GRADIVA

Volume 62 | Issue 9 | September 2023
DOI: 10.5281/zenodo.8398633

# Validity and Reliability: What Are They, How Are They Measured and When Are They Used?

Oke James Ajogbeje

Department of Counselling Psychology, Bamidele Olumilua University of Education,
Science and Technology, Ikere Ekiti, Ekiti State.
ORCID ID: https://0000-0003-4266-3657, Email: ajogbeje.oke@bouesti.edu.ng, ojajogbeje@gmail.com

## Abstract

This paper describes various formats and examines some of the issues associated with validity and reliability measurement methods and their various usages in educational research. In most cases, the accuracy and consistency of research instruments form an important part of research methodology. Some researchers, especially newcomers, are at a loss about choosing and implementing the appropriate and acceptable validity and reliability for their research instruments. A result may be reliable but not applicable to what the researchers have in mind. Reliability is very important in educational research, but it is not enough without validity. Research instruments are reliable only if they are valid. The paper suggests that reliability and validity are applicable to both qualitative and quantitative educational research.

**Keywords:** *Reliability, Validity, Instrument, Educational Research.*

## 1. INTRODUCTION

Research is a process of problem solving, or a way of solving tricky problems to push forward the frontiers of ignorance and knowledge (Bandele, 2004). A systematic investigation or research process involves identifying and defining a problem, formulating hypotheses, generating, organizing, analyzing, and interpreting data, drawing conclusions and taking decisions about such problems (Bandele, 2004). Educational research can be broadly divided into two categories, namely: (i) quantitative research and (ii) qualitative research (Bandele, 2004; Cohen et al., 2017; Oluwatayo, 2012; Ruane, 2016).

Quantitative research explains and understands phenomena through objective measurements and statistical analysis of numerical data, whereas qualitative research explains and understands phenomena from the perspective of the research participants. Oluwatayo (2012), suggests that there is an overlap in educational research and that there are different measurement tools available to researchers conducting quantitative and qualitative research. Educational tools used in research must meet two critical criteria: validity and reliability. Validity and reliability concepts are explained in this paper, along with their measurement methods and their various usages in educational research. Furthermore, the significance of validity and reliability in educational research was discussed in detail.

## 2. CONCEPT OF VALIDITY

Defines validity as a measure what researchers intended measure, everything they want to measure, and only what they want to measure. While Field (2010), argues that validity actually means `measuring a measure', Similarly, Kaplan & Saccuzzo (2017) see validity as evidence of conclusions drawn from the instrument, while McBurney (2007) see validity as the degree of consistency of research findings with reality. The various forms of validity used to

Volume 62 | Issue 9 | September 2023
DOI: 10.5281/zenodo.8398633

evaluate and derive the validity of the collected data or to support the interpretation of test results include content validity, face validity, construct validity, and criterion-related validity.

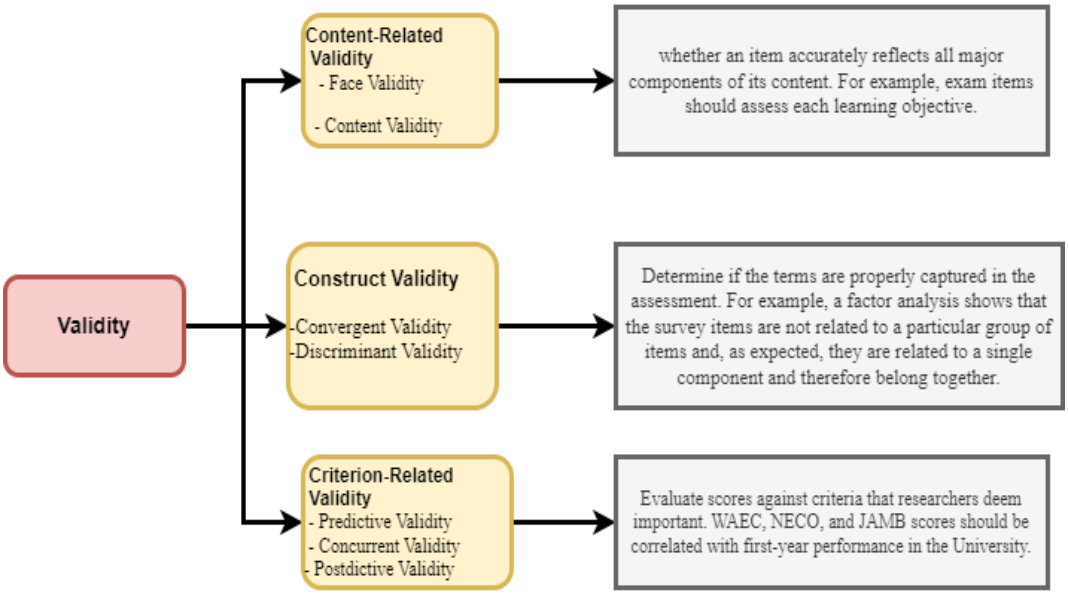


**Figure 1: The Various Forms of Validity Tests**

From the various definitions given above, the researcher is of the opinion that validity now answers to the question. Does the assessment provide meaningful information and relevant inferences based on scores obtained from the research instrument used? Figure 1 shows the various forms of validity testing discussed in this article. Several authors, psychologists, and researchers have identified different types of validity used in educational research (Andrews, 1984; Bandele, 2004; Bowling, 2014; Evans et al., 2021; Joshua, 2005; Kolawole, 2010; Oluwatayo, 2012; Sireci, 1998; Taherdoost, 2016). These validities include face, content, criterion-related, construct, convergent, predictive, and concurrent validity, as well as jury, external, consequential, descriptive, interpretive, internal, and evaluative validity. Other types of validity used in educational research include known-group validity (Taherdoost, 2016), as well as curricular and instructional validity (Kolawole, 2010). The most commonly used validities in educational research are content-related, criterion-related, and construct validity.

## 2.1 Face Validity

Face validity is the degree to which a research tool seems to measure what it purports to measure. In terms of whether the items are relevant, clear, appropriate, and well-defined, it assesses the presentation and applicability of the research instrument. According to Oluwatayo (2012), face validity is evaluated in terms of readability, feasibility, uniformity of style and format, and clarity of the language utilized. Subject experts and judges are usually asked to ascertain the face validity of a research instrument (Taherdoost, 2016). According to Kaplan & Saccuzzo (2017) and Whiston (2005), research instruments can appear valid even if they do not measure what they claim to measure. In addition, instruments developed for children can be criticized for lack of demonstrable validity when applied to adults. Individual performance can be quantified by assessing the suitability of a measurement tool for its intended use with the participation of research professionals. The yes or no, indicating favorable or unfavorable items, may be used as categorical options to test face validity. Data collected are subjected to

Volume 62 | Issue 9 | September 2023
DOI: 10.5281/zenodo.8398633

statistical analysis using Cohen's Kappa Index (CKI) to determine instrument validity and a CKI of 0.60 is recommended as a minimum acceptable value for face validity.

Based on the various definitions and arguments above, face validity is one aspect of validity that researchers conducting quantitative and qualitative studies should be aware of when reporting the validity procedures of a research instrument. Also, simply reporting that a researcher has provided a research tool to a supervisor or subject matter expert is not sufficient to determine face validity. The comments of judges and subject experts on the above criteria must be submitted in order to make the study meaningful.

### 2.2 Content Validity

In order to guarantee that a research instrument is successful in assessing what it is designed to measure, content validity is a critical component in educational research. According to Oluwatayo (2012), content validity refers to how well a research tool covers the range of variables it is intended to measure. Babbie (2007) defined content validity as the degree to which a measure adequately reflects the range of meanings associated with a concept. Similar to this, Cohen et al. (2017) define content validity as a subtype of validity that ensures the components of the research instrument chosen are an accurate representation of the issue under study and are addressed in depth and breadth.

Therefore, maintaining content validity is essential for the precision and dependability of research results. Examining whether a research instrument contains components that are accurate representations or examples of all possible content is the subject of evidence for the validity of content in educational research. Therefore, careful selection of topics is a basic requirement for securing representatives. Messick (1989) suggests that item organization and read-level matching should be considered by test designers as part of content validity.

There is sufficient evidence in the literature that judgments of evidence for the adequacy of content are often based on expert judgments, and there are several ways to combine such expert judgments into content representation indices (Oluwatayo, 2012). Here's how. 1. Multiple judges or group rankings; 2. Statistical method and 3. Table of specifications (TOS) method.

In an assessment or panel approach, researchers consult subject matter experts to assess the consistency or relevance of each element of a content tool(Rubio et al., 2003). Revisions, suggestions, or changes by experts should be reflected in the final presentation of the study. A literature review and follow-up evaluation by an expert judge or panel is required to determine the adequacy of the content.

Although it is vital to have researchers and experts on hand for new instrument validation, it is not always feasible to have a lot of experts in one place for such an assignment. One of the challenges to the validation of survey instruments is when specialists are spread over different geographical areas (Choudrie & Dwivedi, 2005).

However, researchers can send content evaluation tools to experts working in diverse locations by following the steps outlined below (Taherdoost, 2016). a. A thorough review of the literature to glean pertinent information; b. A content validation tool is created; c. Experts in the same field of study should be consulted; d. A content validity ratio (CVR) is determined for each item using a formula created by Lawshe in 1975 (Lawshe, 1975), and e. insignificant components are eliminated at the critical level.

Volume 62 | Issue 9 | September 2023
DOI: 10.5281/zenodo.8398633

$$CVR = \frac{n_e - \left(\frac{N}{2}\right)}{\frac{N}{2}}$$

Where $n_e$ is the number of panelists who have been given the designation "important", N is the overall number of panelists, and CVR is the content validity ratio? Table 1 displays the Lawshe Table for Minimum Content Validity Ratio CVR One-Tailed Test Values. P = .05

**Table 1: Minimum Content Validity Ratio CVR Lawshe Table**

| Number of Panelists | Minimum Value |
|---|---|
| 5 | 0.99 |
| 6 | 0.99 |
| 7 | 0.99 |
| 8 | 0.75 |
| 9 | 0.78 |
| 10 | 0.62 |
| 11 | 0.59 |
| 12 | 0.56 |
| 13 | 0.54 |
| 14 | 0.51 |
| 15 | 0.49 |
| 20 | 0.42 |
| 25 | 0.37 |
| 30 | 0.33 |
| 35 | 0.31 |
| 40 | 0.29 |

The number of panels determines the ultimate score for keeping items based on CVR. Guidelines for valid CVR values for the evaluated items are given in Table 1 above. The most commonly used statistical method is factor analysis. It determines whether instrument items are tied to a single factor, fit within the conceptual space, hang as predicted, or are unrelated to another particular set of items (Sireci, 1998), According to Notar et al., (2004) and Fives & DiDonato-Barnes (2013), a table of specifications (TOS), often referred to as a testing blueprint, is a diagram that helps classroom teachers coordinate objectives, directions, activities, and assessments.

Given the cognitive level and proportional importance of each content area in the learning process, Joshua (2005) compared TOS, which offers instructions for making items, with a test blueprints. Onunkwo (2002) describes the TOS as a two-dimensional diagram that lists the learning objectives to be tested as columns and the learning content as rows. In order to construct a test or research instrument that accurately represents its content and objectives, a TOS must be created.

To construct an effective TOS, Joshua (2005) and Oluwatayo (2012) recommend the following steps:

- Determine the objectives of the content areas to be covered and the total number of items that will constitute the test or instrument.
- Specify the behavioral, affective, or cognitive changes that the researcher intends to evaluate, and decide how many items will be provided at each domain level.

Volume 62 | Issue 9 | September 2023
DOI: 10.5281/zenodo.8398633

- The alignment of the content areas and behavioral domains serves as guidance for the researcher in presenting the instrument's intended purpose(s) in a straightforward manner.
- The researcher must take into account the degree of item complexity, the respondents' proper reading ability, and the type of individuals for whom the test or instrument is being developed (Kane, 2001).

### 2.3 Construct Validity

According to Taherdoost, 2016 and Walden (2012). , the construct validity of a concept, an idea, or an action measures how well it translates into an operational and functional reality. These instruments are frequently utilized in educational research since they are based on logical correlations between variables. Convergence and divergence approaches are two ways to determine construct validity(Kerlinger, 1979)

### 2.4 Convergent Validity

Convergent validity is the degree to which two measures of a construct that are theoretically predicted to be related are actually related. Convergent validity, as defined by Bell et al., (2022), Brock-Utne (1996) and Campbell & Fiske (1959)entails the correlation of instrument results with results from comparable variables. They argued that a high correlation coefficient indicates the validity of new instrument constructs. Cross-correlations from multitrait-multimethod matrices, according to Campbell & Fiske (1959) and Oluwatayo (2012), could be used to support convergence validity. The theory behind each of these validation approaches is that findings from many ways of assessing the same construct should be comparable.

### 2.5 Divergent Validity

When measurements of other components using comparable techniques exhibit relatively minimal cross-correlation, this is known as divergent validity. In other words, the configuration under examination should be distinct from other configurations of a similar nature. According to Taherdoost (2016), divergent validity refers to how a latent variable A differs from other latent variables (B, C, D, etc.). This indicates that considerable disparities between linked observable variables and other constructs within the conceptual framework can be explained by latent variables. Cohen et al., (2017), Koh & Nam, (2005) and Soo Wee & Quazi (2005) measure divergent and convergent validity using factor analysis. Straub et al (2004) had earlier claimed that the obtained factor analysis results were consistent with convergent validity (eigenvalue 1, loading $\geq 0$), divergent validity (loading $\geq 0.40$, cross-loading $\geq 0.40$), and both. Item cross-loadings of 0.40 are the minimum specified value in research studies.

### 2.6 Criterion-related validity

The degree to which an instrument correlates with an outcome and assesses how well one instrument predicts the outcome of another instrument is known as criterion-related validity. Taherdoost (2016), reports that test users can use the test to discriminate between groups or predict future outcomes. Criterion-related validity serves as an alternative to undermining the conceptual meaning or interpretation of the test.(Bowling, 2014) reports that criterion-related validity involves calculating the correlation coefficient between measures of the instrument under construction and measures of other criteria found to be important. Therefore, it is necessary to obtain a high correlation coefficient between the value of one instrument and the value of another existing instrument to be considered reliable. Bandele (2004), Cohen et al.

Volume 62 | Issue 9 | September 2023
DOI: 10.5281/zenodo.8398633

(2017), Evans et al. (2021) and Whiston (2005) distinguish between concurrent and predictive validity as two types of criterion-related validity, and Taherdoost (2016), distinguishes between concurrent, predictive, and postdictive validity as three types of criterion-related validity.

### 2.7 Concurrent Validity

The degree to which the results of an instrument agree with a given measurement on the same construct is called concurrent validity. It deals with the relationship between two simultaneous measures. One is for verification only, and the other is for reference. Kaplan & Saccuzzo (2017) and Kolawole (2010)report that concurrent measurements and baseline measurements should be both valid and provide useful diagnostic information to guide the educational development of learners. In other words, the external criteria are determined almost simultaneously with the test or instrument results and subjected to correlation analysis to obtain concurrent validity coefficients. The coefficients obtained are expected to be high, significant and positive to qualify a test or an instrument as valid (Bandele, 2004). Bandele (2004) further explains that obtaining a high or low validity coefficient depends on several factors such as sample size, and representativeness. For example, how confident researchers are that the standard measures actually measure the same construct or traits as the instrument being tested?

### 2.8 Predictive Validity

The ability of an instrument to predict future events is called predictive validity (Alonge, 2004) . The factor that differentiates predictive validity from concurrent validity is the time lag. The measure of the criterion is assumed to be obtained after a period of time and should not be too close to the characteristics of the instrument being tested for validity. So, predictive validity is fundamentally about the ability and extent of a test or instrument to predict future measurements. For example, if the cognitive entry characteristic (WAEC, NECO, or JAMB) scores of polytechnic students correlate strongly with their National Diploma (ND) or Higher National Diploma (HND) examination results, then one can conclude that the cognitive entry characteristic demonstrated strong predictive validity ( Ajogbeje, Oke James, 2010; Ajogbeje, Oke James et al., 2013; Ajogbeje, Oke James & Adewale, 2012; Ajogbeje, Oke James & Tunde, 2013). By comparing the prediction score to the reference score, the predictive validity ratio is calculated. The best method for explicitly demonstrating predictive validity is through long-term validity studies; however, these studies take a long time and need a very large sample size in order to collect useful aggregate data (Taherdoost, 2016).

### 2.9 Postdictive Validity

Postdictive validity is a type of criterion-related validity that determines how closely a given instrument's outcomes correlate with those of a different previously run instrument or criterion. The criteria for this form of validity include measurements that were acquired in the past.

In general, criterion-related validity coefficients can be determined using three methods: correlation methods, regression methods, and decision theory or group separation methods (Oluwatayo, 2012). The direction and size of the relationship between a measure and a base measure are determined using correlation techniques. Perform the following procedure to validate your device. 1. Select a suitable group for validation study. 2. Manage your instrument. 3. Collect and calculate the correlation coefficient between the instrument's baseline and measurements. If the researcher is interested in concurrent validity, the calculated correlation is the validity coefficient. This validity coefficient is squared to provide the coefficient of determination, which represents the proportion of base measure variance that the instrument

Volume 62 | Issue 9 | September 2023
DOI: 10.5281/zenodo.8398633

takes into account. It tells researchers how much variation is shared between the two variables. However, if the researcher is interested in predictive validity, which involves intervals, then the researcher would have to wait for the appropriate time for the collection of criterion measures and thereafter compute the validity coefficient.

The regression method is based on the assumption that a regression line can be used to describe the relationship between baseline and instrument scores. The instrument's score is plotted relative to the original score, and a regression line is obtained. The resulting regression line can be used to predict test scores using instrument scores. Decision theory, also known as group separation, is primarily concerned with whether subjects who score high on the instrument meet the expected criteria. Novick (1965) report that decision theory helps successful applicants get into college by taking available information and converting it into mathematical form to provide guidelines for selection and placement. Reportedly, this helps determine whether or not a student will graduate from college. The question here is how large the concurrent validity should be. According to Kaplan & Saccuzzo (2017), validity coefficients seldom reach 0.60 and frequently range from 0.30 to 0.40.

## 3. THE RELIABILITY CONCEPTS

Measures of consistency, accuracy, precision, stability, reproducibility, dependability, repeatability, and replication across components of a measuring device are referred to as reliability (Bandele, 2004; HUCK, 2007).. It is focused on how consistently and reliably observations of phenomena produce results (Carmines & Zeller, 1979). An instrument or test is deemed reliable, according Moser & Kalton (2017), if repeated measurements made with the instrument under the same circumstances produce the same findings. Reliability has diverse meanings in qualitative and quantitative research. According to (Bowling, 2014), consistency and reproducibility across time, means, and groups of respondents are comparable to reliability in quantitative research. Within a specified margin of error, the instrument should yield the same results when employing the same procedure on the same sample (Cohen et al., 2017). An instrument must be able to show that it yields comparable results when repeated in similar groups under similar conditions in order to be considered reliable.

In qualitative research, reliability is considered the best match between what researcher records in the data and what occurs in the natural environment under investigation (Oluwatayo, 2012). Brock-Utne (1996) also opined that, in qualitative research, researchers tend to capture different interpretations of the intent and meaning given to situations and events. Bodgan & Biklen (1982) report that qualitative research is about accuracy and completeness rather than consistency. They further explained that two researchers working on a construct or concept in one of those settings could produce different results, even though the two results were found to be reliable. LeCompte et al. (1993) made the case that reliability concept used in quantitative research might not be applicable to qualitative research. Variable manipulation can alter how naturally occurring occurrences occur.

Volume 62 | Issue 9 | September 2023
DOI: 10.5281/zenodo.8398633

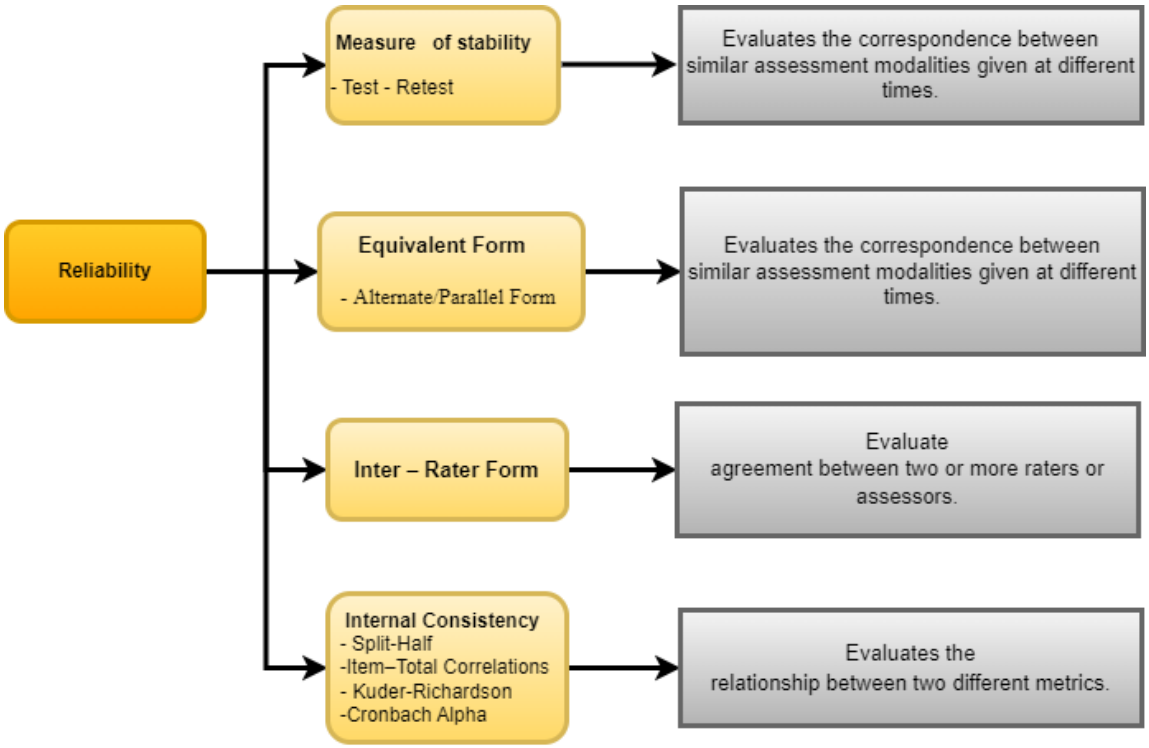


**Figure 2: The Various Forms of Reliability Tests**

It should be emphasized nonetheless that there are four main types of reliability in educational research, including test-retest, equivalence form, inter-rater, and internal consistency (i.e., split-half, Kuder-Richardson, and Cronbach alpha). The main goal of reliability is to establish whether an instrument or a specific technique produces the same results when used repeatedly on the same sample, as is evident in Figure 2 above.

### 3.1 Stability Measurement

When comparing instrument consistency across time between similar samples, this is the method that is most frequently used to estimate reliability coefficients (Cohen et al., 2017). It is assumed that reliable instruments will yield comparable results over time from comparable respondents. Typically, the test-retest method is used to estimate stability metrics.

### 3.2 Reliability Test-Retest

The instrument must be used twice during intervals when the target sample does not change in order for the test-retest to be reliable (Collins, 2007). This form of reliability testing can be performed through questionnaires, interviews, and observational probing techniques. Test-retest reliability presupposes that the measured true score remains constant over a short period of time, as does the relative location of an individual's score in the population distribution (Revelle & Condon, 2019). To assess the reliability of a test-retest, the correlation coefficient of values recorded at two different times is frequently utilized. A higher correlation between the values of the two instruments indicates greater stability or test-retest reliability over time (Shou et al., 2021). Conceptually, test-retest reliability is a good measure of score consistency because it can directly measure consistency from administration to administration. Weir (2005)states that the intraclass correlation coefficient (ICC) can be used to assess test-retest reliability. He further reported that the ICC can be defined as "the between-subject variability divided by (the between-subject variability plus the positive error)".

Volume 62 | Issue 9 | September 2023
DOI: 10.5281/zenodo.8398633

$$Intraclass\ Correlation\ Coefficient\ (ICC) = \frac{\text{Between subjects' variability}}{(\text{Between subjects' variability} + \text{Error})}$$

The ICC increases from 0 to 1, signifying perfect reliability as the error term diminishes. According to Fleiss quoted in Oremus et al. (2012), there is no acceptable general consensus on how to interpret the ICC value. He then provided the following test-retest reliability strength classification based on the ICC value: poor is ICC < 0.40. ICC > 0.75 is regarded as exceptional, whereas 0.40 to 0.75 (0.40 < ICC < 0.75) is regarded as medium to good. The standard error of measurement (SEM) is what determines absolute reliability, according to Bruton et al. (2000). The absolute term divergence becomes more reliable when more measurements are made. The SEM and ICC of an instrument are inversely related. That is, if the SEM is zero, then there is no measurement error, but if the instrument is completely reliable then ICC = 1.0 (Harvill, 1991)..

According to Haynes et al. (2018), the use of test-retest reliability depends either on the construct's temporal dynamics or the duration of the time gap. For example, moods can change in a short amount of time, and if the rate of change varies from person to person, people's ranks in the distribution can also change over time. Applying and interpreting the test-retest reliability of these systems will be difficult as a result. Second, it is assumed that the research instrument's two acts are identical and independent and that individual scores remain consistent over time in order to understand the test-retest reliability coefficient (James & Tunde, 2013). Unfortunately, these assumptions are not achievable under real test conditions. Third, the respondent's memory and practice effects in the first and second examinations may have a negative impact on the independence of dual instrument administrations. These two effects are dependent on the period between instrument administrations and can differ from person to person. The application and interpretation of the test-retest reliability coefficient require a thorough analysis of the theoretical and practical issues involved.

### 3.3 Equivalence Form Reliability

There are two approaches to estimating the reliability of the equivalence form: the alternate or parallel form and the inter-rater form.

### 3.4 Alternate or Parallel-Form Reliability

The alternate or parallel reliability approach attempts to address some of the key issues associated with the test-retest approach, such as: long-term stability and testing methods or test-wiseness. An alternate reliability estimation method creates two sets of instruments that are equivalent in terms of content coverage, test specification, question format, difficulty level, and characteristic scale, and time. A correlation coefficient obtained by calculating the relationship between the scores of two parallel instruments gives the reliability coefficient. As an alternative, two instruments might be used simultaneously on two homogeneous groups (Bandele, 2004; Oluwatayo, 2012).The results obtained are compared using either the Pearson statistic or the t-test statistic, and equivalent results have a correlation coefficient of at least 0.80 (Bowling, 2014)Reliability measures both the stability over time and the consistency of responses to different samples of an item. It also shows the distribution of errors due to content sampling. The capacity of researchers to produce collections of items that reflect the same concept, which is challenging to do, is a significant flaw in this method.

Volume 62 | Issue 9 | September 2023
DOI: 10.5281/zenodo.8398633

### 3.5 Inter-Rater Reliability

Examining inter-rater reliability is one of the best techniques to assess reliability when a measurement involves several ratters or observers. It is a measure of consistency used to assess how well different judges agree on scoring decisions. Inter-rater reliability is essential for research decision-making. Given that it focuses on how much the scores obtained by two or more ratters agree in proportionate amounts, it is the most straightforward sort of reliability to understand and is commonly employed in games and sports. This approach, according to Oluwatayo (2012), is ideal for research teams gathering structured observational or semi-structured interview data. In this situation, each team member must agree on the data to be entered into various categories. He also reported that the issue of reliability of observational data was addressed in the researcher's training and familiarization with the data so that the data could be entered immediately. The simplest level of consensus calculation between participants is to use percentages (Bowling, 2014) by taking the following steps: 1. Count the number of matching ratings. 2. Calculate the total number of reviews. 3. Divide the amount by the appropriate number to get a fraction. 4. Convert fractions to percentages.

Cohen's kappa formula can be used to determine inter-rater reliability (IRR).

$$\kappa = \frac{p_a - p_\varepsilon}{1 - p_\varepsilon}$$

*Where* $P_a$ = the percentage of observations that match and agree, and $P_s$ = the rate of random matches. Alternatively,

$$\kappa = \frac{n_a - n_\varepsilon}{n - n_\varepsilon}$$

Where $n$ = *the* number of subjects, $n_a$ = the number of matches, and $n_s$ is the number of random matches. Inter-rater reliability coefficient is acceptable if the computed IRR coefficient is **≥ 75%**, moderately acceptable if IRR coefficient is 50% < IRR < 75%, and unacceptable if IRR is < 50%.

### 3.6 Internal Consistency Reliability Measurement

Using internal consistency measurements, the uniformity or homogeneity of the items in an instrument is frequently assessed. It refers to the degree to which objects connected to a specific instrument construct employ that construct exclusively (Bowling, 2014). The suitable statistical techniques are item-total correlations, split-half, Cronbach alpha, and Kuder-Richardson 20 and 21. These are used to determine the internal consistency reliability coefficient after the instruments have been given to selected samples only once. The instrument is deemed to have good internal consistency if the items are "related" and evaluate the same concept (HUCK, 2007; Robinson, 2010).

### 3.7 Split-half Reliability

In split-half reliability, the content and overall complexity of research instrument items can be divided into two interconnected components. This is done at random by either assigning all items with the same construct to one of the two groups or by assigning all odd and even items to different groups. The results obtained from one half of the instrument are expected to match the results obtained from the second half of the instrument. Pearson statistic is then used to compare odd-numbered item scores against even-numbered item scores to determine reliability coefficients. The split-half reliability coefficient has the disadvantage of using just

Volume 62 | Issue 9 | September 2023
DOI: 10.5281/zenodo.8398633

half of the items, which lowers the reliability coefficient. However, the Spearman-Brown correction method can be used to provide a reliable estimate of the complete test.

$$\rho = \frac{2r}{1+r}$$

Where $\rho$ is the score ratio in the first half and $r$ is the correlation between the instrument halves. This Spearman-Brown adjustment version performs perfectly when the two halves are of equal length. If $r \neq \pm 1$, we might instead use the following formula:

$$\rho = \frac{r\left(\sqrt{r^2 + 2c(1-r^2)} - r\right)}{c(1-r^2)}$$

Where $\rho$ = the proportion of tests by the first half and $c = 2\rho\ (1–p)$ respectively. When adopting split-half reliability, Bryman & Cramer (1992) indicate that reliability coefficients of 0.80 or higher are appropriate.

### 3.8 Item-Total Correlation Reliability

The correlation between a single item and the entire score excluding that item is measured by item-total correlation reliability. In other words, it describes how closely the scores for each item on an instrument correlate with the score for the entire instrument. As an example, one may calculate the correlation between item 1 and the total of the other 24 items, and so on. Hence, if a researcher has a 25-item research instrument, the total correlations are 25. An item-total correlation reliability test is usually performed to determine if an item within a test or research instrument matches or does not match other items and decide whether to discard or retain the item. Respondents who answered the question correctly are expected to have a higher total score than those who answered the question incorrectly. This relationship tends to indicate how well the question identifies or distinguishes between respondents who are familiar with the material and those who are not. Respondents who have learned what they are taught are expected to perform very well on the item or question and achieve very high overall assessment scores, and vice versa. Therefore, item-total correlation is a useful way for researchers to see if any of the items have no response and assess the performance of the item as it varies in line with the performance of other items across the population. Item-total correlation reliability across elements can be calculated using spreadsheets, statistical software, or manually. Manual calculation includes the following steps:

Step 1: Add the scores for each item to determine each person's overall score.

Step 2: Subtract the first item's score from each person's total.

Step 3: Correlate the score of the first item with the score calculated in step 2 to get the overall item correlation for item 1.

Step 4: For each additional item, follow steps 2 and 3 once more.

According to Streiner, G. I. & Norman (2003), in order to satisfy the reliability and scaling requirements, item-total correlations must be correlated with total scale scores of 0.20 or higher. Point-Biserial correlation coefficients are recommended for instruments having dichotomous response items, such as yes or no, agree or disagree, and true or false. The product-moment correlation coefficient is recommended for questions with two or more responses, such as "strongly agree," "agree," "disagree," and "strongly disagree" (Kline, 2013; Oluwatayo, 2012). According to Clark & Watson (2016), it is possible to evaluate how discriminatory a question is by looking at the item-total correlation's (Point-Biserial) value. (i)

Volume 62 | Issue 9 | September 2023
DOI: 10.5281/zenodo.8398633

A question is bad if its value falls between 0 and 0.19. (ii) Values above 0.4 demonstrate exceptionally strong discrimination; (iii) Values between 0.2 and 0.39 suggest strong discrimination.

### 3.9 Kuder-Richardson $KR_{20}$ and $KR_{21}$ Reliability

Kuder & Richardson (1937) developed a formula for determining the uniformity of items. The Kuder-Richardson KR_20, which is based on the ratio of correct to incorrect answers for each test item, is the most well-known measure of homogeneity. The Kuder-Richardson KR_ (20) formula is given as

$$KR_{\_20} = \frac{k}{k-1}\left\{1 - \frac{\sum p_j q_j}{\sigma^2}\right\}$$

Where $k$ = total number of questions, $p_j$ = the percentage of individuals who answered question j correctly, $q_j$ = the percentage of individuals who answered question j incorrectly, and $\sigma^2$ = the variance scores for all individuals who took the test.

The Kuder-Richardson KR_21 is a shortened form of the Kuder-Richardson KR_20 and is defined as follows when the questions on a test are of comparable difficulty (i.e., the mean score of each question is roughly equal to the mean score of all the questions):

$$KR_{21} = \frac{K}{K-1}\left\{1 - \frac{\mu(K-\mu)}{K\,\sigma^2}\right\}$$

Where k = number of items or questions, $\mu$ = population mean scores, and $\sigma^2$ = variance of the total scores of all the individuals.

Ary, D, Jacobs, L. C. & Razavich (2002) state that the $KR_{21}$ method takes less time than all reliability estimation methods because it uses available information and requires only one test run. The values are between 0 and 1. A high score denotes reliability, whereas an excessively high value (above 0.90) denotes a homogeneous test, which is typically undesirable. Note that KR21, when compared to KR20, often underestimates a test's reliability.

### 3.10 Cronbach Alpha or Alpha Coefficient Reliability

The Cronbach alpha coefficient is the most widely utilized internal consistency measurement in educational research. It evaluates the reliability of multiple-question surveys using the Likert scale. Cronbach's alpha also reveals if the instrument developed by the researcher accurately measures the relevant variable. It is also thought to be the most appropriate measure of reliability when using instruments that assess items using a range of values, such as the Likert scale. This is because we take into account the variance of each item. Cronbach's alpha, according to Ruane (2016), is the best for the assessment of reliability index when the instrument or test items are heterogeneous, i.e. when measuring multiple traits or attributes. Cronbach's alpha (α) coefficient is a well-known formula and it is expressed as:

$$\alpha = \frac{k}{k-1}\left\{1 - \frac{\sum S_y^2}{S_x^2}\right\}$$

Where $k$ = the number of test items, $s_x^2$ = the variance of total score, and $\sum S_y^2$ = sum of the item variance.

Note that there are no absolute rules for internal consistency in educational research. Robinson (2010) and Whitley et al. (2013) agreed on a reliability factor of 0.70 as the minimum

Volume 62 | Issue 9 | September 2023
DOI: 10.5281/zenodo.8398633

internal consistency factor. Straub et al. (2004) recommend that the reliability coefficient be greater than or equal to 0.60 for exploratory or pilot investigations. Hinton et al. (2004) suggest the following interpretation of Cronbach alpha for dichotomous or Likert scale questions: Excellent is defined as 0.90 and above, good as 0.80 to 0.89, acceptable as 0.70 to 0.79, questionable as 0.60 to 0.69, poor as 0.50 to 0.59, and unacceptable as 0.50 and below.

## 4. CONCLUSION

Validity and reliability issues must be dealt with in educational research. To enhance researchers' knowledge and research outcomes, it was discussed how to measure and compute the validity and reliability of research instruments and tests, the concept of validity and reliability, various types of validity and reliability, and how to use validity and reliability in educational research. Valid findings might not, however, be applicable to the application in question. Validity requires reliability, but reliability alone is not enough. The findings must first be at least somewhat reliable in order for them to be considered valid (Thorndike, 2005; Testing, 1999). In summary, reliability is crucial in educational research, but it is insufficient in the absence of validity. Tests are only valid if they are reliable.

### Contributions of the authors:

The author is 100% responsible for the contributions.

**Financing:** The research is auto financed by the researcher.

### Conflicts of interests

The author declares not having any conflicts of interest related to the publication of this article.

### References

1) Ajogbeje, J. O. (2010). Self-concept as predictor of mathematics achievement among secondary school students in Ado-Ekiti, Nigeria. *Nigerian Journal of Guidance and Counselling*, *15*(1).

2) Alonge, M. F. (2004). Measurement and evaluation in education and psychology. *Ado Ekiti: Adedayo Printing (Nigeria) Limited*.

3) Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly*, *48*(2), 409–442.

4) Ary, D.; Jacobs, L. C. & Razavich, A. (n.d.). *Introduction to research in education (6th ed.). Wadsworth Thomson Leaving.*

5) Babbie, E. (2007). Tile practice of social research. *Istanbul Bilgi University Library*.

6) Bandele, S. O. (2004). Educational research in perspective. *Ibadan: Niyi Commercial and Printing Ventures*.

7) Bell, E., Bryman, A., & Harley, B. (2022). *Business research methods*. Oxford university press.

8) Bodgan, R. C., & Biklen, S. K. (1982). *Qualitative Research For Education: An Introduction to Theory and Method. Boston: Ally &Bacon*. Inc.

9) Bowling, A. (2014). *Research methods in health: investigating health and health services*. McGraw-hill education (UK).

Volume 62 | Issue 9 | September 2023
DOI: 10.5281/zenodo.8398633

10) Brock-Utne, B. (1996). Reliability and validity in qualitative research within education in Africa. *International Review of Education*, *42*, 605–621.

11) Bruton, A., Conway, J. H., & Holgate, S. T. (2000). Reliability: what is it, and how is it measured? *Physiotherapy*, *86*(2), 94–99.

12) Bryman, A., & Cramer, D. (1992). Quantitative data analysis for social scientists. *Estudios Geográficos*, *53*(207), 347.

13) Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81.

14) Carmines, E. G., & Zeller, R. A. (1979). Reliability and validity assessment Newbury Park. *Cal.: Sage Publications*.

15) Choudrie, J., & Dwivedi, Y. K. (2005). *Investigating broadband diffusion in the household: towards content validity and pre-test of the survey instrument*.

16) Clark, L. A., & Watson, D. (2016). *Constructing validity: Basic issues in objective scale development.*

17) Cohen, L., Manion, L., & Morrison, K. (2017). *Research methods in education*. routledge.

18) Collins. (2007). *L.M. in Encyclopedia of Gerontology (Second Edition)*.

19) Evans, C., Kandiko Howson, C., Forsythe, A., & Edwards, C. (2021). What constitutes high quality higher education pedagogical research? *Assessment & Evaluation in Higher Education*, *46*(4), 525–546.

20) Field, A. (2010). Discovering Statistics Using SPSS, Sage Publications Inc. *Thousand Oaks, CA*.

21) Fives, H., & DiDonato-Barnes, N. (2013). Classroom test construction: The power of a table of specifications. *Practical Assessment, Research, and Evaluation*, *18*(1), 3.

22) Harvill, L. M. (1991). Standard error of measurement: an NCME instructional module on. *Educational Measurement: Issues and Practice*, *10*(2), 33–41.

23) Hinton, P. R., Brownlow, C., McMurray, I., Cozens, B., & SPSS, E. (2004). Routledge Inc. *East Sussex, England*.

24) HUCK, S. W. (2007). *Reading Statistics and Research, United States of America, Allyn &amp*. Bacon.

25) Ajogbeje, O. J. & Ojo, A. A. (2012). Relationship between senior secondary school student's achievement in mathematical problem solving and intellectual abilities tests. European Scientific Journal. 8(15), 169-179. http://www.eujournal.org/index.php/esj/issue/archive

26) Ajogbeje, O. J. & Borisade, F. T. (2013). Cognitive entry characteristics and semester examination scores as correlates of college students' achievement in mathematics. British Journal of Education, Society and Behavioural Science. 3(4), 478-489. http://www.sciencedomain.org/Journals

27) Ajogbeje, O. J., Borisade, F. T., Aladesaye, C. A. & Ayodele, O. B. (2013). Effect of gender, mathematics anxiety and achievement motivation on college students'

Volume 62 | Issue 9 | September 2023
DOI: 10.5281/zenodo.8398633

achievement in mathematics International Journal of Education and Literacy Studies. 1(1), 15-22.

28) Joshua, M. T. (2005). *Fundamentals of Education Test and Measurement*. Calabar: University of Calabar Press.

29) Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342.

30) Kaplan, R. M., & Saccuzzo, D. P. (2017). *Psychological testing: Principles, applications, and issues*. Cengage Learning.

31) Kerlinger, F. N. (1979). Foundation of behavioural research. *New York: Holt Rhinehart &Winstor.*

32) Kline, P. (2013). *Handbook of psychological testing*. Routledge.

33) Koh, C. E., & "Ted" Nam, K. (2005). Business use of the internet: a longitudinal study from a value chain perspective. *Industrial Management & Data Systems*, *105*(1), 82–95.

34) Kolawole, E. B. (2010). Principles of tests construction and administration. *Lagos: Bolabay Academic Publishing Consultant*, *18*.

35) Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151–160.

36) Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563–575.

37) LeCompte, M. D., Preissle, J., & Tesch, R. (1993). *Ethnography and qualitative design in educational research*. Academic Press,.

38) McBurney, T. L. W. (2007). *Research Methods*, (7th Editio). https://www.amazon.com/Research-Methods-7th-Donald-McBurney/dp/0495092088

39) Messick, S. (1989). Validity. em r. linn (org.), educational measurement.(13-103). *New York, NY: American Council on Education and Macmillan Publishing Company*.

40) Moser, C. A., & Kalton, G. (2017). *Survey methods in social investigation*. Routledge.

41) Notar, C. E., Zuelke, D. C., Wilson, J. D., & Yunker, B. D. (2004). The table of specifications: Insuring accountability in teacher made tests. *Journal of Instructional Psychology*, *31*(2), 115.

42) Novick, M. R. (1965). Psychological Tests: Psychological Tests and Personnel Decisions. Lee J. Cronbach and Goldine C. Gleser. University of Illinois Press, Urbana, ed. 2, 1965. xiv+ 347 pp. Illus. $7.95. *Science*, *148*(3671), 803–804.

43) Oluwatayo, J. A. (2012). Validity and reliability issues in educational research. *Journal of Educational and Social Research*, *2*(2), 391–400.

44) Onunkwo, G. I. N. (2002). Fundamentals of educational measurement and evaluation. *Owerri: Cape Publishers International Ltd*.

45) Oremus, M., Oremus, C., Hall, G. B. C., McKinnon, M. C., & Team, E. C. T. & C. S. R. (2012). Inter-rater and test–retest reliability of quality assessments by novice student raters using the Jadad and Newcastle–Ottawa Scales. *BMJ Open*, *2*(4), e001368.

Volume 62 | Issue 9 | September 2023
DOI: 10.5281/zenodo.8398633

46) Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*(12), 1395.

47) Robinson, J. (2010). *Triandis' theory of interpersonal behaviour in understanding software piracy behaviour in the South African context*. University of the Witwatersrand Johannesburg.

48) Ruane, J. M. (2016). *Introducing social research methods: Essentials for getting the edge*. John Wiley & Sons.

49) Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research*, *27*(2), 94–104.

50) Shou, Y., Sellbom, M., & Chen, H.-F. (2021). Fundamentals of measurement in clinical psychology. In *Reference Module in Neuroscience and Biobehavioral Psychology*. Elsevier.

51) Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, 83–117.

52) Soo Wee, Y., & Quazi, H. A. (2005). Development and validation of critical factors of environmental management. *Industrial Management & Data Systems*, *105*(1), 96–114.

53) Straub, D., Boudreau, M.-C., & Gefen, D. (2004). Validation guidelines for IS positivist research. *Communications of the Association for Information Systems*, *13*(1), 24.

54) Streiner, G. I. & Norman, D. R. (2003). *Health measurement scales; A guide to their development and use*.

55) Taherdoost, H. (2016). Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in a research. *How to Test the Validation of a Questionnaire/Survey in a Research (August 10, 2016)*.

56) Testing, S. for E. and P. (1999). *American Educational Research Association, American Psychological Association, and National Council on Measurement in Education*.

57) Walden, U. (2012). *Educational social psychology. www. experiment-reserch. com*.

58) Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *The Journal of Strength & Conditioning Research*, *19*(1), 231–240.

59) Whiston, S. C. (2005). *Principles and applications of assessment in counselling. California: Brooks*. Cole Publishers.

60) Whitley, B. E., Kite, M. E., & Adams, H. L. (2013). *Principles of research in behavioral science*.