The course code is PSC106, the title of the unit is reliability and validity and the name of the module is Reliability estimates: Kuder Richardson Formulas and Inter-scorer reliability.

I am Dr. Michelle Fernandes from Carmel College of Arts, Science and Commerce for Women.

In this module we'll be studying the methods to estimate internal consistency, that is the Kuder Richardson formulas and Coefficient alpha. The measures of inter-scorer reliability and how do you use.

By the end of this module, you should be able to differentiate between the Kuder Richardson-20 formula and the coefficient alpha formula. Understand how they are used to estimate internal consistency. You should be able to understand the measures of inter scorer reliability and how do you use and interpret a coefficient of reliability.

This is a continuation of the last module, where we spoke about the various reliability estimates. We spoke about split-half reliability test, test and re-test reliability and parallel forms.

Now this is another method wherein you're testing the internal consistency of items. So, in addition to the Spearman brown formula, you can use other methods to measure the internal consistency of reliability, and that would be the Kuder Richardson's formulas and the Cronbach's alpha formulas.

Now, internal item consistency refers to the degree of correlation among all the items on a scale. A measure of inter item consistency is calculated from a single administration of a single form of a test.

So, unlike the test -retest, unlike the parallel form, there's only a single administration and a single form of the test. An index of inter-item consistency, in turn, is useful in assessing the homogeneity of a test. Now a test is set to be homogeneous if they contain items that measure a single trait.

Although it is desirable for tests to be homogeneous because it so readily lends itself to a clearer interpretation, it is often an insufficient tool for measuring multifaceted psychological variables such as intelligence or personality.

Now the Kuder- Richardson's formulas came about because there was a lot of dissatisfaction that was there in among the psychologists due to the split-half methods of estimating reliability and this compelled to researchers Frederic Kuder and M.W Richardson to develop their own methods of estimating reliability.

The most widely known of the many formulas they collaborated on is the Kuder-Richardson Formula 20, also known as the KR 20, and it was so named because it was the 20th formula developed in the series.

Where test items are highly homogeneous, KR-20 and split-half reliability estimates will be similar. The KR-20 is used for determining the inter-item consistency of dichotomus items, primarily those items that can be scored right or wrong such as multiple-choice items. If test items are more heterogeneous, KR-20 will yield lower reliability estimates than the split-half method.

This formula is the KR-20

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

where r_{KR20} stands for the Kuder–Richardson formula 20 reliability coefficient, k is the number of test items, σ^2 is the variance of total test scores, p is the proportion of test takers who pass the item, q is the proportion of people who fail the item, and Σ pq is the sum of the pq products over all items.

An approximation of KR-20 can be obtained by the use of the 21st formula in the series developed by Kuder and Richardson, a formula known as—you guessed it—KR-21. The KR-21 formula may be used if there is reason to assume that all the test items have approximately the same degree of difficulty. Formula KR-21 has become outdated in an

era of calculators and computers. Way back when, KR-21 was sometimes used to estimate KR-20 only because it required many fewer calculations.

Now there are numerous modifications of Kuder–Richardson formulas have been proposed through the years. The one variant of the KR-20 formula that has received the most acceptance and is in widest use today is a statistic called coefficient alpha. You may even hear it referred to as coefficient α−20. This expression incorporates both the Greek letter alpha (α) and the number 20, the latter a reference to KR-20.

Coefficient alpha Developed by Cronbach and subsequently elaborated on by others, coefficient alpha may be thought of as the mean of all possible split-half correlations, corrected by the Spearman–Brown formula. In contrast to KR-20, which is appropriately used only on tests with dichotomous items, coefficient alpha is appropriate for use on tests containing non dichotomous items.

The formula for coefficient alpha is $r_\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\Sigma\sigma_i^2}{\sigma^2}\right)$

where r_α is coefficient alpha, k is the number of items, is the variance of one item, Σ is the sum of variances of each item, and σ2 is the variance of the total test scores.

Coefficient alpha is the preferred statistic for obtaining an estimate of internal consistency reliability. A variation of the formula has been developed for use in obtaining an estimate of test-retest reliability. Essentially, this formula yields an estimate of the mean of all possible test-retest, split-half coefficients. Coefficient alpha is widely used as a measure of reliability, in part because it requires only one administration of the test.

Unlike a Pearson r, which may range in value from −1 to +1, coefficient alpha typically ranges in value from 0 to 1. The reason for this is that, conceptually, coefficient alpha is calculated to help answer questions about how similar sets of data are. Here, similarity is gauged, in essence, on a scale from 0 where there is absolutely no similarity to 1 where it is perfectly identical.

The various measures of inter-scorer reliability, often referred to as scorer reliability, judge reliability, observer reliability, and inter-rater reliability. This is basically the degree of agreement or consistency between two or more scorers or judges or raters with regard to a particular measure.

Now, if the If the reliability coefficient is high, the prospective test user knows that test scores can be derived in a systematic, consistent way by various scorers with sufficient training.

Inter-scorer reliability is often used when coding nonverbal behavior. For example, a researcher who wishes to quantify some aspect of nonverbal behavior, such as depressed mood, would start by composing a checklist of behaviors that constitute depressed mood such as looking downward and moving slowly and so on and so forth.

The simplest way of determining the degree of consistency among scorers in the scoring of a test is to calculate a coefficient of correlation. This correlation coefficient is referred to as a coefficient of inter-scorer reliability.

Now, how do you use and interpret the coefficient of reliability? One frequently asked question would be how high should be the coefficient of reliability be? Reliability is a mandatory attribute in all tests we use. Now, if a test score carries with it life-or-death implications, then the test has to be a very high reliability. If the test does not have life or death implications, then it can have a little lower reliability.

As a rule of thumb, it may be useful to think of reliability coefficients in a way that parallels many grading systems: In the .90s rates a grade of A, if it falls in the range of .80s then it can be rated as a B and so on and so forth. If the scores range from .65 through the .70s then these are reliability coefficients which will not be acceptable.

You may be wondering what is the purpose of reliability coefficient? why do we need to have the reliability coefficients in a psychological test?

Now if a specific test of employee performance is designed for use at various times over the course of the employment period, it would be reasonable to expect the test to demonstrate reliability across time. It would thus be desirable to have an estimate of the instrument's test-retest reliability. For a test designed for a single administration only, an estimate of internal consistency would be the reliability measure of choice.

If the purpose of determining reliability is to break down the error variance into its parts, like scorer error, administration error and so on, then a number of reliability coefficients would have to be calculated. The various reliability coefficients do not all reflect the same sources of error variance.

Thus, an individual reliability coefficient may provide an index of error from test construction, test administration, or test scoring and interpretation. A coefficient of inter-rater reliability, for example, provides information about error as a result of test scoring. Specifically, it can be used to answer questions about how consistently two scorers score the same test items.

With this, we complete the concept of reliability over the course of the three modules. You have studied, the concept of reliability, the various sources of reliability. We also studied how to measure what are the various estimates of reliability? What was the purpose, and how do you determine whether a test has high reliability?

I hope you have understood this module.

These are the books that you can refer to.

I wish you all the very best.

Thank you, students.