Chapter-6

Reliability and Validity of the Test

6.1.0 Introduction

6.2.0 Reliability

   6.2.1 Methods of determining Reliability

6.3.0 Reliability of the present test

   6.3.1 Test-Retest Reliability
   6.3.2 Split-Half Reliability
   6.3.3 Reliability by Cronbach’s $\alpha$ (alpha)
   6.3.4 Reliability by Kuder-Richardson Formula

6.4.0 Validity

   6.4.1 Methods of Validity

6.5.0 The Validity of this test

   6.5.1 Construct Validity

6.6.0 Conclusion
Chapter-6
Reliability and Validity of the Test

6.1.0 Introduction

The process of tool construction was discussed in the previous chapter. Selection of items various try-outs, final run of the tests were discussed.

This chapter deals with some theoretical background of reliability and validity, methods employed for estimation of reliability and validity of the present test.

6.2.0 Reliability

A test constructor has to check the accuracy and precision of the measurement procedure as well as the extent to which the test measures what it tends to measure. The term Reliability refers to the stability or consistency as well as the precision which enter into measurement procedure. The main objective of standardization is to eliminate the source of “error variance” as far as possible.

According to Anastasi, (1982):

"Reliability refers to the consistency of scores obtained by the same individuals when re-examined with the same test on different occasion, or with different sets of equivalent times, or under other variable examining condition."

One cannot expect test results to be perfectly consistent, as there are numerous other factors. If a single test is administered to the same group at two different times, some variation in score is observed due to temporary fluctuations in memory attention, effort, fatigue and similar other factors with a longer time period between tests. Additional variation

in scores may be caused by intervening learning experiences changes in health, forgetting and less comparable testing conditions.

The Reliability of a test is, therefore, found out by administering the same test to the same group again after a reasonable gap of time so that the students forming the group have not remembered some of the items of the test.

6.2.1 Methods of determining Reliability:

Various methods of estimating test reliability are in vogue. They are: The following methods are commonly used to gauge the reliability or self-correlation of the test.

The Test-Retest Method:

The simplest way to measure the reliability of a test is to apply it again to the same group after an interval of time. The results of the two trials are then correlated and the co-efficient of correlation denotes the reliability of the test.

Split-Half Method:

In this method, the test is divided into two equivalent halves and the correlation between the scores of the two halves is calculated which gives the half-test reliability. From the half-test reliability the self-correlation of the whole test is calculated by Spearman-Brown formula.

This method is regarded as the best of all the methods of determining test of Reliability. The greatest advantage of this method is that both the equivalent parts of the test are administered at one occasion eliminating all chance errors occurring in other methods.

The only drawback of this method is the fact that the test can be divided into two equivalent halves in a variety of ways each of which would give a different coefficients of reliability. But that is only possible in the case of tests with items of equal difficulties. A test consisting of graded difficulties can be divided only in one way viz., splitting up in to odd and even items.
Alternate or parallel forms Method:

Instead of giving the same test again, the alternate or parallel form of the test is used in the second trial. If the alternate form is used after a fairly long interval, the practice and confidence effects are, to a large extent, eliminated. This method gives more reliable results than the Test-Retest method.

But then, it is very difficult to prepare a parallel form of a test. So that the investigator don’t use this method or uses very rarely, this method is dropped.

Rational Equivalence Method:

This is the fourth method of estimating test reliability. This method involves to many calculations and therefore can be applied with ease only to tests having a few items. To determine the reliability by this method, Kuder-Richardson formula is used. So this method is also known as Kuder-Richardson method.

Standard Error of Measurement:

The reliability of test can be expressed in terms of standard error of measurement of score. It is determined by the formula is given below:

\[ SE_M = SD \sqrt{1 - r} \]

Here,
S.D. means Standard deviation of the test scores.
\( r \) stands for The reliability co-efficient

Using the above formula standard error of measurement for each reliability coefficient was obtained which are shown in the tables of coefficient of correlation that follow.

Standard Error of Correlation:

The reliability of test can be expressed in terms of standard error of co-relation also. It is determined by the formula given as under.

\[ SE_r = \frac{1-r^2}{\sqrt{N}} \]

Here,
\( r \) denotes Correlation coefficient;
\( N = \) No. of students or sample size
Using the above mentioned formula standard error of measurement for each reliability coefficient was obtained which are shown in following tables all tables of coefficient of correlation.

### 6.3.0 Reliability of the present test

The reliability of the present test is measured in terms of the reliability coefficient.

According to Ebel(1972):

> “The reliability co-efficient for a set of a scores from a group of examinees is the co-efficient of co-relation between that set of scores and another set of scores on a equivalent test obtained independently from the members of the same group.”

#### 6.3.1 Test-Retest Reliability

The present test was administered to 120 (40 students of each class V, VI, VII) standardwise students after a period of 15 days. Two sets of scores were worked out and then coefficient of correlation was computed by using NRTVB software. The calculated value of coefficient of correlation along with its level of significance is presented in Table 6.3.1.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Attempt</th>
<th>N</th>
<th>r</th>
<th>Mean</th>
<th>S.D.</th>
<th>SE_M</th>
<th>SE_R</th>
<th>Level of Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>Test</td>
<td>40</td>
<td>0.8</td>
<td>20.52</td>
<td>4.27</td>
<td>1.706</td>
<td>0.0412</td>
<td>**</td>
</tr>
<tr>
<td></td>
<td>Retest</td>
<td>40</td>
<td>6</td>
<td>20.53</td>
<td>4.30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td>Test</td>
<td>40</td>
<td>0.9</td>
<td>27.60</td>
<td>8.97</td>
<td>2.136</td>
<td>0.0184</td>
<td>**</td>
</tr>
<tr>
<td></td>
<td>Retest</td>
<td>40</td>
<td>4</td>
<td>27.58</td>
<td>9.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VII</td>
<td>Test</td>
<td>40</td>
<td>0.9</td>
<td>24.70</td>
<td>9.34</td>
<td>1.896</td>
<td>0.0154</td>
<td>**</td>
</tr>
<tr>
<td></td>
<td>Retest</td>
<td>40</td>
<td>5</td>
<td>24.65</td>
<td>9.59</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It is observed from the Table 5.3.1 that coefficient of correlation (r) for all the three Grades are significant at .01 level of significance. Hence the Test-Retest reliability of the present three tests is significantly very high. Also the above results show that standard error of measurement ranges from 1.70 to 2.13 and standard error of correlation coefficients ranges from 0.015 to 0.041. It means that the Test-Retest reliability is significantly very high.

---

reliability coefficients of the present three tests are very high. Hence the present three tests are reliable tests.

6.3.2 Split-Half Reliability

The test was administered over 360 students (120 students per each grade) and usual method of dividing the test into two equivalent halves to take all odd items in one half and all even items in the other half to calculate the reliability was used. To compute coefficient of correlation, the data were entered in spreadsheet standard wise. To compute Split-half reliability, the following formula, given by Spearman-Brown, was used. The formula is as follows.

\[
r = \frac{2r \frac{1}{2} l}{1 + 2r \frac{1}{2} l}
\]

Where \( r \) = Reliability for the Whole test

\( \frac{r_{\frac{1}{2} l}}{2} \) = Coefficient of co-relation for the half test.

The computation of coefficient of correlation was done by using NRTVB software. Different necessary values and value of \( r \) is presented in Table 6.2.3 Results are shown in following tables.

<table>
<thead>
<tr>
<th>Coefficient of correlation</th>
<th>( r ) Between two halves</th>
<th>( r ) The Whole Test</th>
<th>SEM</th>
<th>SER</th>
<th>Level of Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r_V )</td>
<td>0.87</td>
<td>0.93</td>
<td>1.447</td>
<td>0.0123</td>
<td>**</td>
</tr>
<tr>
<td>( r_{VI} )</td>
<td>0.85</td>
<td>0.92</td>
<td>2.415</td>
<td>0.0140</td>
<td>**</td>
</tr>
<tr>
<td>( r_{VII} )</td>
<td>0.92</td>
<td>0.96</td>
<td>1.964</td>
<td>0.0716</td>
<td>**</td>
</tr>
</tbody>
</table>

It is observed from the Table 6.2.3 that coefficient of correlation for all the three Grades, values of \( r \) ranges from 0.85 to 0.92 between two halves while it ranges for the entire test from 0.92 to 0.96. They are all significant at .01 level of significance. Also the results show that standard error of measurement ranges from 1.45 to 2.42 and standard error of correlation coefficients ranges from 0.012 to 0.071. Which are as per criteria decided by statistics experts. It means the values of Split-Half reliability of the present three tests are very high, which shows that the present three tests are reliable tests.
6.3.3 Reliability by Cronbach’s $\alpha$ (alpha)

Cronbach’s ‘$\alpha$’ is the coefficient of reliability. It is commonly used as a measure of the internal consistency or reliability of a test score for a sample of examinees. It was first named by alpha by Lee Cronbach in 1951, as he had intended to continue with further coefficients. The measure can be viewed as an extension of the Kuder-Richardson formula (KR20), which is an equivalent measure of dichotomous items. Cronbach has given a formula to find out the reliability by using standard deviations of the scores not only on odd and even items but also the total number of items is as follows.

$$Cronbach's \alpha = 2[1 - (\sigma_{odd}^2 + \sigma_{even}^2)]\sigma_{total}$$

A commonly accepted rule of thumb for describing internal consistency using Cronbach’s alpha is as follows.

<table>
<thead>
<tr>
<th>Cronbach’s alpha</th>
<th>Internal Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha \geq 0.9$</td>
<td>Excellent</td>
</tr>
<tr>
<td>$0.8 \leq \alpha \leq 0.9$</td>
<td>Good</td>
</tr>
<tr>
<td>$0.7 \leq \alpha \leq 0.8$</td>
<td>Acceptable</td>
</tr>
<tr>
<td>$0.6 \leq \alpha \leq 0.7$</td>
<td>Questionable</td>
</tr>
<tr>
<td>$0.5 \leq \alpha \leq 0.6$</td>
<td>Poor</td>
</tr>
<tr>
<td>$\alpha \leq 0.5$</td>
<td>Unacceptable</td>
</tr>
</tbody>
</table>

The present test was administered to 360 students, the data were entered grade wise in spread sheet. The computation of coefficient of correlation was done by using NRTVB software. Different necessary values and value of Cronbach $\alpha$ is presented in Table 6.3.3. $SE_M$, $SE_r$ and their level of significance are also presented in the said table. Results are shown in following tables

### Table 6.3.3
**Grade V to VII**

<table>
<thead>
<tr>
<th>Grade</th>
<th>Cronbach $\alpha$</th>
<th>$SE_M$</th>
<th>$SE_r$</th>
<th>Level of Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>0.87</td>
<td>1.972</td>
<td>0.022</td>
<td>**</td>
</tr>
<tr>
<td>VI</td>
<td>0.84</td>
<td>3.416</td>
<td>0.0269</td>
<td>**</td>
</tr>
<tr>
<td>VII</td>
<td>0.87</td>
<td>3.541</td>
<td>0.0222</td>
<td>**</td>
</tr>
</tbody>
</table>

It is clear from the Table 6.3.3 that coefficient of correlations for all the three Grades range from 0.84 to 0.87. In view to the referencing range of the Cronbach’s $\alpha$ all the three values are high. Additionally all values are significant at .01 level of significance. Moreover the above results show that standard error of
measurement ranges between 1.97 and 3.54 and standard error of co-relation coefficients ranges between 0.0220 and 0.0269. It means that the present three tests are reliable tests.

6.3.4 Reliability by Kuder-Richardson Formula

There are several Kuder- Richardson formulas, out of which, the formula 16 was used to determine the reliability of the test. It gives an estimate of internal consistency. The Kuder- Richardson formula is

\[ r_{xx} = \frac{K \left[ 1 - M_x - \frac{M_x^2}{K} \right]}{(K - 1) (S_x^2)} \]

The present Action oriented test was administered to 360 students of primary schools and required statistical values were calculated by using NRTVB. They are in Table 6.3.4.

<table>
<thead>
<tr>
<th>Grade</th>
<th>K= Number Of items</th>
<th>Mx= Mean of Test</th>
<th>Sx2= Variance on the test score</th>
<th>r</th>
<th>SE_M</th>
<th>SE_R</th>
<th>Level of Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>25</td>
<td>19.78</td>
<td>29.92</td>
<td>0.90</td>
<td>1.7298</td>
<td>0.9260</td>
<td>**</td>
</tr>
<tr>
<td>VI</td>
<td>40</td>
<td>26.37</td>
<td>72.93</td>
<td>0.90</td>
<td>2.7006</td>
<td>0.9260</td>
<td>**</td>
</tr>
<tr>
<td>VII</td>
<td>40</td>
<td>25.64</td>
<td>96.43</td>
<td>0.93</td>
<td>2.5981</td>
<td>0.9211</td>
<td>**</td>
</tr>
</tbody>
</table>

It is observed from the Table 6.2.4 that coefficient of correlation for all the three grades range from 0.90 to 0.93. All of them are significant at .01 level of significance. The results also show that standard error of measurement ranges between 1.73 and 2.70 and standard error of correlation coefficients ranges between 0.9211 and 0.9260. These results evidently show that the values of the reliability of the three tests are significantly high. It means that the present three tests are reliable tests.
6.4.0 Validity

According to John W. Best (1995),

"The validity of a test may be defined as the accuracy with which it measures that which it is intended to measure or as the degree to which it preaches infallibility in measuring what it purports to measure."³

The evaluation of a test does not end with the estimation of the stability and precision of its measurement. It only begins there. A highly reliable test may not measure what it intends to measure. Besides, it is necessary to know how of what is intended, is measured as well as to be sure that nothing else is measured. The question is fundamental with “Assessment” tests but not with the predictor tests. Such tests are more concerned with is termed as “Concept” or “Construct” validity.

6.4.1 Methods of Validity :

Thorndike and Hagen have categorized the three types of validity namely. (1) Congruent validity. (2) Concurrent validity (3) Predictive validity.

Anastasia discusses face validity and factorial validity in addition to content validity and various types of empirical validity. Ross and Stanly mention curricular validity. Gullisken has discussed instruc validity and Moiser differentiated four types of face validity. 1. Validity by assumption 2. Validity by definition 3. The appearance of validity 4. Validity by hypothesis.

- **Concurrent Validity**: It is concerned with the relation of test scores to an accepted contemporary criterion of performance on the variable which the test is intended to measure.
- **Construct Validity**: It is concerned with “What qualities a test measures” and is evaluated, “by demonstrating that certain explanatory constructs account to some degree for performance on the test.
- **Content Validity**: It is concerned with the adequacy of sampling of a specified universe of content.
- **Curricular Validity**: It is determined by examining the content of the test itself and judging the degree to which it is a true measure of the important objectives of the course, or truly a representative sampling of the essential materials of instruction.

• **Empirial Validity**: It refers to the relation between test scores and direct measure of that which the test is designed to predict.

• **Face Validity**: It refers not to what a test necessarily measures but to what it appears to measure.

• **Factorial Validity**: It is the co-relation between that test and the factors common to a group of test or other measure of behavior. Such validity is based on factor analysis.

• **Intrinsic Validity**: It involves the use of experimental techniques other than co-relation with a criterion to provide objective, quantitative evidence that the test is measuring what is ought to measure.

• **Predictive Validity**: It is concerned with the relation of test scores to measures on a criterion based on performance at some later time.

These types of validity are not all distinctly different from each other. In fact, one or two of them are practically identical with one or two others. But enough major differences appears to justify grouping them into two major categories. These concerned with primary or direct validity and those concerned with secondary or derived validity.

### 6.5.0 The Validity of this test

#### 6.5.1 Construct Validity:

In the present research study, the Construct Validity was estimated. There are different methods for examining construct validity. Although factor analysis is a well-known method of construct validity, in the present investigation, the method of Cliff’s consistency Indices ‘C’ was used due to its strongness. In the present study Cliff’s Consistency Indices ‘C’ was calculated using the computer program NRTVB for each test. To find out the validity of this test the investigator has used only one type of validity procedure among the given above.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Cliff’s Consistency ‘C’</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>0.39</td>
</tr>
<tr>
<td>VI</td>
<td>0.63</td>
</tr>
<tr>
<td>VII</td>
<td>0.43</td>
</tr>
</tbody>
</table>
It can be observed from table that the Cliff’s Consistency Indices ‘C’ for the three tests are above average as reported by Raviya (1990, P.260). As per criteria, if the value of ‘C’ is 0.32 or above it, then the test is treated as valid.

The internal consistency of all the three tests were estimated and determined by t test. In each grade difficulty value (D.V.) of each item during piloting, Final run and Re-test were listed out. From that consistency of D.V. between two phases were found by t test consistency of D.V. for three grades V, VI and VII were presented as Appendix–X.

### 6.6.0 Conclusion

The discussion of the reliability and validity of this present test discussed in this chapter. Now, The final form of test is ready for administer to measure the action oriented mathematical achievement of the students for Grade V, VI and VII. The details of Final Run of the test and Data analysis are explained in the following chapter.