Chapter IV

Construction and Standardization of the Mathematics Achievement Test

CHAPTER 4

CONSTRUCTION AND STANDARDIZATION OF THE MATHEMATICS ACHIEVEMENT TEST

4.1 Introduction

The present chapter is a detailed description about the construction and standardization of the mathematics achievement test. The entire process of standardization from item writing to establishing of norms is described. The procedure adopted is presented along with the purpose and the essential descriptive statistics employed. The sample taken for each step is also given in this chapter.

4.1.1 Purpose of the Mathematics Achievement Test

The test was constructed to measure the competencies possessed by eighth standard students. It was a survey of the mathematics achievement. The test items were chosen from seventh standard mathematics test book, of Directorate of Education Govt. of Goa. From eight standard onwards mathematics is further categorized in to algebra, geometry, trignometry. However, the basic fundamentals are taught by seventh standard. Hence, the test was made to measure the extent to which the fundamentals have been achieved before the students enter eighth standard.

The objectives of teaching mathematics at upper primary according to National Curriculum Framework (2000) are that the students should acquire knowledge and understanding of facts, concepts, principles of mathematics needed for daily use, practical geometry, simple mensuration, descriptive preliminary aspects of statistics and fundamentals of algebra. At the secondary stage, the teaching – learning of mathematics should be to further enhance the capacity of the students to employ mathematics in solving day-to-day problems.

4.2 Item Writing and Item Selection

The items to be included in the test were selected on the basis of discussion with mathematics teachers of secondary level, researchers, guide. In order to include all the types of items from the content of subject matter under consideration, content analysis was done. A pool of test items, based on seventh standard textbook published by Directorate of Education, Govt. of Goa, was made. These test items were categorized chapter wise (refer Appendix-A). Test items were taken from arithmetic, algebra, geometry. The test items were tried out on a sample of ten students, taken at random. On the basis of the test scores and further discussion with mathematics teachers, test items based on geometry were dropped. The teachers were of the opinion that students generally score well in geometry. The try out version consisted of hundred items. With the omitting of test items based on geometry, the number of test items were reduced to ninety three. The test items were constructed with the following criteria:

i) each test item has only one outcome

ii) language of the instruction is clear and simple

The test was again given to mathematics teachers, guide, language experts, for adequacy of test items, language of the instructions. Enough space was provided for calculations, in the test paper itself. Blanks were provided for writing the final answer, along with the question. The test is given in Appendix-B.

4.3 Pilot study

The investigator personally administered the test in all the schools. No time limit was prescribed for the test, as this was a power test. Apart from the instructions provided along with the test, investigator oriented the students regarding the purpose and nature of the test, the manner in which the test is to be answered. Investigator had to elaborate upon some of the instructions and the nature of response expected. All the doubts raised by students regarding instructions and response pattern were explained to the students by the investigator. The time taken for the test was about one hour and thirty minutes.

4.3.1 Purpose

The pilot study was done on small sample similar to the larger sample to be tested later. The test was administered on the small sample to obtain the following:

i) the difficulty value of each item

ii) the discriminating power of each test item

iii) the adequacy of the item instruction of the time limit, test format

4.3.2 Sample

The pilot study was conducted on a sample of three hundred and fifty pupils, in six schools. Cluster sampling technique was employed.

Pilot study version consisting of ninety-three items was administered in the six schools personally by the investigator. No time limit was given. However, the students took on an average one hour and thirty minutes.

4.3.3 Scoring and Item Analysis

Item analysis brings to light general areas of weakness requiring attention, reveals ambiguities, technical defects with regard to language of instruction, mode of presentation order of presentation. The test constructed being a norm-referenced test, a suitable procedure of comparing upper and lower twenty-seven percent of the group on the basis of the test performance, was adopted. The responses of the remaining pupils not included in the analysis, were assumed to follow the same trend as those in the upper and lower groups.

Step I : After collecting the test papers from the students, the papers were scored, one mark for the right response and no marks for the wrong response.

Step II: When all the test papers were scored they were arranged in the ascending order of marks. Papers of all the schools were considered together. The upper 27% and lower 27% of the students arranged in the ascending order with regard to their scores, were selected. Hence for the item analysis low scoring students and high scoring students.

Step III: The scores were tabulated against the test items. For each test item responded rightly, entries were made in the table against each score. Thus total number of correct responses for each item was found. The tabulation of the responses facilitates in obtaining estimate of item difficulty, item discrimination power. Although item analysis reveals the general effectiveness of a test item, it is desirable to obtain more precise estimate item difficulty and discrimination power.

4.3.4 Item Difficulty

The item difficulty of a test, item is indicated by the percentage of pupils who get the item right. It was estimated by the formula, instructions were given to the students to avoid copying. In the case of Gujarati medium schools, help was sought from the mathematics teachers of the school, to clarify the doubts regarding the instructions. Item difficulty of a test, is indicated by the percentage of students who get the item right. It was estimated by the formula,

$$\text{Item difficulty} = \frac{R}{T} \times 100$$

R – the number of students who got the item right

T – the total number of student who attempted the item

Higher the item difficulty estimate lower is the difficulty of the item. Items with item difficulty more than seventy percent and less than forty percent were not considered for the final administration. The item difficulty was decided by consulting subject experts, statisticians. Only forty eight item were retained for the final administration. The item difficulty of the items included in the test are given in Table (4.3.3.1).

Table 4.3.3.1

Item Difficulty & Discriminating Power

Item No. (Final Test)	Item No. (Pilot Study)	Item Difficulty %	Discriminating Power
1.a	3.2a	62.43	0.38
1.b	3.2b	60.15	0.46
1.c	3.2c	58.42	0.48
2.a	1.1a	58.01	0.38
2.b	1.1b	57.44	0.42
2.c	1.1c	57.34	0.42
3.a	2.2b	57.22	0.33
3.b	2.2d	56	0.36
4.	1.4a	56	0.38
5.a	2.1b	55	0.38
5.b	2.1a	55	0.57
5.c	2.1c	55.7	0.43
5.d	2.1d	54	0.38
5.e	2.1g	54	0.58
6.a	3.5a	53.35	0.41
6.b	3.5f	53.29	0.48
6.c	3.5d	53.24	0.41
6.d	3.5e	53.22	0.36
6.e	2.5c	53.11	0.35
6.f	3.3c	52.44	0.45
6.g	3.1a	52.19	0.43
6.h	3.1b	52.15	0.44
6.i	3.5g	52.10	0.43
7.a	1.2a	52.09	0.35
7.b	1.2b	52.05	0.43
8.a	9.1a	52.02	0.47

8.b	9.1b	52.01	0.41
9.a	2.6a	51.45	0.43
9.b	2.6b	51.44	0.41
9.c	2.6d	51.43	0.37
9.d	2.6e	51.41	0.42
10.a	9.2a	51.22	0.47
10.b	9.6a	51.22	0.42
10.c	9.6b	51.13	0.41
11.a	9.3	50.05	0.42
11.b	11.5	50.03	0.48
12.a	10.1	49.40	0.43
12.b	10.2	49.33	0.45
12.c	10.4	49.07	0.42
12.d	10.3	49.01	0.42
12.e	10.5	48.11	0.42
13.a	12.5	48.11	0.42
13.b	11.1b	46.10	0.35
13.c	11.1c	45.12	0.35
14.a	11.2	45.07	0.39
14.b	11.3	45.01	0.33
15.	5.2	44.05	0.38
16.	6.1	44.03	0.36

4.3.5 Item Discrimination

The item discrimination power of the item refers to the degree to which it discriminates between students with high and low scores. An estimate of item discrimination was obtained by the formula,

$$\text{Discriminating power} = \frac{R_u - R_L}{\frac{1}{2}T}$$

R_u – number of students of the upper groups who got the item right

R_L – number of students of the lower group who got the item right

The items having discriminating power approximately 0.5 were retained. The item discriminating power of the items included in the test are given in Table (4.3.3.1).

4.4 Administration of the Final Version

The final version of the test was constructed on the basis of the item difficulty. Items were arranged in the increasing order of difficulty. Items with difficulty value between forty percent and seventy percent were included. The test items had discriminating power approximately 0.50. The test was administered to establish the reliability, norms. The test had forty-eight items after dropping test items which were found to be very difficult. The test is given in Appendix-C.

4.4.1 Sample

The test was administered on a sample of three hundred seventy seven in ten schools. Cluster sampling technique was used. The sample consisted of boys and girls.

4.4.2 Test Administration

The investigator administered the test in each of the school. There was no time limit, the test being a power test. There were some clarifications to be made during administration. Sufficient time was provided for completion of the test. On an average it took about forty five minutes. Students were made to sit, one on a bench to avoid copying.

4.4.3 Scoring and Analysis

After the administration was over in each school, the test papers were scored according to scoring key. The scores for each student was tabulated, showing the total score. In the following sections the detailed procedure of establishing reliability, validity and norms are described in detail. The test papers were scored according to the scoring key. The scores were arranged in increasing order.

4.5 Establishing Reliability

Reliability refers to the consistency of measurement. It provides the consistency which makes validity possible and indicates how confident one can be with the results. Unless the measurement can be shown to be reasonably consistent over different occasions or over different samples of the same behavior, little confidence can be placed in the results. However, one cannot expect test results to be perfectly consistent. There are numerous factors other than the quality being measured which may influence the test score. In determining reliability it would be desirable to obtain two tests measured under identical conditions and then compare the results. This procedure is difficult, of course, since the conditions under which evaluation data are obtained can never be identical. As a substitute for this for this ideal procedure several methods of estimating reliability have been introduced.

4.5.1 Split Half Method

The method used in establishing reliability of the test was split-half method. The reliability of the test-scores was estimated from a single administration of a test. The test was administrated on the sample selected for final administration, and then it was divided in half for scoring purposes. To split the test into halves which are most equivalent, the usual procedure is to score the even numbered items and the odd numbered items separately. This provides, two scores for each student, which, when correlated, provides a measure of internal consistency. This coefficient indicates the degree to which the two halves of the test are equivalent. To estimate the reliability of the scores based on the full length test the Spearman Brown formula was applied. This formula is as followed:

$$\text{Reliability on full test} = \frac{2 \times \text{Reliability on } \frac{1}{2} \text{ test}}{1 + \text{Reliability on } \frac{1}{2} \text{ test}}$$

According to Garret (1953) the split-half method is generally regarded as the best of the method for determining test reliability owing to certain advantage. They are (i) data is obtained on one single administration (ii) similar conditions prevail in the administration of the two halves (iii) the coefficient of correlation is more

reliable (iv) the question of practice effect does not arise. The product-moment coefficient of correlation between the scores on even and odd items was calculated. The scores on even and odd items of all the students of the sample were arranged in a bivariate table. To establish the reliability of the scores based on the full-length test the Spearman Brown formula was applied. The reliability was found to be 0.8723.

4.5.2 KR 21 Method

Reliability was also established using KR 21 formula. This method also provides a measure of internal consistency without splitting the test into half. The formula for reliability estimate by KR 21 is –

$$\text{Reliability estimate (KR 21)} = \frac{K}{K-1}\left(1 - \frac{M(K-M)}{KS^2}\right)$$

K = the number of items in the test

M = the mean of the test scores

S = the S.D. of the test scores

Reliability estimate was found to be 0.7965.

4.6 Establishing Validity

Validity refers to the extent to which the results of an evaluation procedure serve the particular uses for which they are intended. Basically, validity is always concerned with specific use to be made of evaluation results. Validity pertains to the results of the test and not to the instrument itself. It is a matter of degree. It is best considered in terms of categories. The results of an arithmetic test may have a high degree of validity for indicating computational skill, but a low degree of validity for indicating mathematical reasoning. Thus, when describing validity, it is necessary to consider the use of be made of the results. The purpose of the test being constructed was to find the achievement in mathematics. Hence it would be beneficial to have a test with high content validity. Content validity is the extent to which a test measures a representative sample of the subject matter content and the behavioral changes under consideration. The focus of content validity is on the adequacy of test items. The test is examined to determine the subject matter content covered and the

responses are intended to make to the content, and this is compared with the domain of achievement to be measured. The procedures used are those of logical analysis and comparison. In order to establish content validity the test items were compared with (refer Appendix – A). The items were found to be adequately covering the content-areas under consideration. The test was also given to mathematics teachers, guide, researchers, for further examination, of the contents, and were found sufficient with regard to the mathematics backwardness in eighth standard students.

4.7 Administration of the Test for Norms

The standardized test was administered to establish norms like mean, median standard deviation, percentile, skewness kurtosis. The sample consisted of five hundred and eighty six students from fourteen schools. The investigator personally administered the test in each school. No time limit was given. It took almost forty five minutes. The test papers were scored using scoring key.

4.8 Establishing Norms

Direct results of tests are referred as raw scores. Raw scores are seldom directly meaningful until they are compared with standards. One type of standard is the performance of larger groups, possibly including students of other schools and other cities. Scores obtained from such large groups of students are called norms. By comparing the average performance in the class with the average performance in other schools and other cities, the teacher would have a better idea of how well her class is progressing. One of the most widely used and easily comprehended methods of describing test performance is that of percentile rank. A percentile rank (percentile score) indicates a pupil's relative position in a group in terms of the percentage of pupils scoring below him. From the table of norms, if one finds that a pupil's raw score of 10.181 equals percentile rank of sixty, one could conclude that sixty percent of the pupils in the reference group obtained a score lower than 10.181. Most commonly, performance is reported in terms of the pupil's relative standing in his own grade or age group.

Another method of indicating a pupil's relative position in a group is by showing how far his raw score is above or below average. The test performance is expressed in terms of mean and standard deviation. Mean and standard deviation were computed for boys, girls. Mean and standard deviation were also computed for the whole sample.

The norms for the whole sample is given in Table (4.8.1). The norms for boys is given in Table (4.8.2). The norms for girls is given in Table (4.8.3).

Table 4.8.1

Norms for Entire Sample

Mean	9.642	Percentile	N=586
Median	8.653	10	2.911
Mode	4.00	20	4.522
Kurtosis	0.435	30	5.864
Skewness	0.778	40	7.224
S.D.	5.951	50	8.653
		60	10.181
		70	12.006
		80	14.454
		90	18.275

Table 4.8.2

Norms for Boys

Mean	9.506	Percentile	N=322
Median	8.944	10	1.964
Mode	7.00	20	4.191
Kurtosis	0.721	30	5.790
Skewness	0.769	40	7.210
S.D.	6.179	50	8.944
		60	10.304
		70	11.900
		80	14.271
		90	18.233

Table 4.8.3

Norms for Girls

Mean	9.763	Percentile	N=264
Median	8.484	10	3.391
Mode	5.00	20	4.75
Kurtosis	0.129	30	5.909
Skewness	0.802	40	7.233
S.D.	5.756	50	8.484
		60	10.00
		70	12.111
		80	14.667
		90	18.300

4.9 Administration of the Standardised Mathematics Achievement Test

The standardised mathematics achievement test was administered to select the sample for diagnosis. The sample consisted of three hundred and fifty eight students from ten schools. The investigator personally administered the test in each school. No time limit was given. It took almost forty-five minutes.

4.10 Scoring and Sample Selection for Diagnosis

The scoring of the test papers were done using the answer key. The descriptive statistics were computed. The mean, median, mode, skewness, kurtosis, S.D., percentile for the entire sample, boys and girls, are given in Table 4.10.1, Table 4.10.2 and Table 4.10.3 respectively. The sample for diagnosis were selected from four schools from among the ten schools. The four schools were chosen randomly. The sample consisted of only those students from the four schools who scored below 5.864 (thirty percentile). The sample for diagnosis consisted of one hundred and sixty students.

Table 4.10.1

Descriptive Statistics for Entire Sample

Mean	7.952	Percentile	N=358
Median	6.815	10	1.583
Mode	5.000	20	3.158
Kurtosis	0.038	30	4.422
Skewness	0.769	40	5.517
S.D.	5.496	50	6.815
		60	8.315
		70	10.008
		80	12.739
		90	16.269

Table 4.10.2

Descriptive Statistics for Boys

Mean	7.919	Percentile	N=168
Median	6.918	10	1.763
Mode	5.000	20	3.302
Kurtosis	0.336	30	4.530
Skewness	0.838	40	5.660
S.D.	5.372	50	6.918
		60	8.300
		70	9.832
		80	12.487
		90	15.354

Table 4.10.3

Descriptive Statistics for Girls

Mean	7.992	Percentile	N=190
Median	6.656	10	1.427
Mode	5.000	20	2.958
Kurtosis	0.407	30	4.302
Skewness	0.697	40	5.351
S.D.	5.653	50	6.656
		60	8.338
		70	10.287
		80	13.200
		90	16.810