Name: Dr.P K Sivakumaran,
Assistant Professor

Subject Code : 18BST46S

Subject Name :
Psychological Statistics

8·4. VALIDITY OF TEST SCORES

As already pointed out, test is like a measuring instrument. In addition to reliability another essential property of the measuring instrument is its validity. A measuring instrument is said to be valid if the measurements made by it are accurate and comparable with those made by a standard instrument. For example, a home made yardstick is valid if its readings are accurate in terms of a standard measuring rod. Similarly a test is said to be valid if the performances measured by it are accurate and comparable with the measurements as otherwise independently obtained. In other words, the validity of a test depends upon the accuracy with which it measures what it is supposed or intended to measure. In fact, validity is a relative term, a test which helps in making one decision may have no value at all for another.

8·4·1. Estimation of Validity. To estimate the validity of a test for measuring a given trait, we make use of some known measures of the trait, called the *criterion variable* and the estimate is given in terms of a coefficient, the *coefficient of validity* which determines the relationship between the scores obtained on the test and the value of the criterion variable. For example, the validity of a writing test may be judged by the correlation between the writing errors (observed score) and the writing speed (criterion variable).

The reliability of a test is determined by obtaining the coefficient of self-correlation between the scores of n individuals on two repetitions of a test or on two parallel forms of a test. On the other hand, validity of a test is determined experimentally by obtaining the coefficient of correlation between the scores of n individuals on the given test (X) and some independent standard test (Y) called *criterion*. One of the most difficult aspects of the validity problems is the choice of a proper and adequate criterion variable and obtaining measures on the variable which are to be compared with the scores on the given test. A criterion may be an objective measure of performance or a qualitative measure such as judgement of characters or excellence of work done. These criteria are most often approximate and indirect, for if reliable criterion were easily available, they would be preferred for use rather than the tests. A high correlation coefficient between X and Y is an evidence of validity provided that—

(*i*) the criterion Y was set up independently, and (*ii*) both X and Y are reliable.

For example, the validity of a typing test may be judged by correlation between the errors (score X) in the matter typed and the speed (criterion Y).

Remarks **1.** It may be pointed out that validity is a highly relative concept. A test may be valid for a particular purpose trait, group or situation and not always. If a test is used for measuring different traits, then we must obtain its validity for different traits separately.

2. Since independent standards (criteria) are difficult to get in mental measurements, the validity of mental tests cannot be determined as accurately as the validity of physical instruments.

8·4·2. Types of Validity. We discuss below different types of validity :

(*a*) ***Predictive Validity***. This type of validity comes up in a test for selecting applicants for different courses of study, or training or jobs. Here the criterion variable is the performance of the recruits at a later period after they have completed their training and have been on the job for a sufficiently long period. A test has high predictive ability if it can predict efficiently the performance of the candidates on a later period. A test for admission to a course or recruitment to a job should have high predictive validity.

(*b*) ***Concurrent Validity.*** This type of validity is used for tests like clinical diagnosis etc., where the criterion variable is also available side by side with the test scores. Thus, here we need not wait for the measure of the trait as in the case of the predictive validity. This device is simpler, time-saving and economical.

(*c*) ***Content Validity.*** This type of validity measures how far the test covers the field of study under investigation or, in other words, how good are the items in the test representative of the totality of all items for that test. The validity of the content of the test is satisfactory when the sampling of items is wide and judicious, and when adequate number of standard questions are utilised. However, it is not possible to express content validity as a validity coefficient.

(*d*) ***Construct Validity.*** Construct validity requires study of test scores in relationship not only to variables that the test is intended to assess, but also to some of those variables that have no relationship to the trait under consideration. The approach includes predictions to other tests that are assumed to measure the same underlying trait as well as those that describe unrelated traits. This validity cannot be expressed in a single measure as the correlation between test scores and criterion scores. Validity in this case is demonstrated by showing that the predictions expected on the basis of theory may be confirmed by the test.

The characteristic of a group of people on which the test is validated affect the criterion-related validity. Difference between group of people on variables like sex, age and personality traits may affect the correlation coefficient between the test and the selected criteria. In a group having a narrower range of test scores, *i.e.*, in a more homogeneous group, the validity coefficient tends to be smaller. Since the size of a correlation coefficient is a function of two variables, a narrowing of the range of either the test variable or the criterion variable will tend to lower the validity coefficient.

8·4·3. Validity and Test Length. We have seen in § 8·3·5 that the homogeneous lengthening of a homogeneous test increases its reliability. Consequently, the lengthening of a test will increase its validity also since the more reliable a test, the more valid it is. The validity of a test increased n times is given by the following formula :

$$r_{nx.y} = \frac{n\, r_{x.y}}{\sqrt{n + n\,(n-1) r_{x.x}}} \qquad \ldots (8\cdot28)$$

where $r_{x.x}$ is the reliability coefficient of the test X and $r_{x.y}$ is the validity coefficient for predicting the criterion Y from the test X.

***Remarks* 1.** *Lengthening needed to achieve given Validity.* If we are interested in knowing how much homogeneous lengthening of a test is needed in other to achieve a desired level of validity, we have to solve (8·28) for n giving

$$1 + (n-1)\, r_{x.x} = \frac{n\, r^2_{x.y}}{r^2_{nx.y}} \quad \Rightarrow \quad 1 - r_{x.x} = n \left[\frac{r^2_{x.y}}{r^2_{nx.y}} - r_{x.x} \right]$$

$$\Rightarrow \quad n = \frac{1 - r_{x.x}}{\left[\dfrac{r^2_{x.y}}{r^2_{nx.y}} - r_{x.x} \right]} \qquad \ldots (8\cdot29)$$

2. *Validity for a Test of Infinite Length.* If n is very large, then on dividing numerator and denominator in (8·28) by n and taking the limit as $n \to \infty$, we get

$$r_{\infty x.y} = \frac{r_{x.y}}{\sqrt{r_{x.y}}} \qquad \ldots (8\cdot30)$$

$r_{\infty x.y}$ may be regarded as an *index of validity* and it is the maximum amount of validity that we can expect by lengthening the test. In particular, if the reliability coefficient of a test X is unity, *i.e.*, if $r_{x.x} = 1$ then from (8·30), we get

$$r_{\infty x.y} = r_{x.y}\,,$$

which implies that in this case lengthening of the test is not effective in increasing its validity. Hence it only pays to increase the length of a test to gain validity if the reliability of a test is far from being perfect.

3. *Validity for increased length of both Test (X) and Criterion (Y).* If the length of test X is increased n times and that of criterion Y is increased m times, then the validity of the lengthened test is given by :

$$r_{nx.my} = \frac{nm\, r_{x.y}}{[\{n + n(n-1)\, r_{x.x}\}\{m + m(m-1)\, r_{y.y}\}]^{1/2}} \qquad \ldots (8\cdot31)$$

In particular, if we take $m = 1$ in (8·31), we obtain the expression given in (8·28).

4. It can be seen, though less clearly, that as a consequence of lengthening a test its validity increases less rapidly than does its reliability.

8·4·4. Comparison between Reliability and Validity. 1. The reliability and validity are the essential properties of a test, both stressing the efficiency of the test. By reliability of a test we mean the consistency or stability of test scores when the test is administered to a number of individuals on different occasions. Validity of a test, on the other hand, is concerned with the accuracy of the test scores when compared with the performances on an independent standard test (criterion). The following example will clarify the difference between reliability and validity. Let us consider a clock which is quite good and accurate but is set ahead by, say, 10 minutes. Its time readings are reliable (consistent), but are not valid as judged by standard time.

2. To be valid, a test must be reliable. A test which is not quite reliable can hardly be valid since the test which correlates poorly with itself cannot correlate well with the measure of any other variable.

Index of reliability, $r_{1\infty} = \sqrt{r_{11}}$ *is sometimes taken as a measure of validity.* If reliability coefficient of a test is 0·81, then $r_{1\infty} = 0 \cdot 9$, which implies that the test measures true ability to the extent of 90%. Thus theoretically a reliable test is valid, though practically it may be invalid as judged by its correlations with various independent criteria as, for example, simple tapping tests or the word cancellation tests.

3. Since the correlation of a test with a criterion is limited by its own index of correlation which provides the maximum correlation the test is capable of a highly valid test cannot be unreliable.

8·5. INTELLIGENCE TESTS AND INTELLIGENCE QUOTIENT

Intelligence tests provide a basic tool or procedure for measuring the intelligence of an individual or group of individuals. Intelligence tests are administered and used

(*i*) for vocational guidance and selection,
(*ii*) for measuring intelligence of young children for grading, and
(*iii*) for detection and diagnosis of mental deficiency.

Intelligence tests, like most other tests in psychology and education, may be verbal or non-verbal, requiring an intelligent manipulation of ideas expressed in words and of objects respectively.

Before administering an intelligence test, we must first test its reliability and validity by applying the techniques discussed in earlier sections. Next step is to compute some *standard* or *norm* in order to assess an individual's score. It was in this context, that **Binet**, in the second edition of his *Intelligence-Test* (1908) introduced the concept of *mental age*. **L.M. Terman** used mental ages in reporting scores on *1916 Stanford Revision* of the *Binet-Simon intelligence scale.* An individual's mental age (M.A.) is the age at which an average person can score as much as the given individual. Thus, for example, M.A. of '*x*' means that an individual who obtains this score has the same mental ability as the typical or average '*x*' years-old child.

The concept of *Intelligence Quotient (I.Q.)* was introduced by **Stern** in 1912 and was first used practically by **Terman** in his 1916 edition of the Stanford Binet. I.Q. is defined as : *"Mental age (M.A.) in months expressed as percentage of the chronological age (C.A.) in months."* Thus

$$\text{I.Q.} = \frac{\text{Mental age}}{\text{Chronological age}} \times 100 = \frac{\text{M.A.}}{\text{C.A.}} \times 100 \qquad \ldots (8 \cdot 32)$$

This quotient will be an extremely useful index if the M.A. of individuals increases constantly as they grow older. But actually this is not so. According to **Terman** and **Merrill**, the increase in mental age begins to slow down typically at the age of 13 years and stops by the age of 16 years. Thus after 16 years as an individual grows older and older, the numerator in (8·32*a*) practically remains constant while the denominator goes on increasing constantly. Consequently I.Q. computed from formula (8·32*a*) will give distorted or bizarre picture of an adult's mental ability. *According to Terman and **Merrill**, during the age of 13 years to 16 years, for every three months increase in chronological age, there is only a two-month increase in mental age.*

I.Q. is regarded as an indicator of an individual's mental and intellectual development. Since M.A. of a child is always rising, I.Q. reflects how relatively fast or slow his development is. Stanford-Binet intelligence tests, if correctly administered and scored, give a reliable measure of child's mental ability.

The following description (in *Table 8·14*) is due to **Terman** and classifies an individual into various categories ranging from idiot to genius in terms of different levels of I.Q.

TABLE 8·14 : TERMAN'S CLASSIFICATION

Intelligence Quotient	*Category*
Below 20—25	Idiots
up to 50	Imbeciles*
50—70	Morons**
Below 70	Feeble Minded
70—80	Mental Deficiency
80—90	Dull
90—110	Average Intelligence
110—120	Superior Intelligence
120—140	Very Superior Intelligence
Above 140	Genius
200	Super-genius

TABLE 8·15 : HERRIL'S CLASSIFICATION

I.Q.	*Category*
140+	Very superior
120—139	Superior
110—119	High average
90—109	Normal
80—89	Low average
70—79	Border line
69 and below	Feeble minded

The classification due to Terman was revised in 1937 to make it more compact, less vivid and less specific. The revised distribution is due to *Merrill and is given in Table 8·15*.

* *Imbecile*. One whose defective mental state (from birth or an early age) does not amount to idiocy but who is incapable of managing his own affairs.

** *Moron*. Somewhat feeble minded person : One who remains throughout life at the mental age of eight to twelve.

***Remarks* 1.** Mental Ratio (M.R) is defined as : $\text{M.R.} = \frac{\text{M.A.}}{\text{C.A.}}$ ...(8·33)

and consequently $\text{I.Q.} = \text{M.R.} \times 100$ ... (8·33*a*)

2. If M.R. > 1, the individual is regarded as mentally advanced, if M.R. < 1, he/she is regarded as retarded and if M.R. =1, the individual is considered of average intelligence.

On analysis the data of the intelligence tests it has been seen that intelligence may be regarded to be normally distributed and that it depends on heredity also. As already stated (*Terman and Merrill*) intelligence grows with age up to 16 years, after which it remains steady. There is, however, no evidence of the relation between sex and intelligence.

3. It may be pointed out that *I.Q. does not give a direct proportional comparison*. For example, an individual, with I.Q. = 150, is not necessarily twice as capable and intelligent as an individual with I.Q. = 75 because I.Q. is not a per cent.