CHAPTER X

VALIDITY

As mentioned out by many psychologist from time to time a test can serve no useful function unless it is valid. McCall[1] says, "that validity is the most important characteristic of a good test".

The validity of a test implies the efficiency with which it measures what it attempts to measure. A test is valid[2] when the results obtained from it about a particular trait of an individual are the same as are obtained through some other reliable method. The validity of a test is more important than its reliability. It is not so difficult to make a test reliable as it is to make it valid. A valid test by itself is reliable to some extent. The problem of making a rest valid seems to be very frustrating as

---

1 McCall : How to Measure in Education. New York, The Macmillan Company. p.119.

2 Garret, H.E. : Statistics in Psychology and Education New York, Longmans Green and Co. p.354.

the independent criterion against which a test can be validated is not available in most of the cases. No test has 'high' or 'low' validity in the abstract. Its validity can be established only in terms of one or more specific criteria. A statement of test validity without reference to the procedures employed in computing such validity is even less meaningful than unqualified report of test reliability. The term validity commonly stands for emperical validity. It has been pointed out by psychometricians that validity, has four main factors. These factors are, face validity, content validity, factorial validity and the emperical validity.

Face Validity

Face validity refers not to what the test nearly measures, but what it appears to measure. The use of this concept of test validity seems the least justificable.

Fundamentally, the question of face validity is not one of validity in the usual sense, but rather one of rapport and public relations.

The psychological tests in which this factor of validity is of some importance, are generally, used for selection of industrial and military personnel. This factor cannot also be ignored when selection is made for civil services personnel. Face validity should never be regarded as a substitute for objectivity determined validity.

Content Validity

Content validity is also known as "Logical Validity" and validity by definition is especially pertinent to the evaluation of achievement tests.

Factorial Validity

The factorial validity of a test in the correlation between that test and the factor common to a group of tests or other measures of behaviour.

Empirical Validity

This type of validity refers to the relation between test scores and a criterion the latter being an independent and direct measure of that which the test is designed to predict.

Empirical validation always involves the comparison of two sets of data on the same person, viz. test scores and criterion measures.

A criterion may be objective measure of performance, or a qualitative measure such as a judgement of the character or excellence of the work done. Intelligence tests were first validated against school grades, ratings for aptitudes by different teachers.

Validity Coefficients

Psychologists and experienced test users, express the validity of test in terms of a validity coefficient showing the correlation between the test and the criterion.

Studies of the Validity of Some Allied Tests of Intelligence

Wechsler places major emphasis upon the argument that Wechsler Bellevue Scale has worked well in practice and that it agrees more closely with clinical judgement of intelligence than do other individual tests.

At a slightly more objective level bisérial correlations are reported between psychiatrists recommendations regarding commitment or non-commitment to institutions for mental defectiveness and I.Qs. on the Wechsler Bellevue and the Standford Binet. These correlations are based on cases examined at Bellevue Hospital over a period of several years.

The correlation cited in the manual are .33 for the Standford Binet and .79 for the Wechsler Bellevue.

Examination of original study in which these correlations are found, however, indicates a possibility of criterion contamination in at least some of the cases since the psychistrist knew the patient score on either Standford Binet or the Wechsler Bellevue or both.

Stanford Binet Scale .- The principle evidence for the validity of the Standord Binet is derived from the item analysis. First the preliminary selection of items on the basis of an agreement M.A. on 1916. Standord Binet insured that the new test measured, essentially the same function as the old.

The items which were significantly more difficult for the feeble minded subjects had mean general factor loading. 60 while those which were easier for the feeble minded subjects only .49. A similar study in which 177 children with I.Q. above 120 were compared with the normal children of the same mental age. This group was selected on the basis of I.Q. differed practically in those in which items are mostly saturated with G factor common to whole test.

Validity of the Present Test

The validity of the present test has been studied in a number of ways.

(1) By finding the correlation between test Score and some independent criterion measure.

(2) Ratings of teachers with the I.Q. of the pupils.

(3) Progressive Matrices.

Cross Validation

In recent years Psychologists have laid great stress on Cross Validation in test construction. Cross validation refers to the fact that the validity of the test should be determined on a different sample of the persons from that on which items were selected. Any validity coefficient computed on the same sample which was used for items selection purpose will capitalize on chance errors within that particular

sample and will consequently be high. Cross validation of the test has been carried out in the following manner. In this study 65 students with the highest score group and 65 with lower score group have been chosen to form the two criterion groups from age group 12 plus to 16 plus.

The test was administered to both these criterion groups. The mean scores and the Standard Deviation values for both the groups were calculated. The results are tabulated in the table given below.

TABLE 20

A. Composition of the Sample

Age	Number of Students getting High Score	Low Score
12 years	10	12
13 years	12	17
14 years	18	12
15 years	9	12
16 years	16	12
Total	65	65

B. Mean and S.D. of two Criterion Groups (Top 65 and bottom 65 scores) S.E.,D and C.R. between the Means

Sr. No.	Criterion Group No.	Mean Scores	S.D.	Difference in Means	Standard Error	C.R.
1) The Criterion group with high score	65	M_1= 107.3	14.9.	11.2	2.61	4.29
2) The Criterion Group with low score	65	M_2= 96.1	10.8			

It is seen from the above table that the difference between mean performance of the two groups on the test is 6 times that of the S.E. This shows that the difference is highly significant.

Correlation of I.Qs. With Teacher's Estimates of Intelligence

The validity of the test is further tested by correlating I.Q.s of 100 pupils with teacher's estimates of their intelligence. To obtain the opinions of teachers in a scientific way a table was prepared based on Five Point Scale, viz.

A B C D E

where A - Very inferior

B - Inferior

C - Average

D - Superior

E - Very superior

The names of the pupils were written age wise and grade wise on one side of the column. The teacher who had a close touch with a grade was entrusted the list of pupils of that grade and age group to mark the columns of the table according to his opinion about the intelligence of those pupils.

Then the I.Qs of those pupils were also divided into five groups as shown in the table given below.

TABLE 21

Correlation of I.Qs. with Teacher's Estimates of Intelligence

A. Composition of the Sample

Age	Grade	Total
12 plus	VI	20
13 plus	VII	20
14 plus	VIII	20
15 plus	IX	20
16 plus	X	20
Total		100

B. Teacher's Estimates

I.Qs.	A	B	C	D	E	Total
129-140			1	4	2	7
118-128			7	8	1	16
107-117	1	12	41	11	1	66
96-106	2	3	3	1		9
85-95		1	1			2
Total	3	16	53	24	4	100

$r = .52$

The coefficient of contingency between two measures is found to be .52. In view of the validity of teacher's estimates, .52 is fairly good coefficient

of correlation showing that test agree with teacher's estimates.

Standard Progressive Matrices

The validity of the present test was further established by correlating the scores of pupils on the present test with the corresponding scores on "Ravens Progressive Matrices".

Progressive Matrices are constructed on the prior assumption that Spearman's principles of noegencies were correct. As such it should provide a test for comparing pupils with respect to their capacities of observation and clear thinking.

The scale consists of 60 problems divided into five sets of 12. In each test the first problem is as nearly as possible self evident.

A sample of 100 students was selected. All students were 15 years of age. These students were administered the present as well as Progressive Matrices test. Product moment correlation was calculated between the two sets of scores.

Table 22 gives the scattered diagram for the two sets of scores. It is seen that product moment coefficient correlation between the two sets of scores is .77.

TABLE 22

Test Scores

X / Y	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100	101-110	111-120	121-130	Total
50-59										1		1
40-49							1	1	2	1	1	6
30-39						7	5	3	1			16
20-29			2	3	8	49	4					66
10-19		2	1	3	1	1	1					9
0-9	1	1										
F	1	3	3	6	9	57	11	4	3	2	1	100

Progressive Matrices of Items

$$r = 0.77$$

All the above studies are adequate to establish the validity of the present test.

$$r = \frac{\Sigma \frac{X'Y'}{N} - \frac{\Sigma fx'}{N} \cdot \frac{\Sigma fy'}{N}}{\sqrt{\left\{\frac{\Sigma fx'^2}{N} - \left(\frac{\Sigma fx'}{N}\right)^2\right\}\left\{\frac{\Sigma fy'^2}{N} - \left(\frac{\Sigma fy'}{N}\right)^2\right\}}}$$

$$\frac{.97 - .06 \times .18}{\sqrt{\left\{.66 - (.18)^2\right\}\left\{2.44 - (0.6)^2\right\}}}$$

$$= \frac{0.9592}{0.63 \times 2.438}$$

$$= \frac{.96}{1.54}$$

$$r = 0.774$$

References

1. Anastasi Anne : Psychological Testing, New York, Macmillan Co. 1955.

2. Bhatia, C.M. : Performance Tests of Intelligence under Indian Conditions, Ch.7, London, Oxford University Press.

3. Cronbach, L.J. : Essentials of Psychological Testing, Ch. 7, New York, Harper & Brothers, 1947.

4. Desai K.G. : The Construction and Standardization of a Battery of Group Tests of Intelligence in Gujarati. Ahmedabad, Bharat Prakashan, 1954.

5. Garret, H.E. : Statistics in Psychology and Education, Ch. 13, New York, Harper & Brothers, 1949.

6. Green, W.A., Jorgensen, A.N. & Gerberich, J.R. : Measurement and Evaluation in Secondary Schools, Ch. 4, New York, Longmans Green & Co.

7. Guilford, J.P. : Psychometric Methods, New York, McGraw-Hill Book Co. Inc.

8. Jordon, A.N. : Measurement in Education, Ch.2, New York, McGraw-Hill.

9. Lindquist, E.F.: Educational Measurement, Ch.16, Washington, D.C. American Council on Education.

10. Mursel, J.L. : Psychological Testing, Ch. 2, New York, Longmans Green.

11. Ross, C.C. : Measurement in Today's Schools, Ch. 4, New York, Prentice Hall Inc.

CHAPTER XI

EVALUATION OF THE PRESENT TEST

Characteristics of a Standardized Test

Psychologically tests are meant to measure the intelligence of an individual. They are popularly called intelligence tests. Standardized test has the following characteristics :

(1) Objectivity
(2) Scorability
(3) Discriminability
(4) Administrability
(5) Economy
(6) Reliability
(7) Validity
(8) Interpretation and Comparability.

1. Objectivity

It is necessary for an intelligence test to be objective. The award for the test items must be, uniform. The examiner's discretion must not have any importance. While considering all the above facts in

mind, the present test has been made as objective as possible.

2. Scorability

The test should be so constructed as can be scored easily. For that purpose a Scoring Key should be prepared for making scoring easy and simple. As such Key was prepared with the following consideration.

The Preparation of the Key. - A key is prepared to evaluate the responses of the testees. Great care was taken to make the Key as accurate as possible. The Scoring Key is given in the Appendix.

Scoring Procedure. - In the scoring procedure exactly the same allotment of one mark for each correct response, as was followed in the second try out and the Pilot Test. No partial credits were allowed for partially correct answers. Ross[1] also adds that, "As a rule, the best procedure in scoring is to give one point scale of credit for each correct response. It is unnecessary to weight the times according to estimated difficulty or importance.

---

[1] Ross, C.C. : Measurement in Today's Schools. New York, Prentice Hall, 1956. p. 156.

3. Discriminability

A good measuring instrument must have this characteristic. The items of the test should be arranged according to their difficulty values. So that the test can discriminate between a poor and a clever student. The items in the present test were arranged so as to give a good discriminating power. The more difficult items came first and the easier came last.

Adequacy. - A test is adequate if its sampling is random, if it is within the children's capacity and if it covers many angles of the students capacity. This seems clear from answers of the children and by the scores obtained by them.

This battery consists of 7 sub-tests. Thus it covers many phases of intelligence and therefore it is adequate.

4. Administrability

When the psychological tests are given to a large group, it becomes a laborious task. Hence the process of administering the test should be as simple as possible. This method of administration has been standardized which means that definite instructions have been worked out usually with appropriate time limits and the like. During the preliminary try out of the test every thing that the pupils asked was noted and the instructions were modified in the light of the information collected from the querries made by the pupils. Thus, these instructions were

modified and made simpler and easier for the grasp of the pupils.

The Printing and Arrangement. - The method of administration cannot be said to be well standardized without taking into consideration the printing and the arrangement of test material. According to Menzel clear printing reduces the possibilities of confusion and misunderstanding on the part of examinee. Great care was taken in getting the test printed and also in its arrangement.

Time Limit. - In the present test time limit was fixed on the criterion that at least 90 per cent of the pupils are able to answer all the items within their power. It was fixed to one hour and twenty minutes after taking due consideration the average of the time taken by a pupil during the different try-outs of the test.

So in short, the time limit was not fixed arbitrarily but was fixed after keeping the capacity of the pupils in view.

5. Economy

A test should be so constructed that it helps to economise time, money and the labour of the pupils and the administrator. No special arrangement for the testing was done which would mean unnecessary burden for the school and also expenditure. The testing was done in as natural an environment set up as possible. All the testing in the schools was done

during the first three periods of the day to avoid the effect of fatigue in the latter half of the day.

6. Reliability

The reliability of the present test has been calculated by the following two methods:

(1) Reliability by Split Half Method

(2) Test-Retest Method

Thus the coefficient of reliability by the Split Half Method is 0.94 while that by Test-Retest Method is 0.90, which shows that variation between the two is very small.

7. Validity

The validity of the present test has been studied in a number of ways by finding the correlations between test score, and some independent criterion measure.

Cross Validity. - As seen from the Chapter on 'Validity' under cross validation, the differences between the mean Score of the two groups is highly significant.

Secondly, correlation between the test results and the teacher's estimates of intelligence is .52 which is quite significant showing thereby that the test agrees with the teacher's estimates.

Thirdly, validity of the present test was further established by correlating scores on "Ravens Progressive Matrices". The product moment coefficient correlation between the two sets of scores is .77. All the above studies are adequate to establish the validity of the present test.

8. Interpretation and Comparability

Tests are to be compared and interpreted. Norm, mean and standard deviation are the means of comparing and interpreting. Hence they are necessary characteristics of a good standardized test. In the present study 5372 pupils of age group 12 plus to 16 plus studying in the different schools of Kashmir were tested. The test booklets were assessed and the total scores of each sub-test were filled in on the top of the booklet. The scores were analysed and the following norms were computed.

Mean, standard deviation and also I.Qs. were found.

TABLE 23

Age wise Distribution of Mean, Standard Deviation and Norm

	Years					
	12	13	14	15	16	17
Mean	42.9	48.1	58.0	64.2	69.1	70.2
Std. Deviation	16.9	17.5	18.5	18.1	19.0	18.5
Norms	38	48	58	64	69	70

This test can well be used by the teachers in the higher secondary schools and also by the Vocational Guidance Bureaus of the Educational Departments.

References

1. Lindquist : "Educational Measurement" Washington D.C. American Council on Education, 1950

2. MC Call, W.A. : Measurement, New York, The Macmillan Company.

3. Ross, C.C. : Measurement in Todays School New York, Prentice Hall, 1954

4. Thorndike R.L. : Measurement and Evaluation in Psychology and Education New York, John Wiley and Sons Inc.