RELIABILITY OF RESEARCH INSTRUMENTS IN EDUCATIONAL RESEARCHES: AN EXPLORATORY PERSPECTIVE

[1]Enetairo, Reuben, [2]Stephen, Bulus Gadzama &
[3]Fauziyya Muhammad Yunusa Ibrahim
Department of Educational Psychology, Guidance and Counselling,
Federal College of Education (T) Bichi, Kano State, Nigeria.
E-mail: **reubeniroro**@yahoo.com
Tel: 08069536570
Stephen, Bulus Gadzama
Department of Educational Psychology, Guidance and Counselling,
Federal College of Education (T) Bichi, Kano State. Nigeria.
E-mail: **stephenbulus105**@gmail.com
Tel: 08065029349
Fauziyya Muhammad Yunusa Ibrahim,
Department of Educational Psychology, Guidance and Counseling,
Federal College of Education Technical Bichi Kano State, Nigeria
myfauziyya@gmail.com
08033446650

Abstract
In the realm of research, reliability stands as a cornerstone of methodological rigor and credibility. It refers to the consistency of a measure or instrument, ensuring that the same results can be replicated under consistent conditions. Consistency is crucial for the validity of research findings, influencing the confidence that researchers and stakeholders can place in the results. The purpose of this study is to delve into the multifaceted concept of reliability to assess the various approaches in determining the reliability of research instruments in educational research. The study used an exploratory research technique and relied on information from previous studies and publications, including journals, textbooks, periodicals, and as well the internet. Various forms of reliability, including test-retest reliability, inter-rater reliability, and internal consistency are discussed to underscore their roles in the educational research process. Methods for assessing reliability, such as Cronbach's alpha, Kuder-Richardson, Intraclass Correlation Coefficient (ICC), and Pearson Correlation Coefficient are examined to provide researchers with practical knowledge for enhancing the reliability of their instruments. The article also addresses common challenges and strategies to overcome these challenges in achieving reliable measurements in education, including best practices and future direction in reliability research. Thus, by emphasizing the importance of reliable research instruments, this article aims to guide educational researchers in developing and implementing tools that yield consistent and trustworthy results, thereby contributing to the advancement of educational theory and practice.

Keywords: Research Instrument, Reliability, Alternate-forms Reliability, Inter-rater Reliability, Internal Consistency.

Introduction

Reliability is a cornerstone of a sound research in education. It ensures that research instruments, such as tests and questionnaires, produce consistent and stable results over time. Reliable instruments are crucial for the validity of research findings, enabling educators and policymakers to make informed decisions based on accurate data. In any research, estimating reliability is critical (Imasuen, 2022). Therefore, the amount to which an investigation, test, or measurement process delivers the same result on multiple testing is referred to as reliability. According to Ali (2021), reliability is the consistency between two measures of similar instruments. It is a measure of the precision of the instrument and its ability to produce similar outcomes under consistent circumstances.

Shodiya and Adekunle (2022) opined that to guide and as well reach study aims, researchers are frequently faced with two difficulties. The first is how can the researcher ensure that research instruments are evaluating whatever he/she want to measure?" "And secondly how sure is the researcher, that he/she will receive the same result if he/she reruns the measurement?" As a result of these notable difficulties or questions, the researcher of this study feel that a critical review of the concept of reliability, as well as assessment tools in the dependability of data gathered through tests or questionnaires is necessary so as to improve on futures educational researches. This is because the importance of reliability in research cannot be over emphasized. Reliable instruments yield consistent results, which are essential for replicating studies and validating findings. Without reliability, the validity of the research is compromised, leading to potentially flawed conclusions and undermining the scientific process (Creswell, & Creswell, 2018).

Methodology of the Study

The study used an exploratory research technique and relied on information from previous studies and publications, including journals, textbooks, periodicals, and the internet. The paper explores most of the pertinent aspect concerning reliability of quantitative research instrument.

Scope and Objectives of the Study

This article aims to provide a comprehensive overview of the reliability of research instruments in educational research. It will explain key concepts, discuss various types of reliability, outline methods for assessing reliability using statistical tools, and present current challenges and best practices to mitigate these challenges.

Reliability in Educational Researches

In educational research, reliability refers to the consistency and dependability of a measurement instrument. However, various authors and sources provide nuanced definitions and insights into reliability. These include:

Tavakol and Dennick (2011) defined reliability as the degree to which an assessment tool produces stable and consistent results over repeated trials and various conditions. It is a measure of the precision of the instrument and its ability to produce similar outcomes under consistent circumstances. Likewise, Middleton (2023**),** defined reliability as the consistency of a method to produce the same results under the same conditions. It is categorized into four main types and each type assesses different aspects of consistency,

such as over time, across different raters, with different forms of a test, and within the test itself as stated below (Scribbr).

- **Test-Retest Reliability**: Consistency over time.
- **Inter rater Reliability**: Consistency across different observers or raters
- **Parallel Forms Reliability**: Consistency across different forms or versions of a test.
- **Internal Consistency**: Consistency within the items of a test.

These definitions and approaches underscore the multifaceted nature of reliability in educational research, covering both quantitative and qualitative methodologies. Ensuring reliability is crucial for the validity and credibility of research findings.

Key Terms and Concepts

- **Consistency:** The degree to which an instrument yields the same results on repeated trials.
- **Stability:** The extent to which the results of an instrument remain unchanged over time.
- **Precision:** The exactness of the measurement produced by the instrument.

Types of Reliability

1. **Test-Retest Reliability:** Test-retest reliability measures the stability of a test over time. It involves administering the same test to the same group of people at two different points in time and correlating the scores to determine consistency, (Lee & Sim, 2023).
2. **Internal Consistency Reliability:** Internal consistency reliability evaluates the extent to which items within a test measure the same construct. Internal consistency (or homogeneity) concerns the reliability within the measuring instrument and it questions how well a set of items (or variables) measures a concept that is being tested (or measured) (Ahmed et al., 2022. In other words, the more interrelated (undimensional) the items are, the higher the calculated reliability coefficient (estimate) (Ekolu, Quainoo, 2019). The estimate is obtained by calculating the average inter correlations among all single items (or variables) in a concept, or a test ((Ahmed et al., 2022) using one or more of the following methods: split-half reliability, Kuder-Richardson coefficient, Cronbach's alpha and inter-item consistency (inter-rater reliability) (Ahmed et al., 2022). However, there is no clarity around the interpretation of reliability estimates but estimates > 0.5 have been considered acceptable in research (Ekolu & Quainoo, 2019).

Internal consistency measures the relationship between many items in a test which are meant to evaluate the same construct. Internal consistency is assessed without having to repeat the test or involve additional researchers. If there's only one data set, it is an excellent technique to measure reliability. The researcher creates a number of questions or ratings which is merged into an aggregate score, ensuring that all of the things truly represent the same thing. If replies to multiple items contradict each other, the test may be untrustworthy.

3. **Parallel-Forms Reliability:** Parallel-forms reliability involves administering two different forms of the same test to the same group of people. The results of the two forms are then correlated to determine the consistency of the measurements (Thompson & Miller, 2023). Parallel-form reliability (or alternate-form reliability) is like test-retest reliability

but with an exception that a different (or an alternate) form of the original test is administered at different times (Ahmed et al., 2022).

According to Heale et al., (2015), stated that the concepts being tested are the same in both versions, but the expressions may be presented differently. As the name implies, two or more versions of the test are constructed that are equivalent in content and level of difficulty, e.g. professors use this technique to create makeup or replacement exams because students may already know the questions from the earlier exam (Ahmed et al., 2022). The measuring instrument used is stable when there is a high correlation between the scores (values) obtained each time the tests are completed (Heale &Twycross, 2015). A low correlation indicates the presence of measurement error, which is construed as using two different scales in both tests (Ahmed et al., 2022).

4. **Split-Half Reliability:** Split-half reliability is a measure of internal consistency that involves splitting a test into two halves and correlating the scores from each half, Brown and Green (2022). The split-half method measures the degree of internal consistency by checking one half of the results of a set of scaled items in a measuring instrument against the other half (Ahmed et al., 2022). It requires only one administration of the measuring instrument (Mohajan, 2017), but the items in the instrument are split in half in several ways, for example, first half and second half, or by odd and even numbered items, to form two new measures testing the same social phenomena.

In contrast to the test-retest reliability, the split-half method is usually measured in the same time (Ahmed et al., 2022). When the results are divided into half, correlations are calculated comparing both halves (Heale &Twycross, 2015). Indeed, strong correlations indicate high reliability, while weak correlations indicate the instrument may not be reliable (Ahmed et al., 2022; Heale & Twycross, 2015). The method demands equal item representation across the two halves of the instrument, otherwise, the comparison of dissimilar sample items will not yield an accurate reliability estimate (Ahmed et al., 2022). In split-half reliability we randomly divide all items that purport to measure the same construct into two sets. The researcher administer the entire instrument to a sample of people and calculates the total score for each randomly divided half.

Methods for Assessing Reliability

According to Green and Yang (2021), several statistical tools and techniques can be used to assess reliability. These including the following:

- **Cronbach's Alpha:** (*Measures internal consistency*).

The Cronbach alpha is used to measure the internal consistency of a set of items/variables measuring a construct/concept. The scores on the items/variables designed to measure the same construct/concept should be highly correlated. Therefore, Cronbach's alpha is a function of the average inter-correlations of items and the number of items in the scale (Ahmed et al., 2022; Mohajan, 2017).

Moreover, it should be noted that having multiple items to measure a construct/concept aids in the determination of the reliability of measurement and, in general, improves the reliability or precision of the measurement (Ahmed et al., 2022). Instruments with questions that have more than two responses can be used in this test (Heale & Twycross, 2015), but the greater the number of items in a summated scale, the higher Cronbach's alpha tends to be (Ahmed et al., 2022). The Cronbach's alpha result is a number between

0 and 1. An acceptable reliability score is one that is 0.7 and higher (Heale & Twycross, 2015).

Table 1. Summary of the Reliability value of 6-item scale using Cronbach Alpha

	Corrected Item-total correlation	Cronbach's alpha if item deleted
Q1	0.830	0.820
Q2	0.682	0.839
Q3	0.746	0.831
Q4	0.494	0.893
Q5	0.700	0.838
Q6	0.682	0.840

Reliability Statistics

Cronbach's alpha	
0.866	

Source: Morrison, J. (2019).

Cronbach's alpha for the scale created from these six survey questions is 0.866. The fourth survey item (Q4) does have the poorest association with other questions, and eliminating it from the measure will enhance reliability by raising Cronbach's alpha to 0.893. However, these tests only apply to instruments with a likert scale; however, the Kuder Richardson reliability test is an option for bivariate rating.

- **Kuder-Richardson:** (*Measures internal consistency*).

According to Sarmah and Hazarika (2012), the Kuder-Richardson method was introduced by Kuder-Richardson, a psychometrist in 1937. The Kuder Richardson method is like the split half method except that it is used to measure the degree of internal consistency of items consisting of only two responses (e.g. yes or no, 0 or 1) in a measuring instrument. The method assumes that all items of a test are of equal or almost equal difficulty and inter correlated (Sarmah & Hazarika, 2012). The common Kuder-Richardson method formula is known to be Kuder-Richardson formula 20 or KR20, which was later simplified to be Kuder-Richardson formula 21 or KR21 *(equation shown below).* Their difference is that KR21 can produce a direct estimation of reliability using a minimal dataset with only the number of test items, mean and variance (Ekolu & Quainoo, 2019). According to Heale et al., (2015), it is calculated by the average of all possible split-half combinations and a correlation between 0 and 1 is generated. Like the split-half method, strong correlations indicate high reliability; while weak correlations indicate the instrument may not be reliable (Kaji & Lewis, 2008). In applying the KR formula, it is assumed that all the test items are of the same level of difficulty. KR21 gives reliability index values lying between 0 and 1, as does Cronbach's alpha (Ekolu & Quainoo, 2019). The Kuder-Richardson Formula 20 is as follows:

Where:

k - Total number of questions.

pj - Proportion of individuals who answered question j correctly.

qj - Proportion of individuals who answered question j incorrectly.

σ 2 -Variance of scores for all individuals who took the test.

The value for KR-20 ranges from 0 to 1, with higher values indicating higher reliability.

The following example shows how to calculate the value for KR-20 in practice. Suppose a questionnaire with 7 items or (questions) was administered to test 10 students and to rate their knowledge about a particular product. The perception was rated on a YES or NO scoring and the scores were rendered in the Excel with 1 indicating correct answer and 0 indicating an incorrect answer.

Table 2. Summary of the Reliability value of 7-item using Kurder-Richardson KR-20

		A	B	C	D	E	F	G	H	I
1		Students	QI	Q2	Q3	Q4	Q5	Q6	Q7	Total Correct
2		1	0	1	1	0	1	1	1	5
3		2	1	1	1	1	0	0	0	4
4		3	1	1	1	1	0	1	1	6
5		4	1	1	0	0	1	1	0	4
6		5	0	1	1	1	1	0	1	5
7		6	1	0	1	0	1	1	0	4
8		7	1	1	0	0	0	0	0	2
9		8	1	1	0	1	0	1	0	4
10		9	0	0	1	1	0	0	0	2
11		10	1	1	1	0	1	0	1	5
12										
13		P	0.70	0.80	0.70	0.50	0.50	0.50	0.40	
14		Q	0.30	0.20	0.30	0.50	0.50	0.50	0.60	
15		PQ	0.21	0.16	0.21	0.25	0.25	0.25	0.27	
16										
17		K	7.0000							
18		∑PQ	1.5700							
19		δ^2	1.6556							
20		KR-20	0.0603							

Source: Zach, V. (2022)

Here are the formulas used in various cells:

B13: =SUM (B2:B11)/10. B14: =1-B13 B15: =B13*B14.

B17: =COUNTA (B1:H1). B18: =SUM (B15:H15).

B19: =VAR.S (I2:I11). B20: = (B17/ (B17-1))*(1-B18/B19)

Thus, the KR-20 value turns out to be 0.0603. This number is so low, it shows that the test is unreliable since acceptable reliability is from 0.7 and above. This means that the items may have to be rewritten or restructured in order to improve the test's reliability.

- **Intraclass Correlation Coefficient (ICC):** (*Assesses inter-rater reliability*).

The more that individual judgment is involved in a rating, the more crucial it is that independent observers agree when applying the scoring criteria (Ahmed et al., 2022). Inter-rater reliability establishes the equivalence of ratings obtained with a measuring instrument when used by different raters (Mohajan, 2017). Therefore, it is used to determine the level of agreement between two or more raters (Heale & Twycross, 2015; Ahmed et al., 2022). On the other hand, intra-rater reliability establishes the equivalence of ratings obtained with a measuring instrument used by a single rater over a period (McHugh, 2012).

Table 3. Percent agreement across multiple data collectors (fictitious data)

Variables	**Rater**					**% of Agreement**
	Mark	Susan	Ton	Ann	Joyce	
1	1	1	1	1	1	1.00
2	1	1	1	1	1	1.00
3	1	1	1	1	1	1.00
4	0	1	1	1	1	0.80
5	0	1	0	0	0	0.80
6	0	0	0	0	0	1.00
7	1	1	1	1	1	1.00
8	1	1	1	1	0	0.80
9	0	0	0	0	0	1.00
10	1	1	0	0	1	0.60
Study Interrater reliability						0.90

	Mark	**Susan**	**Ton**	**Ann**	**Joyce**
Is a rater an outlier	1	1	1	1	1
# of unlike responses	1	1	1	1	1

Source: McHugh, M.L. (2012)

Table 3, which exhibits an overall interrater reliability of 90%, it can be seen that no data collector had an excessive number of outlier scores (scores that disagreed with the majority of raters' scores).

- **Pearson Correlation Coefficient:** (*Used for test-retest and parallel-forms reliability*).

This is because parallel-form reliability (or alternate-form reliability) is like test-retest reliability but with an exception that a different (or an alternate) form of the original test is administered at different times (Ahmed et al., 2022). According to Heale et al. (2015), the

concepts being tested are the same in both versions, but the expressions may be presented differently.

Meanwhile, here is a practical example by considering a group of students who have been given 10 items questionnaire on the concepts of Economics to test how knowledgeable they are about the subject at a particular available time. The reported scores were recorded in the 1st test and after about a week the same group of students were given the same or identical questions, their responses were also recorded exactly the same way in 2nd test. The correlation coefficient calculated from these two sets of scores gives us an indication of stability. The outcome is shown in table 4 below, and the Pearson Product-Moment Correlation Coefficient (PPMC) is obtained as follows.

Table 4: Test-retest scores on some Concepts of Economics

Subject	Test (1st Test) Scores	Retest (2nd Test) Scores
1	2	3
2	0	2
3	2	2
4	3	5
5	4	5
6	1	3
7	2	2
8	6	5
9	1	3
10	2	4
	23	**34**

Table 5: Pearson correlation analysis of Test-retest scores on Economics Concepts

	Mean	Std. Deviation	N	Pearson Correlation	Sig. (2-tailed)
TEST 1	2.20	1.476	10	**0.763***	0.010
TEST 2	3.60	1.350	10		

Source: Author computation (2024). * Correlation is significant at the 0.05 level (2-tailed).

The Pearson **r** is significant at .05 with a 10-person sample size**.** As a result, the reliability is set at .763, which is an acceptable score for this sort of test. The main disadvantage of this strategy is that when the retake is administered too soon, the initial test sensitizes the respondents to the issue, and as a consequence, the respondent will recall and repeat the answers already given. This results in upwardly skewed dependability indicators. Also, attitudes may alter as a result of situational effects prior to the retest. The stability scores are biased downward in these circumstances. This implies that longer the time interval between two successive administrations, the lower the correlation coefficient indicating poor reliability.

Practical Applications in Educational Research

In educational research, reliability assessment is applied to various types of instruments such as standardized tests, teacher-made tests, and observational checklists. Ensuring reliability involves conducting pilot studies, analyzing data, and refining instruments to improve consistency.

Challenges in Ensuring Reliability

Despite its importance, ensuring reliability can be challenging. Common issues in ensuring reliability in educational research include:

- Variability in student responses.
- Differences in teacher scoring or observations.
- Inconsistencies in test administration procedures.
- Changes in the testing environment can all impact reliability.

Researchers must be vigilant in identifying and addressing these potential sources of variability to maintain the reliability of their measurements.

Strategies for Mitigating Reliability Issues

McAlinden, N., et al. (2021) mentioned strategies to mitigate these issues include:

- Standardizing test administration procedures.
- Providing comprehensive training for teachers and raters.
- Conducting pilot studies to identify and address potential sources of error.

Best Practices in Educational Research

Best practices in educational research for ensuring reliability include:

- Using established and validated measurement instruments.
- Continuously monitoring and adjusting data collection procedures.
- Regularly training and calibrating teachers and raters to maintain consistency.

Conclusion

It should be noted that educational researchers do not just presume that an instrument is reliable. This is because studies have always shown that instruments are reliable before going on to make analysis and conclusions from these results thus emphasizing the reliability essential for study validity. Considering several articles reviewed by different researchers on reliability of research instrument, it was observed that some scholars were able to test and measure data credibility through different modes such as validity, reliability and generalizability.

Reliability is a fundamental aspect of research that ensures the consistency, validity, and credibility of study findings. By employing rigorous methods to ensure reliability, researchers can enhance the trustworthiness of their work, facilitating comparability and replication. As research methodologies continue to evolve, the importance of reliability remains paramount in maintaining the integrity and impact of scientific inquiry. Thus, reliability is a critical aspect of researches in education, ensuring the consistency and stability of measurement instruments. By understanding and assessing different types of

reliability, researchers can enhance the robustness of their findings and contribute to the advancement of educational knowledge.

Future Recommendations

Future directions in reliability research will likely involve:

1. Development of more sophisticated tools and techniques for assessing and improving the consistency of measurement instruments.
2. Advancements in technology and data analysis methods will continue to play a significant role in enhancing reliability in educational research.

References

Ahmed, V., Opoku, A., Olanipekun, A., Sutrisna, M. (2022).Validity and Reliability in Built Environment Research A Selection of Case Studies. New York: Routledge.

Ali, A. A. (2021). *Basic components of research in education*. Ahmadu Bello University Press Limited.

Brown, C. A., & Green, T. (2022). Split-half reliability and the Spearman-Brown prophecy formula: Applications and interpretations. *Journal of Research Methods and Statistics in Psychology*, 35(3), 275-289. DOI: 10.1177/08944393211054321

Brown, C. A., & Green, T. (2022). Split-half reliability and the Spearman-Brown prophecy formula: Applications and interpretations. *Journal of Research Methods and Statistics in Psychology*, 35(3), 275-289. DOI: 10.1177/08944393211054321

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.

Ekolu, S.O., Quainoo, H. (2019). Reliability of assessments in engineering education using Cronbach's alpha, KR and split-half methods. Global Journal of Engineering Education, 21(1), 24-29.

Green, S. B., & Yang, Y. (2021). Evaluation of Internal Consistency Reliability for Scales and Subscales Using Cronbach's Alpha and Omega. *Educational Measurement: Issues and Practice*, 40(2), 79-89. DOI: 10.1111/emip.12429

Heale, R., Twycross, A. (2015) Validity and reliability in quantitative studies. EvidenceBased Nursing, 18, 66-67.

Imasuen, K. (2022). Sample Size Determination in Test-Retest and Cronbach Alpha Reliability Estimates. British Journal of Contemporary Education, 2(1), 17-29. DOI: 10.52589/BJCE-FY266HK9

Lee, K., & Sim, H. (2023). Enhancing inter-rater reliability in qualitative research: Practical guidelines. *Qualitative Research Journal*, 23(1), 12-28. DOI: 10.1108/QRJ-12-2021-0117

McAlinden, N., et al. (2021). Enhancing Inter-Rater Reliability in Observational Research: Best Practices and Methodological Insights. *Journal of Behavioral Sciences*, 15(2), 145-160. DOI: 10.3390/jbs15020145

McHugh, M.L. (2012). Interrater reliability: the kappa statistic. Biochemiamedica, 22(3), 276-282.

Middleton, F. (2023). The 4 Types of Reliability in Research | Definitions & Examples. Scribbr. Retrieved from Scribbr (Scribbr).

Mohajan, H.K. (2017). Two criteria for good measurements in research: Validity and reliability. Annals of SpiruHaret University. Economic Series, 17(4), 59-82.

Morrison, J. (2019). Assessing Questionnaire Reliability - Select Statistical Consultants. Select Statistical Consultants;https://select-statistics.co.uk/blog/assessing questionnaire reliability.

Sarmah, H.K., Hazarika, B.B. (2012). Determination of Reliability and Validity measures of a questionnaire. Indian Journal of Education and Information Management, 5(11), 508-517

Shodiya & Adekunle(2022). Reliability of Research Instruments in Management Sciences Research: An Explanatory Perspective. Silesian University of Technology Publishing House

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53-55. DOI: 10.5116/ijme.4dfb.8dfd

Thompson, B., & Miller, D. (2023). Practical approaches to parallel-forms reliability in educational assessment. *Journal of Educational Measurement*, 60(1), 33-49. DOI: 10.1111/jedm.12345

Zach, V. (2022, January 7). Kuder-Richardson Formula 20 (Definition & Example) - Statology. Statology; www.statology.org. https://www.statology.org/kuder-richardson-20/. Reliability of research instruments…

Zhang, L., & Wang, J. (2022). Internal consistency reliability: Advances in assessing Cronbach's alpha. *Educational and Psychological Measurement*, 82(4), 601-618. DOI: 10.1177/00131644211034590