Quadrant II – Transcript and Related Materials

Programme: Bachelor of Arts (Third Year)

Subject: Psychology

Course Code: PSC106

Course Title: Psychological Testing

Unit: 03 (Reliability and Validity)

Module Name: Reliability estimates: Kuder Richardson Formulas and Inter-scorer reliability

Name of the Presenter: Michelle Fernandes (Ph.D)

Notes

Other Methods of Estimating Internal Consistency

In addition to the Spearman–Brown formula, other methods used to obtain estimates of internal consistency reliability include formulas developed by Kuder and Richardson (1937) and Cronbach (1951). **Inter-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. An index of inter-item consistency, in turn, is useful in assessing the **homogeneity** of the test. Tests are said to be homogeneous if they contain items that measure a single trait.

The more homogeneous a test is, the more inter-item consistency it can be expected to have. Because a homogeneous test sample a relatively narrow content area, it is to be expected to contain more inter-item consistency than a heterogeneous test. Test homogeneity is desirable because it allows relatively straightforward test-score interpretation. Testtakers with the same score on a homogeneous test probably have similar abilities in the area tested. Testtakers with the same score on a more heterogeneous test may have quite different abilities.

The Kuder–Richardson formulas Dissatisfaction with existing split-half methods of estimating reliability compelled G. Frederic Kuder and M. W. Richardson (1937; Richardson & Kuder, 1939) to develop their own measures for estimating reliability. The most widely known of the many formulas they collaborated on is their **Kuder–Richardson formula 20,** or KR-20, so named because it was the 20th formula developed in a series. Where test items are highly homogeneous, KR-20 and split-half reliability estimates will be similar. However, KR-20 is the statistic of choice for determining the inter-item consistency of dichotomous 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.

The following formula may be used:

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

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 testtakers who pass the item, q is the proportion of people who fail the item, and $\Sigma\, pq$ is the sum of the pq products over all items. For this particular example, k equals 18. Based on the data in Table 5–3, Σpq can be computed to be 3.975. The variance of total test scores is 5.26. Thus, r_{KR20} = .259.

An approximation of KR-20 can be obtained by the use of the 21st formula in the series developed by Kuder and Richardson. 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.

Coefficient alpha Developed by Cronbach (1951) and subsequently elaborated on by others (such as Kaiser & Michael, 1975; Novick & Lewis, 1967), **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 nondichotomous 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. 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.

Measures of Inter-Scorer Reliability

Variously referred to as *scorer reliability, judge reliability, observer reliability,* and *inter-rater reliability,* **inter-scorer reliability** is the degree of agreement or consistency between two or more scorers (or judges or raters) with regard to a particular measure.

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). Accordingly, each subject would be given a depressed mood score by a rater. Researchers try to guard against such ratings being products of the rater's individual biases or idiosyncrasies in judgment. This can be accomplished by having at least one other individual observe and rate the same behaviors. If consensus can be demonstrated in the ratings, the researchers can be more confident regarding the accuracy of the ratings and their conformity with the established rating system.

Using and Interpreting a Coefficient of Reliability

"How high should the coefficient of reliability be?" Perhaps the best "short answer" to this question is: "On a continuum relative to the purpose and importance of the decisions to be made on the basis of scores on the test." Reliability is a mandatory attribute in all tests we use. However, we need more of it in some tests, and we will admittedly allow for less of it in

others. If a test score carries with it life-or-death implications, then we need to hold that test to some high standards—including relatively high standards with regard to coefficients of reliability. If a test score is routinely used in combination with many other test scores and typically accounts for only a small part of the decision process, that test will not be held to the highest standards of 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 (with a value of .95 higher for the most important types of decisions), in the .80s rates a B (with below .85 being a clear B−), and anywhere from .65 through the .70s rates a weak, "barely passing" grade that borders on failing (and unacceptable).

The Purpose of the Reliability Coefficient

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.

References:

Cohen, R. J., & Swerdlik, M. E. (2018). Psychological testing and assessment: An introduction to tests and measurement. (9th ed.). New Delhi: McGraw-Hill Education.