CHAPTER VII

VALIDITY AND CRITERION

7.1 THE CONCEPTS

The simplest meaning with which the term validity is generally used in psychological literature is the quality of a test or some such measuring device by virtue of which it measures what it is supposed to measure. To borrow an example from physical sciences, if a scale measures weight and not height or any other attribute, it is supposed to be a valid measure of weight. It is so simple so far as the physical sciences are concerned. In social sciences, however, neither the attribute being measured is so clearly defined, nor its measuring device is sensitive and perfect enough to measure only that particular attribute, disregarding the others. Moreover, the science of psychology is

still young and its concepts have not reached a mature and stable stage acceptable to all. The attributes to be measured are, therefore, always defined operationally depending upon the purpose of the investigators. There is no wonder that under such circumstances much disagreement prevails as regards the nature of attributes and the ways to measure them. The terms validity and criterion have gained special significance in the field of psychological measurement due to this fact. Though theoretical, some discussion of the meaning and concept of validity to remove such prevalent confusion would not be out of place.

National Association of Directors of Educational Research defined the validity of a test as "the correspondence between the ability measured by the test and ability otherwise objectively defined and measured."[1] The second part of the

1 John A. Long, et al., The Validation of Test Items, Bulletin of The Department of Educational Research (University of Toronto, 1935), No.3, p. 7.

definition which consists in the objective definition and measurement is referred to as the criterion. The test should show correspondence with such a criterion.

Monroe's[2] definition is, "Under the head of validity we inquire into the degree of constancy of the functional relation existing between the scores yielded by the test and the abilities specified as being measured in the statement of its function." And according to Barthelmess[3] "The term validity as applied to an intelligence test battery may be defined as the amount of agreement between the test's differentiation among individuals and the actual differentiation in intelligence among these individuals. The same definition applies to a subtest or to a single element."

---

2 W.S.Monroe, An Introduction to the Theory of Educational Measurement (Boston: Houghton Mifflin Co.,1923), As cited in Ibid. p.7.

3 H.M.Barthelmess, The Validity of Intelligence Test Elements, Contribution to Education No. 505 (New York: Bureau of Publications, Teachers College, Columbia University, 1931), As cited in Ibid. p.7.

That the same definitions are still in use, can be seen from one given by Freeman[4] in his latest revision of his book on psychological testing. According to him, "An index of validity shows the degree to which a test measures what it purports to measure, when compared with accepted criteria."

All of them require that there should be an objectively measured criterion against which the functioning of a new scale should be compared. But the perennial problem in psychological measurement which is already mentioned is to get the objectively measured and acceptable criterion in terms of the original function to be measured. More often than not the validity and even the reliability of such criteria has been questioned.

To find a way out of this complex confusion, different types of validity have been defined.

---

4 F.S.Freeman, Theory and Practice of Psychological Testing (New York: Holt, Rinehart and Winston, 3rd Edn., 1962), p.88.

Face Validity: According to Mosier[5], "the term 'face validity' implies that a test which is to be used in a practical situation, should.....appear practical, pertinent and related to the purpose of the test.... it should not only be valid, but it should also appear valid. This.....is not validity in any usual sense..... (but is) an additional attribute of the test which is highly desirable in certain situation." In the earlier days of test development, this was used as a first step. An investigation[6] on judging the face validity of tests showed that face validity did exist, but there were wide individual differences and subjectivity involved in these judgments. Even though more sophisticated procedures for validating tests have been evolved, the first judgments are often based on the face value of the test characteristics.

---

5 C. I. Mosier, "A Critical Examination of the Concept of Face Validity." Educ.psychol.Meas., VII: 191-206, 1947.

6 Sidney Adams, "Does Face Validity Exist?" Educ. psychol. Meas., X, 2: 320-28, 1950.

Content Validity: "Content validity is evaluated by showing how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn"[7]. Cureton[8] has found it difficult to distinguish this from face validity and uses it synonymously with curricular validity in case of educational achievement tests. Assumption underlying the use of content validity according to Lennon[9] are:

(i) The area of concern to the tester can be conceived as a meaningful, definable universe of responses.

(ii) A sample can be drawn from this universe in some purposive, meaningful fashion.

---

7 Technical Recommendations for Psychological Tests and Diagnostic Techniques, American Psychological Association, Supplement to Psychol. Bull., LI, No.2, 1954, p.12.

8 E.E.Cureton, "Validity", Chapter 16, in E.F.Lindquist, (Ed.),Educational Measurement (Washington D.C.: American Council on Education, 1951).

9 R.T.Lennon, "Assumptions Underlying the Use of Content Validity." Educ.psychol. Meas., XVI, 3: 294-304, 1956.

(iii) The sample and the sampling process can be defined with sufficient precision to enable the test user to judge how adequately performance on the sample typifies performance on the universe.

Even though the meaning and basis of content validity are clear and defendable, in most of the cases there is no way to represent its extent in the exact mathematical language. Moreover, it is open to the same criticism as face validity is, that it depends upon the subjective judgments of the test maker or the experts whom he might consult. The test based on this requires to be verified whether it stands the empirical tests subsequently carried out, and to be modified if needed. But according to Ebel,[10] "all types of validity are based ultimately on the content validity of some measurement procedures."

Concurrent Validity: Concurrent validity is

10 R. L. Ebel, "Obtaining and Reporting Evidence on Content Validity." Educ.psychol. Meas., XVI, 3: 269-282, 1956.

an index of correspondence between the test scores and other measure of the same attribute such as rating, school grade, or clinical diagnosis. The two are obtained at more or less same time or within a short period. To explain it in the words of Ebel in the light of his principles, the test is validated on the basis of certain assumption regarding the content validity of the alternative procedure which is presumably more valid but less convenient.

Predictive Validity: According to Freeman[11] "the predictive validity of a test is the extent to which it is efficient in forcasting and differentiating behaviour or performance in a specified area under actual working and living conditions". It is evaluated by finding the degree of correspondence between the test performance and the actual behaviour at a certain future stage, or in Ebel's terms, some future measure of the forcasted behaviour, with the necessary assumptions regarding the content validity

---

11 F.S.Freeman, Op.Cit. p.89.

of such a measure.

Factorial Validity: By this method "instead of validating the total, undifferentiated instrument against external criterion, an effort is made to identify the component psychological elements and to establish their relative independence, and finally to correlate these elements separately against external criteria."[12] This is based on the assumption that a score on a test is made up of different components, some of which are due to elements common to the test and the criterion, some of which are due to elements specific to the test above, and some relatively small due to some undefined chance factors. Factorial validity refers to the components common with the criterion measures, and is arrived at by laborious statistical analysis.

Construct Validity: "Construct validity depends upon the degree to which the test items individually and collectively sample the range or class of activities or traits, as defined by the

12 Ibid.

mental process or the personality trait being tested."[13] The evaluation of construct validity requires a logical as well as an empirical attack. Theoretical evidence supporting the hypothetical construct is gathered, or rather the particular construct is hypothesized only on the basis of such theoretical evidence. It is a logical deduction. Empirical data leads either to its acceptance or revision depending upon the kind of evidence. As such, construct validation is not a unitary method but a combination of all the validation procedures. Over and above, its purpose is more subtle; it aims at theory construction. It is not generally resorted to by test makers, unless, the nature of the attributes being measured as well as the validating criteria are too ill-defined or the very purpose is construction of theory rather than a test.

Over and above these types certain authors have

13 Ibid.

advocated use of the terms intrinsic validity[14] , and operational validity.[15] Gulliksen's discussion of intrinsic validity refers to content validity of prediction scales. He maintains that too much depend-ence on criterion is unwarranted. The judgments of experts regarding the content validity can be equally faithful and one should primarily insist upon them. Further evidence regarding empirical validity may be gathered by factorial studies of the test as well as that of the criterion. In the case of intrinsic correlational validity, the basic factors giving rise to correlations should be studied as to whether they involve commonness of sampled behaviour or performance. Such a validity will be stable over a long period, and unaffected by coaching. Such an approach is quite akin to the fundamental approach of construct validation.

Many abilities or traits are generally defin-ed in operational terms, and when a test measures

---

14 H.Gulliksen, "Intrinsic Validity". American Psychologist, V, 10: 511-517, 1950.

15 F.S.Freeman, Op.Cit, pp. 88-90.

this, the validity of such a measure is strictly subject to the operational definition adopted. This is called the operational validity, as against the validity of the ability or trait in its true parameter.

The face validity, content validity and partly construct validity and even factorial validity depend upon the logical analysis of the definition of an ability, attribute or trait. It is also customary to speak of logical validity or definitional validity to denote this type broadly. It depends upon the test maker's own judgment as well as those of other experts in the field. No external criterion is needed for this purpose and in most of the cases, there is no exact quantitative measure or index of this type of validity, except in factorial studies.

The concurrent validity, predictive validity and partly construct validity and factorial validity are external criterion oriented types. The problems pertaining to the selection of reliable and valid

criteria are many and discussed at length by Brogden and Taylor.[16] These are classified by them as:

(1) Criterion deficiency - omission of pertinent elements from the criterion.

(2) Criterion contamination - introducing extraneous elements into the criterion.

(3) Criterion scale unit bias - inequality of scale units in the criterion.

(4) Criterion distortion - improper weighing in combining criterion elements.

Validation of tests against external criteria becomes a difficult task due to such biases and inadequacies in the criteria themselves. The test maker has to decide, under these circumstances, what procedures of validation would suit his purpose best and what criteria - internal or external or both - should he adopt.

---

16 H.E.Brogden and E.K.Taylor, "The Theory and Classification of Criterion Bias". Educ.psychol. Meas., X, 2: 159-186, 1950.

This much theoretical discussion has been essential, because, the subsequent treatment of the validity and criteria in the present test should be viewed in this context only.

7.2 CONCEPTS RELEVANT TO THE PRESENT WORK

The purpose of the present work, as has already been stated was to construct a valid and reliable measure of some of the basic dimensions of behaviour. The area of measurement chosen was defined in the first chapter. The factors on which the work began were three, but one had to be given up on the basis of obtained results, as mentioned in the Chapter VI. The measurement of the two factors was to be provided for by the instrument thus constructed, for the purpose of primarily understanding the dynamics of individual's interaction with his environment. Such an understanding, combined with that of other significant aspects might help to understand one for the counselling purposes and to make predictions about his behaviour in various walks of life. Prediction of behaviour is,

however, a complex process. It is governed by the law of multiple causation which implies that any particular behaviour is a result of multiple causes. This instrument can provide only one piece of information required for prediction, and hence, it cannot have independent predictive value apart from that of a battery of tests, of which it might be a part.

On the basis of this point of view, it can be said that the problem of predictive validity did not concern the present worker. It was left to a future stage when instruments for measurement of other aspects of behaviour would be available and studies in prediction of behaviour would use these instruments to find out their effectiveness. It would be essentially a work of theory construction in behaviour prediction and such instruments would have their place in providing the empirical data for that purpose.

In the present work, therefore, the main concern is with the content validity and the

concurrent validity of the Inventory. Its factorial structure could be worked out, but it is a laborious task and in the absence of electronic computational devices it is almost impossible. Due to this reason only such studies have not yet come into vogue in this part of the country in particular and in India in general.

7.3 CRITERIA

The criteria chosen for validation were based on concurrent validity principle. It was thought more desirable to find the discriminating value of each item in terms of the independent criteria for the two scales, therefore, two sets of criteria were selected.

Criterion for the Scale IE (Introversion-Extraversion): The criterion groups were drawn from the college population and the teaching and clerical staff personnel. The students in the classes consisting of not more than forty, were asked to name five students who were extraverts and five who

were introverts. Definition and characteristics of extraverts and introverts as well as the instructions for their guidance were given in written form (Appendix R). Students of sixty classes cooperated in this preliminary selection procedure. It was observed as expected, that the nominations were far from unanimous. In extreme cases the same student was nominated as extravert as well as an introvert. Such students were dropped from further consideration. From the rest, those who were nominated by at least fifty per cent of the students were selected. However, maximum number selected from any single class was five each for extravert and introvert categories.

(1) Total number of students who were nominat-ed as:

Extraverts - 536

Introverts - 494

(2) Total number of students who were nominated, by at least 50 per cent of the group, as:

Extraverts - 213

Introverts - 237

(3) Total number of students who were nominated as both: . . . . . - 44.

(4) Total number of students provisionally selected:

Extraverts - 180

Introverts - 192

The next step consisted in getting these provisionally selected individuals rated by two of their close associates on a five point scale (A B C D E) on the trait continuum, introversion-extraversion. The instructions regarding the procedure of rating were given to the raters in writing and also the definition of the scale (Appendix S).

Those cases, who were provisionally selected as extraverts, and who were rated as either E by both the raters, or D by both the raters, or E by one and D by another rater, were finally selected to be included in the criterion group of extraverts. Those

cases, who were provisionally selected as introverts, and who were rated as either A by both the raters, or B by both the raters, or A by one and B by another, were finally selected to be included in the criterion group of introverts.

(5) Number of students finally selected for the criterion groups:

Extraverts - 103

Introverts - 96

The same rating scale with minor modification was used with the members of the teaching and the clerical staff of the educational institutions. They comprised of persons from high schools, colleges and university. They were asked to rate all the persons in their school, department or section on the scale according to the directions given.

The criteria for anyone to be selected as extravert were:

(i) that he should have been rated either

E by at least two of his associates, or E by one and D by another, and

(ii) he should not have been rated as A or B by anyone.

The introverts were also selected in the corresponding manner.

(6) Number of cases selected for the criterion groups from the non-students population:

Extraverts - 14

Introverts - 18

(7) Total number in the criterion groups - steps (5) + (6):

Extraverts - 117

Introverts - 114

Criterion for the Scale NN (Normal-neuroticism): For the selection of normal group, the instructions and definition were prepared (Appendix T). The group was selected by the same process as

used in the case of IE criterion groups, and from the same population. The rating procedure was also followed as for the same scale (Appendix U).

The results are given below:

(1) Total number of students who were nominated as normals: 283

(2) Total number of students who were nominated by at least 50 per cent of the group: 96

(3) Total number of students selected (from 2) after rating by two close associates: 54

(4) Number of cases selected from teaching and clerical personnel: 12

(5) Total number in the criterion group of normals: 66

It was not expected to find the extreme neurotics in sufficient number among the college going population. It was, therefore, decided to select cases of serious emotional maladjustments or

neuroticism from the general population, as well as from the college going people. An approach was also made to the practitioners in the medical profession to help to locate cases of neurotic illness rather than those of real physical sickness. It was decided, first, to locate such cases and then to get them rated by their close associates or by those who knew them well to pass such a judgment. Further the cases were interviewed by the present investigator while he administered the Inventory to them in person and tried to obtain further evidence regarding their status on the scale of neuroticism. The specimen of the instructions given for locating such cases is given in the Appendix V. The rating scale used for rating on this trait is given in the appendix and it is continuous with the rating scale for normality (Appendix W).

(1) The total number of cases which were located as neurotics or emotionally maladjusted were:

(a) Students:	72
(b) Non-students:	35
Total	107

These cases were rated by two persons who knew them very well. A five point scale was used, extending from normal to neurotic. Directions and definition of the scale were given to the raters. If one was rated by both raters as E, or if he was rated by both as D, or if he was rated by one as E and by another D on the scale, he was selected to be included in the criterion group of the neurotics.

(2) Number of cases selected after rating on the scale - normal-neuroticism:

Students:	43
Non-students:	16

(3) Total number of cases in the criterion group of the neurotics, therefore, were: 59

(4) Total number in the criterion groups:

Normals: 66

Neurotics: 59

Finally, when the tests were administered to these criterion groups, a few cases in each group were eliminated due to non-availability of the person, incompleteness of the forms filled or carelessness in filling in the answers as judged by the author while administering the Inventory. The number of cases in each of the criterion groups after this administration were:

Extraverts: - 96

Introverts: - 102

Normals : - 60

Neurotics : - 52

Scoring keys used in scoring these Inventories are given in Appendices J and K.

7.4 SUMMARY

The validity of a test is a degree to which it measures what it purports to measure. Whether

it does so or not is checked against external criteria. There are different types of validity. They are: face validity, content validity, concurrent validity, predictive validity, factorial validity, construct validity, intrinsic validity, operational validity and so on. The concepts relevant to the present work are: content validity and concurrent validity. The construction of the items as described in the Chapter V, and the definitions of the scales as discussed in the first chapter speak for the content validity of the test. For concurrent validation, external criteria for both the scales were selected by rating procedures. Utmost care was taken to obtain reliable ratings. The second form of the Inventory was administered to these criterion groups and was scored. The data was used for the cross-validation of individual items as described in the next chapter.

REFERENCES

1. Adams, Sidney K., "Does Face Validity Exist?" Educ.psychol.Meas., X,2, 320-28, 1950.

2. American Psychological Association, Technical Recommendations for Psychological Tests and Diagnostic Techniques. Supplement to psychol. Bull./No. 2, 1954. / LI,

3. Barthelmess, H.M., The Validity of Intelligence Test Elements. Contributions to Education No. 505. New York: Bureau of Publications, Teachers College, Columbia University, 1931.

4. Brogden, H.E. and E.K.Taylor, "The Theory and Classification of Criterion Bias". Educ. psychol. Meas., X, 2, 159-186, 1950.

5. Cureton, E.E., 'Validity' Chap.XVI in Educational Measurement, E.F.Lindquist (Ed.), Washington D.C.: American Council on Education, 1951.

6. Ebel, R.L., "Obtaining and Reporting Evidence on Content Validity". Educ.psychol.Meas., XVI, 3, 269-282, 1956.

7. Freeman, F.S., Theory and Practice of Psychological Testing. 3rd Edn. New York: Holt, Rinehart and Winston, 1962.

8. Gulliksen, H., "Intrinsic Validity". Amer.Psychologist, V, 10, 511-517, 1950.

9. Lennon, R.T., "Assumptions Underlying the Use of Content Validity". Educ.psychol.Meas., XVI, 3,294-304,1956.

10. Long, John A., et al., The Validation of Test Items. Bulletin No.3 of the Department of Educational Research, University of Toronto, 1935.

11. Monroe, V.S., An Introduction to the Theory of Educational Measurement. Boston: Houghton Mifflin Co., 1923.

12. Mosier, C.I., "A Critical Examination of the Concept of Face Validity". Educ.psychol.Meas., VII, 191-206, 1947.