DOI: https://doi.org/10.52131/pjhss.2022.1003.0274

Pakistan Journal of Humanities and Social Sciences

Volume 10, Number 3, 2022, Pages 1132–1146
Journal Homepage:
https://journals.internationalrasd.org/index.php/pjhss

Assessment and Achievement Standards of Mathematics Students at Elementary Level in Punjab

Muhammad Arshad Khan[1], Mubashrah Jamil[2]

[1] Ph.D. Scholar, Institute of Social and Cultural Studies, Bahauddin Zakariya University Multan Pakistan. Email: arshadmanais@gmail.com
[2] Department of Education, Bahauddin Zakariya University Multan Pakistan. Email: mubashrahjamil@bzu.edu.pk

ARTICLE INFO

Article History:

Received:	August 04, 2022
Revised:	September 29, 2022
Accepted:	September 29, 2022
Available Online:	September 30, 2022

Keywords:
Development
Standardized Achievement Test
Elementary Level
Mathematics Perceptions
Preparation

ABSTRACT

This research deals with the Development of a Standardized Achievement Test in Mathematicsat Elementary Level. It contains a review of current literature dealing with the decline in mathematics achievement, mathematics assessment, concept development, and the effects of standardized testing. A survey was conducted in elementary schools throughout Punjab Pakistan. The purpose of the survey was to assess teachers' perceptions of how preparation for the major annual standardized achievement test affects the pacing, sequence, and presentation of their mathematics curricula. 1500 from grades 8th completed a limited response questionnaire. The results indicate that a majority of elementary-school students prepare for the standardized achievement test by covering all testable skills by testing time. However, most students feel that preparation for this test has a negative impact on their mathematics programs.

Corresponding Author's Email: arshadmanais@gmail.com

1. Introduction

Testing is a useful technology that is solely based on specialized, skilled and technical experts. Because of their objectivity and qualification, tests provide a mechanism for ensuring that schools and teachers are working on right direction. Popham (2004) says that different educational purposes required differing education tests, and these tests differ in their uses While one-of-a-kind sorts of checks and checks can be standardized on this way, the time period is by and large related to huge-scale checks administered to huge populations of college students, which include a more than one-desire take a look at given to all of the 8th grade public college students in a specific state (Parkay, O'Bryan, & Hennessy, 2014).

In addition to the acquainted more than one desire format, standardized checks can consist of true-fake questions, short-solution questions, essay questions, or a combination of query types. While standardized checks have been historically supplied on paper and finished the usage of pencils and lots of nonetheless are, they're more and more being administered on computer systems related to on-line packages. While standardized checks may also are available in loads of forms, more than one-desire and true-fake codes are broadly used for huge-scale trying out conditions due to the fact computer systems can rating them quickly, consistently, and inexpensively (Schwartz, 2014). In contrast, open-ended essay questions want to be scored via way of means of people the usage of a not unusual place set of suggestions or rubrics to sell regular critiques from essay to essay a much less green and greater time-in depth and high priced alternative this is additionally taken into consideration to be greater subjective (Wardrop, 2011). While standardized checks are a chief supply of discussion within side the Pakistan, many take a look at specialists and educators recall them to be a honest and goal technique of assessing the instructional fulfillment of college students, particularly due to the fact the standardized format, coupled with automated scoring, reduces the capability for favoritism, bias, or subjective critiques. On the opposite hand, subjective human judgment enters into the trying out method at numerous degrees e.g., within side the choice and presentation of questions, or within side the problem depend and phraseology of each questions and answers (Copeland, 1970).

Standardized checks can be used for a huge style of academic purposes. For example, they will be used to decide a younger child's readiness for kindergarten, perceive college students who want special-schooling offerings or specialized instructional assist, area college students in one-of-a-kind instructional packages or direction levels, or award diplomas and different academic certificates. Bennett (2011) Achievement checks are used to assess college students or workers understanding, comprehension, information and/or functionality in a specific area (Dede & Freiberg, 1987). Achievement checks are designed to degree the information and abilities college students discovered in college or to decide the instructional development they've remodeled a length of time. The checks can also be used to assess the effectiveness of a faculties and instructors, or perceive the correct instructional placement for a pupil i.e., what guides or packages can be deemed maximum suitable, or what varieties of instructional assist they will want. Achievement checks are backward-searching in that they degree how nicely college students have discovered what they have been anticipated to learn. Pressure on educators to compete also can come from parents, colleagues, and administrators, in addition to from numerous local, state, and federal institutions (Jurgens, 1987). Following were the objectives of the study (i) Standardization of an achievement test for the students of class VIII in the subject of Mathematics and (ii) to determine the achievement of students of class VIII toward the subject of Mathematics. So, standardized tests is a psychological tool with which can estimate future performance of a person on the basis of his/her performance or interests in past.

These tests are also useful in determining learning readiness, individualizing instruction, organizing class room groups, identifying underachievers, diagnosing learning problems (Gronlund & Linn, 2000), prediction of success (Bennett, 2011), evaluating to perform on tasks or react to different situations, identifying standard ability, unique insight into talents, measuring potential ability of a person for performing (Digumarti, 1994). So, measurement of standard is a crucial one. Limited work in this area has been done in different institutions and universities of Pakistan. In 1960-61, Group Scholastic Standardized Test for class 10th, 11th and adults was adapted from California Test of Mental Maturity by Talib and Hussain. In 1972, Institute of Education and Research (I. E. R.), University of Punjab, Lahore standardized three scale of scholastic standard scale for children. Later on, in 1984, Khurshed worked on the 'Development of Science Standardized Test".

The most educator of Educational Mathematics in the definition of education, it knows that a series of regular activities in order to create desired changes in behavior of learners (Aud et al., 2011). According to this definition cannot claim that learning has been made without the measurement and evaluation of changes. Nowadays Measurement and evaluation of educational activities accounted for a significant part of it .Research carried out shows that in each teaching session between 5 to 15 minutes of class time is spent to measuring and evaluating (Parkay et al., 2014). Measurement and evaluation not only provides some of information about characteristics of students to teachers, but also can affect students' learning styles and strategies and so affect their level and speed learning (Bloom, 1971). There are two general approaches in the learning process: rote learning approach that puts the emphasis on memorization of unrelated facts and deep approach to learning that involves; exploring the deliberate and active for fundamental principles and concepts, and problem solving (Zarei, Mirhashemi, & Pasha Sharife, 2013). The teacher-made tests, often, simple elements, surface and material unrelated to curriculum content are emphasized and largely ignore the more complex and deeper knowledge, so the students employ rote learning approach (Hooman, 1993).

Although apparently it is thought that the assessment is end of the educational activities of teacher, but the reality today is that often assessment and measurement determines the teacher training activities and students learning largely by their performance on achievement tests is shown (Hogan, 2003). Note that the teacher assessment can have a significant impact on the training -learning process (Cizek, 1993), so it is better teachers try to choose well and variety learning objectives for their students and assessment methods appropriate to that goals use. In this regard, Woolfolk and Karlberg (2004) said If the tests determine what teachers actually teach and what students learn - that it truly is – so the way of improving education, is the direct way but uphill: to assess important and fundamental abilities and habits. Theoretical foundations of development and standardization of achievement tests based on psychometric principles and procedures were constructed. Today two methods, the classical test model and Item-Response Theory (IRT), for constructing tests and interpreting scores have served measurement specialists well (Gulliksen, 1950; Hambleton, 1989). However, in recent years, due to limitations of classical theory and advent of computers and software, its application reduced and use of IRT is prevalent.

2. Literature Review

The procedure of obtaining simple expertise begins off evolved with essential training. Without essential training, none of one's goals can be viable. The essential training procedure brings people to a stage of essential competence for fixing problems, adapting to social values and making use of installed social rules. Salma (2010) said "Children study plenty spontaneously, and it's miles handiest through cautious statement to this spontaneous procedure that we are able to increase a valid idea of training and of the college curriculum". Elementary stage examinations in Pakistan are the maximum complete shape of testing, generally given on the give up of the time period and one or instances at some stage in the semester, a check is extra restricted in scope that specialize in extra precise factors of the path fabric.

Airasian and Madaus (1983) mentioned "Standardized check is a check, administered and scored in a steady manner. These checks are designed in this kind of manner that the questions, situations for administering, scoring procedures, and interpretations are steady". The study explores the issues of validity and reliability generally is considered as important factors for figuring out the exceptional of any standardized check. "The fulfillment check focuses upon examinees attainments at a given factor in time". According to Suter (2011) said that "Elementary college stage is an vital stage of formal training wherein not unusual place simple expertise and abilities are taught, that are required for all residents within side the society". Johnson (2006) found "expert and practitioner institutions regularly have positioned those issues inside broader contexts whilst growing requirements and making universal judgments approximately the exceptional of any standardized check as an entire inside a given context".

Leung, Fung, and Farver (2018) discovered "The fulfillment checks are served as a device to degree modern expertise degrees for the motive of putting college students in a person surroundings in which they've the risk to increase at a tempo this is appropriate for his or her abilities". Morales (2019) discovered that "Measurement is critical to the development of an exceptional scholar assessment, even within side the case of classroom-designed or non-standardized assessments. Measuring variables is one of the vital steps within side the studies procedure".

Ebel and Frisbie (1972) described the check in very complete manner "check is a way of measuring the expertise, ability, feeling, intelligence and flair of a person or group". Morales (2009) described "Achievement checks degree someone's accomplishment in a topic or task. One tool can also additionally serve each purpose, appearing as a flair check to forecast destiny performances and a fulfillment check to display beyond and gift getting to know".

Aggarwal (2019) certainly defined "Elementary training is just like the first step we absorb life; we are able to by no means be capable of run if we do now no longer discover ways to walk". Education is the maximum vital asset we have, due to the fact our expertise is the sort of wealth that we are able to by no means lose irrespective of what, and the extra we proportion it the extra it increases". Austin and Garber (2013) defined that Standardized or none standardized, measures that how plenty of the fabric has been mastered and examine the scholar modern status". These checks are used to decide what a scholar has found out including vocabulary, reading, math ability, etc. "Achievement checks are used to assess college students or laborers understanding, comprehension, expertise and functionality in a selected region.

Airasian and Madaus (1983) said "Knowledge and abilities received within side the essential training, is the simple expertise and ability to be received at different instructional degrees, for that reason different instructional degrees is likewise primarily based totally at the essential training". Therefore, this essential training stage influences negatively or definitely now no longer handiest the academic machine of the society. Wiersma and Jurs (1985) recognized "A fulfillment check is meant to degree what the scholar has found out or what abilities the scholar has mastered". Carpenter and Moser (1984) described "Achievement checks because the sort of cap potential check that describes what someone has found out consequently is known as a fulfillment check".

Harkness (2020) said that "They are generally norm referenced checks that degree the pupil's stage of fulfillment in numerous content material and ability areas". Copeland (1970) mentioned "Achievement checks are examinations which can be designed to decide the diploma of expertise and skill ability exhibited through a person in a unique region or set of areas". Moreover, they may be extraordinarily essential for the scholars, for their meant both to make the scholars by skip and fail the check. The study found that Achievement check are nicely suitable and it offer educators with goal remarks as to, how plenty college students are getting to know and understanding". "Achievement check rankings are frequently utilized in an academic machine to decide what stage of coaching for which a scholar is prepared.

Harkness (2020) High fulfillment rankings typically suggest a mastery of grade-stage fabric, and the readiness for superior coaching. Low fulfillment rankings can suggest they want for remediation or repeating a path grade. Results of fulfillment check additionally offer music to the counselor to manual and offer treatment to the scholars in pleasant viable manner. This research is based on the relationship students of 8th class of government's schools from Punjab which is purely based on education involved in the mathematics because of assessment and performance of students has been taken in this research through the standardization of an achievement test and the data has been taken from the Pakistan economic survey, Ministry of education, Pakistan statistical analysis department and population welfare department of Pakistan. Therefore many researches on standardization of an achievement test but the combination that I have made in this research has never been made by any researcher as I have selected the 80 items to analyze the performance of 8th class students from the elementary and high schools of Punjab with a unique sample size and number of items. I have completed the research with unique strategy and combination which has made my research novel because I made five different sets of items to analyze the performance of above said limits from both male and female from rural and urban areas of Punjab which made my research different from other researchers and I also have shown the different way to present the research with different item setting.

3. Methods

The primary data was taken from 17500 participants through a survey from government elementary school students of 8th class. This was through a multiple choice type questions from the participants according to their specifications. The primary data was collected through test. The secondary data was taken from the different websites, newspapers, articles, and some government departments so that there was no ambiguity in presenting the data because the data was pure. Special attention has been paid in this regard that the data was taken from the relevant department like, Pakistan economic survey, Ministry of education, Pakistan statistical analysis department and population welfare department of Pakistan.

3.1 Sampling

In this research there was an adoption of survey method. The population was the students of 8th class of government's schools from Punjab. At first stage 8750 math students from different government elementary schools and at second stage, 8750 math students of 8th class from different government elementary schools were also selected to conduct survey from different schools of different areas of the Punjab. Then it was divided into two equal groups of items (40 items in each group) and groups were named as group A and group B. The total participants of the survey were 17500. This survey was conducted to check the standardization of an achievement test from government elementary school students. There was an application of quantitative approach. The record of schools and education department were examined and data regarding government elementary school students of 8th class were taken. The sample and population table is given below:

Table 1: Sampling and Population

	Population size						Sample size			
District	Schools	Male	Female	Students (male)	Students (female)	Total Students	Male Schools	Female Schools	Students (male)	Students (female)
Multan	160	105	55	2500	2000	4500	20	16	290	250
Khanewal	180	109	71	2625	2250	4875	21	18	270	270
Vehari	155	81	74	2375	2375	4750	19	19	260	250
Bahawalpur	149	82	67	2500	2500	5000	20	20	270	240
Bahawalnagar	188	103	85	2625	2500	5125	21	20	250	240
Rahimyarkhan	212	130	82	2875	2375	5250	23	19	270	230
D G Khan	117	75	42	2375	1625	4000	19	13	260	250
Layyah	150	73	77	2625	2500	5125	21	20	250	240
Attock	160	91	69	2500	2500	5000	20	20	240	240
Bhakkar	148	84	64	2625	2750	5375	21	22	250	260
Chakwal	155	88	67	2500	2375	4875	20	19	240	230
Chiniot	146	83	63	2875	2375	5250	23	19	270	230
Faisalabad	248	141	107	3000	2250	5250	24	18	280	240
Gujranwala	211	120	91	2375	2625	5000	19	21	250	250
Gujrat	195	111	84	2500	2750	5250	20	22	270	260
Hafizabad	165	94	71	2625	2375	5000	21	19	250	230
Jhang	204	116	88	2750	2500	5250	22	20	260	240
Jhelum	187	106	81	2500	2625	5125	20	21	270	250
Kasur	147	84	63	2625	2125	4750	21	17	250	240
Khushab	139	79	60	2750	2875	5625	22	23	260	270
Lahore	346	197	149	2875	2750	5625	23	22	260	260
Lodhran	174	99	75	2375	2500	4875	19	20	270	240
Mandi Bahaudin	155	89	66	2500	2375	4875	20	19	260	230
Mianwali	135	77	58	2250	2250	4300	18	18	250	220
Narowal	174	98	76	2375	2000	4375	19	16	230	260
Nankana Sahib	115	66	49	2500	2250	4750	20	18	240	220
Okara	167	95	72	2375	2500	4875	19	20	260	240
Pakpatan	127	72	55	2500	2375	4875	20	19	240	230
Rajanpur	124	71	53	2625	2625	5250	21	21	250	250
Rawalpindi	297	169	128	2875	2250	5125	23	18	270	240
Sahiwal	184	105	79	2625	2250	4875	21	18	250	220
Sargodha	213	121	92	2875	2500	5375	23	20	270	240
Sheikhupura	175	100	75	2750	2750	5500	22	22	260	260
Sialkot	143	81	62	2625	2375	5000	21	19	250	230
Toba Tek Singh	129	73	56	2500	2250	4750	20	18	240	240
Grand total	6074	3468	2606	90750	84250	175000	726	674	9010	8490
	Population Size = 175000						Sample Size = 17500			

(Author's Calculations)

3.2 Development of the Test

By applying Rasch model, items were arranged in order to their difficulty. The analysis was deal with up to the extent of item person interaction and probability curves i.e. ICC and PCC (item character curve and person character curve). After the refinement of items through both methods, the selected items were written on items cards. These files were become the basis for the development of standardized achievement test of Mathematics at elementary level in 35 districts of Punjab.

Table 2: The Test after incorporation of suggested changes

Sr. #	Nature of the test	Before Pilot Study	After Pilot Study	Types of items	Options	Time
1	Thinking	18	16	Multiple Choice	4	25 mins
2	Reasoning	18	16	Multiple Choice	4	25 mins

3	Conceptual Understanding	18	16	Multiple Choice	4	25 mins
4	Procedural Knowledge	18	16	Multiple Choice	4	25 mins
5	Problem Solving	18	16	Multiple Choice	4	25 mins

(Author’s Calculations)

3.3 Collection of Data

The data collected through the administration of the test from the male and female students of class VIII of both urban and rural areas of 35 districts of Punjab. Different steps were taken for the sake of data collections which are as under;

- The test was developed on the basis of multiple choices with four options.
- The test was presented for final approval before the researchers and specialists of different universities of Punjab such as Islamia University Bahawalpur, Bahauddin Zakariya University Multan, Education University Multan Campus, Punjab University Lahore, etc.
- After the suggestions and recommendations some modifications were made in the statements of test and considered final with 80 statements divided into two groups of 40 statements each.
- The test was administered in government elementary and high schools of 35 districts from Punjab in class VIII and the test sheets were distributed among the students and collected after the specified time mentioned in the test sheet and data was gathered in almost one year from 17500 male and female students of 35 districts of Punjab from both urban and rural areas.
- When the data was collected, it was presented in tabulated and graphical form with the interpretations and explanations to elaborate the results.

3.4 Split Half Method of Computing Reliability

There were 80 items in the test. So, the following methods were used to determine the reliability of the test.

- Split – half reliability
- KR # 20 reliability method
- KR # 21 reliability method

A questionnaire based on Likert scale was developed with the help of experts to calculate the face validity and opinions were collected from students of 8th class through the test. To find content validity, content of the test was analyzed on the basis of model given by Barrett (2004) and (Digumarti, 1994). To find predictive validity, the results of standardized test of math were compared with the achievements of students in class VIII in the subjects of Math through correlations. Standardized profile gives the information about of relationship of one standardization test to another and indicates potential of types of work tasks, occupational choices (Vision, 2006). Math Standardized profile of a student was prepared as a model.

Table 3: Reliability of the Test (Split Half Method)

Class	Test	Value of “r”
VIII	1	0.85
VIII	2	0.80
VIII	3	0.64
VIII	4	0.45
VIII	5	0.35

The table 4 shows that the maximum value of correlation i.e., 0.85 between odd and even scores of the test 1 of the class VIII while the minimum value of correlation is 0.35 has been seen in test 5.

Table 4: Internal Reliability of the Test

Class	Test	Value of "r"
VIII	1	0.83
VIII	2	0.73
VIII	3	0.58
VIII	4	0.42
VIII	5	0.39

Table 5 shows that the range of Value of "r" for class VIII is 0.39 to 0.83

Table 5: KR # 20 and KR # 21 Reliability Test, Class VIII, (Age group 11 to 14+), Total no. of the candidates = 17500, NO. of items = 16

Item #	No. of correct Responses	No. of incorrect Responses	p	q	p x q
1	11321	6179	.65	.35	.23
2	10578	5922	.60	.40	.24
3	12695	4805	.72	.28	.20
4	13214	4286	.75	.25	.19
5	11256	6244	.64	.36	.23
6	11358	6142	.65	.35	.23
7	10875	6625	.62	.38	.24
8	11621	5879	.66	.34	.22
9	10852	6648	.62	.38	.24
10	11133	6367	.64	.36	.23
11	10247	7253	.59	.41	.24
12	11159	6341	.64	.36	.23
13	11457	6043	.65	.35	.23
14	12121	5379	.69	.31	.21
15	10612	6888	.61	.39	.24
16	11337	6163	.65	.35	.23

Now $n = 16$, $\alpha = 1.76$ $\sum p_x q = 3.63$ $\alpha^2 = 3.1$
Applying KR # 20 Formula r= -.43
Applying KR # 21 Formula r= -.43

The table 6 is showing the results for test 1, Kudar and Richardson reliability # 20 and 21 which have been calculated after applying the KR reliability # 20 and 21 formulas and the results are same in the application of both formulas. The age group of the participants was from 11 years to 14+ years as described in the demo graphs of the sample of the current research. For the test 1 there were 17500 sample sizes was set aside for all 16 categories. After the calculation of the correct and incorrect responses which are shown in the table 6. The p value was calculated for correct responses and the q value was calculated for incorrect responses, then the indices were made for both values which were calculated 3.63. Then α and α^2 were also calculated to check the further reliability.

Table 6: Standard Error of Measurement

Tests	SD	KR # 20	SE_{Means}
1	1.76	-.43	2.10
2	1.76	-.43	2.10
3	1.74	-.43	2.14
4	1.76	-.43	2.10
5	1.81	-.43	1.97

In table 7 the small value of SE (Means) ranging from 1.97 to 2.14 for class VIII which is the indication of reasonable consistency.

Table 7: Significance of difference between mean scores

S #	Category	N	Mean	SD	Z
1	Male (Urban) Vs. Male (Rural)	5290 3720	66.1 46.5	13.22 9.3	16.89
2	Male (Urban) Vs. Female (Urban)	5290 4917	66.1 61.5	13.22 12.3	19.14
3	Male (Urban) Vs Male (Urban+Rural)	5290 5290+3720	66.1 112.6	13.22 22.52	26.80
4	Male (Urban) Vs Male+Female(Urban)	5290 5290+4917	66.1 127.6	13.22 25.52	29.05
5	Male (Rural) Vs. Female (Rural)	3720 3573	46.5 44.66	9.3 8.9	13.67
6	Male (Rural) Vs. Male (Urban+Rural)	3720 5290+3720	46.5 112.6	8.9 22.5	23.81
7	Male (Rural) Vs. Male+Female(Rural)	3720 3720+3573	46.5 91.16	9.3 18.23	20.65
8	Female (Urban) Vs. Female (Rural)	4917 3573	61.46 44.66	12.3 8.9	15.9
9	Female (Urban) Vs. Female(Urban+Rural)	4917 4917+3573	61.46 106.12	12.3 21.22	25.14
10	Female (Urban) Vs. Male+Female(Urban)	4917 5290+4917	61.46 127.6	12.3 25.52	28.36
11	Female (Rural) Vs. Female(Urban+Rural)	3573 4917+3573	44.66 106.12	8.9 21.22	22.61
12	Male(Urban+Rural) Vs. Female(Urban+Rural)	5290+3720 4917+3573	112.6 106.12	22.52 21.22	32.81
13	Male+Female(Urban) Vs. Male+Female (Rural)	5290+4917 3720+3573	127.6 91.16	25.52 18.23	32.81

1. As the CV (16.89) greater than TV (1.96), the mean performance of male urban was better than that of male students in rural areas
2. As the CV (19.14) greater than TV (1.96), the mean performance of male urban was better than that of female students in urban areas
3. As the CV (26.80) greater than TV (1.96), the mean performance of male urban+rural was better than that of male students in urban areas
4. As the CV (29.05) greater than TV (1.96), the mean performance of male+female urban was better than that of male students in urban areas
5. As the CV (13.67) greater than TV (1.96), the mean performance of male rural was better than that of female students in rural areas
6. As the CV (23.81) greater than TV (1.96), the mean performance of male urban+rural was better than that of male students in rural areas
7. As the CV (20.65) greater than TV (1.96), the mean performance of male+female rural was better than that of male students in rural areas
8. As the CV (15.9) greater than TV (1.96), the mean performance of female urban was better than that of female students in rural areas
9. As the CV (25.14) greater than TV (1.96), the mean performance of female urban+rural was better than that of female students in urban areas
10. As the CV (28.36) greater than TV (1.96), the mean performance of male+female urban was better than that of female students in urban areas
11. As the CV (22.61) greater than TV (1.96), the mean performance of female urban+rural was better than that of female students in rural areas
12. As the CV (32.81) greater than TV (1.96), the mean performance of male urban+rural was better than that of female students in urban+rural areas
13. As the CV (32.81) greater than TV (1.96), the mean performance of male+female urban was better than that of male+female students in rural areas

3.5 Correlation Co-efficient Components

To compare the internal reliability of the test the scores obtained by the students were transferred into percentages. The percentages on the tests were calculated with the results of the newly developed, test and the correlation coefficient were determined

$$\mathbf{r} = \frac{\sum xy}{\sqrt{(\sum x2)(\sum y2)}}$$

The research study was based on a statistical design. Calibrations of intelligence tests were done through item analysis. The procedures adopted for item analysis were of two types Traditional and Rasch Model. The latest method of item analysis and rasch calibration was adopted to testify the results derived by the traditional method. It was in fact the confirmatory technique leading to the standardization of the test.

Table 8: Correlation Coefficient Components, (Class VIII)

	Test 1	**Test 2**	**Test 3**	**Test 4**	**Test 5**
Test 1	1.00	.45	.74	.39	.62
Test 2	.45	1.00	-.12	-.009	-.56
Test 3	.74	-.14	1.00	.007	.19
Test 4	.39	-.009	.007	1.00	.26
Test 5	.62	-.56	.19	.26	1.00

The table 8 shows that all the tests are correlated with each other but test 1 and test 5 are highly correlated with all other sets as their ranges are very much clear to decide their correlation. The correlation between the results of all the tests were (divided into two halves i.e., odd or even) calculated by Pearson Product Formula (r). The internal reliability coefficient for class VIII, test number 1 was 0.85, for test number 2 was 0.80, for test number 3 was 0.64, for test number 4 was 0.45 and for test number 5 was 0.35

3.6 Item Analysis

Item analysis is basically related to construction of any informal test such as quizzes and exams made by teacher for the sake of student's knowledge and abilities generally. Items can be analyzed statistically through main three properties either qualitative or quantitative.

Facility Index	(F %)
Discrimination Index	(D)
Power of Discrimination	(Ø)

3.6.1 Facility index

$$F\% = \frac{N_R}{N_T} \times 100$$

Where F% = Facility Index, NR= Number of the students, they attempted the item correct, NT = Total number of students, Criteria: The standard value of facility index is from 30% to 80%

3.6.2 Discrimination Index

$$D = \frac{N_h - N_l}{n}$$

Where D = Discrimination Index, NH = Number of the students in A group, they attempted the items correct, NL = Number of the students in B group, they attempted the items correct, n = number of students in each group, Criteria: The standard value of discrimination index is 0.20 or above

3.6.3 Power of Discrimination

$$Ø = \frac{ad - bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$$

Where Ø = Power of discrimination, a = number of students giving answer correct in high achievers, b = number of students giving answer in-correct in high achievers, c = number of students giving answer correct in low achievers, d = number of students giving answer in-correct in low achievers, Criteria: The standard value of discrimination index is 0.20 or above

Table 9: Item Analysis Sheet, Group A (Class VIII)

Item #	Groups	Responses		D	Ǿ
		Correct	Incorrect		
1	A	3741	634	.36	.36
	B	554	3821		
2	A	3821	554	.35	.35
	B	723	3652		
3	A	3852	523	.33	.34
	B	963	3412		
4	A	3924	451	.36	.36
	B	788	3587		
6	A	3652	723	.34	.35
	B	678	3697		
7	A	3674	701	.33	.34
	B	723	3652		
8	A	3652	723	.33	.34
	B	734	3641		
9	A	3641	734	.33	.33
	B	688	3687		
10	A	3625	750	.36	.35
	B	410	3965		
11	A	3621	754	.35	.34
	B	522	3853		
12	A	3975	400	.40	.39
	B	414	3961		
13	A	3963	412	.38	.39
	B	630	3745		
14	A	3965	410	.38	.39
	B	624	3751		
15	A	3932	443	.37	.37
	B	643	3732		
15	A	3856	519	.35	.35
	B	773	3602		
16	A	3850	525	.35	.36
	B	754	3621		

Table 9 indicates the values D and phi (**Ǿ**) with respect to the students of class VIII. The value of D falls between .33 to .40 and the value of phi (**Ǿ**) also falls among .33 to .39 which represents that item number 4 should be improved. Table 10 indicates the values D and phi (**Ǿ**) with respect to the students of class VIII. The value of D falls between .27 to .38 and the value of phi (**Ǿ**) also falls among .28 to .38 which represents that item number 12 should be improved. Table 11 indicates the value of F with respect to the students of class VIII. It falls among 59 % to 75 % which represents that none of the item is rejected, for the investigation of the suitability of multiple-choice items in the test. I scrutinized the behavior of each of distracters, none of the distracter ought to be re-examined as they were attracting both groups almost equally. None of the items ought to be replaced as they are more attractive to the members of the A group than to the B group.

Table 10: Item Analysis Sheet, Group B (Class VIII)

Item #	Groups	Responses		D	Ǿ
		Correct	Incorrect		
1	A	3412	963	.30	.30
	B	761	3614		
2	A	3514	861	.34	.35
	B	530	3845		
3	A	3369	1006	.27	.28
	B	954	3421		
4	A	3578	797	.33	.33
	B	690	3685		

6	A	3648	727	.34	.33
	B	630	3745		
7	A	3956	419	.35	.36
	B	851	3524		
8	A	3589	786	.35	.35
	B	480	3895		
9	A	3487	888	.31	.32
	B	721	3654		
10	A	3652	723	.35	.35
	B	534	3841		
11	A	3896	479	.37	.36
	B	620	3755		
12	A	3588	787	.33	.33
	B	665	3710		
13	A	3479	896	.31	.31
	B	724	3651		
14	A	3621	754	.32	.32
	B	763	3612		
15	A	3685	690	.32	.33
	B	855	3520		
15	A	3874	501	.38	.38
	B	535	3840		
16	A	3534	841	.29	.30
	B	948	3427		

Table 11: Item Analysis Sheet, Class VIII

S #	Groups	1	2	3	4	Omitted	Total Correct Answers	Total Incorrect Answers	F% (Total Correct Answers
1	A	6810	874	890	925	12	12457	5017	71
	B	5647	912	876	540	14			
2	A	845	6830	905	811	14	12354	5118	70
	B	870	5524	823	864	14			
3	A	7009	790	811	784	14	12598	4873	72
	B	5589	810	845	833	15			
4	A	5775	1046	1011	1102	12	11246	6229	64
	B	5471	1078	1002	990	13			
5	A	1103	5880	952	1047	14	11201	6267	64
	B	985	5321	1057	1123	18			
6	A	5595	1144	966	1059	19	11023	6443	63
	B	5428	1183	1078	1013	15			
7	A	5842	1114	1025	1011	12	11058	6417	63
	B	5216	1121	1156	990	13			
8	A	5761	987	963	1015	14	11248	6223	64
	B	5487	1104	988	1166	15			
9	A	5810	1132	970	1014	15	11256	6214	64
	B	5446	1115	1021	962	15			
10	A	5869	1203	945	852	14	11235	6239	64
	B	5366	1014	963	1262	12			
11	A	1023	5822	970	1014	13	11324	6150	65
	B	1041	5502	985	1117	13			
12	A	1256	5072	1025	1012	12	10287	7189	59
	B	1241	5215	1104	1551	12			
13	A	5614	1309	1100	1087	10	10698	6781	61
	B	5084	1205	1108	972	11			
14	A	7069	774	854	925	11	12549	4931	72
	B	5480	889	914	575	9			
15	A	6238	952	990	986	9	11457	6036	65
	B	5219	987	1002	1119	8			

16	A	1062	6148	940	1045	7	11332	6151	65
	B	1002	5184	965	1137	10			

4. Findings

1. The correlation between the results of all the tests were (divided into two halves i.e., odd or even) calculated by Pearson Product Formula (r). The internal reliability coefficient for class VIII, test number 1 was 0.85, for test number 2 was 0.80, for test number 3 was 0.64, for test number 4 was 0.45 and for test number 5 was 0.35
2. Using Kudar and Richardson's formula, KR # 20 reliability (r) was -.43 and it was calculated for all five test at the same for class VIII students of both genders form rural and urban areas of 35 districts of Punjab.
3. Using Kudar and Richardson's formula for KR # 21 the results were same like KR # 20 and were calculated at -.43.
4. The value of standard error was also calculated and was ranging from 1.97 to 2.14 for class VIII which is the indication of reasonable consistency.
5. Item analysis was calculated on the basis of each test, picking up the total correct attempts. The difficulty (facility indices) of the item for class VIII, its values are for test 1 (58 % to 75 %), for test 2(58 % to 71 %), for test 3(58 % to 72 %), for test 4 (58 % to 70 %) and for test 5(59 % to 75 %)
6. Item analysis was also conducted on the basis of each test (16 items) picking up 25% of cases from the top (Upper) and 25% from the bottom (Lower). The discriminatory index (D) was calculated with the help of formula.

Its values for class VIII are for test 1 (.27 to .38), for test 2 (.30 to .39), for test 3 (.31 to .40), for test 4 (.32 to .39) and for test 5 (.33 to .40). Item difficulty was determined on the basis of correct and incorrect responses given by each test. Item discrimination was calculated with the help of Phi-Coefficient (Ø) formula. Its values for class VIII were, For test 1 (.28 to .38), for test 2 (.30 to .39), for test 3 (.31 to .40), for test 4 (.32 to .40) and for test 5 (.33 to .39)

The value of Z was calculated with the help of formula to find the significant difference between the mean performance of male and female, urban and rural students.

1. The number of the students in case of male (urban) was 5290, while in case of male (rural) 3720. Mean sore of male (urban) and male (rural) was 66.1 and 46.5 and the standard deviation was 13.22 and 9.3 respectively.
2. The number of the students in case of male (urban) was 5290, while in case of female (urban) 4917. Mean sore of male (urban) and female (urban) was 66.1 and 66.5 and the standard deviation was 13.22 and 12.3 respectively.
3. The number of the students in case of male (urban) was 5290, while in case of male (urban+rural) 5290+3720. Mean sore of male (urban) and male (urban+rural) was 66.1 and 112.6 and the standard deviation was 13.22 and 22.52 respectively.
4. The number of the students in case of male (urban) was 5290, while in case of male+female (urban) 3720. Mean sore of male (urban) and male+female (urban) was 66.1 and 12.6 and the standard deviation was 13.22 and 25.52 respectively.
5. The number of the students in case of male (rural) was 3720, while in case of female (rural) 3573. Mean sore of male (rural) and female (rural) was 46.5 and 44.66 and the standard deviation was 9.3 and 8.9 respectively.
6. The number of the students in case of male (rural) was 3720, while in case of male (urban+rural) 5290+3720. Mean sore of male (rural) and male (urban+rural) was 46.5 and 112.6 and the standard deviation was 8.9 and 22.5 respectively.
7. The number of the students in case of male (rural) was 3720, while in case of male+female (rural) 3720+3573. Mean sore of male (rural) and male+female (rural) was 46.5 and 91.16 and the standard deviation was 9.3 and 18.23 respectively.
8. The number of the students in case of female (urban) was 4917, while in case of female (rural) 3573. Mean sore of female (urban) and female (rural) was 61.44 and 44.66 and the standard deviation was 12.3 and 8.9 respectively.
9. The number of the students in case of female (urban) was 4917, while in case of female (urban+rural) 4917+3573. Mean sore of female (urban) and female (urban+rural) was 61.46 and 106.12 and the standard deviation was 12.3 and 21.22 respectively.

10. The number of the students in case of female (urban) was 4917, while in case of male+female (urban) 5290+4917. Mean sore of female (urban) and male+female (urban) was 61.46 and 127.6 and the standard deviation was 12.3 and 25.52 respectively.
11. The number of the students in case of female (rural) was 3573, while in case of female (urban+rural) 4917+3573. Mean sore of female (rural) and female (urban+rural) was 44.66 and 1.6.12 and the standard deviation was 8.9 and 21.22 respectively.
12. The number of the students in case of male (urban+rural) was 5290+3720, while in case of female (urban+rural) 4917+3573. Mean sore of male (urban+rural) and female (urban+rural) was 112.6 and 106.12 and the standard deviation was 22.52 and 21.22 respectively.
13. The number of the students in case of male+female (urban) was 5290+4917, while in case of male+female (rural) 3720+3573. Mean sore of male+female (urban) and male+female (rural) was 127.6 and 91.16 and the standard deviation was 25.52 and 18.23 respectively.

5. Conclusions

Mathematics performs an essential position to increase thinking, reasoning, and trouble fixing competencies that permit people to end up excellent citizens. Mathematics may be outstanding from different topics because of its ordinary language, symbols, and summary principles. Students face problems in studying arithmetic, a number of which can be attributed to coaching, specifically with an unmarried trainer coaching the concern. An unmarried trainer cannot address all troubles efficiently due to time, energy, expertise, methods, and shortage of interplay with college students individually. AT is a coaching technique wherein or extra instructors collaboratively plan, organize, present, and examine their coaching. It has exclusive settings like one trainer coaching and one assisting, teaming, and parallel coaching. Literature suggests that the Standardized fulfillment Test technique is extra high-quality than different arithmetic coaching procedures in phrases of college students' studying.

In the context of Pakistan, arithmetic is taught predominantly with the aid of using one trainer. Moreover, arithmetic instructors do now no longer collaborate with colleagues to speak about principles or methodologies of coaching which ends up in low fulfillment of college students on this concern. Keeping in view the significance of Standardized fulfillment Test, the targets of this observe have been to: study the effect of standardization on eighth grade college students' fulfillment in arithmetic, study Standardized fulfillment Test's effect on content material strands of arithmetic (algebra and geometry), study Standardized fulfillment Test's effect on mathematical competencies (conceptual expertise, procedural expertise, and trouble fixing), and discover the ideals of college students approximately arithmetic and coaching of arithmetic in standardized settings.

The nature of observe changed into in particular centered on quantitative aspects; the usage of experimental studies. A test changed into performed on eighth grade college students the usage of the Solomon Four Group experimental studies layout. This layout includes 4 organizations and randomly assigns the topics to the organizations. Observe changed into delimited to eighth grade within side the concern of arithmetic. All the scholars of eighth grade reading within side the public colleges of district of Punjab; Pakistan changed into the populace of this observe. I confronted trouble within side the choice of a public faculty as a pattern because of reasons. The first motive changed into the dearth of willingness of the headmasters. Most of the heads of public colleges refused to permit the test due to random undertaking of college students into 4 organizations. The 2nd motive changed into the dearth of availability of arithmetic instructors, every with M.Sc. (Mathematics) and a B.Ed. Finally, one public faculty changed into decided on from the districts of Punjab. All to be had college students reading within side the eighth grade, i.e. 118 participated within side the test? I assigned 118 college students to 4 organizations randomly via SPSS-16. Two volunteer arithmetic instructors (every with M.Sc. in Mathematics and a B.Ed.) from the sampled faculty participated on this observe. I, having the identical qualifications, additionally took element within side observe as a co-trainer. I held schooling classes for the 2 arithmetic instructors over days, with hours every day. I defined the studies targets to the 2 arithmetic instructors, and shared information of the syllabus to be protected and the time table of durations with the aid of using subjects and dates. I requested each the arithmetic instructors to be everyday and punctual all through the test. Further, I mentioned with the co-trainer of observe approximately standardization separately.

Achievement checks and trying out are a part of education, business, and the law of professions within side the United States and are growing in use internationally. The improvement of high-quality, handy fulfillment checks calls for tremendous expertise of a content material area – which include arithmetic, language arts, or science – and the layout of take a look at gadgets or responsibilities which can be truthful and legitimate measures of essential expertise and capabilities in a given content material area. The choice and sound use of fulfillment checks additionally calls for tremendous capabilities to make certain they're suitable for the motive meant and to keep away from poor consequences. There are vast sources and an abundance of records on fulfillment checks and sound trying out practices. With suitable interest and schooling, fulfillment checks can offer customers treasured records approximately learners' fulfillment development and status.

Standardized Achievement Test is extra powerful than the conventional coaching of arithmetic on the eighth grade degree in enhancing educational fulfillment of the scholars within side the mathematical proficiencies of conceptual expertise and procedural expertise. Standardized Achievement Test did now no longer enhance college students' trouble fixing capacity drastically extra than the conventional approach of coaching arithmetic. Standardized Achievement Test is extra powerful than the conventional coaching on the eighth grade degree in enhancing educational fulfillment of college students within side the mathematical content material strands of algebra and geometry. In the content material strand of algebra the Standardized Achievement Test is higher than the conventional approach of coaching in enhancing the instructional fulfillment of college students in conceptual expertise and procedural expertise. However, it did now no longer enhance their trouble fixing capacity drastically extra than the conventional approach of coaching arithmetic. In geometry Standardized Achievement Test changed into extra powerful than the conventional approach of coaching arithmetic in enhancing the instructional fulfillment of college students handiest in trouble fixing however changed into now no longer drastically extra powerful than the conventional approach of coaching in case of conceptual expertise and procedural expertise capacity. The college students' ideals approximately arithmetic and coaching of arithmetic within side the collaborative placing may be modified undoubtedly the usage of Standardized Achievement Test. Through using Standardized Achievement Test in arithmetic lecture room at eighth grade, the ideals of college students may be modified approximately the usefulness of arithmetic in ordinary life, usefulness of co-coaching in arithmetic class, and powerful studying of mathematical principles via doing activities.

6. Recommendations

In the illumination of findings of the research, the under given recommendations are:

1. A parallel form of this test should be constructed to provide for congruent validity.
2. A follow up study of the present sample should be under taken so as to calculate correlation between their performance on this test at present and later to find out the predictive validity of the test.
3. Fresh version of this test after recommended replacements and modifications may be administered to different samples of different grades from Pakistan in order to compare the results of students from different backgrounds.
4. Teachers are suggested to make use of the "Rasch Model" for the calibration of their tests in addition to the traditional methods of item analysis and test calibration.

References

Aggarwal, J. (2019). *Theory & Principles Of Education*: Vikas Publishing House.

Airasian, P. W., & Madaus, G. F. (1983). Linking testing and instruction: Policy issues. *Journal of Educational Measurement, 20*(2), 103-118.

Aud, S., Hussar, W., Kena, G., Bianco, K., Frohlich, L., Kemp, J., & Tahan, K. (2011). The Condition of Education 2011. NCES 2011-033. *National Center for Education Statistics*.

Austin, G. R., & Garber, H. (2013). *The rise and fall of national test scores*: Academic Press.

Bennett, W. (2011). *A Report on Elementary Education in America* Retrieved from Washington: https://eric.ed.gov/?id=ED270236

Bloom, B. S. (1971). Handbook on formative and summative evaluation of student learning.

Carpenter, T. P., & Moser, J. M. (1984). The acquisition of addition and subtraction concepts in grades one through three. *Journal for research in Mathematics Education, 15*(3), 179-202. doi:10.5951/jresematheduc.15.3.0179

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*(2), 93-106. doi:10.1111/j.1745-3984.1993.tb01068.x

Copeland, R. W. (1970). How Children Learn Mathematics, Teaching Implications of Piaget's Research.

Dede, C., & Freiberg, H. J. (1987). *The long-term evolution of effective schools.* Paper presented at the The Educational Forum.

Digumarti, B. R. (1994). *Scientific aptitude*: APH Publishing.

Ebel, R. L., & Frisbie, D. A. (1972). Essentials of educational measurement.

Gronlund, N. E., & Linn, R. L. (2000). *Measurement and Evaluation in Teaching.* Retrieved from New York:

Gulliksen, H. (1950). Intrinsic validity. *American Psychologist, 5*(10), 511.

Hambleton, R. K. (1989). Principles and selected applications of item response theory.

Harkness, J. (2020). *Mastery. In B. Fischer (Ed.), Today's Education*. Retrieved from Washington, DC:

Hogan, P. C. (2003). *The mind and its stories: Narrative universals and human emotion*: Cambridge University Press.

Hooman, J. (1993). *A compositional approach to the design of hybrid systems.* Paper presented at the International Hybrid Systems Workshop, International Hybrid Systems Workshop.

Johnson, L. (2006). The sea change before us. *Educause Review, 41*(2), 72.

Jurgens, P. P. (1987). Effects of standardized achievement tests on mathematics education.

Leung, A. N. M., Fung, D. C.-L., & Farver, J. M. (2018). A cyberbullying intervention for Hong Kong Chinese college students. *Applied Research in Quality of Life, 13*(4), 1037-1053. doi:10.1007/s11482-017-9572-1

Parkay, F., O'Bryan, S., & Hennessy, M. (2014). *Quest for quality*: University Publications New York.

Popham, W. J. (2004). Curriculum, instruction, and assessment: Amiable allies or phony friends? *Teachers College Record, 106*(3), 417-428.

Salma, W. (2010). *Elementary Education*. Retrieved from Islamabad:

Schwartz, B. I. (2014). Chinese communism and the rise of Mao. In *Chinese Communism and the Rise of Mao*: Harvard University Press.

Suter, W. N. (2011). *Introduction to educational research: A critical thinking approach*: SAGE publications.

Vision, C. (2006). Retrieved on March 16, 2007 from http://www. careervision. org/Consulting. *FAQs. htm*.

Wardrop, J. L. (2011). *Standardized testing in the schools: Uses and roles* (0818501731). Retrieved from

Wiersma, W., & Jurs, S. G. (1985). *Educational measurement and testing*: Allyn & Bacon.

Woolfolk, A., & Karlberg, M. (2004). *Pedagogisk psykologi*: Tapir akademisk forlag.

Zarei, H. A., Mirhashemi, M., & Pasha Sharife, H. (2013). The Relationship between Thinking Styles and Academic Adjustment among the Students of Islamic Azad University-Khoy Branch. *The Journal of Modern Thoughts in Education, 8*(1), 30-19.