CHAPTER VI

RELIABILITY

6.1 Sampling
6.2 Procedure and analysis
6.3 Comparative studies
6.4 Summary

6.1 Sampling

The second administration, described in the last chapter was done for the reliability and validity studies, besides for knowning the innate properties of the tests such as distribution of Scores and Intercorrelation between tests. As is stated earlier, the sample consisted of 170 students in all, drawn from all the classes of four schools. The number of students who took all the 8 tests was 72, on which the intercorrelation matrix was based.

It was also reported in the last chapter, that the final sample for CSA was 101, after discarding students of 2 schools. It was discovered that pupils in these schools had either faked the results or had not observed the proper time limit of 3 minutes for each part. The later was probably more likely. Thus remained an odd number of 101 students, on whom the reliability studies for the CSA was based; for all other tests, the sample was 170.

The reliability sample was similar to that of the item analysis study. Both were drawn on same criteria, viz, representativeness and average quality of the schools, medioire socio-economic status of the parents of most of the students, and more variability in the occupation of the parents.

6

6.2 Procedure and analysis

The internal consistency was studied by the well-known split-half technique, corrected by the Spearman-Brown Prophecy formula. All answersheets were scored both for odd and even numbered items. Table 24 shows the various coefficients obtained as also some other statistics, the Standard Error of Measurement.

TABLE 24

Descriptive Statistics and Split-Half Reliability* Results for Tests

Test	N	r	M	σ	SE measmt.
VR	170	.81	23.4	7.25	3.19
AR	170	.90	23.0	10.75	3.44
SR	170	.70	18.2	3.18	1.75
MR	170	.75	35.00	6.3	3.15
CSA	101	.99	41.4	6.93	.69
NA	170	.90	19.15	6.85	2.19
LU-sp	170	.92	59.8	9.00	8.01
LU-gr	170	.91	31.6	6.65	6.36

* All coefficients, except for CSA, are derived by split-half technique, corrected by S-B formula. For CSA, equivalent Form of Reliability was obtained.

It may be noted that the reliability coefficient for the CSA was obtained by alternate forms method. As this was primarily a speeded test, it was not proper to estimate the 'r' by split-half technique as in such cases, this technique tends to give highly spurious and inflated results. It is not necessary to delve into the theory and the rationale for it, which may be found in several leading books on psychometric techniques and testing.[1]

Fortunately, the CSA test has two equal parts consisting of 100 items each. The time limit for each ~~item~~ is 3 minutes. The reliability coefficient was the correlation between the scores of both the parts. No correction was necessary for halving the tests, as only the second part is scored.[2] Thus the reliability coefficient reported in the manual also refers to only one part which scored.

All the tests except Space Relations and Mechanical Reasoning, have satisfactory reliability coefficients, i.e. above .90. The low reliability estimates for Space Relations and Mechanical Reasoning can be attributed to several factors of which one of the most important is the Narrow Range of Ability. It is a known fact that the range or spread of the ability in the group tested affect the reliability estimates. The coefficient increases as the group becomes more

[1] An unusually clear discussion can be found in Gullickson, Theory of Mental Tests and Guilford, Psychometric Methods.

[2] G.K.Bennett et al, Manual, p. 18.

heterogenous and vice versa. Again, a detailed discussion about this is not appropriate here as this is given in most of the standard statistical text books. It may be pointed out here, that reliability coefficient is a correlation coefficient and therefore is affected by all factors which affect a correlation coefficient.

An excellent and very lucid classification of the effect of the range of talent on reliability is given by Wesman.[1] He has shown that, "it is not the smaller number of cases which brings about the lower coefficient. It is the narrower range talent which is responsible."[2] The same view is expressed by different psychologists, for example Thorndike.[3] Viewing from this angle, it is found that the distribution for these two tests does not cover the entire range of possible scores from minimum to maximum. Although in case of other tests, practically all scores are represented in the range, the range of scores for these tests cover only about 60% of the middle scores. The extreme scores are not represented. For example, the range of MR is from 21 to 40 while the maximum possible score is 68. Similarly, the range for SR extends from 10-33 while the maximum possible score is 60. It is clear that the sample consisted of pupils with a narrow range of ability

[1] Wesman, Reliability and Confidence, Test Service Bulletin, The Psychological Corporation, New York, No. 44.

[2] ibid.,

[3] R.L.Thorndike, Personnel Selection, p. 19.

as pupils with middle scores probably clustered together. Given different sample or samples whether wider variability of scores, it may fairly be assumed that the coefficients would improve.

In statistical books we have a formula estimating the reliability from another group of different variability than the one on which the study was originally based. The formula quoted by Guilford[1] reads:

$$r_{uu} = 1 - \frac{\sigma_k^2 (1 - r_{kk})}{\sigma_u^2}$$

where r_{uu}= reliability coefficient for the population in which it is unknown
σ_u = SD in the population for which reliability is known
r_{kk}= known reliability
σ_k= SD in the population for which reliability is unknown

The same formula is mentioned in a slightly different form by Mcnomar[2] and Thorndike.[3]

For the practical application of this formula, we may assume, for example, a more heterogenous sample, resulting in a wider variability of scores, and covering the entire range. Taking the risk, inherent in all such wide assumptions, if we further assume that while Means remain the same as in the present sample, the S.Ds. would vary, and the distributions are 12-6 ± 35.00, and 6.00 ± 18.2 for MR and SR respectively. Applying the above mentioned correction formula, quoted by Guilford, the 'r's

[1] J.P.Guilford, Psychometric Methods, p. 392.
[2] Q. Mcnomar, Psychological Statistics, p. 159.
[3] R.L.Thorndike, Personnel Selection, p. 99.

obtained would be .94 and .92 respectively which appear to be quite satisfactory.

Another reason for the low reliabilities of these two tests may be the unfamiliarity of the school children with situations where the abilities demanded in these tests are usually manifested. There are very few such learning experiences and opportunities available to them in the ordinary environment. With another sample with different type of subjects, therefore, a significant improvement in reliability coefficients may be expected.

Low reliabilities in themsleves do not make a battery unfit for use. Considering the limiting factors affecting the size of the reliability coefficients, writers generally agree that low reliability coefficients may be tolerated in several cases, especially in "early stages of experimentation... and can then be built up into more reliable instruments before publication."[1]

According to Guilford, "any reliability better than chance is justified for use and for research purposes; lower reliabilities can be tolerated than may be needed for diagnosis and prediction."[2] He further says, "for some purposes, even a test of low reliability adds enough to prediction to justify its use particularly when used in a battery along with other tests."[3]

[1] Wesman, *op. cit.*

[2] J.P.Guilford, *Psychometric Methods*, p. 388.

[3] *ibid.*, p. 389.

According to Flanagan, "the reliability coefficient for a specific test in a multi-factor test battery in which the scores are combined to make predictions is not crucial (though) it is desirable that all of the tests have atleast a moderate degree of reliability."[1] It is interesting to note in this connection that the median 'r' of the 14 tests in the FACT battery is 0.76 only. It may also be appropriate to quote Nunnaly, according to whom, "if a predictor test has a high correction with its criterion, reliability is no problem... The test constructor is concerned with measurement error when a test fails to predict its criterion."[2]

By these rather elaborate quotations the investigator wished to emphasize that though low reliability coefficients are a matter of concern, they are not always unacceptable. Though a high 'r' would almost always mean that a trust could be placed on the test, a low 'r' does not always mean the opposite. At best, this indicates the need of further experimentations and studies, as Wesman has remarked 'before publication' of tests.

Standard Error of Measurement.-- To a considerable extent SE measmt. is a better indicator of the trust that can be placed in a test. A large SE measmt. naturally indicates a variability of the error and so lower these figures, lower

[1] J.C.Flanagan, "The Flanagan Aptitude Classification Test" in Use of Multifactor Test in Guidance, p. 72.

[2] Nunnaly, Tests and Measurement p. 111.

they would appear to be, which means in other words, higher reliability. Table 24 also shows the SE measmt. for various tests. It is obvious that the figures are satisfactory, which in turn testify the reliability of the tests. The low errors indicate the amount of confidence which could be placed in the results of the tests.

Inter-item consistency.-- This was obtained by using Kuder-Ruhardson formula 21, from the table Dietrich.[1] Table 25 shows the KR 21 reliability coefficient thus obtained.

TABLE 25

KR-21 Reliability of the Tests under Adaptation

Test	N*	M	S.D.	S.D.^{2}	KR-21 rel.
VR	70	23.35	7.25	52.56	.66
AR	50	23.00	10.75	115.58	.90
SR	60	18.21	3.18	10.11	-
MR	68	35.00	6.30	39.69	.66
NA	40	19.95	6.85	46.92	.75
LU-sp	60	31.60	6.65	44.89	.61
LU-gr	100	59.80	9.00	81.00	.77

SOURCE: Paul Dietrich, Short cut Statistics, quoted in Adams, Measurement and Evaluation in Education, Psychology and Guidance, P. 89

Split-half 'r' are higher than KR-21 (or other) estimates (Psychometric Methods, p. 377)

* no. of items.

[1] quoted in Adams, Measurement and Evaluation in Education, Psychology and Guidance, p. 89.

The K-R formula gives lower bound estimates of the reliability. The reliability coefficients found by split-half technique are usually higher than those obtained by the use of KR-21 formula. There is also a difference in the assumptions in two cases. In KR-21, it is assumed that the items are of equal difficulty. It is a reliability estimate obtained essentially from a single administration, while the split-half technique assumes two halves as just equivalent. While split-half technique is in essence, the parallel form reliability, obtained through a single administration. As assumption in both are hardly met perfectly, the one overestimates, while the other underestimates the actual reliability.

6.3 Comparative studies

1. Comparision with original tests.-- Table 26 shows the two 'r's. The American 'r's are those for grade 10, boys, as the average age group of our class IX boys, is similar to it.

It is evident from the Table 26 that except for the MR and SR, the reliability coefficient as obtained in the present investigations are comparable to the original study. It may be noted that even for the American study the 'r' for the Mechanical Reasoning is the lowest in the series.

The present study, as reported earlier, was based on Form L, wherein the formulas of the three tests-Space Relations,

TABLE 26

Comparision of the Present Reliability Results with the Results on the Original Tests*

Test	Original** r	Obtained r
Verbal Reasoning	.90	.81
Abstract Reasoning	.90	.90
Space Relations	.93	.70
Mechanical Reasoning	.85	.75
Clerical Speed and Accuracy	.87	.99
Numerical Ability	.90	.90
Language Usage-sp	.92	.92
Language Usage-gr	.88	.91

* These coefficients for both the studies-original and present-are obtained by the split-half technique except for the CSA, where alternate from reliability was computed.

** Average reliability coefficients for Form A, boys, as given in G.K.Bennett et al, Manual, p. 66.

Language Usage-spelling and Verbal Reasoning-were changed. As the reliabilities of Form L were not available upto the time of writing this report, the original reliabilities reported in the Table are for the Form A, as reported in the Manual, including the three tests which have been changed. It may be expected, however, that the reliabilities may not be much different.

It may be useful here to compare the general range of reliabilities of the DAT tests with those of other similar Multi-factor test batteries.

Table 27 shows the comparative reliabilities of the various American multi-factor test batteries, and the method used for obtaining the coefficient.

TABLE 27*

Reliabilities of some important Multi-factor Test Batteries

Battery	Range of coefficients	Method
DAT		
Boys	.85-.93	Split-half
Girls	.71-.92	Split-half
Boys: clerical speed and accuracy	.77-.93	Alternate forms
Girls: clerical speed and accuracy	.84-.91	Alternate forms
FACT		
Grade 9	.52-.86	Sep~~e~~arately timed halves
Grade 12	.65-.86	Split-half
Grade 9 and 12	.83-.93	Split-half; combined "occupational scores"
GATB	.70-.95	Test-retest after an interval; andequivalent forms close together
Guilford-Zimmerman	.89-96	Split-half
	.88-.92	Alternate forms
	.74-.94	Kuder-Richardson formula
Holzinger-Crowder	.76-.95	Alternate forms
	.88-.95	Split-half
MAT	.75-.94	Kuder-Richardson formula
PMA	.87-.96	Split-half
	.72-.90	Sep~~e~~arately timed halves

* quoted from Freeman, Psychological Testing, p. 427.

It is evident that the results obtained in the present investigation (where the range is .70-.94 by split-half technique, and by KR 21 formula, besides .99 for CSA by parallel f form methods) are satisfactory and comparable to the various well known studies in this field.

2. Comparison with some Indian studies.-- As reported earlier, there have not been many Indian studies in this field of Differential testing. In one institute, the reliability coefficient for the AR test was found as 0.90, which is very similar to the one in the present investigation. In another study by Verma, the range of coefficient was .60 to .93. The Table 28 shows the reliability coefficients in this battery.

TABLE 28*

Reliability Results on Verma's Differential Prediction Battery

Test	'r'
Numerical	.73
Verbal	.70
Inductive	.93
Deductive	.60
Spatial	.79
Perceptual Speed	.75
Finger Dexterity	.81
Role Memory	.87

* quoted from Verma, Manual.

The Table 28 appears to support the contention that for practical purposes, a reasonably low reliability in a limited sample is acceptable. The results obtained by the present investigator were invariably higher than those by Verma, except in Spatial tests. Any comparison, however, between the two coefficients of spatial ability is not possible as Verma's report did not specify the distribution characteristics of the sample.

6.4 Summary

The reliability coefficients were obtained by the split-half technique (odd-even items), for all tests except for the Clerical Speed and Accuracy test, which is a highly speeded test. For the latter, parallel form reliability was obtained by comparing the two parts of the same test; Spearman-Brown correction was not applied as in operational use only one part i.e. part II is scored. Inter-item consistency by the use of KR-21 formula was also obtained. S.E. of measurement have also been reported for various tests.

The reliability coefficients obtained were above .90 for all tests, except for the Mechanical Reasoning Test and Space Relations Test, where they are .75 and .70 respectively by split-half technique. The low reliability estimates for these two tests were explained. These might have been caused by several factors such as the narrow range of ability in the sample, and the possible unfamilaarity of the students with tasks involving manifestation of such abilities. Views of several scholars have been reported to show that a reasonably low reliability is not entirely unexpected and does not come in the way of the use of tests for practical purposes. The writer has suggested, however, that further studies may be made on different samples in different occupational areas.

In section 3 some comparative figures of reliability co-efficients in India and America are presented through various tables. The investigator felt that the results obtained by him were comparable and atleast were equally satisfactory.