Reliability and Validity

Learning Objectives

After reading this chapter, you should be able to:

- Understand the concepts of reliability and validity.
- Explain how to estimate reliability using different methods.
- List the factors affecting reliability.
- Discuss how to devise a reliable scale.
- Describe the different types of validity.
- Enumerate the steps to be taken to improve validity.

INTRODUCTION

There are several ways to measure every attribute (say mathematical proficiency) and particular set of items (how good are the test-takers in trignometry?) can be constructed to measure it. How does one evaluate the effectiveness of these items that are supposed to measure the selected attribute or answer the specific question?

The main questions to face while evaluating or measuring an object are:

- Does the same measurement process yield the same results?
- Is what we intend to measure being really measured?

The main factors to consider while evaluating a test, an object or item are: reliability and validity.

Reliability takes in stability and consistency. Does repeated application of a tool lead to a reliable tool?

Validity indicates the degree of accuracy of the measurement. To apply these concepts to teaching, tools used for measurement need to be reliable as well as valid.

Validity and reliability compared

The relationship between validity and reliability is shown in Figure 7.1.

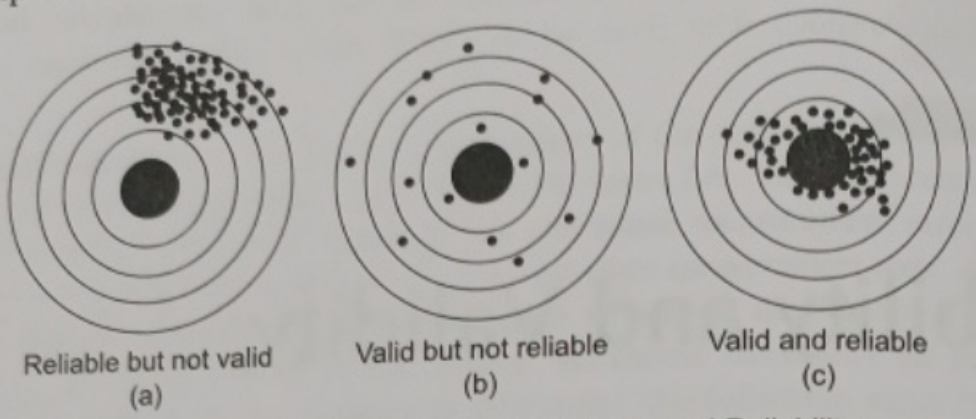

Figure 7.1 Relationship between Validity and Reliability.

- Figure 7.1(a), (b) and (c) indicates how reliability and validity differ using an example in target shooting.
- The target shooting in Figure 7.1(a), indicates how reliability is necessary, but not sufficient condition for validity.
- Figure 7.1(c) is a measure that has both high validity and high reliability giving consistent results on repeated trials.
- Figure 7.1(b) shows how it is also possible to have one that is unreliable and invalid: inconsistent and not on target.
- Finally, it is not possible to have a measure that has low reliability and high validity.

7.1 TESTING RELIABILITY FOR SOCIAL SCIENCES AND EDUCATION

Unlike the physical sciences where microscopes and spectrometers are used for measurement, the behavioural science uses achievement and psychometric tests, questionnaires and the like. Reliability can be considered as a prerequisite for validity. For example, if a self-concept test gives very different scores for the same student essentially under the same conditions, then these scores cannot be considered as a measure of the student's self-concept. Similarly, if the items on a self-concept test are not correlated with each other, then they cannot all be measuring self-concept, and an aggregate of them cannot be a very good measure of self-concept.

A reliability coefficient, which gives the correlation between two or more variables, is used to determine the reliability of a test. This indicates the degree and direction of a relationship between two or more variables. Reliability is the state wherein consistent scores are obtained over repeated testing. Measures that are high in reliability should exhibit all stability and consistency.

Reliability, in simple terms, describes the ***stability*** and ***consistency*** of a test (Figure 7.2).

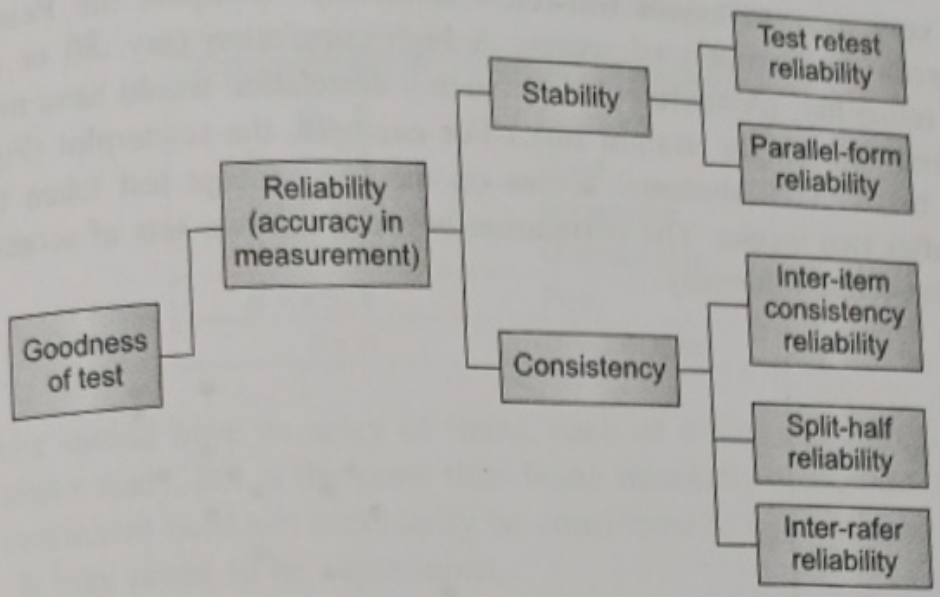

Figure 7.2 Reliability of Test.

Stability of a test

Stability is an aspect of reliability. **Stability** refers to the consistency of the scores over time. It is a feature of reliability which is computed by correlating the two test scores of a group of students taken at different times. This definition clearly focuses on the measurement instrument and the obtained test scores in terms of test-retest stability.

Consistency of a test

Internal consistency refers to correlations between different items on the same test. It indicates the degree of consistency between the various items which purport to measure a particular attribute.

7.1.1 Stability

It is of the following two types:

Test-retest reliability

The test-retest method is the simplest method for testing reliability, and involves testing the same subjects at a later date, ensuring that there is a correlation between the results. An educational test retaken after a month should yield the same results as the original. When a measurement tool is administered multiple times, posing the same questions and using the same procedures of administration, if the results are consistent the measure has test-retest reliability. If the test is reliable, the scores of each student on the first occasion should be similar to the scores on the second. Test-retest reliability refers to the test's consistency when administered under different occasions. The same test is given to a group of subjects on at least two separate occasions and reliability found. Test-retest reliability is the degree to which scores are consistent over time. Examples in education where test-retest could be used are studies on memory, maturation, learning.

It appears very easy to assess test-retest reliability. Compute the Pearson's r for the correlation between the two sets of scores. A high correlation (say .80 or above) indicates good test-retest reliability. (A **scatterplot** for such a correlation would have most of the points falling pretty close to a single straight line.) For example, the scatterplot (Figure 7.3) shows the relationship between 20 students' scores on the self-concept test taken first on Monday and then again after two weeks. The correlation between the two sets of scores is +.87, which indicates good test-retest reliability.

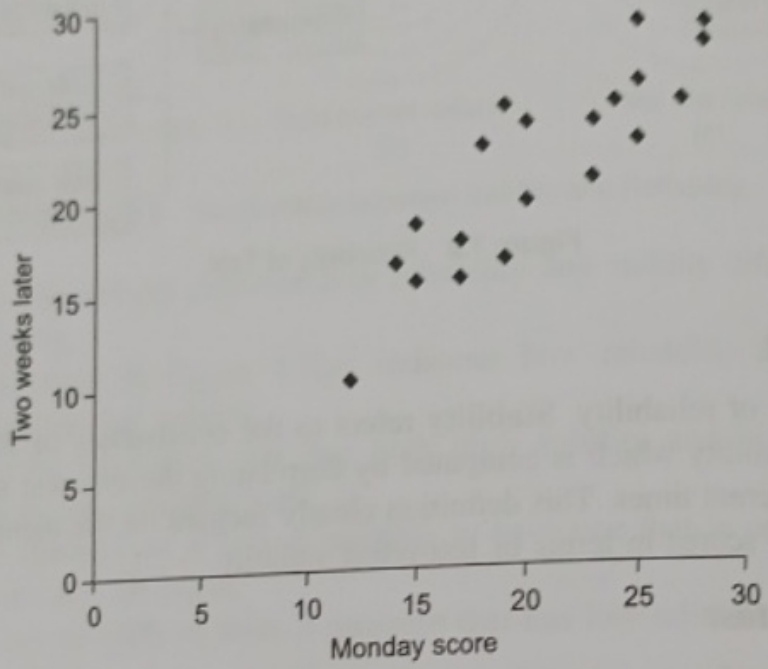

Figure 7.3 Scatterplot—Self Concept.

Parallel forms/Equivalent-forms/Alternate-forms reliability

Two tests that are identical in every way except for the actual items included. Used when it is likely that test takers will recall responses made during the first session and when alternate forms are available. Correlate the two scores. The obtained coefficient is called the *coefficient of stability* or *coefficient of equivalence*. But, the problem is difficulty of constructing two forms that are essentially equivalent. Both of the above require two administrations. Using different pre and post-tests helps minimise errors due to memory effect, to ensure this, the two tests must be parallel or equal in what they measure. To determine parallel forms reliability, a reliability coefficient is calculated on the scores of the two measures taken by the same sample.

7.1.2 Consistency

It includes the following:

Inter-item consistency reliability

Internal consistency is generally measured with Cronbach's alpha, which is a pairwise correlations between items. The values of internal consistency varies from zero to one. To interpret the Cronbach's alpha values Table 7.1 is used.

Table 7.1 Interpretation of Cronbach's alpha

Cronbach's alpha	Internal consistency
$\alpha \geq .9$	Excellent
$.9 > \alpha \geq .8$	Good
$.8 > \alpha \geq .7$	Acceptable
$.7 > \alpha \geq .6$	Questionable
$.6 > \alpha \geq .5$	Poor
$.5 > \alpha$	Unacceptable

A reliable test should have an array of items, each of which contributes a unique aspect of the attribute under study, but at the same time being internally consistent. An item having a high reliability coefficient need not necessarily be considered to contribute to the measurement of the attribute. It may prove to be superfluous.

The variance of the sum of two items is equal to the sum of the two variances from which twice the value of the covariance is subtracted. The most common index of reliability, is Cronbach's coefficient *alpha* (α). Therefore, coefficient *alpha* will be equal to zero. All perfectly reliable items will have a reliability coefficient of 1. Conceptually, Cronbach's alpha is the mean split-half correlation for all possible ways of splitting the items in half. The items could be split into the even items and the odd items; in a 10-item measure the first half (items 1-5) and the second half (items 6–10), or even items 1, 3, 4, 9, and 10 vs. items 2, 5, 6, 7, and 8 ... and so on. Splitting the items into halves and taking the mean of these split-half correlations, gives Cronbach's alpha.

(Adapted from Cronbach, L.J., 1990, *Essentials of Psychological Testing*, 5th ed., Harper and Row, New York.)

Split-half reliability

The split-half method has the advantage that the test need be administered only once. This is very time saving when one has long tests. Generally, the test is split into two halves—odd number questions and even number questions. Spearman–Brown prophecy formula is also called the *split-half reliability* which is used for finding internal consistency reliability.

The **Spearman–Brown split-half coefficient** can be used to find the reliability of the sum scale. Internal consistency can be computed by finding the **split-half correlation**. This is the correlation between two scores, one based on one half of the items and the other based on the other half of the items. Imagine that 100 students have taken the Science Aptitude test comprising 30 items. Two Science Aptitude scores can be computed: one based on items 1, 3, 5, 7, and 9, ...29 and the other based on items 2, 4, 6, 8, and 10...30. The correlation between the two scores can then be computed.

Kuder Richardson formula (KR20)

This is used when the test has dichotomous items, that is yes/no or right/wrong. For Likert type items or other type's Spearman–Brown formula can be used. The formula for KR20 is:

$$r_{KR20} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\Sigma pq}{\sigma^2}\right)$$

where

r_{KR20} = Kuder Richardson formula 20

k = total number of test items

Σ = sum

p = proportion of the test takers who pass an item

q = proportion of the test takers who fail an item

σ^2 = variation of the entire test.

Example of use of KR20

Santhoshi was a Maths teacher in a reputed school in Cochin. After finishing a unit in Arithmetic she wanted to check how well the students had learned the basic concepts in that unit. She administered a 10 item arithmetic test to 15 children. Table 7.2 gives the test scores of the 15 students in her class. Santhoshi marked a 1 if the student answered the item correctly and a 0 if the student answered incorrectly.

Table 7.2 Scores on a Maths Test—KR20 Formula

Roll No.	Maths Problem Number									
	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	1	1	1	1
2	1	0	0	1	0	0	1	1	0	1
3	1	0	1	0	0	1	1	1	1	0
4	1	0	1	1	1	0	0	1	0	0
5	0	0	0	0	0	1	1	0	1	1
6	0	1	1	1	1	1	1	1	1	1
7	0	1	1	1	1	1	1	1	1	1
8	0	0	1	1	0	1	1	0	1	0
9	0	1	1	1	1	1	1	1	1	1
10	0	0	1	1	0	1	0	1	1	1
11	0	0	1	1	0	0	0	0	0	1
12	1	1	0	0	0	1	0	0	1	1
13	1	1	1	1	1	1	1	1	1	1
14	0	1	1	1	0	0	0	0	1	0
15	0	1	1	1	1	1	1	1	1	1
No. of 1s	6	8	12	12	7	11	10	10	12	11
Proportion passed (*p*)	.40	.53	.80	.80	.47	.73	.67	.67	.80	.73

Substituting values of k, p and q into the KR20 formula,

$$r_{KR20} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\Sigma pq}{\sigma^2}\right)$$

$$= \left(\frac{10}{10-1}\right)\left(1 - \frac{2.05}{5.57}\right)$$

$$= 1.11 * 0.63 = 0.70$$

This test was administered for a single unit in arithmetic. The KR20 is sensitive to measurement error introduced by content sampling. In most cases Cronbach's alpha is preferred for estimating reliability of a test.

Inter-rater reliability

One way to determine whether the observations are reliable or not, is to have two or more observers rate the same subjects and then correlate their observations. When a single concept is measured using multiple items, the inter-rater reliability is employed. In such cases, answers to a set of questions designed to measure some single concept (for example, altruism) should be associated with each other. If the same event is observed by two different individuals using the same rating scale, one would expect their scores to be in the same range. If it does not match it implies that the inter-rater reliability has to be checked. Table 7.3 gives the rating of science projects by four experienced raters. A 5 point scale was used. The disparity in rating is clear in student #2, Student #4. Student #1 and Student #5 seem to be rated similarly by the four raters.

Table 7.3 4 Experienced Raters of Science Projects by 10 Students

Classification on a 5 point scale by 4 experienced raters of science projects by 10 students				
	Rater 1	Rater 2	Rater 3	Rater 4
1	5	5	5	5
2	3	4	5	2
3	5	4	4	3
4	3	2	3	2
5	4	4	4	4
6	3	2	4	4
7	2	2	2	3
8	5	4	5	3
9	3	3	3	4
10	3	3	2	3

Table 7.4 gives the relationship of test forms and testing sessions required for reliability procedures.

Table 7.4 Test Forms and Testing Sessions Required for Reliability Procedures

Testing sessions required	Test forms required	
	One	Two
One	Split half Kuder Richardson Cronbach's Alpha	Equivalent (Alternative) Form
Two	Test-retest	

7.2 DESIGNING A RELIABLE SCALE

To construct a more valid sum scale it is necessary to add more items to the test. However, a long test or questionnaire could lead to respondent fatigue, administrative constraints besides time management.

Measures of reliability refer to a statistic to measure or describe the consistency of a test item or a scale. The proportion of true score variability that is depicted across test-takers in relation to the total observed variability is called *index of reliability*. Kuder-Richardson's Formula is an estimate of reliability that is essentially equivalent to the average of the split-half reliabilities computed for all possible halves.

7.2.1 Sum Scales

Assuming that the error component is random, it can be assumed that the different error components in the responses of the students will cancel out, giving a mean value of zero for the error component across the items. This results in the sum of the items depicting the true score remaining the same across the items. Hence, the inclusion of more items will reflect the true score, which in turn, will be depicted in the sum scale.

A more reliable measure of the concept under study can be obtained by inclusion of more items.

7.2.2 Internal Consistency

Steps for designing a reliable test

Step 1 ***Constructing items.*** Keeping the attribute under study in mind, a plethora of items are constructed. This is a creative process and a lot of brain storming generally goes into this step.

Step 2 ***Selecting items of appropriate difficulty level.*** Items which exhibit extreme means and zero or near zero variances are eliminated.

Step 3 ***Choosing internally consistent items.*** Items chosen to reflect the attribute under consideration will help enhance the reliability of the scale.

Shown in Table 7.5 are the results of a reliability analysis using SPSS for 6 items. The three right most columns give the correlation between the specific item and the total sum score (exclusive of the particular item), the squared multiple correlation between the specific item and the rest of the items, and the internal consistency of the scale (coefficient *alpha*) if the specific item is deleted. The overall alpha value is 0.6534. Therefore, the alpha values in the last column should be around 0.6534. As item 2 has a value of 0.6905 it can be deleted to increase the reliability of the tool.

Step 4 ***Getting back to Step 1.*** Items not consistent with the rest of the scale are deleted and only items consistent with the overall scale are retained. Construction of a reliable scale involves repeated inclusion of items, testing for consistency and then including or deleting depending on the effectiveness of the item with respect to the overall scale.

Table 7.5 Results of Reliability Analysis of 6 Items

Variable	Scale Mean if Item deleted	Scale Variance if Item deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
Item 1	14.8650	16.6141	.5187	.3881	.6406
Item 2	*14.1684*	*16.7370*	*.5190*	*.3846*	*.6905*
Item 3	14.8114	19.7360	.2999	.1665	.6439
Item 4	14.6429	17.8086	.5067	.3744	.6337
Item 5	14.2159	18.7190	.4768	.2567	.6480
Item 6	14.5235	16.8324	.3913	.4543	.6455

Reliability Coefficients: 6 items

Alpha: .6534 Standardised item alpha: .6214

7.3 THREATS TO RELIABILITY

Observer reliability

Errors can arise during the measurement process. These errors are referred to as observer error. In spite of accurate measurement procedures being followed, an observer can make judgmental errors.

Situational reliability

While the test is being administered, situational changes can occur. For example, the school may be having dress rehearsals for school day, or examinations may be going on in one of the floors. These kind of situations can lead to errors. These errors caused by environmental changes can lead to discrepancies in measurement.

Subject reliability

Subjects can change during the course of measurements. Fatigue can set in; subjects can get bored with the process of measurement. These variations can result in measurement errors.

Instrument reliability

The instrument used may be poorly worded, too lengthy or not well. The instrument have extreme response style, may otherwise encourage midpoint responding. There may be leading questions or the direction of wording may affect the response.

7.4 TEST VALIDITY

Validity has long been recognised as an important aspect of testing and psychological assessment (Standards for Educational and Psychological Testing, 1999). Before drawing inferences from test scores, the validity of the test scores have to be ascertained by the test-taker. Cronbach maintained that rarely is a validity coefficient higher than 0.60.

The unified theory of Messick's (1989) suggested an all encompassing canopy of construct validity under which all facets of validity fall (Figure 7.4).

Messick's Facets of Validity Framework		
	Test interpretation	Test use
Evidential basis	Construct validity	Construct validity + Relevance/utility
Consequential basis	Value implications	Social consequences

Adapted from "Validity" by S. Messick in Educational Measurement (3rd ed., P. 20) edited by R.L. Linn, 1989. New York American Council on Education and National Council on Measurement in Education.

Figure 7.4 Messick's Unified Theory on Validity.

- **Content validity** examines how well the test items measure the construct under consideration. The content validity is found using a panel of experts who carefully review each item in the test. Sometimes, the blueprint of the test is examined to see how well the objectives and the content are matched, whether there is adequate representation of all areas of content and so on.
- **Substantive validity** studies are concerned with the basic concepts forming the core of the construct under consideration. Collection of evidence should indicate clearly that the basic concepts underlying the core of the construct are being measured.
- **Structural validity** is studied by analysing the interrelationships of the various sub-constructs assessed by the test. It further involves drawing inferences based on the relationship of each of these sub-construct with the main construct and drawing.
- **Generalisability** is concerned with the confidence with which the scores obtained on the test can be generalised over different samples and populations.
- **Criterion validity** as well as predictive validity express the extent to which the scores correlate to a criterion, which is an external standard. This facilitates prediction of future performance of the test-taker.
- **Consequential validity**, is according to Messick, is a study of the effect of use of invalid scores for making decisions.

Validity points to the strength of the results obtained and whether they can be taken to be accurate assessment of whatever is being measured.

7.4.1 Categories of Validity

There are many ways of gathering information to check the validity of a test score. The five main sources of evidence are:

1. Test content, which includes an analysis of the content to be tested.
2. Response processes, which refers to the different ways in which the students respond to the questions.
3. Internal structure of the test, where the relationship to other variables is considered.

4. Relation to other variables.
5. Consequences of testing.

Validity has been generally classified into three categories—***content validity***, ***criterion-related*** and ***construct-related***. Each category has to be considered to evaluate the degree of validity of any set of scores or any instrument used for testing. Figure 7.5 shows the different categories and the sub-categories of validity.

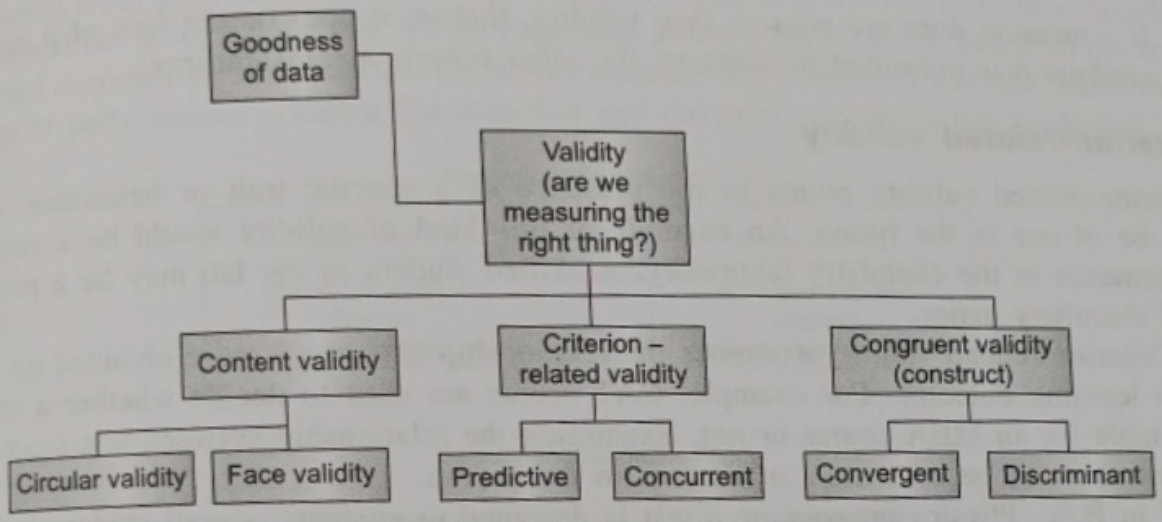

Figure 7.5 Types of Validity.

Content validity

Content validity is a measure of all aspects of a construct. Taking fatigue as the construct, different aspects of it would be yawning, sleeping at odd hours, indifferent to things usually enjoyed, etc. However, if there are other aspects of fatigue which have been left out in the test to measure this construct, the test would have low content validity.

Example of content validity

To test knowledge on Indian History it is not fair to have most questions limited to the history of Asia.

An example would be intelligence. Intelligence has various dimensions like analytical ability, logical reasoning, verbal comprehension and so on. Any test to check intelligence should represent all these aspects of intelligence, like verbal ability, spatial reasoning, analytical proficiency and other aspects of intelligence to establish its content validity. Content-related evidence of validity comes from the judgements of people who are either experts in the testing of that particular content area or are content experts. In contrast, because these two groups may approach a test from different perspectives, it is important to recognise the valuable contributions made by both. The content validity can be obtained using either or both ***face validity*** and ***curricular validity***.

Curricular validity: Curricular validity is the degree to which the content of the test and the objectives of the curriculum are in tune with each other. Curricular validity can be the key determiner on whether a student makes the grade or not. A panel of experts, including

educators, subject-matter experts and teachers assess the suitability of the test to measure the attainment of the stated objectives of the subject.

Face validity: Face validity is the level to which a test appears to measure an attribute as assessed by all concerned. The test should appear sound in testing whatever it is supposed to. This assessment is vital as it facilitates use of the test by laypersons for a variety of purposes.

If a measure does not possess face validity, that is, it does on the face of it measure the attribute it is purported to measure, the other criteria are insignificant.

Criterion-related validity

Criterion-related validity points to the presence of a specific trait or behaviour, which may be of use in the future. An example of this kind of validity would be a student's performance in the chemistry laboratory. A skilled student in the lab may be a potential good chemistry major.

Criterion-related validity examines the relationship between a score obtained on a test and a learning outcome. For example, CAT scores are used to decide whether a student is suitable for an MBA course or not. Examining the relationship between test scores and the criterion can be a measure of success in the course.

If in B.Sc. Physics programme a test is designed to evaluate overall student learning throughout over the academic year. The correlation of the score obtained with a standardised measure of ability in this discipline can then be found. A high correlation value indicates that the tool is good. The criterion-related validity can be obtained using either or both ***predictive validity*** and ***concurrent validity***.

Concurrent validity: It is easier to understand concurrent validity with a simple example. Wechsler Scale for Adult Intelligence is considered as the standard construct of intelligence in its field. Therefore, any new construct (termed device) we attempt to evolve in the same field has to necessarily have positive correlation to concur with this. In general, one should ensure that the new construct evolved has positive correlation to concur, i.e., directly vary with an earlier established and recognised construct in the same field. Conversely, if in case there is an existing construct which measures just the opposite of the characteristic under study, then the construct now evolved should have negative correlation with the existing one, i.e., conversely or indirectly vary with the existing one. In this way it is possible to demonstrate concurrence or validity of the currently attempted test to a test already established as valid.

Concurrent validity indicates the extent to which the scores on a test correlate with another similar established test. For example, replacing a long winded test with a simpler test, obtaining scores from both the tests.

Predictive validity: It refers to the "power" or usefulness of test scores to predict future performance tests need predictive validity in order to be of use for screening and selection purposes (Figure 7.6). The GATE score is used by most university screening committees as predictors of future success in college. The GMAT is used for prediction of success in business management courses. Predictive validity is a key factor when using these tests

Predictive validity is computed by a correlation coefficient comparing GATE scores, for example, and college grades. If they are directly related, then it can be used to make a prediction regarding college grades based on GATE score. A criterion-related validation study can serve two purposes: as prediction of future behaviour or a concurrent measure of current behaviour. Finding predictive validity is recommended when standardised test scores are used for admission in courses offered by the University or college. The same placement study used to determine the predictive validity of a test can be used to determine an optimal cut score for the test. When expecting a future performance based on the scores obtained currently by the measure, the scores obtained are correlated with the performance. The later performance is called the *criterion* and the current score is the *prediction*.

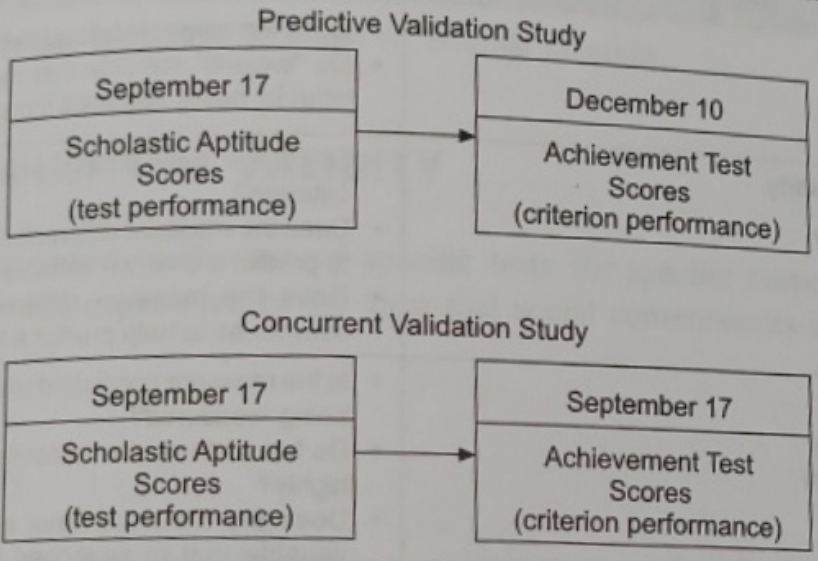

Figure 7.6 Studies of Types of Assessment—Criterion Relationships (based on time difference only).

Construct validity

Construct validity also called ***congruent validity*** is the term given to a test that measures a construct accurately. Construct validity is the extent to which a test measures what it is supposed to measure. If the construct is honesty, then a test is designed to assess the level of honesty of the person, giving several situational questions. To assess construct validity data will have to gathered through several sources of evidence. For example, to assess "honesty", would require not only answering the situational questions, but also observation of behaviour in other real life situations like in the classroom, the playground and so on.

Construct validity refers to ***"the degree to which a test measures what it claims, or purports, to be measuring."*** In other words it occurs "whenever a test is to be interpreted as a measure of some attribute or quality which is not operationally defined."

Convergent validity: Construct validity is made up of convergent and discriminant validity. If theoretically two measures are postulated to be related, and are in fact found to be related, they are said to have convergent validity.

Discriminant validity: It is the reverse. If theoretically two measures are postulated to be unrelated, and are in fact found to be unrelated, they are said to have discriminant validity.

Discriminant validity, indicates that two tests do not correlate strongly with each other if they are not measures of similar skills or knowledge. Discriminant validity is an important pointer of construct validity. For example, a test of arithmetic should basically measure constructs related to arithmetical concepts and not reading and vocabulary skills.

Validity refers to the extent to which the item truly measures what it intends to measure (Table 7.6).

Table 7.6 Summary of Validity on Target

Validity	Description
Content validity • Face validity • Curricular validity	• Does the measure adequately measure the concept? • Do "experts" validate that the instrument measures what its name suggests it measures?
Criterion-related validity • Concurrent validity • Predictive validity	• Does the new measure agree with an external criterion? • Does the measure differentiate in a manner that helps to predict a criterion variable? • Does the measure differentiate individuals in a manner as to help predict a future criterion?
Construct validity • Convergent validity • Discriminant validity	• Is the measure consistent with the theoretical concept being measured? • Do two instruments measuring the concept correlate highly? • Does the measure have a low correlation with a variable that is supposed to be unrelated to this variable?

7.5 THREATS TO SCORE VALIDITY

Two serious threats to validity of score are when the construct is not suitable represented in the test and when construct irrelevance exists. Examples of these two threats are:

1. To measure problem solving ability in mathematics, if the item requires a great deal of reading, then the reading ability mars the assessment of the student's problem solving ability.
2. In multiple choice questions, students can often arrive at the correct option not because they know the concept, but because of elimination of other options.

7.6 IMPROVING VALIDITY

Following are to be considered for improving validity:

- Statements of goals and objectives should be clear, concise and operational.
- Behaviour expected of students should be written in attainable terms.

- Assessment measures should correspond to the goals and objectives stated.
- Face validity can be obtained by forming a panel of teachers from cooperating schools.
- Pilot study will help check ease of administration, length of test, wording and other such features of the test.

7.7 SENSITIVITY

The sensitivity of a scale helps assess changes in attitudes or other hypothetical constructs which may are being studies. The ability of an instrument to accurately measure changes in stimuli or responses is sensitivity. A dichotomous choice such as "Yes or No" or "True or False" does not cover the spectrum of range of attitude changes.

7.8 CONSEQUENTIAL VALIDITY

Consequential validity refers to the use of specific tests for specific purposes which benefit society. However, several experts are of the view that social consequences does not fall in the domain of validity at all.

CONCLUSION

Reliability and validity of the data collected is very important, as they facilitate good decisions. Reliability refers to consistency of scores, whereas validity implies accuracy. This chapter takes look at the different types of reliability and validity which will prove useful for a teacher in r day-to-day interactions with her students while assessing them.

Review Questions

1. What is the role of reliability in testing and assessment?
2. Define validity and reliability. How do the two differ?
3. When is the reliability coefficient used?
4. How would you increase the reliability of a test you administer?
5. What is meant by validity coefficient? What are the factors that affect validity?
6. List the several factors which could affect the validity of your class test. Describe how you could enhance the validity of your test.