Review, 6*(2), 97–108. Retrieved from https://ojs.pssr.org.pk/journal/article/view/117

Gyamfi, A. (2022). Application of Classical Test Theory (CTT) in the Validation of Teacher Made Mathematics Multiple Choice Test (MMCT) Items. *Asian Journal of Advanced Research and Reports, 16(11),1-1* Retrieved from https://journalajarr.com/index.php/AJARR/article/view/434.

Mahjabeen, W., Alam, S., Hassan, U., Zafar, T., Butt, R., Konain, S., & Rizvi, M. (2017). Difficulty index, discrimination index and distractor efficiency in multiple choice questions. *Annals of PIMS-Shaheed Zulfiqar Ali Bhutto Medical University, 13*(4), 310-315 Retrieved from https://www.researchgate.net/publication/323705126_Difficulty_Index_Discrimination_Index_and_Distractor_Efficiency_in_Multiple_Choice_Questions

Mohd Arof, K. Z., Ismail, S., & Saleh, A. L. (2018). Critical Success Factors of Contractor's Performance Appraisal System in Malaysian Construction Industry. *Indian Journal of Public Health Research & Development, 9(11)* Retrieved from https://www.researchgate.net/publication/329479092_Critical_success_factors_of_contractor's_performance_appraisal_system_in_Malaysian_construction_industry

Panjaitan, R. L., Irawati, R., Sujana, A., Hanifah, N., & Djuanda, D. (2018, March). Item validity vs. item discrimination index: a redundancy?. In Journal of Physics: Conference Series (Vol. 983, No. 1, p. 012101). IOP Publishing Retrieved by https://iopscience.iop.org/article/10.1088/1742-6596/983/1/012101/meta.

Radmehr, F. (2023). *Toward a theoretical framework for task design in mathematics education* Retrieved from https://www.researchgate.net/publication/369742475_Toward_a_theoretical_framework_for_task_design_in_mathematics_education.

Retrieved From: https://www.researchgate.net/publication/340432627_The_Role_of_Classical_Test_Theory_to_Determine_the_Quality_of_Classroom_Teaching_Test_Items

Vincent, W., & Shanmugam, S. K. S. (2020). The role of classical test theory to determine the quality of classroom teaching test items. *Pedagogia: Jurnal Pendidikan, 9(1), 5-34.*

Zippert, E. L., Douglas, A. A., Tian, F., & Rittle-Johnson, B. (2021). Helping preschoolers learn math: The impact of emphasizing the patterns in objects and numbers. *Journal of Educational Psychology, 113(7), 1370* Retrieved from https://files.eric.ed.gov/fulltext/ED610851.pd