International Journal of Scientific Research and Engineering Development-– Volume 4 Issue 5, Sep- Oct 2021

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

ESTIMATING RELIABILITY INDEX USING PRACTICAL EXAMPLES: PERSPECTIVE FROM THE CLASSROOM ASSESSMENT AND EVALUATION EXPERIENCE

By

[1]EyongEmmanuel Ikpi

[1]Department of Educational Foundations and Childhood Education

Faculty of Education Cross River University of Technology Calabar, Nigeria

ABSTRACT

The study focused onestimating reliability index with practical examples: a perspective from the classroom assessment and evaluation experience. The crux of the study was based on the fact that most students do not seem to have a better understanding of the types of reliability that can be adopted in a practical situation and possible ways of estimating the different reliability indexes. The study used different scores generated from respondents to practically show how the different methods of reliability can be determined. The statistical techniques were Pearson's Product Moment Correlation and split-half reliability index and the reliability was determined for all the types of reliability methods. The study concludes on the role and relevance to teachers, students, and researchers who lack the mathematical skills how to determine the reliabilities could be determined. The study also pinpoints the need for scholars (teachers) to effectively validate their instruments before administration. This act will promote a high level of trust in the overall output of the academic undertakings.

Keywords Estimating, Reliability, Index, Classroom, evaluation, and Experience

Background to the study

In the context of the school setting, students' data (examination scores) cannot be takenfor granted to achieve the overall goal of education. The goals that are not achieved can reduce the consistency of the overall output of the system. Determining a reliable method of estimating the overall consistency of the data is a serious issue in most schools. Schools interested in establishing a culture of data are advised to come up with a plan before going off to collect it. Thus, anunderstanding of reliability help measures consistency in the school. One of the basic aims of every academic exercise is to produce reliable (consistent) outcomes on the part of the teacher and learner. When the learning process is not reliable, it can lead to distrust, and abuse of the general aims and objectives of education. By way of definition, reliability is the degree to which an assessment tool produces stable and consistent results. It is the degree

to which a measurement technique can be depended upon to secure consistent results upon repeated application. Messick (1995) observed that research requires dependable measurement. Measurements are dependable to the extent that they are repeatable and any random influence which tends to make measurements different from occasion to occasion or circumstance to circumstance is a source of measurement error. Reliability is the degree to which a test consistently measures whatever it measures. There are several ways of estimating the reliability of an instrument they are test-retest, split half, equivalent forms (parallel form), inter-rater reliability, Kuder Richardson K-R20 and K-R21, Cronbach alpha reliability, etc.

Types of Reliability and way of estimating reliability index

Test-retest Reliability

Test-retest reliability is the degree to which scores are consistent over time. It indicates score variation that occurs from testing session to testing session as a result of errors of measurement. In this method, a test is administeredto a single group ofexaminees as a pretest after an interval of two to three weeks the same test is re-administered to the same respondents (post-test). Typically, the two separate administrations areonly a few days or a few weeks apart and the time of administration should be short enough so that theexaminees' skills in the area being assessed have not changed through additionallearning. The relationship between the examinees' scores from the two differentadministrations is estimated, through statistical correlation with Pearson's Product Moment Correlation (PPMC) to determine how similar thescores are. The reliability coefficient obtained with the two administrations is termed the coefficient of stability.

Practical application of test re-test reliability: If for instance, a teacher gave a test to a group of seven students in Mathematics after two weeks the same test is re-administered on the same set of respondents and the following scores were obtained.

First administration (pre-test): 20, 15, 16, 10, 8, 10, 7

Second administration (post-test): 15, 10, 10, 8, 12, 11, 13

Required: Use the appropriate reliability to determine the coefficient and the degree of consistency of the tests.

Solution

To calculate the reliability we need to apply Pearson's Product Moment Correlation to determine the coefficient of stability. The formula to accomplish this exercise is thus:

$$= \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2)\,(n\sum y^2 - (\sum y)^2)}}$$

Where

r= Pearson's Product Moment Correlation coefficient

n= Number of cases in the distribution

X and Y= the raw score for the first and second tests.

$\sum X$, and $\sum Y$= sum of scores for X and Y.

$\sum X^2$ and $\sum Y^2$ = each of the scores squared and then summed.

$\sum XY$ = the product of X and Y scores summed.

Thus we need to reproduce the scores in a tabular form to obtain odd and even items.

Table 1: Reliability estimate with test-retest reliability

S/N	X (1st Admin.)	X^2	S/N	Y (2st Admin.)	Y^2	$\sum xy$
1.	20	400	2.	15	225	300
3.	15	225	4.	10	100	150
5.	16	256	6.	10	100	160
7.	10	100	8.	8	64	80
9.	8	64	10.	12	144	96
11.	10	100	12.	11	121	110
13.	7	49	14.	13	169	91
$n_1 = 7$	$\sum x = 86$	$\sum x^2$ 1194	$n_2 = 7$	$\sum y = 79$	$\sum y^2$ = 923	$\sum xy = 987$

Using Pearson's Product Moment correlation

$$r = \frac{7 \times 987 - (86)(79)}{\sqrt{(7 \times 1194 - (86)^2)\,(7 \times 923 - (79)^2)}}$$

$$= \frac{6909 - 6794}{\sqrt{(8358 - 7396)\,(6461 - 6241)}}$$

$$= \frac{115}{\sqrt{(211640))}}$$

$$= \frac{115}{460.04}$$

r = 0.249

r ≈ 0.25

The value of the coefficient indicates a weak relationship between the two tests.

Split-Half Reliability Method

In this case, the administration is just one administration, especially when the test is very long. The most commonly used method is to split the test into two halves using the odd and even strategy. A correction formula will be applied with Pearson's Product Moment Correlation (PPMC). After obtaining the correlation coefficient with PPMC, the Spearman-Brown Prophecy formula will then be used to step up the co-efficient. Also known as correction for attenuation. The co-efficient obtained with Split-half reliability is called the coefficient of internal consistency. The computational formula for spit half reliability is $r_{tt} = \frac{2roe}{1+ troe}$

Where

$r_{tt=}$ split-half reliability index

o= odd items

e= even item

Recall that before applying the formula we need calculate with Pearson's Product Moment correlation. Take, for example, A class teacher administrated a test to 14 students in Mathematics which was scored over 30, and obtained the following scores.

Table 2: Students' performance scores in Mathematics

S/N	Score (s)
1	20
2	15
3	15
4	10
5	16
6	10
7	10
8	8
9	8
10	12

11	10
12	11
13	7
14	13

To estimate (calculate) the coefficient of internal consistency we need to calculate the PPMC first using the conventional algebraic formula as thus;

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2)\,(n\sum y^2 - (\sum y)^2)}}$$

Thus, we need to reproduce the scores in Table 3 to obtain our odd and even items.

Table 3: Odd and even items for the test instrument

Odd items	Scores (x)	X^2	Even items	Scores (y)	Y^2	∑xy
1.	20	400	2.	15	225	300
3.	15	225	4.	10	100	150
5.	16	256	6.	10	100	160
7.	10	100	8.	8	64	80
9.	8	64	10.	12	144	96
11.	10	100	12.	11	121	110
13.	7	49	14.	13	169	91
$N_1 = 7$	∑x = 86	$\sum x^2$ 1194	$N_2 = 7$	∑y= 79	$\sum y^2 = 923$	∑xy = 987

Using Pearson's Product Moment correlation

$$r = \frac{7 \times 987 - (86)(79)}{\sqrt{(7 \times 1194 - (86)^2)\,(7 \times 923 - (79)^2)}}$$

$$= \frac{6909 - 6794}{\sqrt{(8358 - 7396)\,(6461 - 6241)}}$$

$$= \frac{115}{\sqrt{(962)\,(220)}}$$

$$= \frac{115}{\sqrt{211640}}$$

$= \frac{115}{460.04}$

r = 0.249

r ≈ 0.25

Applying Spearman's Brown Prophecy formula to step up the coefficient of internal consistency, thus:

$r_{tt} = \frac{2roe}{1+ troe}$

$r_{tt} = \frac{2 \times 0.25}{1+ 0.25}$

$r_{tt} = \frac{0.50}{1.25}$

$r_{tt} = 0.40$

Equivalent-Forms/Alternate-forms/Parallel Forms Reliability

In Nigeria and other parts of the world, most examinations conducted in most cases are developed to accommodate alternate or parallel questions to help reduce test malpractices among test takers. These parallel forms are all designed to match the test blueprint, which is constructed to be similar in average item difficulty. Onunkwo (2002) noted that parallel form reliability is estimated by administering both forms of the test, (say form A and form B) to the same group of examinees. While the time between the two test administrations should be short, it does need to be long enough so that examinees' scores are not affected by maturation or fatigue. The examinees' scores on the two test forms are correlated to determine how similar the two test forms function. A parallel form reliability estimate is a measure of how consistent examinees' scores can be across test forms. The two forms (tests) are identical in every way except for the actual items included. After correcting the two scores they obtained a coefficient with Pearson's Product Moment correlations called the coefficient of equivalence.

Practical example: A teacher gave a test to 14 students in Mathematics in two forms (form A and form B) comprising 7 items in the test.

Table 4: Showing the instrument of the two forms (odd and even items) of the test instruments.

S/N	Form A (x= score)	x^2	S/N	Form B (y= score)	y^2	∑xy
1.	20	400	2.	15	225	300

3.	15	225	4.	10	100	150
5.	16	256	6.	10	100	160
7.	10	100	8.	8	64	80
9.	8	64	10.	12	144	96
11.	10	100	12.	11	121	110
13.	7	49	14.	13	169	91
$n_1 = 7$	$\sum x = 86$	$\sum x^2$ 1194	$n_2 = 7$	$\sum y = 79$	$\sum y^2$ = 923	$\sum xy = 987$

Using Pearson's Product Moment correlation to determine the co-efficient of equivalent becomes:

$$r = \frac{7 \times 987 - (86)(79)}{\sqrt{(7 \times 1194 - (86)^2)\,(7 \times 923 - (79)^2)}}$$

$$= \frac{6909 - 6794}{\sqrt{(8358 - 7396)\,(6461 - 6241)}}$$

$$= \frac{115}{\sqrt{(962)\,(220)}}$$

$$= \frac{115}{\sqrt{211640}}$$

$$= \frac{115}{460.04}$$

$r = 0.249$

$r \approx 0.25$

Inter-rater Reliability

All the methods for estimating reliability discussed so far are intended to be used for objective tests. When a test includes performance tasks or other items that need to be scored by human raters, then the reliability of those raters must be estimated. Thus, the reliability method asks the question, "If multiple raters scored a single examinee's performance, would the examinee receive the same scores? Inter-rater reliability provides a measure of the dependability or consistency of scores that might be expected across raters.Inter-rater reliability, inter-rater agreement, or concordance is the degree of agreement among raters. It gives a score of how much homogeneity or consensus a score is when rated by several raters (judges). For matched scores, rate one (1) while for unmatched scores rate as zero (0). Sum the matched scores and divide by the total number of respondents to arrive at the coefficient of consistency. The formula is the

number of matched questions divided by the total number of scores in the distribution. For example, 11 undergraduate students' thesis was rated by two raters (rater one and rater two) to determine their coefficient of internal consistency. Their various rating was presented in Table 5.

Table 5: Inter-rater Reliability (IRR)

S/n	Rater 1	Rater 2	Match
1	6	6	1
2	5	4	0
3	6	4	0
4	3	3	1
5	4	0	0
6	5	5	1
7	3	0	0
8	4	4	1
9	2	2	1
10	5	5	1
11	7	7	1

Table 5 revealed the scores of the two raters who rated the thesis. For serial number 1, rater one rated 6, and rater 2 also rated the thesis 6, since the two scores (rate) are similar, their match will be 1. For serial number two, rater 1 rated 5, and rater 2 rated 4. Since their ratings are different (5 and 4), their match was 0. Worthy of note is that, when tied scores occur, rate as zero. The total of the student's thesis rates are presented below:

The formula is IRR is thus:

TM/TS

Where

IRR= Interrater reliability

TM = Total Matched

TS = Total scores in the distribution

Therefore,

Match = 7

Total = 11

IRR = $\frac{7}{11}$

= 0.64

= 64%

Kuder-Richardson (K-R20 and K-R21)

Kuder-Richardson formulas 20 and 21 (also known as K-R20 and K-R2, are the most frequently reported internal consistency estimates of the K-R20 (Kuder & Richardson, 1937). They provide a sound under-estimate (that is a conservative or safe estimate) of the reliability of a set of test results. However, the K-R20 can only be applied if the test items are scored dichotomously (i.e., right or wrong). The **Kuder and Richardson Formula K-20** test check the internal consistency of measurements with dichotomous choices. In this method, a correct question scores 1, and an incorrect question scores 0. The test statistic is$PKR - 20 = \frac{K}{K-1}\left(1 - \frac{1\text{pjqj}}{\sigma^2}\right)$

where

k = number of questions

p_j = number of people in the sample who answered question j correctly

q_j = number of people in the sample who didn't answer question j correctly

σ^2 = variance of the total scores of all the people taking the test

Values range from 0 to 1. A high value indicates reliability while too high a value (over .90) indicates a homogeneous test. *Example 1:* A questionnaire with 11 questions is administered to 12 students. The results are listed in the upper portion of Figure 1. Determine the reliability of the questionnaire using Kuder and Richardson Formula 20. On the other hand, KR-21 is similar, except it's used for a test where the items are all about the same difficulty. The formula for estimating the test reliability using Kuder-Richardson Formula 6 calculator.

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

Where:

k - Number of questions

μ - Population means score

σ^2 - Variance of the total scores of all the people

PKR-21 - Reliability of the test

Cronbach's Alpha

Cronbach's alpha, developed by Lee Cronbach (1951) is a way to measure the reliability or internal consistency of a psychometric instrument. For example, a classroom teacher might administer a questionnaire to his/her students to determine students' learning outcomes in Mathematics. High-reliability index for the test would mean that the test is consistently measuring students' outcomes in Mathematics. A low-reliability index would mean it is measuring something else, or possibly nothing at all. Cronbach's alpha is most commonly used to see if questionnaires with multiple Likert scale questions are reliable. These questions are designed to measure latent variables. A latent variable is a hidden or unobservable variable, like a person's conscientiousness, neurosis, or openness. These variables are notoriously difficult to measure; Cronbach's alpha will tell you if the test you have designed is accurately measuring the latent variable you are interested in.

Calculating and Interpreting Cronbach Alpha Reliability

A Cronbach alpha estimate (often symbolized by the lowercase Greek letter ∞) should be interpreted just like other internal consistency estimates, that is, it estimates the proportion of variance in the test scores that can be attributed to true score variance. Put more simply, Cronbach alpha is used to estimate the proportion of variance that is systematic or consistent in a set of test scores. It can range from 0.00 (if no variance is consistent) to 1.00 (if all variance is consistent) with all values between 0.00 and 1.00 also being possible. For example, if the Cronbach alpha for a set of scores turns out to be .90, you can interpret that as meaning that the test is 90% reliable, and by extension that it is 10% unreliable (100% - 90% = 10%). However, when interpreting Cronbach's alpha, you should bear in mind at least the following two concepts:

1. Cronbach alpha provides an estimate of the internal consistency of the test, thus (α) alpha does not indicate the stability or consistency of the test over time, which would be better estimated using the test-retest reliability strategy, and (β) alpha does not indicate the stability or consistency of the test across test forms, which would be better estimated using the equivalent forms reliability strategy.
2. Cronbach alpha will be higher for longer tests than for shorter tests most comfortably used for a questionnaire that has sub-variables (Brown 1998 & 2001), and so alpha must be interpreted in light of the particular test length involved. It is worth mentioning the Differences and Similarities between Kuder-Richardson K-r20 and K-r21 and Cronbach's alpha reliability. While Cronbach alpha can also be applied when test items are scored dichotomously, alpha has the advantage over

K-R20 of being applicable when items are weighted (Kuder & Richardson, 1937). Hence, Cronbach alpha is more flexible than K-R20 and it is often the appropriate reliability estimate for language test development projects and language testing research.

Way on how to ensure the reliability of a research instrument

The reliability of any research instrument is certainly creating a strong research design, choosing appropriate methods and samples, and conducting the research carefully and consistently.

i. The research design used must be accurate. In doing this, the method and measurement technique adopted in the study must be of high quality and targeted to measure exactly what you want to know. For example, to collect data on a personality trait, you could use a standardized questionnaire that is considered reliable and valid. If the questionnaire is an adapted questionnaire, it should be based on established theories or findings of previous studies, and the questions should be carefully and precisely worded.

ii. Adopt the accurate sampling technique (sampling methods) to select your subjects or respondents that will constitute the study.

iii. To ensure high validity and reliability (valid generalization), do not fail to define the characteristics of the population of your research. E.g. people from a specific age range, socio-economic status, gender, geographical location, or profession). Also, the defined population must be the true representativeness of the population (Cozby, 2001).

iv. Apply your methods consistently by planning the methods/methodology carefully to make sure you carry out the same steps in the same way for each measurement. For example, if you are conducting interviews or observations, clearly define how specific behaviours or responses will be counted, and make sure questions are phrased the same way each time.

v. When you collect your data, keep the circumstances as consistent as possible to reduce the influence of external factors that might create variation in the results. For example, in an experimental setup, make sure all participants are given the same information and tested under the same conditions (Moskal&Leydens, 2000).

Applying Reliability in your Research Project, Thesis, or Dissertation

As a scholar in the field of educational measurement and evaluation, it is expedient to explain some major section (s) of any research undertaking in that validity and reliability are applied. Showing that you have taken them into account in planning your research and interpreting the results makes your work more credible and trustworthy. Thus:

i. *Literature review:* This section adopts validity and reliability in terms of what have other researchers done to devise and improve methods that are reliable and valid. That is, the gap in knowledge advancement and how your study intends to fill these gaps identified from the literature of other scholars.

ii. *Methodology:* How did you plan your research to ensure the reliability and validity of the measures used? This includes the chosen sampling techniques and sample size, suitability, and measuring techniques.

iii. *Results:* In the results section, if the validity and reliability are determined, clearly present the index in line with your findings.

iv. *Discussions:* In this part, discuss in detail how valid and reliable the findings/results of your study were. Were they consistent, and did they reflect true values? What are the surprises in the findings? Were the findings in conformity with expectations? Etc.

CONCLUSION

In education, the reliability of any instrument administered to the student must be estimated to foster academic excellence and consistency in the learning outcome of students. Any test tool that fails the test of validity is doomed to failure and should not be encouraged to be employed for any ramifications. Test developers in the academic environment should ensure that any test meant to be used in assessing a learner's ability should be subject to a reliability test, this will help to maintain a high level of integrity in the academic and research setting. Test that lack reliability can pose a problem to the teacher and the learners at the same time, teachers with very good knowledge and understanding of reliability can satisfactorily take cornet decisions that can foster learners in all the domains of learning. In conclusion, an understanding of reliability and its application in the classroom context allows educators (teachers) to make decisions that improve the lives of their students both academically and socially, as these concepts teach educators how to quantify the abstract goalsof the teaching undertaking in their various schools.

Classroom Assessment and Evaluation Experience

The role of classroom evaluation cannot be undermined in the school setting. This is because, without evaluation, the worth or value of a progrmme cannot be ascertained. Reliability is the surest way of determining the degree of consistency of an instrument. In this regard, teachers should have a better understanding of how to practically apply reliability using different approaches depending on the nature of the data. It is therefore advisable for teachers at the classroom level, who may want to establish the reliability of the instrument given to students to create clear instructions for each question by presenting questions that capture the material taught (possibly from the course outline or syllabus or scheme of work). Put in another way, teachers are expected to phrase each question clearly so that students have a better understanding of the expected goals of the course. similarly, teachers are towrite items that discriminate between good and poor students and are of an appropriate difficulty level. In this case, good planning of the test and writing the items well ahead of the time the test is to be given.Also, teachers in the classroom should endeavor to always seek feedback regarding the clarity and thoroughness of the assessment from students. This will help to improve the reliability of the classroom test. Rightly stated by Eyong(2019) to improve reliability the in a classroom assessment and evaluation requires five (5) simple steps teachers must do viz;confirmthat learners are conversant are familiar with the assessment and evaluation, ensure that the students are well informed of the timing of the examination, after all, lesson, revise with the students the concepts to be covered in the assessment in each subject and finally, ensure that there is no setting interaction in the examination conduct that is, have a consistent environment for the students.

REFERENCES

Cozby, P.C. (2001). Measurement Concepts. *Methods in Behavioral Research* (7th ed.). California: Mayfield Publishing Company.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.). *Educational*

Eyong, E.I. (2019). Principles and Applications of Educational Test, Measurement, and Evaluation (Psychological assessment Instruments and Techniques in Action).Aganzo Printing Press. Muri Road, Kaduna. Kaduna State, Nigeria.

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

Measurement (2nd ed.). Washington, D. C.: American Council on Education.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741–749.

Moskal, B.M., &Leydens, J.A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation, 7(10),* 12-18.

Onunkwo, G.I. (2002). Fundamentals of Educational Measurement and Evaluation. The University of Port Harcourt, Port Harcourt, Nigeria.

