Course work on reliability and validity

War, Intelligence

Reliability and Validity

I. Reliability of a Test

The reliability of a test is defined as its degree of consistency, stability, accuracy and predictability hence the repeatability of its scores. Generally, it refers to the extent to which the ultimate results of a given test are consistent over time thereby presenting a degree of accuracy representative of the total population considered under the study. Reliability is quantified in terms of reliability coefficient. The reliability coefficient give insight on the significance of the scores (Waltz, Strickland, & Lenz, 2010). A high reliability coeffiecient implies that the findings are considerably plausible, while a low reliability coefficient implies that the findings are less plausible (Waltz, Strickland, & Lenz, 2010). By calculating the reliability coefficient, the reliability of various psychological tests, for instance, Stanford Binet and the California Achievement Test can be asserted.

A. Types

Rubin and Babbie (2009) identify four types of reliability including inter-observer reliability also known as Inter-rater, Parallel-forms, Internal Consistency and Test-retest Reliabilities. Inter-rater/Inter-Observer reliability refers to the variation in measurements taken by different researchers but using the same methods and/or instruments. Researches involving humans as part of the measuring procedure often results into variations in consistency since humans are innately distractible, unpredictable and inconsistent over repeated tasks. This type of reliability can be estimated by calculating the percentage of agreement between the raters involved.

Comparatively, the Test-Retest reliability refers to the variation in measurements taken by a single method/instruments or person on the same item under standardized conditions, (Craig & Nemeroff, 2002). This is perhaps the most common type of reliability in most psychological tests. The Stanford Binet and the California Achievement Test are in most case tested and retested to aver initial finds as a means of testing for consistency. For Stanford Binet test, the findings can further be asserted by employing the split-half or the interscorer method. For California Achievement Test, a non-referenced test, in which results from selected individuals are use to affirm the reliability of the results obtained from the takers of the test, is in most cases preferred.

The reliability type is, therefore, estimated when same tests are administered to the same sample on two different occasions with the time between the two measurements being of critical importance. Internal Consistency reliability, however, assesses the degree of consistency in results across items within the same test, (Hambleton & Swaminathan, 1985). A single measurement instrument is administered to a group of people in one test with the motive of estimating reliability. Finally, the Parallel-Forms reliability assesses the consistency of the results of two tests that are constructed in the same way from comparable content domain. Forms used within this reliability type are often not only considered equivalent measures, but independent of each other, as well.

B. Validity and its Various Forms

Correspondingly, a test's validity refers to the extent to which the evidence and the theory support the interpretations of a test score with respect to its intended purpose or simply, the extent to which a test correctly measures the construct it claims to or it is designed to measure, (Hopkins, 2000). It is, therefore, a determination of how well a given test measures what it claims or is designed to measure. Five types of evidence inclusive of; the content, response processes, internal structure, relation to other variable and consequences of testing are critical in determining a test's validity. There exist multiple types of validity including content, criterion, face and construct validity. Face validity, measured either on a single item or all items of a test, indicates how well the items reveal the meaning or the purpose of the test or the test itself, (Rupert, Kozlowski, Hoffman, Daniels, & Piette, 2001). Face validity, perhaps the most common type of validity, is normally verified in light of the construct questions (McDavid & Hawthorn, 2006). This can be undertaken by an untrained personnel inclusive of the test takers, consumers among others.

Content validity, on the other hand, refers to the adequacy of sampling content across the trait or construct under investigation, (Craig & Nemeroff, 2002). Based upon the existing literature of a particular trait, a test is only content valid if all aspects of the given concept are represented within the test. In most cases, trained personnel institute content validity of a research by paying close attention to the content by means of a subjective measure that cannot be quantified statistically.

Criterion Validity refers to the relation between scores in a test with the external measure of performance. The validity type is simply estimated theoretically by relating two or more measures and assessing their relation by determining their correlation coefficient, (Halligan & Bouckaert, 2008). Also known as empirical validity, Criterion validity makes use of objectives and empirical evidence to assert the validity of measures, measures that are in most cases demonstarated though two strategies; predictive and concurrent (Ruane, 2005). Just as their names suggest, predictive and concurrent strategies/validity test whether research findings can be used to predict other related expectations and the concurrence of the findings with already documented findings respectively (Ruane, 2005).

Finally, Construct validity demonstrates the extent to which a test captures a given theoretical trait or constructs and how it overlaps with other aspects of validity. The measure is founded on the fact that, nearly all tests are based upon a given conceptual framework or theory that clearly delineates the meaning of the construct and the relationship with other variables.

II. Measuring Reliability and Validity

Many different approaches and methods are often combined to present the overall picture of a test’s validity and reliability. Since the different general classes of reliability estimates/measure reliability differently, the measuring procedure vary. To determine the Inter-Observer/Inter-Rater reliability, a researcher can divide measurements into categories then assign raters which check off which category an observation falls. The researcher can thereafter determine the level of agreement between individual raters by calculating the percentages. For example, in a case of 50 observations rated by two different raters and involving three categories, for each observation a rater could check any of the three categories. Suppose out of the 50 observations, 42 were checked within the same category, then the percentage of agreement becomes 84%, (Hopkins, 2000).

Test-Retest reliability is measured based upon the assumption that there exists no substantial change in the construct under observation during both occasion. Generally, scores by either the same person or results from a given observation are compared with those obtained at a different time. The reliability coefficient is the then calculated by correlating the scores thereby determining the extent to which generalizations can be made on the test scores in varying situations. As noted above, when estimating the Parallel-Forms Reliability, parallel forms must be created.

The process, therefore, entails the creation of a given set of questions, randomly dividing them into sets, then administering the sets as stipulated by the test. The correlation arising between the two parallel forms is, therefore, used to estimate reliability. Researchers can also measure Internal Consistency reliability by determining the average inter-item correlation. This is achieved by using all items within the instrument to measure the same construct, computing the correlation between every pair of items, then averaging the correlation. For example, suppose we have six items in the instrument, then a total of 15 different pairs of items or correlations will result. Averaging the interim correlation or simply determining the mean will result to the reliability coefficient, (Rupert et al., 2001).

Validity and Reliability Coefficients

Comparatively, varying forms of validity are measured differently, with the simplest being face validity, which simply measures, how well an item reveal either the meaning or the purpose of the test or the item itself. For example, given a test item, ‘ I have been thinking of killing myself’ a researcher can infer that the item measures suicidal ideation. Content validity is measured by checking the operationalization against the content domain of a given construct, (Rupert et al., 2001). The approach is based on the assumption that the researcher has clear, detailed description of and information about the given content domain. Criterion related validity is determined by checking the performance of the researcher’s operationalization against pre-set criterion. For example, suppose the researcher’s criterion is prediction, then predictive validity is the assessment of the operationalization ability to theoretically predict a given phenomena.

Good Level of Reliability and Validity

The good level of both validity and reliability are placed at 90% reference range of the difference between a given subject’s value for the criterion and the practical measures. However, estimates of 80% or . 80 are regarded as moderate to high while estimates below . 60 or 60% are regarded as unacceptably low. Exemplification is a test designed to estimate a Bod Pod, a test measuring body fat using both the DEXA scan and X-ray absorptiometry. Since both the test express the units of body fat as a percentage of the body mass with the criterion set that 90% of the test subjects would have a body fat of between 2. 0 and 5. 4% BF, their results can be compared against each other while thereby estimating both reliability and validity.

III. Validity versus Reliability

Although Validity and reliability remains two most influential and fundamental components of any test, the former is more salient than the latter. Craig and Nemeroff (2002) argue that, a reliable instrument that does not measure what it purports to measure is essentially worthless. Within every attribute, it is often critical to make a distinction between the attribute being measured and the variables that are directly used in the process, (Martin, 2011). Majority of researchers affirm that, a test can have validity without reliability, however, a valid test is almost always reliable since, in order to be reliable, a test must be valid in the first place.

An illustration is the case in which a new theoretical test on intelligence is based on the assumption that ‘ the larger a human’s brain is, the more intelligent he or she is’. The assumption implies that, the larger one’s brain is corresponding to the larger one’s head is and, therefore, the levels of intelligence can be measured by estimating the circumference of peoples’ head. The test is evidently, invalid since comparing the head of a person to his or her level of intelligence is absurd. However, how reliable is the test? A plethora of researchers argue that, the test is reliable since reliability is not a measure of truthfulness, but a measure of consistency.

The test is reliable since it is consistent owing to the fact that, each individual’s score will either be the same or extremely close to his or her other scores within other time periods. The test is, therefore, reliable without being valid but can the test be valid without being reliable? In reverse, the participants would have similar scores every time they took the test since intelligence does not often change. Then the test would be valid, and since intelligence never varies significantly over time, it would be reliable, therefore, proving that a test that is valid is necessarily reliable.

Validity is, therefore, more important since a measuring instrument must be both valid and reasonably reliable. Although the validity of a test is undermined when it is not reliable, an invalid test is not useful at all. Bouckaert and Halligan compares the two noting that, whereas reliability is only pertinent, validity is critical, (Knapp, 2009). Validity is critical since if a test is valid, the resulting scores are able to predict significant kinds of performance. Psychometric test experts generally agree that a test can be reliable, that is stable and consistent, but not valid; however, a test cannot be valid without being reliable.

Construct Validity: Most Important?

Comparing the four forms of validity, construct validity is considered to be the most important type of validity. The validity type assesses the extent to which a given measuring instrument accurately measures the theoretical construct or the trait it is designed to measure. It, therefore, forms the basis of a given test since it relates theory to practical and is the overall validity or the extent to which a test actually serves its purpose. Examples of the theoretical construct or traits often estimated by construct validity, which includes verbal fluency, depression, intelligence, anxiety, neuroticism, and scholastic aptitude.

Construct validity’s purpose of correlating performance of a given test with the performance on the established test or with test subjects who have varying levels of traits the test claims to measure is the most critical. Others' role are relatively less significant; for example, content validity plays the role of enabling experts to assess the test so as to establish whether the items are representative of what is being measured while face correlates performance on a test with a concurrent behaviour. Criterion also only serves to correlate the performance construct and the future behaviour.

IV. A Psychological Test: Reliability and Validity

Psychologists should not use a test that cannot be proven to have both validity and reliability since these forms the essential elements in determining the quality of a test. Although there exists other elements of testing standards such as errors of measurement, validity and reliability are the backbone during the construction, evaluation and documentation of a given test. Besides, psychology professionals and practitioner associations have developed standards that determine the overall judgement about the relevance, quality and admissibility of a given test. A test lacking on the two critical aspects would be lacking in fairness and would further compromise the rights and responsibilities of the research subjects, (Halligan & Bouckaert, 2008).

Barlow and Durand (2004) contend that, a psychological test must be scientifically consistent since the majority of decisions made from the tests have lifelong implications. Standardization achieved by checking a test's reliability is necessary since the tests have multiple applications; in legal matters, school issues and disability issues to name but a few. Besides, for individuals undergoing the particular psychological test such as; fire fighters, law enforcement officers, medical/psychology experts, airline pilots, and nuclear power facility workers, the risks are often too great to administer a test giving false results as a result of compromised validity and reliability standards.

V. Ensuring Adequate levels of Validity and Reliability

Prior to the use of a test, a psychologist must first ensure that he or she is fully aware of the professional and ethical standards of excellent practice that would affect the way and the process of administering the test. The psychologist must also ensure that he/she possesses the relevant knowledge, skills and understanding related to the testing process. Prior to testing, the professional must check whether the tests are supported by evidence of both reliability and validity for their intended purpose, (Rupert, Kozlowski, Hoffman, Daniels, and Piette, 2001). Besides, evidence that support the probable inferences that may be drawn from the scores of the test must be provided. The psychologist must also ensure that the evidence is accessible to both the test user and the available independent scrutiny and evaluation, (Friesen, 2010).

The psychologist must also consider the various social, institutional, cultural, political and linguistic differences that exist between varying assessment testing. The laws of the country the testing is taking place, and the differences relating to individual versus groups assessment must also be put into consideration. The test setting, whether clinical, educational or work related and the primary recipients of the test results inclusive of the test takers, test developers, employers or the parents or guardian must also be put into considerations. Of critical importance is the individual’s attributes inclusive of his or her socioeconomic status, race, language, gender and educational background. In most cases, the gender role socialization influences the identities of men and women thereby defining their behaviour. The differences must, therefore, be put into consideration as they would result into varying test results, (Reis, Duobse, Ainsworth, Macera & Yore, 2005).

VI. Ethical, Legal and Confidentiality Issues

The psychologist should provide clients with information about the nature of the test, rights, risks and the required obligations of him or her, (Thompson, 2004). The psychologist must not endorse or lend credence to any inappropriate use or interpretation of given assessment results besides; serious concerns arise on the compromise of the effective use of a given psychological test. Therefore, a given practitioner is not allowed to misuse the test results by either publishing or disclosing the contents to unauthorised or unqualified persons, (Meincke, 1999). Ethically, psychologists must neither allow nor permit testing by unqualified persons through employment, sponsorship or supervision, (APA, 2011; Rubin & Babbie, 2007).

The practitioner is further legally bound to protect the identity of his or her clients besides the law sets the limits of privileged communication, which must be given to the client. Generally, access to psychological test protocols and results must be under the control of the psychologist besides test documents in most cases are deemed exempt documents even in cases of the public interest especially when; disclosure may invalidate the utility of the test results and impair the ability of the psychologists to properly perform their duties, (APA, 2011). Generally, unauthorized disclosure constitutes a breach of the contractual arrangement between the psychologist and the client.

VII. Psychological Tests as Applicable to Different Settings

Psychological tests are administered in variant settings with some of the most common ones being; the intelligence tests administered educational settings, occupational tests administered in employment settings and aptitude tests which attempt to measure certain type of achieved knowledge. Psychological tests are used to measure intelligence with the most common one being the IQ test, which originated in about 1916 when Lewis Terman, a Stanford University psychologist, revised the intelligence scale that had been designed by Alfred Binet and Theodore Simon, (Devlin, Daniels & Roeder, 1997).

The test, which compares an individual’s mental age to that of his or her chronological age, is constructed with the median score of 100 and a standard deviation of 15. Research indicates that IQ tests have high levels of statistical reliability implying that the majority of test-takers often have uniform scores irrespective of their age or the time the tests are taken, (Devlin, Daniels & Roeder, 1997; Flynn, 1984)). The test is further regarded as having sufficient statistical validity owing to the multiple methods such as; visual, verbal, abstract reasoning problems, spatial imagery, memory, general knowledge and vocabulary.

Aptitude tests are usually used within the educational or employment settings and are designed to measure how much an individual know about a given subject or the capacity one has to master material in a particular area. In most cases, aptitude tests reveal a high degree of internal consistency reliability besides, the content validity when measuring knowledge or job competencies are often. Barlow and Durand (2004) also note that, convergent validity of the majority of aptitude tests is high since they often measure similar traits or constructs.

Comparatively, occupational tests are designed for population based wide-scale surveillance of occupational suitability. The test serves to match individual interests to the interests of other occupying other careers and is based on the assumption that, the interest of humans often matches to the things people do or say, (Knapp, 2009). Budding professionals are compared to others who have worked in a broad range of occupations. Test retest reliability coefficient often lies between . 55 and . 91 dependent upon the specific occupation, for example, the coefficients of heavy labor work stood at . 71, (Paterson & Uys, 2005). Generally, the test-retest reliability and validity of occupational, psychological tests are often predictive of the general occupational atmosphere within a given region, age group, or social status, (Passalacqua & Cervantes, 2009).

References

APA, (2011). Ethical Principles of Psychologists and Code of Conduct: 2010 Amendments.

Web. http://www. apa. org/ethics/code/index. aspx

APA, (2011). Rights and Responsibilities of Test Takers: Guidelines and

Expectations. Web.

http://www. apa. org/science/programs/testing/rights. aspx#

Barlow, D. H. and Durand, V. M. (2004) Abnormal Psychology: An Intergrative Approach.

New Jersey: Thomson Publishers.

Craig head, E. W. and Nemeroff, C. B. (2002). The concise Corsini encyclopedia of

Psychology and behavioral science. New York: Wiley Publishers.

Devlin, B.; Daniels, M.; Roeder, K. (1997). “ The Heritability of IQ.” Nature, 388(6641): 468-71.

Flynn, J. R. (1984). " The mean IQ of Americans: Massive gains 1932 to 1978.”

Psychological Bulletin 95 (1): 29–51.

Friesen, B. K. (2010) Designing and Conducting Your First Interview Project. New York:

John Willey and Sons.

Hambleton, R. K., & Swaminathan, H. (1985). Item Response Theory: Principles and

Applications. Boston: Kluwer-Nijhoff.

Halligan, J. and Bouckaert, G. (2008). “ Performance Management.” Management Review,

17(1), 31-43.

Hopkins, W. G. (2000). Measures of Validity. Sportsci. org. Web. Retrieved July 04 2011

From: http://www. sportsci. org/resource/stats/valid. html

Knapp, T. R. (2009). The Reliability of Measuring Instruments.

Knapp, T. R. (2001). “ Reporting the reliability of research instruments.” Nurse Author &

Editor, 11 (3), 1-2, 4. (3)

Martin, R. L. (2011) Reliability vs. Validity. Web. Businessweek. com

McDavid, J. C., & Hawthorn L. R. L (2006). Program evaluation & performance measurement: an introduction to practice. London: Sage Publications.

Meincke, S. (1999). International Guidelines for Test-Use. The Council of the International

Test Commission.

Passalacqua, S. and Cervantes, J. M. (2009). Understanding Gender and Culture Within the

Context of Spirituality: Implications for Counsellors. Web. http://family-marriage-counseling. com/mentalhealth/understanding-gender-and-culture-within-the-context-spirituality. htm

Paterson, H. and Uys, K. (2005) Critical Issues in Psychological Test Use in the South

African Workplace. SA Journal of Industrial Psychology, 31 (3), 12-22

Reis, J. P., Duobse, K. D., Ainsworth, B. E., Macera, C. A. and Yore, M. M. (2005).

Reliability and Validity of Occupational Physical Activity Questionnaire.” Med Sci Sports Exerc, 37(12): 2075-83.

Ruane J. M. (2005). Essentials of research methods: a guide to social research. Malden, MA: Blackwell Publishing.

Rubin, A. and Babbie, R. E. (2007). Essential Research Methods for Social

Work. New York:

Brooks Cole.

Rupert, P. A.; Kozlowski, N. F.; Hoffman, L. A.; Daniels, D. D.; and Piette, J.

(2001).

“ Practical and Ethical Issues in Teaching Psychological Testing.” Professional

Psychology: Research and Practice, 30(2), 209-214.

Swanswick, T. (2010) Understanding Medical Education: Evidence, Theory

and Practice.

Wiley-Blackwell.

Thompson, B. R. (2004). Exploratory and Confirmatory Factor Analysis:

Understanding

Concepts and Applications. American Psychological Association.

Waltz, C. F., Strickland, O. & Lenz, E. R. (2010). Measurement in Nursing and

Health Research. New York, NY: Springer Publishing Company, LLC.