SILESIAN UNIVERSITY OF TECHNOLOGY PUBLISHING HOUSE

SCIENTIFIC PAPERS OF SILESIAN UNIVERSITY OF TECHNOLOGY 2022
ORGANIZATION AND MANAGEMENT SERIES NO. 166

RELIABILITY OF RESEARCH INSTRUMENTS IN MANAGEMENT SCIENCES RESEARCH: AN EXPLANATORY PERSPECTIVE

Olayinka Abideen SHODIYA[1*], Tijani Abideen ADEKUNLE[2]

[1] Department of Business & Finance, Crescent University, Abeokuta, Nigeria; olayinkashodiya@yahoo.com, ORCID: 0000-0002-5551-3335
[2] Department of Business Administration, Lagos State University, Ojo, Lagos State; abadeen080@yahoo.com, ORCID: 0000-0002-8530-5959
* Correspondence author

Purpose: The reliability of test scores is the extent to which they are consistent across different occasions of testing, different editions of the test, or different raters scoring the test taker's responses. The purpose of this study is to assess the various approaches in determining the reliability of research instruments in management sciences research.
Design/methodology/approach: The study used an exploratory research technique and relied on information from previous studies and publications, including journals, textbooks, periodicals, and the internet.
Findings: Consequence upon several articles reviewed on the subject matter 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 internal consistency, inter rater, alternate form and reliability coefficient.
Practical implications: The paper explored all of the pertinent concerns surrounding quantitative research instrument reliability and reviewed test reliability which include but not limited to: "alternate-forms reliability," "inter-rater reliability," "internal consistency," "reliability coefficient," "classification consistency," with illustrations.
Originality/value: Popular and commonly used reliability assessment approaches in Nigeria and in the field of management are the use of Cronbach alpha and Test-retest reliability tests for instrument reliability. Despite these, there are different types of reliabilities which are less reported in the field of management in the Nigeria academia. Based on this, the study reviewed various approaches and types of reliability test commonly utilised in Management sciences.

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

JEL CODE: M30.

1. Introduction

Questionnaire surveys are a valuable technique for gathering information from respondents in a range of situations, including self-reported outcomes in management research. Research always utilise surveys to gauge something, as such, surveys may be thought of as a measuring tool. Surveys can evaluate behaviours, attitudes, and views in the same manner that thermometers measure temperature and potential of hydrogen metres detect acidity. Surveys are frequently used to assess more sophisticated and varied human behaviours or qualities, referred to as constructs. Because they are complicated and varied, they are better assessed by asking a series of linked questions about various facets of the construct of interest. Individual replies to these questions can then be used to generate a score or scale measure along a continuum. In any research, estimating reliability is critical (Imasuen, 2022). To attain the research aim, we are generally faced with the question of whether we can be certain that when the repeated measurements are made, we will receive the same result. The amount to which an investigation, test, or measurement process delivers the same result on multiple testing is referred to as reliability. If a test is completely reliable, there really is no measurement error; everything we see is the true score (Imasuen, 2022). In every research, estimating reliability and validity is critical. To reach the study aim, the researcher is frequently faced with two difficulties. The first is how can the researcher ensure that research instruments are evaluating whatever he/she want to measure?" "How sure is researcher that he/she will receive the same result if he/she reruns the measurement?" As a result, the researcher of this study feel that a critical review of the idea, as well as assessment tools in the dependability of data gathered through tests or questionnaires, is necessary to improve management sciences research.

Statement of the Problem

In business and management research, utilising data at face value without screening for potential errors and bias or measuring dependability cannot be trusted (Flintermann, 2014). Several academics have sought to build tools and procedures for gauging reliability in order to boost researchers' trust in the use of quantitative data. The most popular and commonly used reliability assessment approach in Nigeria and in the field of management sciences as far as researcher knowledge is concerned are the use of Cronbach alpha and Test-retest reliability tests for instrument reliability (Imasuen, 2022). Despite this, there are different types of reliabilities which are less reported in the field of management in the Nigeria academia. Based on this, the study carryout a review of the various approaches and types of reliability test commonly utilised in Management sciences.

Objective of the Study

To assess the various approaches in determining the reliability of research instruments in management sciences research.

Methodology of the Study

The study used an explanatory research technique and relied on information from previous studies and publications, including journals, textbooks, periodicals, and the internet. The paper explores all of the pertinent concerns surrounding quantitative research instrument reliability.

Reliability in Management Research

In quantitative management sciences research, measurements of social concepts are carried out by using measuring instruments (i.e. questionnaire). The measuring instrument is reliable when it yields consistently the same or comparable results over repeated measures (Ahmed et al., 2022). That is, regardless of who performs the measurement, and the occasion and condition under which measurement was carried out, the results produced by the measuring instrument is consistent (or comparably consistent) (Mohajan, 2017). Therefore, reliability in management sciences is regarded as the accuracy of a measuring instrument in quantitative management sciences research (Heale, Twycross, 2015). Therefore, for the management sciences researcher, the challenge of reliability is to develop measuring instruments to obtain the true values of measured concepts to reduce error in measurement process. This requires the testing of reliability of measuring instruments (Heale, Twycross, 2015). The three attributes of reliability that are often tested are: stability, homogeneity or internal consistency and equivalence.

Stability

Stability refers to the ability of a measure to remain the same over time without controlling the testing conditions or respondent themselves (Mohajan, 2017). Therefore, a perfectly stable measuring instrument will produce the same results when administered time after time to collect data (Bannigan, Watson, 2009) and this is obtained by performing the test-retest reliability method.

Internal consistency

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). According to Kimberlin et al. (2008), the assumption of internal consistency is that items (or variables) measuring the same concept should correlate, and therefore, the coefficient of internal consistency provides an estimate of the reliability of measurement. 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).

Equivalence

Equivalence establishes the extent to which the measuring instrument collects information in a consistent manner. According to Heale et al. (2015), equivalence is established by evaluating the consistency among (1) responses of multiple users of an instrument (inter-rater reliability) and (2) among alternate forms of an instrument (parallel-form or alternate-form reliability). Often, observational instruments or rating scales are developed to evaluate the behaviours of subjects who are being directly observed. However, any measure that relies on the judgments of raters or reviewers requires evidence that any independent, trained expert would come to the same conclusion (Ahmed et al., 2022). 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 being assessed (Mohajan, 2017).

2. When to apply reliability testing for instrument

2.1. During a new scale or measure development

In Psychometric analysis, the researcher must assess whether the new scale has construct reliability. Once a new scale of measurement there is an important need to test to see if it is reliable; that is, to see if the scale items are internally consistent (Badenes-Ribera, Silver, Pedroli, 2020). Scale development and validation of scores is not a job to be taken on lightly. Development is a rigorous process which is based on item generation and content validation using expert feedback and pre-testing. In fact, it may take numerous iterations for the scale to be economically feasible and yet convey the appropriate construct (Badenes-Ribera, Silver, Pedroli, 2020). Reliability is usually done after item generation where items through pilot testing, in a larger sample after scale or measure has been established and follow-up when tested in another study location.

2.2. Pre-testing before a main study

A pilot survey is essentially a copy and trail of the main survey. The goal of doing a pilot study is to identify any flaws in the measurement device. It is concerned with whether the respondents decode the information intended to be measured very well before administering it to a larger sample to avoid wastage or to reduce number of items. The key advantage of pilot testing is that it allows the researcher to spot problems before launching the complete survey. The purpose of pilot testing is to determine the reliability as part of the validity for of each question. Items with poor reliability are removed at this stage (Kimberlin, Winterstein, 2008).

2.3. During main Cross-sectional studies and large survey to eliminate response bias detect measurement errors

Reliability implies consistency but not accuracy. Self-reports of behavior are particularly subject to problems with social desirability biases. Subjects may provide responses that are socially acceptable or that are in line with the impression they want to create. In addition, self-report questions may elicit an estimation of behavioral frequency rather than the recall and count response desired by the researcher (Kimberlin, Winterstein, 2008).

2.4. Repeated studies

Part of Reliability is that a condition where a measurement process yields consistent scores (given an unchanged measured phenomenon) over *repeat* measurements. A repeated measurements design is a type of study design in which several measures from same variable are performed with the same or matched participants under variable circumstances or over two different time periods. In longitudinal research, for example, repeated measures are gathered to analyse change over time. Therefore, every trial includes the assessment foe consistency over time (Badenes-Ribera, Silver, Pedroli, 2020).

2.5. When a scale or measurement is adapted or adopted

Whenever a measure is adopted, the validity and reliability research from previous studies on that instrument may be applied to the present study, such that a new validity is not established but requires reliability evidence. Adopting an instrument connects the study to all prior research studies that utilised the same instrument by showing that the measure has the same consistency level as the previous studies. However, when an instrument is modified, it has been drastically altered, and earlier reliability and validity results will no longer apply to the current investigation. Thus, while adopting or altering an existing scale, dependability is achieved (Korb, 2013).

3. Types of reliability

There are four categories of dependability. Each of the four broad groups of reliability estimations evaluates dependability in a different way. They are as follows:

1. Internal Consistency Reliability: This term is used to describe the consistency of outcomes across items in a test.
2. Test-Retest Reliability: Used to analyse a measure's consistency from one time to the next.
3. Inter-Rater or Inter-Observer Reliability: This term refers to how well various raters/observers estimate the same phenomena.
4. Parallel-Forms Reliability: A measure of the consistency of the outcomes of two tests built in the same fashion from the same content domain.

4. Internal consistency reliability tools

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. This is carried out in three-ways which include:

4.1. Average Inter-item Correlation

The average inter-item correlation employs all of our instrument's items that are meant to assess the same construct. As shown in Figure 1, the analyst will first calculate the significant relation amongst each pair of items. For instance, if there are six things, there will be 15 potential item pairs generated (i.e., 15 correlations).The average inter-item correlation is summation of all these correlations. The researcher discovers an average inter-item correlation of .90 in the illustration, with participant correlations ranging from .84 to .95.

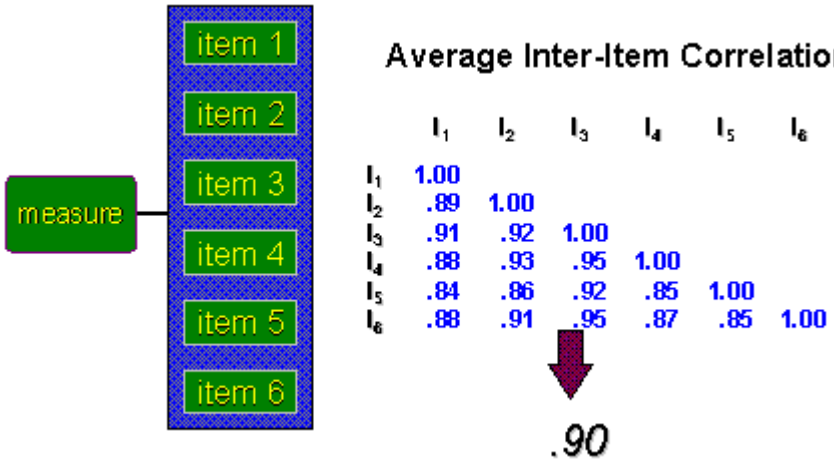

Figure 1. Average total correlation of 6 -item scale.

Source: Retrieved July 21, 2022, from https://conjointly.com/kb/types-of-reliability/.

4.2. Split-half

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 (Ahmed et al., 2022). 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 in 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. The split-half reliability estimate, as shown in the figure, is simply the correlation between these two total scores. In the example it is .87.

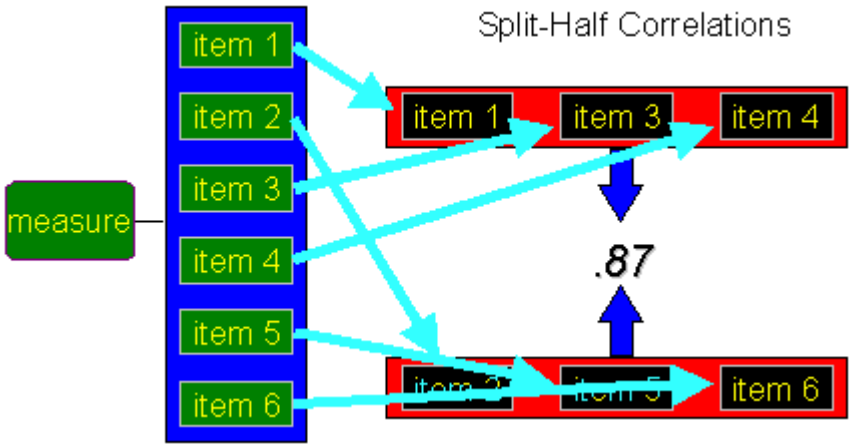

Figure 2. Split half-reliability for 6 -item scale.

Source: Retrieved July 21, 2022, from https://conjointly.com/kb/types-of-reliability/.

4.3. Cronbach alpha

The Cronbach alpha is used to measure the internal consistency of a set of items/variables measuring a construct/concept. Therefore, it measures the degree to which the different items/variables, especially those that each yield numerical response (Lam et al., 2010), but measuring the same construct/concept attains consistent results (Ahmed et al., 2022). The scores on the items/variables designed to measure the same construct/concept should be highly correlated (Ahmed et al., 2022). 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). Of note is 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). Most analytic tools will also automatically calculate the value of Cronbach's alpha if a question or survey item in the scale is eliminated. These values can indeed be examined to determine if the scale's reliability can be improved by discarding any one of the questionnaire items, as shown in the example below.

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

Reliability statistics

Cronbach's alpha
0.866

	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

Source: Morrison, J. (2019, May 30). *Assessing Questionnaire Reliability - Select Statistical Consultants*. Select Statistical Consultants; https://select-statistics.co.uk/blog/assessing-questionnaire-reliability/.

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 another questions, and eliminating it from the measure will enhance reliability, 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.

4.4. Kuder-Richardson

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:

$$KR\text{-}20 = (k / (k-1)) * (1 - \Sigma p_j q_j / \sigma^2)$$

where:

k - Total number of questions.

p_j - Proportion of individuals who answered question j correctly.

q_j - 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 questions was administered a test to 10 students to rate their knowledge about a particular product. The perception was rated on a yes or no scoring and the scores is rendered the in Excel, with 1 indicating a 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	**Student**	**Q1**	**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**	0.7	0.8	0.7	0.5	0.5	0.5	0.4	
14	**q**	0.3	0.2	0.3	0.5	0.5	0.5	0.6	
15	**pq**	0.21	0.16	0.21	0.25	0.25	0.25	0.24	
16									
17	**k**	7.0000							
18	**Σpq**	1.5700							
19	**σ^2**	1.6556							
20	**KR-20**	0.0603							

Source: Zach, V. (2022, January 7). *Kuder-Richardson Formula 20 (Definition & Example) - Statology*. Statology; www.statology.org. https://www.statology.org/kuder-richardson-20/.

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).

The KR-20 value turns out to be 0.0603. Because this number is so low, it shows that the test is unreliable. This means that the items may have to be rewritten or restructured in order to improve the test's reliability.

5. Test-retest reliability method

The test-retest reliability refers to the temporal stability of test from one measurement session to another (Ahmed et al., 2022). It is obtained by administering the same test twice, or more over a period ranging from few weeks to months, on a group of individuals (respondents) (Mohajan, 2017) under similar circumstances (Heale, Twycross, 2015). The procedure is to administer the test to a group of respondents and then administer the same test to the same respondents later (Ahmed et al., 2022). Thereafter, a statistical comparison is made between participant's test scores (values) for each of the times they have completed it to provide an indication of the reliability of the instrument (Heale, Twycross, 2015). For example, construction workers may be asked to complete the same questionnaire about safety satisfaction twice in three months so that test results can be compared to assess stability of scores. The correlation coefficient calculated between two sets of data, and the higher the coefficient, the better the test-retest reliability (and stability) (Mohajan, 2017). Test-retest reliability is defined by the correlation between scores (values) on the identical tests given at different times (Ahmed et al., 2022) and this leads to some limitations. For instance, when the interval between the first and second test is too short, respondents might remember what was on the first test and their answers on the second test could be affected by memory. Alternatively, when the interval between the two tests is too long, maturation happens – which is the changes in the subject factors (measured variables) or respondents that occur over time and cause a change from the initial measurements to the later (Ahmed et al., 2022). During the time between the two tests, the respondents could have been exposed to things which changed their opinions, feelings or attitudes about the behaviour under study (Ahmed et al., 2022). Ideally, the interval between administrations should management sciences long enough that values obtained from the second administration will not management sciences affected by the previous measurement but not so

distant that learning or a change in health status could alter the way subjects respond during the second administration.

Consider a group of students who have been asked to describe how knowledgeable they are about a particular available at the time. The reported responses were recorded using the following scale, 0 = Not at all, 1 = Somewhat knowledgeable, 2 = Very knowledgeable, and so on. Later, the same group was asked the identical questions, and their responses were recorded exactly the same way. The correlation coefficient calculated from these two sets of scores gives us an indication of stability. The outcome is shown in the table below, and the product-moment correlation coefficient is obtained as follows.

Table 3.
Test-retest scores on job performance

Subject	Test scores	Retest scores
1	1	2
2	0	3
3	2	2
4	4	5
5	3	5
6	2	3
7	1	2
8	5	6
9	1	4
10	1	4
	20	36

Source: Author computation (2022).

Table 4.
Pearson correlation analysis of Test-retest scores on job performance

	Mean	Std. Deviation	N	Pearson Correlation	Sig. (2-tailed)
Test scores	2.0000	1.56347	10	.746*	.013
Retest scores	3.6000	1.42984	10		

* Correlation is significant at the 0.05 level (2-tailed).

Source: Author computation (2022).

The Pearson r is significant at .05 with a 10-person sample size (a table value of .632 is required for r to be significant). As a result, the reliability is set at .746, 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 sensitises the responders to the issue, and as a consequence, the respondent will recall and repeat the answers already given. This results in upwardly skewed dependability indicators. Second, 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.

6. 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). The researcher formed a matrix wherein the columns depicted the different raters as well as the rows depicted variables whereby the raters had obtained data to find the estimate of percent agreement (Table 5). The data collectors' scores for each variable were stored in the cells of the matrix. Table 5 provides an illustration of this procedure. In this example, five raters measured their rankings for variables one through ten. To calculate the % agreement, the researcher deducted the number of incorrectly scored questions from the total number of zeros. The number of zeros divide it by the number of variables offers a measure of agreement among the raters. In Table 5, the agreement is 90%. This suggests that 10% of the data acquired in the research is incorrect. This metric is immediately translated as the percentage of accurate data. The number 1.00 - percent agreement might be interpreted as the percentage of wrong data. In other words, if the percent agreement is 90, 1.00-0.90 = 0.10, and 10% is the quantity of data that misrepresents the study findings.

Table 5.
Percent agreement across multiple data collectors (fictitious data)

Var#	Raters					% Agreement
	Mark	Susan	Tom	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

Is a rater an Outlier?	Mark	Susan	Tom	Ann	Joyce
#of unlike responses:	1	1	1	1	1

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

Table 5, 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).

7. Parallel-form reliability

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. 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).

Example of parallel form reliability: To calculate parallel form's reliability, first administer the two different tests to the same participants in a short period of time (perhaps with one week of each other). Then calculate the total score for each variable on the two separate tests.

Table 6.
Parallel form reliability of sales person job performance and sale performance

Participants	Sales person job performance	Sales Performance
1	67.00	68.00
2	53.00	56.00
3	67.00	61.00
4	55.00	59.00
5	46.00	42.00
6	59.00	57.00
7	52.00	51.00
8	59.00	55.00
9	38.00	54.00
10	41.00	44.00
11	40.00	54.00

Total scores for Sales person job performance scores were correlated with another performance rating, sales performance. This was calculated using the Pearson's Product Moment Correlation between sales person job performance and sales performance.

Table 7.
Pearson correlation analysis of Parallel scores on job performance and sales performance

Variable	Variable	Statistic				
		Correlation	**Count**	**Lower C.I.**	**Upper C.I.**	**Notes**
Language Proficiency	Sales Performance	.720	11	.211	.922	Significant
Missing value handling: PAIRWISE, EXCLUDE. C.I. Level: 95.0						

This is the parallel form's reliability coefficient was 0.720 for sales person job performance and sales performance.

Reliability Method in Research Study

Table 8.
Showed Reliability of Research Instrument by Previous Researcher(s) Useful for Further Research

S/N	Author(s)	Year	Title	Methodology	Remarks
2.	Taherdoost H.	2016	Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire/ Survey in a Research	Research Instrument, Questionnaire, Survey, Survey Validity, Questionnaire Reliability, Content Validity, Face Validity, Construct Validity, and Criterion Validity	This study review article explores and describes the validity and reliability of a questionnaire/survey and also discusses various forms of validity and reliability tests
3.	Ibiyemi, A., Yasmin Mohd Adnan, Md Nasir Daud, Segun Olanrele & Abiodun Jogunola (2019)	2019	A content validity study of the test of valuers' support for capturing sustainability in the valuation process in Nigeria	Content validity Face validity	The study presents the content domain of the valuers' perception of sustainability reporting in Nigeria for the purpose of identification and eliciting the character. It carried out the content validity index (i-CVI), the scale content validity index (s-CVI) and the content validity ratio (CVR).The paper argued for consistent and explicit content validation in sustainability research to avoid probable chance effects. Content validation helps to provide reliable data for causal model development of the knowledge management (KM) requirements for the integration of sustainability into real estate valuation.

Cont. table 8.

4.	Taherdoost. H.	2022	What are Different Research Approaches? Comprehensive Review of Qualitative, Quantitative, and Mixed Method Research, Their Applications, Types, and Limitations	Research methodology; Research approach; Qualitative research; Quantitative research; Mixed methods approach; Research design	This study provides a comprehensive review of qualitative, quantitative, and mixed-method research methods. Each method is clearly defined and specifically discussed based on applications, types, advantages, and limitations to help researchers identify select the most relevant type based on each study and navigate accordingly
5.	Bertea, P.E & Zaiţ, A.	2013	Scale Validity in Exploratory Stages of Research	Construct validity Content Validity Ratio Q-sorting	The paper draw the attention on alternative methods for scale validation that should be used in the exploratory phase. The role of these methods is to improve validity of results of the further confirmatory phases of research. The Lawshe (1975)content validity ratio and the Q-sorting procedure for testing construct validity are applied in the process of developing a scale for perceived risk
6.	Nnorom, G.K, Asikhia, O.U, Magaji, N, Makinde, O.G, Akpa, V.O & Obianwu, N.E		Contextual Factors and Organizational Performance: A Validity and Reliability Approach	Construct Validity Confirmatory Factor Analysis Convergent validity Discriminant validity	This study validated an instrument to aid research efforts in the area of contextual factors and organizational performance. After an initial questionnaire administration, the data was tested using validity and reliability tools. It was established that scale was fit for application in other studies as all scientific conditions were met.
7.	Ursachi, G., Horodnic, I.A., Zait, A.	2015	How Reliable are Measurement Scales? External Factors with Indirect Influence on Reliability Estimators	Research methods, instruments, validity, scale reliability	The study investigates role of external factors influence a largely used reliability estimator - Cronbach Alpha. Several scales commonly used in marketing researches were tested, using a bootstrapping technique. Results show that important differences in the values of Cronbach Alpha are possible due to indirect influence from external factors - respondents' age, gender, level of study, religiousness, rural/urban living, survey type and relevance of the research subject for the participants to the survey.

Source: Researcher (2022).

Discussion of Findings

Consequence upon several articles reviewed on the subject matter 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 generalisability. The concept of reliability and generalisability have been identified and redefined for its usefulness for improving quantitative research study. Researchers assess their measurements using two independent criteria: reliability and validity. Test-retest reliability, internal consistency, and consistency between researchers are all examples of dependability (interrater reliability) (Ahmed et al., 2022).

8. Conclusion

Management scientists do not just presume that instrument is reliable. Instead, 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. Over time, reliability represents consistency and replicability. Furthermore, reliability is seen as the degree to which a test is devoid of measurement errors, because the greater the number of measurement mistakes, the less trustworthy the test. Researchers are concerns on how far the same test would generate the same findings if given to the similar populations under the same settings. This enables researchers and management scientists to conduct valid comparisons. The more inaccuracies identified in an evaluation, the less reliable it is, and vice versa. The study conclude that reliability is an important factor in assessment, and it is presented as an aspect that contributes to validity rather than as an aspect that is opposed to validity.

References

1. Ahmed, V., Opoku, A., Olanipekun, A., Sutrisna, M. (2022).*Validity and Reliability in Built Environment Research A Selection of Case Studies*. New York: Routledge.
2. Badenes-Ribera, L., Silver, N.C., Pedroli, E. (2020). Editorial: Scale Development and Score Validation. *Frontier Psychology, 11,* 799. doi: 10.3389/fpsyg.2020.00799.
3. Bannigan, K., Watson, R. (2009). Reliability and validity in a nutshell. *Journal of Clinical Nursing*, *18*(23), 3237-3243.
4. Bertea, P.E Zaiț, A. (2013). Scale Validity in Exploratory Stages of Research. *Management & Marketing*, *11(1),* 36-49.

5. 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.
6. Flintermann, B. (2014). *The quality of market research reports: The case of Market line Advantage and the automobile industry.* Retrieved from http://essay.utwente.nl/66122/1/Flintermann_MA_MB.pdf, 29.05.2017.
7. Heale, R., Twycross, A. (2015) Validity and reliability in quantitative studies. *Evidence-Based Nursing*, *18*, 66-67.
8. Ibiyemi, A., Mohd Adnan, Y., Daud, M.N., Olanrele, S., Jogunola, A. (2019). A content validity study of the test of valuers' support for capturing sustainability in the valuation process in Nigeria. *Pacific Rim Property Research Journal, 25(3),* 177-193. doi:10.1080/14445921.2019.1703700.
9. 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.
10. Kaji, A.H., Lewis, R.J. (2008). Assessment of the reliability of the Johns Hopkins/Agency for Healthcare Research and Quality hospital disaster drill evaluation tool. *Annals of Emergency Medicine*, *52(3),* 204-210.
11. Kimberlin, C.L., Winterstein, A.G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy, 65(23),* 2276-2284. doi:10.2146/ajhp070364.
12. Korb, K.A. (2013). *Educational Research Steps*. Educational Research Steps, http://korbedpsych.com/R00Steps.html.
13. Lam, P.T., Chan, E.H., Poon, C.S., Chau, C.K., Chun, K.P. (2010). Factors affecting the implementation of green specifications in construction. *Journal of Environmental Management*, *91(3),* 654-661.
14. McHugh, M.L. (2012). Interrater reliability: the kappa statistic. *Biochemiamedica*, *22*(3), 276-282.
15. Mohajan, H.K. (2017). Two criteria for good measurements in research: Validity and reliability. *Annals of SpiruHaret University. Economic Series*, *17(4),* 59-82.
16. Morrison, J. (2019). *Assessing Questionnaire Reliability - Select Statistical Consultants*. Select Statistical Consultants. select-statistics.co.uk. https://select-statistics.co.uk/blog/assessing-questionnaire-reliability/.
17. Nnorom, G.K., Asikhia, O.U., Akpa, V.O., Magaji, N., Obianwu, N.E. (2020). Contextual Factors and Organizational Performance:A Validity and Reliability Approach. *The International Journal of Business & Management*, *8(6),* 1-8.
18. 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.

19. Taherdoost, H. (2016). Validity and reliability of the research instrument; How to test the validation of a questionnaire/survey in research. *International Journal of Academic Research in Management, 5(3),* 28-36.
20. Taherdoost, H. (2022). What are Different Research Approaches? Comprehensive Review of Qualitative, Quantitative, and Mixed Method Research, Their Applications, Types, and Limitations. *Journal of Management Science & Engineering Research, 5*(1), 53-63. doi: https://doi.org/10.30564/jmser.v5i1.4538.
21. Ursachi, G., Horodnic, I.A., Zait, A. (2015). How Reliable are Measurement Scales? External Factors with Indirect Influence on Reliability Estimators. *Procedia Economics and Finance, 20,* 679-686. doi: 10.1016/s2212-5671(15)00123-9.
22. Zach, V. (2022). *Kuder-Richardson Formula 20 (Definition & Example) - Statology*. Statology, https://www.statology.org/kuder-richardson-20/.