



Munich Personal RePEc Archive

## **Two Criteria for Good Measurements in Research: Validity and Reliability**

Mohajan, Haradhan

Assistant Professor, Premier University, Chittagong, Bangladesh.

1 October 2017

Online at <https://mpra.ub.uni-muenchen.de/83458/>

MPRA Paper No. 83458, posted 24 Dec 2017 08:48 UTC

# **Two Criteria for Good Measurements in Research: Validity and Reliability**

**Haradhan Kumar Mohajan**

Premier University, Chittagong, Bangladesh

Email: [haradhan1971@gmail.com](mailto:haradhan1971@gmail.com)

## **Abstract**

Reliability and validity are the two most important and fundamental features in the evaluation of any measurement instrument or tool for a good research. The purpose of this research is to discuss the validity and reliability of measurement instruments that are used in research. Validity concerns what an instrument measures, and how well it does so. Reliability concerns the faith that one can have in the data obtained from the use of an instrument, that is, the degree to which any measuring tool controls for random error. An attempt has been taken here to review the reliability and validity, and threat to them in some details.

**Keywords:** Validity and reliability, errors in research, threats in research.

**JEL Classification:** A2, I2.

## **1. Introduction**

Reliability and validity are needed to present in research methodology chapter in a concise but precise manner. These are appropriate concepts for introducing a remarkable setting in research. Reliability is referred to the stability of findings, whereas validity is represented the truthfulness of findings [Altheide & Johnson, 1994].

Validity and reliability increase transparency, and decrease opportunities to insert researcher bias in qualitative research [Singh, 2014]. For all secondary data, a detailed assessment of reliability and validity involve an appraisal of methods used to collect data [Saunders et al., 2009]. These provide a good relation to interpret scores from psychometric instruments (e.g., symptom scales, questionnaires, education tests, and observer ratings) used in clinical practice, research, education, and administration [Cook & Beckman, 2006]. These are important concepts in modern research, as they are used for enhancing the accuracy of the assessment and evaluation of a research work [Tavakol & Dennick, 2011]. Without assessing reliability and validity of the research, it will be difficult to describe for the effects of measurement errors on theoretical relationships that are being measured [Forza, 2002]. By using various types of methods to collect data for obtaining true information; a researcher can enhance the validity and reliability of the collected data.

The researchers often not only fail to report the reliability of their measures, but also fall short of grasping the inextricable link between scale validity and effective research [Thompson, 2003]. Measurement is the assigning of numbers to observations in order to quantify phenomena. It involves the operation to construct variables, and the development and application of instruments or tests to quantify these variables [Kimberlin & Winterstein, 2008]. If the better mechanism is used, the scientific quality of research will increase. The variables can be measured accurately to present an acceptable research. Most of the errors may occur in the measurement of scale variables, so that the scales development must be imperfect for a good research [Shekharan, & Bougie, 2010]. The measurement error not only affects the ability to find significant results but also can damage the function of scores to prepare a good research. The purpose of establishing reliability and validity in research is essentially to ensure that data are sound and replicable, and the results are accurate.

## **2. Literature Review**

The evidence of validity and reliability are prerequisites to assure the integrity and quality of a measurement instrument [Kimberlin & Winterstein, 2008]. Haynes et al. (2017) have tried to create an evidence-based assessment tool, and determine its validity and reliability for measuring

contraceptive knowledge in the USA. Sancha Cordeiro Carvalho de Almeida has worked on validity and reliability of the 2<sup>nd</sup> European Portuguese version of the “*Consensus auditory-perceptual evaluation of voice*” (II EP CAPE-V) in some details in her master thesis [de Almeida 2016]. Deborah A. Abowitz and T. Michael Toole have discussed on fundamental issues of design, validity, and reliability in construction research. They show that effective construction research is necessary for the proper application of social science research methods [Abowitz & Toole 2010]. Corey J. Hayes, Naleen Raj Bhandari, Nirranjan Kathe, and Nalin Payakachat have analyzed reliability and validity of the medical outcomes study short form-12 version 2 in adults with non-cancer pain [Hayes, et al. 2017]. Yoshida, et al. (2017) have analyzed the patient centered assessment method is a valid and reliable scale for assessing patient complexity in the initial phase of admission to a secondary care hospital. Roberta Heale and Alison Twycross have briefly discussed the aspects of the validity and reliability in the quantitative research [Heale & Twycross 2015].

Moana-Filho et al. (2017) show that reliability of sensory testing can be better assessed by measuring multiple sources of error simultaneously instead of focusing on one source at a time. Reva E. Johnson, Konrad P. Kording, Levi J. Hargrove, and Jonathon W. Sensinger have analyzed in some detail the systematic and random errors that are often arise [Johnson et al., 2017]. Christopher R. Madan and Elizabeth A. Kensinger have examined the test-retest reliability of several measures of brain morphology [Madan et al., 2017]. Stephanie Noble, Marisa N. Spann, Fuyuze Tokoglu, Xilin Shen, R. Todd Constable, and Dustin Scheinost have obtained results on functional connectivity brain MRI. They have highlighted the increase in test-retest reliability when treating the connectivity matrix as a multivariate object, and the dissociation between test–retest reliability and behavioral utility [Noble et al., 2017]. Kilem Li Gwet has explored the problem of inter-rater reliability estimation when the extent of agreement between raters is high [Gwet, 2008]. Satyendra Nath Chakrabarty has discussed an iterative method by which a test can be dichotomized in parallel halves, and ensures maximum split-half reliability [Chakrabarty, 2013]. Kevin A. Hallgren has computed inter-rater reliability for observational data in details for tutorial purposes. He provides an overview of aspects of study design, selection and computation of appropriate inter-rater reliability statistics, and interpreting and reporting results. Then he has included SPSS and R syntax for computing Cohen’s kappa for

nominal variables and intra-class correlations for ordinal, interval, and ratio variables [Hallgren 2012].

Carolina M. C. Campos, Dayanna da Silva Oliveira, Anderson Henry Pereira Feitoza, and Maria Teresa Cattuzzo have tried to develop and to determine reproducibility and content validity of the organized physical activity questionnaire for adolescents [Campos et al., 2017]. Stephen P. Turner has expressed the concept of face validity, used in the sense of the contrast between face validity and construct validity, is conventionally understood in a way which is wrong and misleading [Turner, 1979]. Jessica K. Flake, Jolynn Pek, and Eric Hehman indicate that the use of scales is pervasive in social and personality psychology research, and highlights the crucial role of construct validation in the conclusions derived from the use of scale scores [Flake et al. 2017]. Burns et al. (2017) has analyzed the criterion-related validity of a general factor of personality extracted from personality scales of various lengths has explored in relation to organizational behavior and subjective well-being with 288 employed students.

### **3. Research Objectives**

The aim of this study is to discuss the aspects of reliability and validity in research. The objectives of this research are:

- To indicate the errors the researchers often face.
- To show the reliability in a research.
- To highlight validity in a research.

### **4. Methodology**

Methodology is the guidelines in which we approach and perform activities. Research methodology provides us the principles for organizing, planning, designing and conducting a good research. Hence, it is the science and philosophy behind all researches [Legesse, 2014]. Research methodology is judged for rigor and strength based on validity, and reliability of a research [Morris & Burkett, 2011]. This study is a review work. To prepare this article, we have used the secondary data. In this study, we have used websites, previous published articles, books,

theses, conference papers, case studies, and various research reports. To prepare a good research, researchers often face various problems in data collection, statistical calculations, and to obtain accurate results. Sometimes they may encounter various errors. In this study we have indicated some errors that the researchers frequently face. We also discuss the reliability and validity in the research.

## 5. Errors in a Research

Bertrand Russell warns for any work “*Do not feel absolutely certain of anything*” [Russell, 1971]. Error is common in scientific practice, and many of them are field-specific [Allchin, 2001]. Therefore, there is a chance of making errors when a researcher performs a research is not certainly error free.

### 5.1 Types of Errors

When a researcher runs in research four types of errors may occur in his/her research procedures [Allchin, 2001]: Type I error, Type II error, Type III error, and Type IV error.

**Type I error:** If the null hypothesis of a research is true, but the researcher takes decision to reject it; then an error must occur, it is called Type I error (false positives). It occurs when the researcher concludes that there is a statistically significant difference when in actuality one does not exist. For example, a test that shows a patient to have a disease when in fact the patient does not have the disease, it is a Type I error. A Type I error would indicate that the patient has the virus when he/she does not have, a false rejection of the null hypothesis. Another example is, a patient might take an HIV test, promising a 99.9% accuracy rate. This means that 1 in every 1,000 tests could give a Type I error informing a patient that he/she has the virus, when he/she has not, also a false rejection of the null hypothesis.

**Type II error:** If the null hypothesis of a research is actually false, and the alternative hypothesis is true. The researcher decides not to reject the null hypothesis, and then it is called

Type II error (false negatives). For example, a blood test failing to detect the disease it was designed to detect in a patient who really has the disease is a Type II error.

Both Types I and II errors were first introduced by Jerzy Neyman and Egon S. Pearson [Neyman & Pearson, 1928]. The Type I error is more serious than Type II, because a researcher has wrongly rejected the null hypothesis. Both Type I and Type II errors are factors that every scientist and researcher must take into account.

**Type III Error:** Many statisticians are now adopting a third type of error, a Type III, which is, where the null hypothesis was rejected for the wrong reason. In an experiment, a researcher might postulate a hypothesis and perform research. After analyzing the results statistically, the null is rejected. In 1948, Frederick Mosteller first introduced Type III error [Mitroff & Silvers, 2009]. The problem is that there may be some relationship between the variables, but it could be for a different reason than stated in the hypothesis. An unknown process may underlie the relationship.

**Type IV Error:** The incorrect interpretation of a correctly rejected hypothesis is known as Type IV error. In 1970, L. A. Marascuilo and J. R. Levin proposed Type IV error. For example, a physician's correct diagnosis of an ailment followed by the prescription of a wrong medicine is a Type IV error [Marascuilo & Levin, 1970].

We have observed that a research is error free in the two cases: i) if the null hypothesis is true and the decision is made to accept it, and ii) if the null hypothesis is false and the decision is made to reject it.

Douglas Allchin identifies taxonomy of error types as [Allchin, 2001]: i) material error (impure sample, poor technical skill, etc.), ii) observational error (instrument not understood, observer perceptual bias, sampling error, etc.), iii) conceptual error (computational error, inappropriate statistical model, miss-specified assumptions, etc.), and iv) discursive error (incomplete reporting, mistaken credibility judgments, etc.).

## 5.2 Errors in Measurement

Measurement requires precise definitions of psychological variables such as, intelligence, anxiety, guilt, frustration, altruism, hostility, love, alienation, aggression, reinforcement, and memory. In any measure, a researcher is interested in representing the characteristics of the subject accurately and consistently. The desirable characteristics of a measure are reliability, and validity. Both are important for the conclusions about the credibility of a good research [Waltz et al., 2004]. The measurement error is the difference between the true or actual value and the measured value. The true value is the average of the infinite number of measurements, and the measured value is the precise value. These errors may be positive or negative. Mathematically we can write the measurement error as;

$$\Delta x = x_r - x_i \quad (1)$$

where  $\Delta x$  is the error of measurement,  $x_r$  is the real untrue measurement value, and  $x_i$  is the ideal true measurement value. For example, if electronic scales are loaded with 10 kg standard weight, and the reading is 10 kg 2 g, then the measurement error is 2 g.

Usually there are three measurement errors occur in research [Malhotra, 2004]: i) gross errors, ii) systematic error, that affects the observed score in the same way on every measurement, and iii) random error; that varies with every measurement. In research a true score theory is represented as [Allen & Yen, 1979];

$$X = T + E_r + E_s \quad (2)$$

where  $X$  is the obtained score on a measure,  $T$  is the true score,  $E_r$  is random error, and  $E_s$  is systematic error. If  $E_r = 0$  in (2), then instrument is termed as reliable. If both  $E_r = 0$  and  $E_s = 0$  then,  $X = T$  and the instrument is considered as valid.

**5.2.1 Gross errors:** These occur because of the human mistakes, experimenter's carelessness, equipment failure or computational errors [Corbett et al., 2015]. Frequently, these are easy to recognize and the origins must be eliminated [Reichenbacher & Einax, 2011]. Consider a person using the instruments take the wrong reading. For example, the experimenter reads the 50.5°C reading while the actual reading is 51.5°C. This happens because of the oversights. The



experimenter takes the wrong reading. Hence, the error occurs in the measurement. This error can only be avoided by taking the reading carefully. Two methods can remove the gross error as: i) the reading should be taken very carefully, and ii) two or more readings should be taken by the different experimenter, and at a different point for removing the error.

**5.2.2 The systematic errors:** These influence all examinee's scores in a systematic way. These occur due to fault in the measuring device. These can be detected by correcting the measurement device [Taylor, 1999]. The systematic errors can be classified as: i) instrumental errors, ii) environmental errors, iii) observational errors, and iv) theoretical errors (figure 1).

***Instrumental errors:*** These occur due to manufacturing, calibration or operation of the device. These may arise due to friction or hysteresis [Swamy, 2017]. These include loading effect, and misuse of the instruments. In order to reduce the gross errors in measurement, different correction factors must be applied, and in the extreme condition instrument must be recalibrated carefully. For example, if the instrument uses the weak spring, then it gives the high value of measuring quantity.

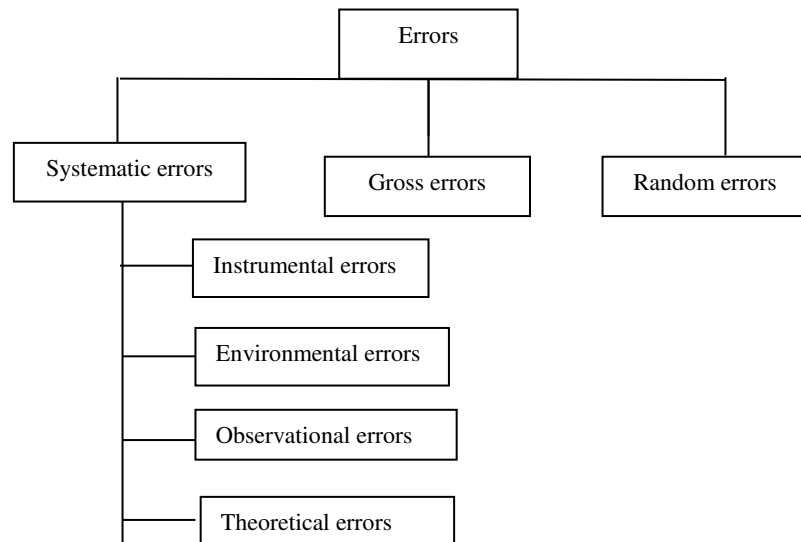
***Environmental errors:*** These occur due to some external conditions of the instrument. External conditions include pressure, temperature, humidity, dust, vibration, electrostatic or magnetic fields [Gluch, 2000]. In order to reduce these errors a researcher can try to maintain the humidity and temperature constant in the laboratory by making some arrangements, and ensuring that there shall not be any external electrostatic or magnetic field around the instrument.

***Observational errors:*** These types of errors occur due to wrong observations or reading in the instruments particularly in case of energy meter reading [Allchin, 2001]. The wrong observations may be due to parallax. To reduce the parallax error highly accurate meters are needed with mirrored scales.

***Theoretical errors:*** These are caused by simplification of the model system [Allchin, 2001]. For example, a theory states that the temperature of the system surrounding will not change the readings taken when it actually does, then this factor will begin a source of error in measurement.

**5.2.3 Random Errors:** After calculating all systematic errors, it is found that there are still some errors in measurement left [DeVellis, 2006]. These errors are known as random errors (figure 1). These are caused by the sudden change in experimental conditions, also for noise, and tiredness in the working persons. These errors are either positive or negative [Taylor, 1999]. Examples of the random errors are; changes in humidity, unexpected change in temperature, and fluctuation in voltage during an experiment. These errors may be reduced by taking the average of a large number of readings.

If both systematic and random errors are occurred in a research, it is considered as total measurement error [Allen & Yen, 1979]. Systematic errors are found for stable factors which influence the observed score in the same way on every occasion of measurement. But, random error occurs due to transient factors which influence the observed score differently each time [Malhotra, 2004]. If the random error is zero then research is considered as reliable. If both systematic error and random error are zero then research is considered as valid [Bajpai & Bajpai, 2014]. To minimize overall error, random errors should be ignored, whereas systematic errors should result in adaptation of the movement [Johnson et al., 2017].



**Figure 1:** Structure of errors occurs in measurement.

### **5.3 Evaluation of the Quality of Measures**

Key indicator of the quality of a measure is the proper measurement of reliability and validity of the research. In a standard research, any score obtained by a measuring instrument is the sum of both the 'true score', which is unknown, and 'error' in the measurement process. If the error margins are low and reporting of results of a research are of high standards, no doubt the research will be fruitful. If the measurement is very accurate then a researcher will find a true score [Kimberlin & Winterstein, 2008]. Actually, the foundation of a good research is the trustworthiness (reliability and validity) of the data to make decisions; otherwise a good decision cannot be made.

In quantitative research it is possible for a measurement to be reliable but invalid; however, if a measurement is unreliable, then it cannot be valid [Thatcher, 2010; Twycross & Shields, 2004].

## **6. Reliability**

The reliability refers to a measurement that supplies consistent results with equal values [Blumberg et al., 2005]. It measures consistency, precision, repeatability, and trustworthiness of a research [Chakrabarty, 2013]. It indicates the extent to which it is without bias (error free), and hence insures consistent measurement cross time and across the various items in the instruments (the observed scores). Some qualitative researchers use the term 'dependability' instead of reliability. It is the degree to which an assessment tool produces stable (free from errors) and consistent results. It indicates that the observed score of a measure reflects the true score of that measure. It is a necessary, but not sufficient component of validity [Feldt & Brennan, 1989].

In quantitative research, reliability refers to the consistency, stability and repeatability of results, that is, the result of a researcher is considered reliable if consistent results have been obtained in identical situations but different circumstances. But, in qualitative research it is referred to as when a researcher's approach is consistent across different researchers and different projects [Twycross & Shields, 2004].

It is a concern every time a single observer is the source of data, because we have no certain guard against the impact of that observer's subjectivity [Babbie, 2010]. Reliability issues are most of the time closely associated with subjectivity, and once a researcher adopts a subjective approach towards the study, then the level of reliability of the work is going to be compromised [Wilson, 2010].

The coefficient of reliability falls between 0 and 1, with perfect reliability equaling 1, and no reliability equaling 0. The test-retest and alternate forms are usually calculated reliability by using statistical tests of correlation [Traub & Rowley, 1991]. For high-stakes settings (e.g., licensure examination) reliability should be greater than 0.9, whereas for less important situations values of 0.8 or 0.7 may be acceptable. The general rule is that reliability greater than 0.8 are considered as high [Downing, 2004].

Reliability is used to evaluate the stability of measures administered at different times to the same individuals and the equivalence of sets of items from the same test [Kimberlin & Winterstein, 2008]. The better the reliability is perform, the more accurate the results; which increases the chance of making correct decision in research. Reliability is a necessary, but not a sufficient condition for the validity of research.

## 6.1 Types of Reliability

Reliability is mainly divided into two types as: i) Stability, and ii) Internal consistency reliability. Stability: It is defined as the ability of a measure to remain the same over time despite uncontrolled testing conditions or respondent themselves. It refers to how much a person's score can be expected to change from one administration to the next [Allen & Yen, 1979]. A perfectly stable measure will produce exactly the same scores time after time. Two methods to test stability are: i) test-retest reliability, and ii) parallel-form reliability.

***Test-retest reliability:*** The reliability coefficient is obtained by repetition of the same measure on a second time, is called the test-retest reliability [Graziano and Raulin, 2006]. It assesses the external consistency of a test [Allen & Yen, 1979]. If the reliability coefficient is high, for

example,  $r = 0.98$ , we can suggest that both instruments are relatively free of measurement errors. If the coefficients yield above 0.7, are considered acceptable, and coefficients yield above 0.8, are considered very good [Sim & Wright, 2005; Madan & Kensinger, 2017].

The test-retest reliability indicates score variation that occurs from testing session to testing session as a result of errors of measurement. It is a measure of reliability obtained by managing the same test twice over a period of time ranging from few weeks to months, on a group of individuals. The scores from Time 1 and Time 2 can then be correlated between the two separate measurements in order to evaluate the test for stability over time. For example, employees of a Company may be asked to complete the same questionnaire about employee job satisfaction two times with an interval of three months, so that test results can be compared to assess stability of scores. The correlation coefficient calculated between two set of data, and if it found to be high, the test-retest reliability is better. The interval of the two tests should not be very long, because the status of the company may change during the second test, which affects the reliability of research [Bland & Altman, 1986].

***Parallel-forms reliability:*** It is a measure of reliability obtained by administering different versions of an assessment tool to the same group of individuals. The scores from the two versions can then be correlated in order to evaluate the consistency of results across alternate versions. If they are highly correlated, then they are known as parallel-form reliability [DeVellis, 2006]. For example, the levels of employee satisfaction of a Company may be assessed with questionnaires, in-depth interviews and focus groups, and the results are highly correlated. Then we may be sure of the measures that they are reasonably reliable [Yarnold, 2014].

**Internal Consistency Reliability:** It is a measure of reliability used to evaluate the degree to which different test items that probe the same construct produce similar results. It examines whether or not the items within a scale or measure are homogeneous [DeVellis, 2006]. It can be established in one testing situation, thus it avoids many of the problems associated with repeated testing found in other reliability estimates [Allen & Yen, 1979]. It can be represented in two main formats [Cortina, 1993]: i) The inter-item consistency, and ii) Split-half reliability.

***Inter-rater reliability:*** It is the extent to which the way information being collected is being collected in a consistent manner [Keyton et al., 2004]. It establishes the equivalence of ratings obtained with an instrument when used by different observers. No discussion can occur when reliability is being tested. Reliability is determined by the correlation of the scores from two or more independent raters, or the coefficient of agreement of the judgments of the raters. It is useful because human observers will not necessarily interpret answers the same way; raters may disagree as to how well certain responses or material demonstrate knowledge of the construct or skill being assessed. For example, levels of employee motivation of a Company can be assessed using observation method by two different assessors, and inter-rater reliability relates to the extent of difference between the two assessments. The most common internal consistency measure is Cronbach's alpha ( $\alpha$ ), which is usually interpreted as the mean of all possible split-half coefficients. It is a function of the average inter-correlations of items, and the number of items in the scale. It is widely used in social sciences, business, nursing, and other disciplines. It was first named alpha by Lee Joseph Cronbach in 1951, as he had intended to continue with further coefficients. It is typically varies between 0 and 1, where 0 indicates no relationship among the items on a given scale, and 1 indicates absolute internal consistency [Tavakol & Dennick 2011]. Alpha values above 0.7 are generally considered acceptable and satisfactory, above 0.8 are usually considered quite good, and above 0.9 are considered to reflect exceptional internal consistency [Cronbach, 1951]. In the social sciences, acceptable range of alpha value estimates from 0.7 to 0.8 [Nunnally & Bernstein, 1994].

***Split-half reliability:*** It measures the degree of internal consistency by checking one half of the results of a set of scaled items against the other half [Ganesh, 2009]. It requires only one administration, especially appropriate when the test is very long. It is done by comparing the results of one half of a test with the results from the other half. A test can be split in half in several ways, for example, first half and second half, or by odd and even numbered items. If the two halves of the test provide similar results this would suggest that the test has internal reliability. It is a quick and easy way to establish reliability. It can only be effective with large questionnaires in which all questions measure the same construct, but it would not be appropriate for tests which measure different constructs [Chakrabartty, 2013].

It provides a simple solution to the problem that the parallel form faces. It involves, administering a test to a group of individuals, splitting the test in half, and correlating scores on one half of the test with scores on the other half of the test [Murphy & Davidshofer, 2005]. It may be higher than Cronbach's alpha only in the circumstances of there being more than one underlying responses dimension tapped by measure, and when certain other conditions are met as well.

## **7. Validity**

Validity is often defined as the extent to which an instrument measures what it asserts to measure [Blumberg et al., 2005]. Validity of a research instrument assesses the extent to which the instrument measures what it is designed to measure (Robson, 2011). It is the degree to which the results are truthful. So that it requires research instrument (questionnaire) to correctly measure the concepts under the study (Pallant 2011). It encompasses the entire experimental concept, and establishes whether the results obtained meet all of the requirements of the scientific research method. Qualitative research is based on the fact that validity is a matter of trustworthiness, utility, and dependability [Zohrabi, 2013]. Validity of research is an extent at which requirements of scientific research method have been followed during the process of generating research findings. It is a compulsory requirement for all types of studies [Oliver, 2010].

In quantitative research validity is the extent to which any measuring instrument measures what it is intended to measure [Thatcher, 2010]. But, in qualitative research it is when a researcher uses certain procedures to check for the accuracy of the research findings [Creswell, 2014]. It is not a property of the instrument, but of the instrument's scores and their interpretations. It is best viewed as a hypothesis for which evidence is collected in support of proposed inferences [Messick, 1989]. Lee J. Cronbach and Paul E. Meehl first introduced the issue of validity in quantitative research in the mid 20<sup>th</sup> century in relation to the establishment of the criteria for assessing psychological tests [Cronbach & Meehl, 1955].

In research, validity has two essential parts: a) internal (credibility), and b) external (transferability). Internal validity indicates whether the results of the study are legitimate because

of the way the groups were selected, data were recorded or analyses were performed. It refers to whether a study can be replicated [Willis, 2007]. To assure it, the researcher can describe appropriate strategies, such as triangulation, prolonged contact, member checks, saturation, reflexivity, and peer review. External validity shows whether the results given by the study are transferable to other groups of interest [Last, 2001]. A researcher can increase external validity by: i) achieving representation of the population through strategies, such as, random selection, ii) using heterogeneous groups, iii) using non-reactive measures, and iv) using precise description to allow for study replication or replicate study across different populations, settings, etc.

It alarmed with whether a researcher measures the right concept or not [Shekharan & Bougie, 2010]. Validity requires that an instrument is reliable, but an instrument can be reliable without being valid [Kimberlin & Winterstein, 2008].

### **7.1 Types of Validity**

Validity test is mainly divided into four types as [Creswell, 2005; Pallant, 2011]: i) content validity, ii) face validity, iii) construct validity, and iv) criterion-related validity (figure 2).

**Content Validity:** It is the extent to which the questions on the instrument and the scores from these questions represent all possible questions that could be asked about the content or skill [Creswell, 2005]. It ensures that the questionnaire includes adequate set of items that tap the concept. The more the scale items represent the domain of the concept being measured, the greater the content validity [Shekharan & Bougie, 2010]. With it is the interested in assessing current performance rather than predicting future performance. It is related to a type of validity in which different elements, skills and behaviors are adequately and effectively measured [DeVellis, 2006; Messick, 1995]. There is no statistical test to determine whether a measure adequately covers a content area, content validity usually depends on the judgment of experts in the field. The unclear and obscure questions can be amended, and the ineffective and non-functioning questions can be discarded by the advice of the reviewers. For example, if we want to test knowledge on Bangladeshi Geography it is not fair to have most questions limited to the geography of Dhaka, the capital city of Bangladesh. Another example is, in arithmetic operations, the test problem will be content valid if the researcher focuses on addition,



subtraction, multiplication and division, but will be content invalid if the researcher focuses on one aspect of arithmetic alone, addition (say) [Thatcher, 2010].

To effectively evaluate content validity, L. Crocker and J. Algina suggest the four steps procedures as [Crocker and Algina, 2010]: i) identify and outline the domain of interest, (ii) gather resident domain experts, (iii) develop consistent matching methodology, and (iv) analyze results from the matching task. Content validity can be grouped into two types: i) face validity, and ii) logical validity [Allen & Yen, 1979].

**Face Validity:** It is considered as a basic and minimum index of content validity, but it is determined after the test is constructed [Allen & Yen, 1979]. The concepts of content evidence and face validity bear superficial resemblance, but they are in fact quite different. Face validity refers to the degree to which a test appears to measure what it claims to measure [Leedy & Ormrod, 2004]. It is a global answer as a quick assessment of what the test is measuring. It is the simplest and least precise method of determining validity which relies entirely on the expertise and familiarity of the assessor concerning the subject matter [Nwana, 2007]. It ascertains that the measure appears to be assessing the intended construct under study. It is usually used to describe the appearance of validity without empirical testing [Cook & Beckman, 2006]. So, it is normally considered to be the weakest form of validity [Hashim et al., 2007]. For example, estimating the speed of a car based on its outward appearance (guesswork) is face validity.

If the test is known to have content validity, face validity can be assumed, but face validity does not ensure content validity. The stakeholders can easily assess face validity. Although this is not a very scientific type of validity, it may be an essential component for enlisting motivation of stakeholders. If the stakeholders do not believe the measure is an accurate assessment of the ability, they may become detached with the task. Therefore, it looks as if it is indeed measuring what it is designed to measure. Unlike content validity, face validity does not depend on established theories for support [Fink, 1995].

**Criterion-Related Validity:** It is used to predict future or current performance. It correlates test results with another criterion of interest [Burns et al., 2017]. It deals with relationship between scale scores, and some specific measurable criterion. It tests how the scale differentiates

individuals on a criterion it is expected to predict [Pallant, 2011]. That is, when we are expecting a future performance based on the scores obtained currently by the measure, correlate the scores obtained with the performance [Messick, 1989]. For example, a hands-on driving test has been shown to be an accurate test of driving skills. The test can be repeated by the written test to compare validity of the test. It can be established by; i) the concurrent validity, and ii) the predictive validity.

**Concurrent Validity:** It is the degree to which the scores on a test are related to the scores on another, already established as valid, designed to measure the same construct, test administered at the same time or to some other valid criterion available at the same time. It is necessary when a test for assessing skills is constructed with a view to replacing less efficient one in used [Denga, 1987]. It is established by correlating one question with another that has previously been validated with standard setting [Okoro, 2002]. It examines the validity of a tool on a highly theoretical level [Messick, 1989]. Example, a new simple test is to be used in place of an old troublesome one, which is considered useful; measurements are obtained on both at the same time.

**Predictive Validity:** It is often used in program evaluation studies, and is very suitable for applied research. It is a test constructed and developed for the purpose of predicting some form of behavior [Allen & Yen, 1979]. It indicates the ability of the measuring instrument to differentiate among individuals with reference to a future criterion. Test that are constructed to pick applicants who are most likely to be successful subsequently in their training while rejecting those applicants who are most likely to be failures if given admission [Nwana, 2007]. Logically, predictive and concurrent validation are the same, the term concurrent validation is used to indicate that no time elapsed between measures.

The higher the correlation between the criterion and the predictor indicates the greater the predictive validity. If the correlation is perfect, that is 1, the prediction is also perfect. Most of the correlations are only modest, somewhere between 0.3 and 0.6.

**Construct Validity:** It is especially important for the empirical measures and hypothesis testing for the construction of theories. Researchers create theoretical constructs to better understand,

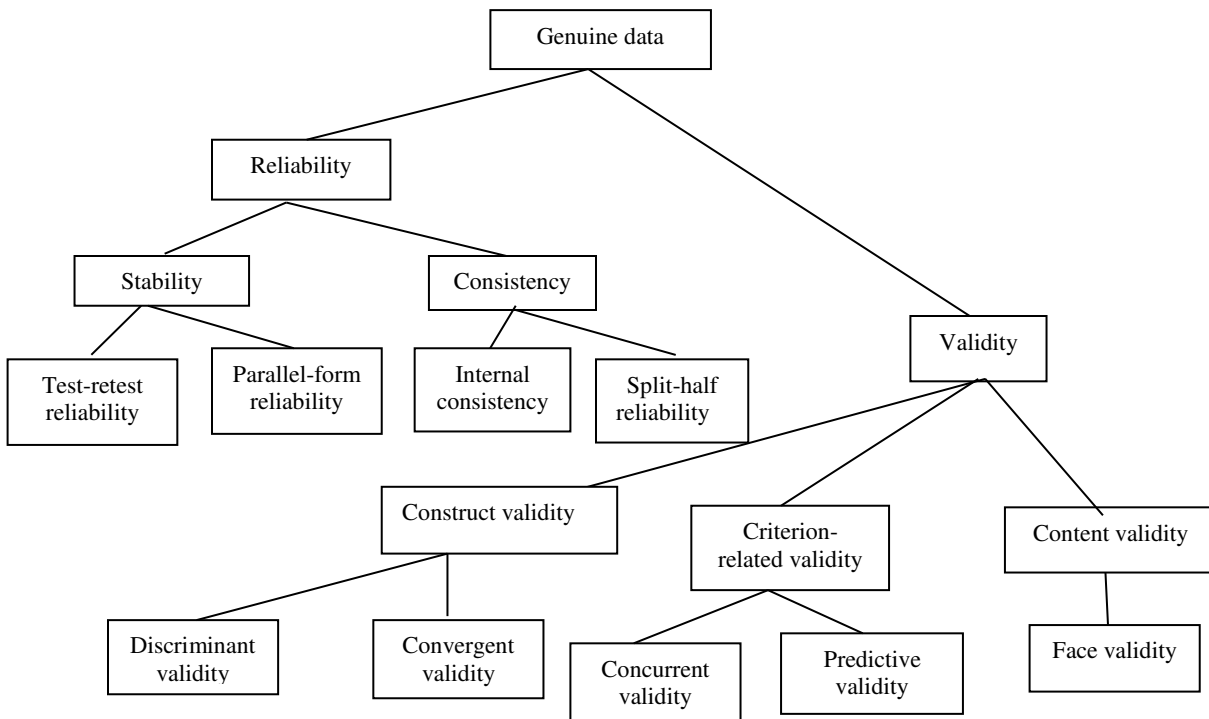
explain, and predict behavior [Thatcher, 2010]. It involves testing a scale in terms of theoretically derived hypotheses concerning the nature of underlying variables or constructs [Pallant, 2011]. The term ‘construct validity’ was first formulated by a sub-committee (P. E. Meehl and M. C. Challman) of the American Psychologists Association’s Committee on Psychological Tests [Cronbach & Meehl, 1955]. A construct needs to be both operationalized and syntactically defined in order to measure it effectively. The operationalization of the construct develops a series of measurable behaviors that are hypothesized to correspond to the latent construct. The construct syntactically involves establishing hypothesized relationships between the construct of interest and other related behaviors [Crocker & Algina, 1986; DeVellis, 2006]. It pertains to a specific use of a scale, and can often be context or population dependent [Kane, 2013].

It is a judgment based on the accumulation of evidence from numerous studies using a specific measuring instrument. It is used to ensure that the measure is actually measure what it is intended to measure, and not other variables [Twycross & Shields, 2004]. Using a panel of experts familiar with the construct is a way in which this type of validity can be assessed [Kane, 2013]. The experts can examine the items and decide what that specific item is intended to measure. The process of validating the interpretations about that construct as indicated by the test score is construct validation. It is used to refine a theory, for making predictions about test scores in various settings and situations [DeVellis, 2006]. It is evaluated through convergent and discriminate validity. Construct validity of the instrument is checked by correlation analysis, factor analysis, and the multi-trait, multi-method matrix of correlations [Pett et al., 2003]. For example, a researcher inventing a new IQ test might spend a great deal of time attempting to ‘define’ intelligence to reach an acceptable level of construct validity. It is divided into two categories: i) convergent validity, and ii) discriminant validity [Huck, 2007].

***Convergent validity:*** It refers to the extent to which scores on a measure share a high, medium or low relationship with scores obtained on a different measure intended to assess the similar construct [Messick, 1995]. It is established when the scores obtained with two different instruments measuring the same concept are highly correlated. It is the degree to which two variables measured separately bear a relationship to one another [Straub, 1989]. It is the actual

general agreement among ratings, gathered independently of one another, where measures should be theoretically related [Campbell, 1959].

**Discriminant validity:** It is established when, based on theory, two variables are predicted to be uncorrelated, and the scores obtained by measuring them are indeed empirically found to be so, that is, to differentiate one group from another. It is the lack of a relationship among measures which theoretically should not be related [Messick, 1995; Sperry, 2004]. For example, surveys that are used to identify potential high school drop-outs would have discriminant validity if the students who graduate score higher on the test than students who leave before graduation [Campbell, 1959].



**Figure 2:** Structure of reliability and validity. Source: Bajpai and Bajpai (2014).

To ensure validity of a research following points are measurable:

- Appropriate time scale for the study has to be selected;
- Appropriate methodology has to be chosen, taking into account the characteristics of the study;

- The most suitable sample method for the study has to be selected;
- The respondents must not be pressured in any way to select specific choices among the answer sets.

There are some ways to improve validity as follows:

- Make sure a researcher's goals and objectives are clearly defined and operationalized.
- Match the assessment measure to the goals and objectives of research.
- The researcher looks over the assessment for troublesome wording, or other difficulties.
- If possible, compare the measure with other measures, or data that may be available.

It is possible to have a high degree of reliability with a low level of validity, but for a research instrument to be valid it must also be reliable [Keller, 2000]. Therefore, reliability is a sub-component of validity, and must first be attained if validity is to be achieved [Willis, 2007].

## **8. Threats to Validity and Reliability**

The multiple factors can create risks to the validity and reliability of the findings of a researcher. Error is one of them. Researchers thus must be careful of the sources of errors in plans and implementation of their studies. The major sources of research errors can be obtained from the careless of researcher, the subjects participating in the study, the social context, and the methods of data collection and analysis [Lillis, 2006]. Errors of measurement that affect reliability are random errors, and errors of measurement that affect validity are systematic or constant errors. Threats to the validity and reliability of a research exist at almost every turn in the research process. It can never be totally eliminated, so a researcher needs to try his best to minimize the threats as much as possible. A common threat to internal validity is reliability.

Threats to reliability may occur for lack of clear and standard instructions, not all alternatives are provided, the questions are not presented in the proper order, measurement instruments describe items ambiguously so that they are misinterpreted, the questionnaire is too long or hard to read, and the interview takes too long time [Kerlinger, 1964; Fink and Kosecoff, 1985].

Threats to the internal validity may occur throughout the research process. The threats to internal validity are insufficient knowledge during data collection, analysis and/or interpretation. During data collection, possible threats to internal validity are instrumentation issues, order bias, and researcher bias in the use of techniques [Tashakkori & Teddlie, 1998; Ongwuegbuzie, 2003]. The external validity of a quantitative study may threaten in population, time and environmental validity [Ryan et al., 2002]. External validity is seriously threatened, if biases or other limitations exist in the accessible population [Howell, 1995].

Instrumentation issues occur when scores yielded from a measure lack the appropriate level of consistency, or do not generate valid scores. Order bias threat occurs if the effect of the order of the intervention conditions cannot be separated from the effect of the intervention conditions. Researcher bias threat is a personal bias in favor of one technique over another. Errors in statistical testing, illusory correlation, and causal error are some threats during data analysis and interpretation [Ihantola & Kihn, 2011]. For example, a table clock that is always five minutes fast is reliable because it is always five minutes fast; however, it is not valid because when compared to a standard format such as the GMT, it is not correct.

## **9. Conclusion**

In this study we have tried to show that reliability and validity of instrumentation are important considerations for researchers in their investigations. To perform a good research validity and reliability tests are needed to take very carefully. We have highlighted on the research errors that are arisen in measurements. In the study we have observed that a valid tool must be reliable, but a reliable tool may not necessarily be valid. We have also included the threat to reliability and validity when a researcher tries to do a good research.

## **References**

Abowitz, D. A., & Toole, T. M. (2010). Mixed Method Research: Fundamental Issues of Design, Validity, and Reliability in Construction Research. *Journal of Construction Engineering and Management*, 136(1), 108-116.

Allchin, D. (2001). Error Types, *Perspectives on Science*, 9(1), 38-58.

Allen, M. J., & Yen, W. M. (1979). Introduction to Measurement Theory. Monterey, CA: Brooks/Cole Publishing Company.

Altheide, D. L., & Johnson, J. M. (1994). Criteria for Assessing Interpretive Validity in Qualitative Research. In N. K. Denzin & Y. S. Lincoln (Eds.). *Handbook of Qualitative Research*, pp. 485-499. Thousand Oaks, CA: SAGE.

Babbie, E. R. (2010). The Practice of Social Research. Belmont, CA: Wadsworth.

Bajpai, S. R., & Bajpai, R. C. (2014). Goodness of Measurement: Reliability and Validity. *International Journal of Medical Science and Public Health*, 3(2), 112-115.

Bland, M. J., & Altman, D. (1986). Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. *The Lancet*, 327(8476), 307-310.

Blumberg, B., Cooper, D. R., & Schindler, P. S. (2005). *Business Research Methods*. Berkshire: McGrawHill Education.

Burns, G. N., Morris, M. B., Periard, D. A., LaHuis, D., Flannery, N. M., Carretta, T. R., & Roebke, M. (2017). Criterion-Related Validity of a Big Five General Factor of Personality from the TIPI to the IPIP. *International Journal of Selection and Assessment*, 25, 213–222.

Campbell, D. T. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, 56(2), 81–105.

Campos, C.M.C., da Silva Oliveira, D., Feitoza, A. H. P., & Cattuzzo, M. T. (2017). Reliability and Content Validity of the Organized Physical Activity Questionnaire for Adolescents, *Educational Research*, 8(2), 21-26.

Chakrabartty, S. N. (2013). Best Split-Half and Maximum Reliability. *IOSR Journal of Research & Method in Education*, 3(1), 1-8.

Cook, D. A., & Beckman, T. J. (2006). Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. *The American Journal of Medicine*, 119, 166.e7-166.e16.

Corbett, N., Sibbald, R., Stockton, P., & Wilson, A. (2015). *Gross Error Detection: Maximising the Use of Data with UBA on Global Producer III (Part 2)*. 33<sup>rd</sup> International North Sea Flow Measurement Workshop 20<sup>th</sup> – 23<sup>rd</sup> October 2015.

Cortina, J. M. (1993). What is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology*, 78(1), 98–104.

Creswell, J. W. (2005). *Educational Research: Planning, Conducting and Evaluating Quantitative and Qualitative Research* (2<sup>nd</sup> Ed.). Pearson Merrill Prentice Hall.

Creswell, R. (2014). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. USA: SAGE Publications.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Philadelphia: Harcourt Brace Jovanovich College Publishers.

Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, 16, 297–334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct Validity in Psychological Tests, *Psychological Bulletin*, 52, 281-302.



de Almeida, S. C. C. (2016). Validity and Reliability of the 2<sup>nd</sup> European Portuguese Version of the “*Consensus Auditory-Perceptual Evaluation of Voice*” (II EP CAPE-V). Master Thesis. Health Science School of Polytechnic Institute of Setúbal, Portugal.

Denga, D. I. (1987). *Educational Measurement, Continuous Assessment and Psychological Testing*. Calabar Rapid Educational Publishers Ltd.

Devillis, R. E. (2006). Scale Development: Theory and Application. *Applied Social Science Research Method Series*. Vol. 26 Newbury Park: SAGE Publishers Inc.

Downing, S. M. (2004). Reliability: On the Reproducibility of Assessment Data. *Med Education*, 38, 1006-1012.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.). *Educational Measurement* (3<sup>rd</sup> Ed.). New York: American Council on Education and Macmillan.

Fink, A. (Ed.) (1995). *How to Measure Survey Reliability and Validity*. Thousand Oaks, CA: SAGE.

Fink, A., & Kosecoff, J. (1985). *How to Conduct Surveys?* Newbury Park, CA: SAGE Publications.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research: Current Practice and Recommendations. *Social Psychological and Personality Science*, 8(4), 1-9.

Forza, C. (2002). Survey Research in Operations Management: A Process-based Perspective. *International Journal of Operations and Production Management*, 22 (2), 152-194.

Ganesh, T. (2009). Reliability and Validity Issues in Research. *Integration and Dissemination Research Bulletin*, 4, 35-40.

Gluch, P. (2000). Costs of Environmental Errors (CEE): A Managerial Environmental Accounting Tool or a Symptom of Managerial Frustration? *Greener Management International*, 31, 87-100.

Graziano, A. M., & Raulin, M. L. (2006). *Research Methods: A Process of Inquiry* (6<sup>th</sup> Ed.). Boston, MA: Allyn & Bacon.

Gwet, K. L. (2008). Computing Inter-Rater Reliability and its Variance in the Presence of High Agreement. *British Journal of Mathematical and Statistical Psychology*, 61, 29-48.

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1), 23-34.

Hashim, N. H., Murphy, J., & O'Connor, P. (2007). Take Me Back: Validating the Wayback Machine as a Measure of Website Evolution. In M. Sigala, L. Mich and J. Murphy (Eds.). *Information & Communication Technologies in Tourism*, pp. 435-446, Wien: Springer-Verlag.

Haynes, M. C., Ryan, N., Saleh, M., Winkel, A. F., & Ades, V. (2017). Contraceptive Knowledge Assessment: Validity and Reliability of a Novel Contraceptive Research Tool. *Contraception*, 95, 190–197.

Heale, R., & Twycross, A. (2015). Validity and Reliability in Quantitative Studies. *Evid Based Nurs*, 18(4), 66-67.

Howell, D.C. (1995). *Fundamental Statistics for the Behavioral Sciences* (3<sup>rd</sup> Ed.). Duxbury Press, Belmont, California.

Huck, S. W. (2007). *Reading Statistics and Research* (5<sup>th</sup> Ed.). New York, NY: Allyn & Bacon.

Ihantola, E. -M., & Kihn, L. -A. (2011). Threats to Validity and Reliability in Mixed Methods Accounting Research. *Qualitative Research in Accounting and Management*, 8(1), 39-58.

Johnson, R. E., Kording, K. P., Hargrove, L. J., & Sensinger, J. W. (2017). Adaptation to Random and Systematic Errors: Comparison of Amputee and Non-Amputee Control Interfaces with Varying Levels of Process Noise. *PLoS ONE*, 12(3): e0170473.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, 50, 1–73.

Keller, A. (2000). Electronic Journals: A Delphi Survey. *INSPEL*, 34(3-4), 187-193.

Kerlinger, H. (1964). *Foundations of Behavioral Research*. Holt, Rinehart and Winston, Inc., New York.

Keyton, J., King, T., Mabachi, N. M., Manning, J., Leonard, L. L., & Schill, D. (2004). *Content Analysis Procedure Book*. Lawrence, KS: University of Kansas.

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and Reliability of Measurement Instruments Used in Research. *American Journal of Health-System Pharmacists*, 65(1), 2276-2284.

Last, J. (Ed.) (2001). *International Epidemiological Association, A Dictionary of Epidemiology* (4<sup>th</sup> Ed.). New York: Oxford University Press.

Leedy, P. D., & Ormrod, J. E. (2004). *Practical Research*, (8<sup>th</sup> Ed.). Upper Saddle River, N.J: Prentice Hall.

Legesse, B. (2014). *Research Methods in Agribusiness and Value Chains*. School of Agricultural Economics and Agribusiness, Haramaya University.

Lillis, A. (2006), Reliability and Validity in Field Study Research. In Z. Hoque (Ed.), *Methodological Issues in Accounting Research: Theories and Methods*, pp. 461-475. Piramus, London.

Madan, C. R., & Kensinger, E. A. (2017). Test–Retest Reliability of Brain Morphology Estimates. *Brain Informatics*, 4, 107–121.

Malhotra, N. K. (2004). *Marketing Research: An Applied Orientation* (4<sup>th</sup> Ed.). New Jersey: Pearson Education, Inc.

Marascuilo, L. A., & Levin, J. R. (1970). Appropriate Post Hoc Comparisons for Interaction and Nested Hypotheses in Analysis of Variance Designs: The Elimination of Type-IV Errors, *American Educational Research Journal*, 7(3), 397-421.

Messick, S. (1989). *Validity*. In R. L. Linn (Ed.). *Educational Measurement* (3<sup>rd</sup> Ed.). New York: American Council on Education and Macmillan.

Messick, S. (1995). Standards of Validity and the Validity of Standards in Performance Assessment. *Educational Measurement: Issues and Practice*, 14(4), 5-8.

Mitroff, I. I., & Silvers, A. (2009). *Dirty Rotten Strategies: How We Trick Ourselves and Others into Solving the Wrong Problems Precisely*. Stanford Business Press.

Moana-Filho, E. J., Alonso, A. A., Kapos, F. P., Leon-Salazar, V., Gurand, S. H., Hodges, J. S., & Nixdorf, D. R. (2017). Multifactorial Assessment of Measurement Errors Affecting Intraoral Quantitative Sensory Testing Reliability. *Scandinavian Journal of Pain*, 16(6), 93-98.

Morris, E., & Burkett, K. (2011). Mixed Methodologies: A New Research Paradigm or Enhanced Quantitative Paradigm, *Online Journal of Cultural Competence in Nursing and Healthcare*, 1(1): 27–36.

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological Testing: Principles and Applications* (6<sup>th</sup> Ed.). Upper Saddle River, N.J.: Pearson/Prentice Hall.

Neyman, J., & Pearson, E. S. (1928). On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I. *Biometrika*, 20A(1/2), 175-240.

Noble, S., Spann, M. N., Tokoglu, F., Shen, X., Constable, R. T., & Scheinost, D. (2017). Multifactorial Assessment of Measurement Errors Affecting Intraoral Quantitative Sensory Testing Reliability. *Cerebral Cortex*, 27(11), 5415–5429.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3<sup>rd</sup> Ed.). McGraw-Hill: New York.

Nwana, O. C. (2007). *Textbook on Educational Measurement and Evaluation*. Owerri: Bomaway Publishers.

Okoro, O. M. (2002). *Measurement and Evaluation in Education*. Obosi: Pacific Publisher Ltd.

Oliver, V. (2010). *301 Smart Answers to Tough Business Etiquette Questions*. Skyhorse Publishing: New York, USA.

Onwuegbuzie, A. J. (2003). Expanding the Framework of Internal and External Validity in Quantitative Research. *Research in the Schools*, 10(1), 71-90.

Pallant, J. (2011). *A Step by Step Guide to Data Analysis Using the SPSS Program: Survival Manual*, (4<sup>th</sup> Ed.). McGraw-Hill, Berkshire.

Pett, M., Lackey, N., & Sullivan, J. (2003). *Making Sense of Factor Analysis: The Use of Factor Analysis for Instrument Development in Health Care Research*. Thousand Oaks, CA: SAGE Publications.

Reichenbacher, M., & Einax, J. W. (2011). *Challenges in Analytical Quality Assurance*. Springer-Verlag Berlin Heidelberg.

Robson, C. (2011). *Real World Research: A Resource for Users of Social Research Methods in Applied Settings*, (2<sup>nd</sup> Ed.). Sussex, A. John Wiley and Sons Ltd.

Russell, B. (1971). A Liberal Decalogue. In *The Autobiography of Bertrand Russell*. 3. 1944–1967, pp. 60–101. London: George Allen & Unwin.

Ryan, B., Scapens, R. W., & Theobald, M. (2002). *Research Method & Methodology in Finance & Accounting* (2<sup>nd</sup> Ed.). Thomson, London.

Saunders, M., Lewis, P., & Thornhill, A. (2009). *Research Methods for Business Students*, (5<sup>th</sup> Ed.). Harlow, Pearson Education.

Shekharan, U., & Bougie, R. (2010). *Research Methods for Business: A Skill Building Approach* (5<sup>th</sup> Ed.). New Delhi: John Wiley.

Sim, J., & Wright, C. (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation and Sample Size Requirements. *Physical Therapy*, 85(3), 257–268.

Singh, A. S. (2014). Conducting Case Study Research in Non-Profit Organisations. *Qualitative Market Research: An International Journal*, 17, 77–84.

Sperry, L. (2004). *Assessment of Couples and Families: Contemporary and Cutting Edge Strategies* (1<sup>st</sup> Ed.). New York, NY: Routledge.

Straub, D. W. (1989). Validating Instruments in MIS Research. *MIS Quarterly*, 13(2), 147-169.

Swamy, P. A. V. B., Hall, S. G., Tavlas, G. S., & von zur Muehlen, P. (2017). On the Interpretation of Instrumental Variables in the Presence of Specification Errors: A Reply, 5(32), 1-3.

Tashakkori, A., & Teddlie, C. (1998). *Mixed Methodology: Combining Qualitative and Quantitative Approaches*. Thousand Oaks, CA: SAGE.

Tavakol, M., & Dennick, R. (2011). Making Sense of Cronbach's Alpha. *International journal of Medical Education*, 2, 53-55.

Taylor, J. R. (1999). *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. University Science Books.

Thatcher, R. (2010). Validity and Reliability of Quantitative Electroencephalography. *Journal of Neurotherapy*, 14, 122-152.

Thompson, B. (2003). Understanding Reliability and Coefficient Alpha, Really. In B. Thompson (Ed.). *Score Reliability: Contemporary Thinking on Reliability Issues*. Thousand Oaks, CA: SAGE.

Traub, R. E., & Rowley, G. L. (1991). An NCME Instructional Module on Understanding Reliability. *Educational Measurement: Issues and Practice*, 10(1), 37-45.

Turner, S.P. (1979). The Concept of Face Validity. *Quality and Quantity*, 13(1), 85-90.

Twycross, A., & Shields, L. (2004). Validity and Reliability-What's it All About? Part 2: Reliability in Quantitative Studies. *Paediatric Nursing*, 16 (10), 36.

Waltz, C., Strickland, O., & Lenz, E. (2004). *Measurement in Nursing and Health Research*. New York: Springer Publishing.

Willis, J. (2007). *Foundations of Qualitative Research: Interpretive and Critical Approaches*. SAGE Publications.

Wilson, J. (2010). *Essentials of Business Research: A Guide to Doing Your Research Project*. SAGE Publications.

Yarnold, P. R. (2014). How to Assess the Inter-Method (Parallel-Forms) Reliability of Ratings Made on Ordinal Scales: Emergency Severity Index (Version 3) and Canadian Triage Acuity Scale. *Optimal Data Analysis*, 3(4), 50-54.

Yoshida, S., Matsushima, M., Wakabayashi, H., Mutai, R., Murayama, S., Hayashi, T., Ichikawa, H., Nakano, Y., Watanabe, T., & Fujinuma, Y. (2017). Validity and Reliability of the Patient Centred Assessment Method for Patient Complexity and Relationship with Hospital Length of Stay: a Prospective Cohort Study. *BMJ Open*, 7 (e016175), 1-8.

Zohrabi, M. (2013). Mixed Method Research: Instruments, Validity, Reliability and Reporting Findings. *Theory and Practice in Language Studies*, 3(2), 254-262.