CHAPTER - 5

RELIABILITY AND VALIDITY OF
THE D-I INVENTORY

5.1 Reliability And Validity

Reliability and validity are the essential characteristics of a standardized psychological test. This is because whenever anything is measured there is always a possibility of a chance error. Reliability of a test tells us how far it is free from the elements of chance error in measurement. The reliability of a test is its ability to yield consistent results from one set of measures to another. On the other hand, validity of a test is concerned with its truthfulness to measure a particular construct or behaviour. In other words, validity deals with the question as to whether the test really measures what it is supposed to measure. To go a step further, validity refers to the extent to which differences in scores on a test reflect true differences among individuals, groups, or situations on the characteristic which the test seeks to measure, or true differences in the same individual, group, or situation, from one occasion to another, rather than constant or random errors (Jahoda, Cook, et al, 1961). These authors go on further to say: "To the extent that the scores on a measuring instrument are influenced by random errors, they are not accurate indicators of an individual's position in terms of the characteristic one is trying to measure. In other words, to the extent that a measure is unreliable, it lacks validity." According to Hopkins (1961), reliability co-efficients are estimates of the consistency of measurement by methods that are maximally
similar, while validity is concerned with agreement or convergence among scores that are obtained by quite different methods. Validity studies enable us to make judgements about the individuals examined or the mental functions included in the test, whereas reliability is concerned with the accuracy of such decisions. A test can be valid only to the extent it is reliable. The relative position of a test score which results from measurement errors is a matter of concern to the interpreter of these scores, and hence the reliability of scores is an important aspect of validity. Thus, reliability and validity together go to make a test sound and useful.

Keeping in view these considerations reliability and validity of the D-I Inventory, the test developed to identify disciplined and indisciplined students, were determined as discussed below.

5.2 Reliability Of The D-I Inventory

Different methods of estimating reliability focus on different sources of variation in scores. Some are concerned with the stability of individual's position from one administration of the measure to another. In other words, they are concerned primarily with the fluctuations in the characteristic being measured or with change in transient personal or situational factors. This estimate of reliability is often designated as the stability co-efficient, i.e., it measures the stability of subjects' responses over a period of time. Among other measures of reliability may be mentioned those that tend to determine the equivalence of individuals' position on different
instruments intending to measure the same characteristic or on two equal halves of the same measuring instrument.

But as was mentioned at the outset, each of the given methods for estimating reliability is applicable only under a particular situation. Much is determined by the nature of the test. So far as the D-I Inventory is concerned, it is not susceptible to split-half method of determining reliability. This is because it cannot be split into two halves of an identical nature. It is to be recalled that the D-I Inventory is a forced-choice type of test in which each pair has been equated with respect to social desirability. As such, in the event of splitting the inventory into two halves it would have been very difficult to equate the halves in terms of content. Equivalent or parallel forms reliability was out of question, since there is no other equivalent instrument with a view to differentiate the indisciplined from disciplined students. In fact, in this respect, D-I Inventory is a pioneering venture in the field.

Under such conditions, the only available method to determine the reliability of the present test was the test-retest yielding a measure of the temporal stability of test scores. For this purpose the Inventory was administered to a group of 160 college students on two occasions with a time interval of 45 days. The coefficient of correlation between the scores obtained on two administrations of the test came out to be .68 which is reasonably high keeping in view the fact that the Inventory contained relatively small number of items, i.e., 50. This is especially encouraging in view of
the remarks made by Cronbach (1960) concerning reliability of forced-choice instruments. Cronbach says: "Reliability is decreased by shifting to the forced-choice form, because choice is made difficult. The forced-choice instrument is ordinarily a purer measure of the criterion relevant qualities in the test, because irrelevant verbal response habits are eliminated. Changing to the forced-choice form may or may not raise predictive validity because the loss in reliability may offset the gain in relevance." The D-I Inventory which yielded a coefficient of .68 is therefore quite reliable instrument.

Freeman (1963) has assigned two closely related but somewhat different meanings to reliability: (1) reliability as the extent of internal consistency, and (ii) reliability as the extent of consistency of results upon testing and retesting. The second type of reliability, according to Freeman, establishes the predictive value of the test. Since the reliability study of the D-I Inventory was of the second type and the correlation obtained was fairly high it implies that the test has, as Freeman says, predictive accuracy. This implies that identification of students as disciplined and indisciplined on the basis of D-I Inventory is fairly accurate.

5.3 Validity Of D-I Inventory

Validity has two major aspects - reliability and relevance. For a test to be valid, that is, to provide a sound basis for judgements, it must measure 'something' with reasonable reliability, and that 'something' must either be
a sample of the behaviour we wish to measure or it must have demonstrating relevance to that behaviour (Adams, 1964). A test must be reliable before it is valid. A reliable test may not be valid but a valid test must be reliable. Having determined the reliability of the D-I Inventory we now devote our attention to the validity aspect of the Inventory.

Though the validity of a test depends upon both its relevance and its reliability, the effectiveness of a test is not a function of the reliability of the criterion scores. Validity is defined in terms of the correlation between the actual test scores and the "true" criterion scores. A "true" score is that part of any actual score which is not due to error of measurement. Since we do not usually know an individual's true position on the variable we are attempting to measure there is no direct way of determining the validity of the measure. In the absence of any direct knowledge of individual's true position on the variable being measured, the validity of an instrument is judged by the extent to which its results are compatible with other relevant evidence. What constitutes relevant evidence depends upon the nature and purpose of the measuring instrument. This leads to the determination of a suitable criterion against which the individual's scores on the test can be correlated. Fundamentally, all procedures for determining test validity are concerned with the relationship between the performance on the test (test scores) and other independently observable facts about the behaviour under consideration (criterion scores). The APA Technical Recommendations (1954) have classified the specific
techniques for investigating these relationships under four heads: (i) content validity (ii) Predictive validity (iii) concurrent validity, and (iv) construct validity. The type of validity to be determined in the case of a test is decided upon by the nature and purpose of the test.

D-I Inventory is a forced-choice type of test which aims at the identification of indisciplined and disciplined students. Here a distinction has to be made between prediction and identification. Prediction involves forecast of future behaviour. Predictive validity indicates the effectiveness of a test in predicting some future outcome. For this purpose test scores are checked against a direct measure of the subject's subsequent performance, technically known as the criterion. In identification, however, independent observation of the present behaviour of the individual is the criterion and it does not involve any follow-up of the criterion behaviour. Here, let it be made clear that identification also has an implicit purpose of prediction. The D-I Inventory identifies the potentially indisciplined students as well. As such, the later (follow-up) behaviour of the subjects will be very helpful in judging the validity of this Inventory. Hence, predictive validity will be very useful, but since it involves follow-up, it is outside the purview of the present study. Attempt would, of course, be made to determine the predictive validity of the test at a later stage. For the present, however, the purpose of the D-I Inventory is not prediction but identification. Therefore, concurrent validity suits our purpose best. Concurrent validity is investigated when the test is proposed as a
substitute for some other information; this information is then the criterion. Designers of new tests frequently establish concurrent validity for their instruments by comparing them to established tests. According to Freeman (1963), concurrent validity is one of the newer terms. In this case the new test evolved is validated by correlating it with some present source of information. This source of information might have been obtained shortly before or after the test was given. Thus, the criterion measure is gathered at almost the same time. Such concurrent validation, as distinguished from predictive validation, is used to determine whether one measure can substitute another measure which is not regarded as satisfactory.

As such, the problem of validating D-I Inventory is essentially a problem of establishing its concurrent validity. It has to be borne in mind that the D-I Inventory is the first-ever test to identify the indisciplined and disciplined students. So, the question of using any established test as the criterion does not arise. Under such circumstances, teachers' ratings of students' behaviour was the only plausible criterion against which scores of the students on the Inventory could be correlated. For obtaining teachers' ratings, a 4-point rating scale was prepared as described at length in chapter 3 (pp.58 - 59). To eliminate rater-bias, these ratings were obtained from more than two teachers who had the opportunity to watch these students closely, and a student was put as disciplined or indisciplined only if at least two teachers agreed in describing a student as disciplined or indisciplined.
The particular instances of indiscipline were arranged on the rating scale in terms of the gravity of misbehaviour involved. This was done to render the task of rating easy for the raters and at the same time to make these ratings objective as far as possible. The rating scale thus prepared did not leave any scope for exercising discretion on the part of the rater.

The sample used for validation consisted of 150 college students who were given the Inventory. Ratings were obtained from the teachers immediately after the testing session. Item-wise validity for all the 50 items in the test was determined by means of point-biserial correlations. Simple correlation was not possible in this case because of the dichotomous nature of the item responses as well as the criterion, the ratings given by teachers. Phi-Coefficient method was also tried but was discarded later owing to certain operational difficulties. The point-biserial correlations for each of the 50 items in the Inventory are being reported below in Table 5.1.

### Table 5.1

<table>
<thead>
<tr>
<th>Items</th>
<th>Point-biserial correlations</th>
<th>Level of Significance</th>
<th>Items</th>
<th>Point-biserial correlations</th>
<th>Level of Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A</td>
<td>.45</td>
<td>.01</td>
<td>3A</td>
<td>.68</td>
<td>.01</td>
</tr>
<tr>
<td>1B</td>
<td>.49</td>
<td>.01</td>
<td>3B</td>
<td>.70</td>
<td>.01</td>
</tr>
<tr>
<td>2A</td>
<td>.20</td>
<td>.01</td>
<td>4A</td>
<td>.45</td>
<td>.01</td>
</tr>
<tr>
<td>2B</td>
<td>.25</td>
<td>.01</td>
<td>4B</td>
<td>.50</td>
<td>.01</td>
</tr>
</tbody>
</table>
Table 5.1 Continued

<p>| | | | | | | | | | | | | | | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5A</td>
<td>.75</td>
<td>.01</td>
<td>16B</td>
<td>.24</td>
<td>.01</td>
<td>16A</td>
<td>.55</td>
<td>.01</td>
<td>16B</td>
<td>.60</td>
<td>.01</td>
<td>16A</td>
<td>.65</td>
<td>.01</td>
<td>16B</td>
<td>.70</td>
<td>.01</td>
<td>16A</td>
<td>.78</td>
<td>.01</td>
<td>16B</td>
<td>.60</td>
<td>.01</td>
<td>16A</td>
<td>.78</td>
</tr>
<tr>
<td>7A</td>
<td>.09</td>
<td>Insignificant</td>
<td>17B</td>
<td>.05</td>
<td>Insignificant</td>
<td>18A</td>
<td>.20</td>
<td>.01</td>
<td>18B</td>
<td>.25</td>
<td>.01</td>
<td>18A</td>
<td>.30</td>
<td>.01</td>
<td>18B</td>
<td>.25</td>
<td>.01</td>
<td>18A</td>
<td>.32</td>
<td>.01</td>
<td>19A</td>
<td>.04</td>
<td>Insignificant</td>
<td>19B</td>
<td>.05</td>
</tr>
<tr>
<td>9A</td>
<td>.80</td>
<td>.01</td>
<td>20A</td>
<td>.02</td>
<td></td>
<td>20B</td>
<td>.03</td>
<td></td>
<td>21A</td>
<td>.08</td>
<td></td>
<td>21B</td>
<td>.05</td>
<td></td>
<td>22A</td>
<td>.29</td>
<td>.01</td>
<td></td>
<td>22B</td>
<td>.35</td>
<td>.01</td>
<td></td>
<td>23A</td>
<td>.09</td>
<td>Insignificant</td>
</tr>
</tbody>
</table>

Significance of correlations reported in Table 5.1 was studied in terms of the levels of significance, as laid down by Garrett (1960). According to this, a correlation of .169 is significant at 5% level, while a correlation of the value of .208 is significant at 1% level. On observing
Table 5.1 we find that 35 of the 50 correlations are significant at either of these two levels. This establishes a fairly high validity for the test, since 70% of the items have correlated well with the criterion accepted. It means, in other words, that the Inventory reasonably fulfils the requirements of a valid test. Moreover, it has to be kept in mind that D-I Inventory is a first-ever attempt at measuring student indiscipline, and this singular attempt is marred by the non-availability of a sound criterion against which to validate subjects' scores. It is expected that with a larger and more heterogeneous cross-sample for validation the correlations in the case of items which have failed to reach the level of significance will improve, thus adding to the validity of the Inventory.