Spearman- Brown formula for computing test reliability having two halves is

$R_{xy} = \frac{2\times r_{xy}}{1+r_{xy}}$

Where,

R_{xy} = Stepped up reliability coefficient of first and second half

2 is the number of parts and

r_{xy} = The correlation coefficient between two parts X and Y.

If the score are expressed in the ranks then, correlation coefficient is calculated by

$r_{xy} = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$, Which is called Spearman's Rank correlation Coefficient?

If the scores are in the numeric scale then, correlation coefficient is calculated by,

$$r_{xy} = \frac{n\Sigma XY - \Sigma X \times \Sigma Y}{\sqrt{nY^2-(\Sigma Y)^2}\sqrt{n\Sigma X^2 -(\Sigma X)^2}}$$

Which is called Karl Pearson's coefficient of correlation, where, X represent the scores of the first half and Y the scores on the second set.

Example

A test-score is divided in two halves as the scores on the odd numbered questions and the scores on the even numbered questions. The correlation coefficient between them is obtained as 0.72, what is the reliability coefficient of the whole test.

Solution

Here, Correlation coefficient (r_{xy}) = 0.72

n = 2

Reliability coefficient, $R_{xy} = \frac{2\times r_{xy}}{1+ r_{xy}} = \frac{2\times 0.72}{1+0.72} = 0.8272$

The reliability coefficient is 82.72%; the dependability of the whole score seems to be very good.

d) Rational Equivalence Method (Kuder-Richardson Method)

Two forms of a test are defined as equivalence when corresponding items are interchangeable and inter item correlation is same for both forms. Kuder-Richardson method is the method of

obtaining reliability by using the internal consistency between the measures (questions) of the same scaling (test). The reliability coefficient for this method is obtained by the following two formulae. Kuder-Richardson's first formula for reliability is denoted by KR_1, is computed as

$$KR_1 = R_w = \frac{n}{n-1}\left[1 - \frac{\Sigma pq}{\sigma^2}\right]$$

Where, R_w = The reliability coefficient of the whole test

n - The number of items in the test

σ -Standard deviation of the test score

p-the proportion of the answering test item correctly

q = (1 - p)

Note 1: If the values of p for each test is equal then Σpq = npq

2: If the values of p for 'n' tests are $p_1, p_2 \ldots\ldots p_n$, then $\Sigma pq = p_1q_1 + p_2q_2 + \ldots\ldots\ldots\ldots + p_nq_n$

Example

In a test there are 60 questions. The proportion of answering each question correctly is 70%, if the standard deviation is 10 what is the reliability coefficient?

Solution

Number of questions (n) = 60,

Proportion of correct answers (p) = 0.7 q = 1 - p = 1 - 0.7 = 0.3

Σ pq = npq

S.D. (s) = 10, therefore, the reliability coefficient is given by

$$KR_1 = R_w = \frac{n}{n-1}\left[1 - \frac{\Sigma pq}{\sigma^2}\right]$$

$$KR_1 = \frac{60}{60-1}\left[1 - \frac{60\times0.7\times0.3}{\sigma^2}\right]$$

Hence reliability coefficient is 88.88%

Validity

Introduction

A scale possesses validity when it actually measures what it claims to measure. In other words, a scale is said to be valid if it measures what is expected to measure.

Interpretation of test scores ultimately involves predictions about a subject's behavior in a specified situation. If a test is an accurate predictor, it is said to have good validity. Before validity can be demonstrated, a test must first yield consistent, reliable measurements. In addition to reliability, psychologists recognize three main types of validity

Types of Validity

There are mainly the following types of validity:

1. **Content Validity**

Content validity is the representativeness or adequacy of the unit selected of the content such as the substances, the matter, the topic of the measuring instrument etc. Content Validity is also known as logical validity. A test has content validity if the sample of items in the test is representative of all the relevant items that might have been used. Words included in a spelling test, for example, should cover a wide range of difficulty.

It is related to the objective of the study, it tests the objective of the course of action.

Let, U be the universe of the item; S is the subset of U i.e., $S \subset U$ and x be the item such that x is the element in U then it must be an element of S, i.e., $x \in U \Rightarrow x \in S$

The standardized achievement test is used for the content validity measure.

In Psychometrics, Content Validity refers to the extent to which a measure represents all facets of a given social concept. For example, a depression scale may lack content validity if it only assesses the affective dimension of depression but fails to take into account the behavioral dimension.

Content validity is related to face validity, although content validity requires more rigorous statistical tests than face validity, which only requires an intuitive judgment. &intent validity is most often addressed in academic and vocational testing, where test items need to reflect the knowledge actually required for a given topic area or job skill. In clinical settings, content validity refers to the correspondence between test items and the symptom content of a syndrome.

2. **Criterion-related Validity**

Criterion-related validity refers to a test's accuracy in specifying a future or concurrent outcome. A common approach, called criterion related validity is to correlate measures with a criterion measure known to be valid. for example, an art-aptitude test has predictive validity if high scores are achieved by those who later do well in art school. The concurrent validity of a new intelligence test may be demonstrated if its scores correlate closely with those of an already well-established test.

The criterion related validity is based on the four decisive factors: (i) an external criterion (ii) regular and future behavior (iii) logical analysis and (iv)empirical method.

The criterion-related validity is related to the ability to predict some outcomes or estimate the existence of some current conditions.

• It is used to predict the criterion on the basis of some measure.

• It reflects the success of measures for empirical estimating process

In this type of validity, the proposed criterion must possess the following quality:

- Freedom from bias (criterion should give each subject/matter an equal opportunity to score)

- Reliable (criterion should be stable and reproducible)

- Relevance (criterion should be defined in terms of proper measure). Availability (information specified by the criterion must be available)

The Criterion- related Validity broadly classified as: (a) Predictive validity and (b) Concurrent Validity

(a) Predictive validity:

It refers to the usefulness of the test in prediction some future performances on the criterion. It is concerned with how well the scale can forecast a future criterion. When the criterion measure is collected later the goal is to establish it is called the predictive validity.

(b) Concurrent Validity:

It refers to the usefulness of a test in closely relating to the other measure of known validity. It is concerned with the performances that how it can describe a present criterion. When the criterion measure is collected at the same time as the measure being validated the goal is to establish concurrent validity. Concurrent Validity is demonstrated where a test correlates well with a measure that has previously been validated. For example, if a test measuring job satisfaction gives similar results to those gathered using a job satisfaction which has been validated in past investigations the new measurement has concurrent validity.

The validity coefficient is measured in terms of the correlation coefficient (r) of the scores of the different tests. We say, for $0.9 \leq$ r ≤ 1 there is very high validity; for $0.8 \leq r \leq$ 0.9 a high validity; for $0.6 \leq r \leq$ 0.8 a satisfactory validity: for $0.4 \leq r \leq 0.6$ a moderate validity; for $0.0 \leq r \leq 0.4$ a poor validity; and for $r < 0.0$ a negative validity

3. Construct Validity In social science, construct validity refers to whether a scale measures the unobservable social construct that it claims to measure. It is the validity, most complex and abstract because of the complexity of the social parameter. A scale is said to posses construct

validity to the degree that it confirms to predicted association with other theoretical postulates. The essence of construct validity is its dependence on theory and the examination of observed association is a test of theory as valid scale. It is based on psychological trait and quality Construct validity is generally determined by investigating what psychological traits or qualities a test measures; that is, by demonstrating that certain patterns of human behavior account to some degree for performance on the test. A test measuring the trait "need for achievement," for instance, might be shown to predict that high scorers work more independently, persist longer on problem-solving tasks and do better in competitive situations than low scorers.

A construct is not restricted to one set of observable indicators or attributes. It is common to a number of sets of indicators. Thus, construct validity can be evaluated by statistical methods that show whether or not a common factor can be shown to exist underlying several measurements using different observable indicators. For determining construct validity, we associate a set of other proposition with the result received from the use of our measuring instrument. If measurements on our devised scale correlated (associated) in a predicted way with the other propositions, we can conclude that there is construct validity.

Scaling

Scores and Scales

Scores

The number of points somebody gets for correct answers in a test is said to be scores. In other words the value of parameter in the observed phenomenon is termed as score. From an experiment or from a test what we obtained as the observation is called the raw score. The raw score is the simple numerical count of responses such as the number of correct answers on an intelligence test. The usefulness of the raw score is limited however, because it does not convey how well someone does in comparison with others taking the same test. Suppose, one attempts to answer"20 IQs having 1 point each and answered 16 correctly then the raw score is 16. One tries to measure the length of 5 pencils and found to be the lengths of 10, 11, 10, 14, 12 in cm; then the raw scores of the measuring phenomenon are 10, 11, 10, 14, and 12 cm.

Scale

Scale is a predefined sequence of scores in ascending values that can map an item to it. Scale is a set of all the different levels of symbols or numerals or something so constructed, from the lowest to highest, that these can be assigned by rule to objects or to items or to the individuals or to their behavior to whom it is applied. In general concept, scale is also known as a quantifying appliance used to indicate the systematized numerals of the measuring instrument.

Scaling of the Scores

From a set of scores of the test we can construct a sequence of levels of the values that can be used as an extent for the test of that phenomena, this method of leveling is said to be scaling. For the purpose of scaling it is always desired to make the scores in an array. After that the scores are converted to the percentile points and then to a scale of required form.

The raw scores obtained in test can be converted to different auxiliary scores in relation to the distribution of the raw scores or according to the distribution of parent population. Such scores which are modified/improved/developed from the raw scores are called derived scores. From the derived scores of the same form in sequence can be used to create a continuous structure of the numerals which is the scale required to be constructed.

There are different types of derived scores widely used in the measurement of the phenomenon or the attitudes. The percentile scores, s-score (z-score) and T-score are such derived scores and are use to compare the strength and credibility of the measures. From these scores we can construct the standard scales namely percentile scale, sigma scale (z-scale) and T scale, respectively by arranging the scores lowest to highest.

Difficulties in scaling

In social phenomenon, following are the reasons that create difficulties of scaling in social sciences.

- Social complexity (intricacy) [social phenomena are complex and such complexity cannot be measured]
- Abstractness (nonfigurative) of the social phenomena
- Heterogeneity of the social values, customs and norms.
- Changing nature of human behavior Absence of universal measuring of social values
- Laboratory method cannot be applied in social phenomena.

Scales used in Social and Physical sciences

The main scales used in the measure of social/physical characteristics are:

(i) Point scale
(ii) (Social distance scale
(iii) Rating scale
(iv) Ranking scale
(v) Thurstone scale

(i) Point scale

In this type of scale words or situations representing the criteria are selected and one point (marks or number) is given for each criteria. Attitude of a person can determined by the use of all the three following method effectively.

Method 1: The respondent is asked to tick one that is representing or favorable to him /her. The scores are counted and result is derived**.**

Method 2: In the second method the respondent is asked to cross the one point or situation which is not favorable to him /her. A point is given to each and every word that has not been crossed. The attitude of the respondent is then determined by counting no of points.

Method 3: In the third method of point scale the respondent is asked to cross, on which points he /she is agree or not.

(ii) Social distance scale

The social distance scale is developed by Emary S. Borgadus to measure the social distances (it is commonly known as Borgadus scale). The social distance may be defined as the proximity and favoritism; for example the cultural distance from one race to the other, custom from one ethnic group to the other etc. To measure a person's (respondent's) attitude how far from the given cause situation the Borgadus scale can be used. Borgadus developed a scale to measure the nearness of liking between two social groups using several items or statements which show the varying relationship of social distance of Americans with other races as English, Korean, Swedish and poles.

(iii) ***Rating scale***

When the character to be measured is not dichotomous in this case the rating scale is used. Rating scale consists of a set of figures that can match to the individual or items to be measured. The response or the opinions of the respondent's attitude is rated in three to six points in continuum (range). The intensity of the attitude is measured by using equal or unequal type intervals. An example of three point rating scale is:

Very good .....satisfactory.....poor

A five points rating scale is

Strongly-Agree.....Agree .....Not-Decided. .....Dis-agree.....Strongly-Disagree

(iv) Ranking scale

The ranking scale is similar to rating scale applied to a set of objects or individuals with the preference or liking. In this scale the situations are placed in such a way that, everybody who inspects it knows that one likes the one better than the other. Ranking scales is determined in comparison to a few cases known as stimuli. The item

obtaining first preference scores 1, the second as 2, third as 3 and so on. 'The smaller the score the greater the preferences' is the principle of ranking scale.

(v) Thurstone scale

American psychologist Louis L. Thurstone proposed that intelligence was not one general factor but a small set of independent factors of equal importance. He called these factors primary mental abilities. To identify these abilities, he developed a plan to conduct study amongst 250 college students, identified factors and developed a scale of measuring aptitude using factor analysis. The scale so developed is known as Thurstone scale. In educational and psychological experiments, it is used as a main type of scale used to measure the attitude. The statements are collected and arranged in continuum from most favorable to least favorable with neutral point (zero). It is one type of point scale having neutrality point at the central location.

Most- Favored.....neutral .....Least— Favored