European Journal of Statistics and Probability, 12 (1) 54-78, 2024
Print ISSN: 2055-0154(Print),
Online ISSN 2055-0162(Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Algorithms for Reliability Estimate as a Test – Quality Indicator

Thomas Olabode Abe, (Ph.D)
Counseling Psychology Department, College of Education,
Bamidele Olumilua University of Education Science and Technology Ikere

Samuel Oluwaseyi Olofin (Ph.D)
Department of Science Education, Faculty of Education, Ekiti State University, Nigeria

doi: https://doi.org/10.37745/ejsp.2013/vol12n15478 Published April 18, 2023

Citation: Abe T.O. and Olofin S.O. (2024) Algorithms for Reliability Estimate as a Test – Quality Indicator, *European Journal of Statistics and Probability*, 12 (1) 54-78

ABSTRACT: *This study examined the algorithms for reliability estimate as a test-quality indicator. It was discussed along the methods of estimating reliability such as: Test-retest as measures of stability, Equivalent or Alternative –forms reliability as measures of equivalence and stability, while the measures of internal consistency are Split-half, Kuder-Richardson 20 and 21, Coefficient alpha, Hoyt's analysis of variance, Scorers (Judge) reliability and Inter-rater reliability. Also, using the reliability coefficient as a Test quality indicator was addressed and variables that affect reliability estimate are itemize as test length, test content, test difficulty, item discrimination, group heterogeneity, student motivation, students testwiseness, time limit and security precautions. Therefore, this paper recommends that, in order to demystify the course at any level of our educational system, specialist in the field of Tests, Measurement and Evaluation should be strictly allowed to handle the course and every professional teachers should be abreast to the procedural ways of estimating reliability of test in the classroom examination as a quality indicator in the teachers' made test.*

KEY WORDS: algorithms, reliability, estimate, test quality and indicator.

INTRODUCTION

Reliability in statistics and psychometrics is the overall consistency of a measure (William, 2006). A measure is said to have a high reliability if it produces similar results under consistent conditions." It is the characteristic of a set of test score that relates to the amount of random error from the measurement process that might be embedded in the scores. Scores that are highly reliable are accurate, reproducible and consistent from one testing occasion to another. That is, if the testing process were repeated with a group of test takers, essentially the same results would be obtained.

Various kind of reliability Co-efficient, with value ranging between 0.00 (much error) and 1.00 (no error) are usually used to indicate the amount of error in the score (DavidShofer & Murphy, 2005; Abe 2005; Cortina 1993).

Reliability of a test is the degree to which a test is consistent, stable, dependable or trustworthy in measuring what is supposed to measure (Bandele & Abe, 2017). Agwubike and Momoh (1995) posited that, reliability refers to the degree of consistency between two sets of scores or (observations) obtained with the same instrument. While Uzoagulu (1998) and Alonge (1989 & 2004) argued that reliability of a test instrument is the consistency of the test in measuring whatever it purports to measure. It is also concerned with the repeatability of a rest. We can also look at reliability of a test in terms of inter cohesiveness across forms of the same test. Theoretically, reliability is defined as the ratio of the true variance to the variance of the observed scores or simply as the proportion of Test score differences (Brown, 1970; Oyerinde 1986; Agwubike & Momoh, 1995; Alonge, 1989 & 2004 and Cecil, Ronald & Victor 2011). This show that reliability can be expressed as

$\mathbf{r_{tt}} = \mathbf{V_t} + \mathbf{V_0}$ or

$\mathbf{r_{tt}} = \mathbf{S^2t} + \mathbf{S^2_0}$ ----------(1)

$\mathbf{V_0}$ **or** $\mathbf{S^2_0}$ = Variance of the observed scores

Where $S^2t + S^2_0$ = Variance of the true scores

This equation according to Brown (1970) can be related to the basic equation that depicts the relationship between obtained scores, true score and error scores. Bandele and Abe (2017) argued that, relating the basic equation to the definition of reliability we have

$$X_o = X_T + X_E \text{.....(2)}$$

Where: $\mathbf{X_o}$ = Score obtained by the testee

$\mathbf{X_T}$ = The testee true score

$\mathbf{X_E}$ = Error associated with the score

Which may be positive or negative which Abe (2006) and Bandele and Abe (2017) refer to as leniency or severity errors that necessitated the need for moderation of school based assessment scores at secondary school level. That is to say, when X_E is positive, it shows that, the true score is over estimated or leniently graded and when it is negative, it connotes that, the true score is under estimated or severely graded (Abe, 1995; Abe & Gbore 2006; Abe & Alonge, 2010; Bandele & Abe 2017)

Now, when equation 2 above is applied to a set of scores it can be written on terms of variances as show below:

$$V_0 = V_t + V_e \text{----------} (3)$$

This equation leads to the theoretical deviation of the concept of reliability. The equation

$V_0 = V_t + V_e$ which indicate the relationship between true score and error variance and their additive relationship to form total variance of test scores, is a key to the development of test theory concept of reliability. If the equation is divided through each side by the total variance V_0, the following equation is the result:

$$\frac{V0}{V0} = \frac{Vt}{V0} + \frac{Ve}{V0}$$

$$1 = \frac{Vt}{V0} + \frac{Ve}{V0} \text{---------- (4)}$$

The equation indicate that the ratio of the true score variance to the total variance plus the ratio of the error variance to the variance equal 1. The ratio of the true score variance to the total variance or that proportion of the variance in the obtained score which may be attributed to the true scores, forms the basic definition of reliability.

By Substituting r_{tt} for $\frac{Vt}{V0}$ in equation 4 , we have Realibilty equal 1 minus the error variance divided by the obtained variance

$$r_{tt} = 1 - \frac{Ve}{V0} \text{ ---------- (5)}$$

The proportion of the total variance which is true score variance may be obtained by simply subtracting the proportion of the variance which is error from 1, The two basic theoretical equations for the reliability coefficient are :

$$r_{tt} = \frac{Vt}{V0} : \quad r_{tt} = 1 - \frac{Ve}{V0}$$

Hypothetical illustration: The table below is showing the means, Variance and standard deviations of true and error components of an observed measure.

	X_t	X_e	X_0
	5	-3	2
	15	+3	18
	16	-3	13
	21	-2	19
	27	+2	29
	29	0	29
	33	+10	43

	36	-4	32
	37	-2	35
	43	-1	42
$\sum$X = Total	**262**	**0**	**262**
$\bar{X}$	**26.2**	**0.0**	**26.2**
$\sum X^2$	**1255.6**	**1.55**	**1411**
V	**125.6**	**154.5**	**141.1**
SD	**11.2**	**3.9**	**11.9**

Therefore, from equation

$r_{tt} = \frac{Vt}{V0}$ Where V_t = 125.6 and V_0 = 141.1

$$r_{tt} = \frac{125.6}{141.1} = 0.89$$

Using

: $r_{tt} = 1 - \frac{Ve}{V0}$

$= 1 - \frac{15.5}{141.1} = 1 - 1.1$

$= 0.89$

The above illustration is concerned with the test theory. There are no practical method for directly measuring a testee's true -score on a test. Based on this assumption, to estimate the reliability of a set of scores, it is necessary to use certain computational methods or procedures derived from test theory

Alonge (1989 & 2004) and Uzoagulu (1998) argued that, from theoretical expression reliability r_{tt} can range from plus one when there is no error in the measurement is virtually all error in measurement = then

$$\mathbf{r_{tt}} = \mathbf{1 - \frac{0}{Ve}} = \mathbf{1 - 1 = 0}$$ Since Ve = 0

When there is no error in measurement but when **= 0** then

$\mathbf{r_{tt} = 1 - \frac{0}{Ve} = 1 - 0 = 1.}$ In practice, no instrument is perfectly reliability of an instrument; it is pertinent to note the following: Reliability estimate is strictly a statistical concept.

Negative reliability Coefficients are possible empirically

The r_{tt} must be computed from some data.

the greater the value of the coefficient the higher the reliability of the test

The method to be used to determine the r_{tt} will depend on conditions of the test administration.

Statistically, the term "Estimation" refers to process by which we obtain from a sample some statistic or a value, which helps to determine its corresponding parameter of the population from what the sample was drawn. While an estimate is a statistic obtained from a sample that enables the researchers to make a projection(s) about its corresponding population parameter (Spiegel, 1970, Ugwuja, 2004).

The goal of estimating reliability is to determine how much of the variability in test scores is due to errors in measurement and how much is due to variability in true scores (Davidshofer & Murphy; 2005).

Methods of Estimating Reliability

Test-retest reliability(Measures of Stability): assesses the degree to which test scores are consistent from on test administration to the next measurements are gathered from a single rater who uses the same method or instruments and the same testing conditions (Cortina, 1993). While Cecil, Ronald & Victor (2011) argued that Test-retest reliability is sensitive to measurement error due to time sampling and is an index of the stability of scores over time. One important consideration when calculating and evaluating test-retest reliability is the length of the interval between the two test administrations. The test-retest approach does have significant limitations, the most prominent being carry-over effects from the first to second testing. The Product-Moment Method and Rank difference Method is used in calculating reliability co-efficient.

Equivalent or Alternative – forms Reliability (Measures of equivalence and Stability): It is also called Inter-method reliability which test scores are consistent when there is a variation in the methods or instruments used. It assesses the consistency of results across items with a test. Based on Simultaneous administration that is primarily sensitive to measurement error due to content sampling. It equally based on delayed administrations sensitive to measurement error due to content sampling and time sampling but cannot differentiate the two types of error (Cortina, 1993 & Cecil, Ronald & Victor, 2011).

Split- Half Reliability (Measures of internal Consistency): It can be calculated from one administration of a test and reflect Error due to content sampling.

Computation of a Co-efficient of split-half reliability generally entails three steps.

Step 1: Divide the test into equivalent halves.

Step 2: Calculate a Pearson between Scores on the two halves of the test.

Step 3: Adjust the half – test reliability using the spearman Brown formula as

$\mathbf{R_{xx}} = \frac{\boldsymbol{nrxy}}{\mathbf{1-(n-1)}\boldsymbol{rxy}}$ and

$rsb = rxx = \frac{2\ x\ Realiability\ on\ half\ test}{1+Realiabilty\ on\ haif\ test}$

Where r_{xx} is equal to the reliability adjusted by the spearman – Brown Formula r_{xy} is equal to the Pearson r in the original – length test and n is equal to the number of items in the revised version divided by the number of items in the original version and n = factor by which the test length is increased.

Where r_{xx} or r_{sb} = estimated reliability of the whole test.

r_{hh} or = Reliability of half Test

For example if the correlation Coefficient between the halves of the test is 0.50 the reliability of the whole test is:

$\mathbf{r_{xx}} = \frac{2(0.50)}{1+0.50} = \frac{1.0}{1.50} = 0.51$

Method II Rulon

Rulon (1939) developed the formula:-

$\mathbf{r_{xx}} = \mathbf{1} - \frac{\boldsymbol{sd^2\ d}}{\boldsymbol{Sd^2 t^1}}$ **or** $\mathbf{1} - \frac{\boldsymbol{v^1\ d}}{\boldsymbol{v^1\ t}}$

Where:

Vd or SD^2d = Variance of the difference between each students' score on both halves.

Vt or SD^2t = Variance of total Score.

The split half method gives an indication of the extent to which the items represent the universe of attributes or adequacy of the content sampled.

Example:

Suppose the correlation between old and even halves of your mid-term in the course was 0.74, calculate using Spearman – Brown Formula

Solution:

$\mathbf{r_{xx}} = \frac{2\ X\ 0.74}{1+0.74} = \frac{1.48}{1.74}$

$r_{xx} = 0.85$

Co- efficient Alpha to Estimate split half Reliability, Calculate or using co- efficient Alpha to estimate split half Reliability when the variances to the two halves of the test are unequal.

The reliability Coefficient of 0.85 estimates the reliability of the full test when the odd – even halves correlation at 0.74. This demonstrates that the uncorrelated split half reliability coefficient present an under estimated of the reliability of the full test.

Table 1: Showing the----

Half – Test Coefficients and Corresponding

Full – Test Coefficients corrected with the Spearman Brown Formula

Half-Test Correlation	**Reliability R_{xx}**
0.50	0.61
0.55	0.71
0.60	0.15
0.65	0.29
0.70	0.82
0.75	0.86
0.80	0.89
0.85	0.92
0.90	0.95
0.95	0.95

Source: Cecil, Ronald & Victor, 2011.

Calculate or Using Co-efficient, Alpha to Estimate split Half Reliability when the variances for the two Halves of the Test are Unequal.

Formula $\alpha = 2\left[\frac{Sx^2 - [Sy1^2 + Sy2^2]}{Sx^2}\right]$

e.g if $Sx^2 = 11.5$, $Sy1^2 = 4.5$, $Sy2^2 = 3.2$

$Sy1^2 + Sy2^2 = 4.5 + 3.2 = 7.7$

By Substitution

$\alpha = 2\left[\frac{11.5 - [7.7]}{11.5}\right]$

$= 2\left[\frac{3.8}{11.5}\right]$

$\alpha = 0.66$

Flanagan (1937)

$$\alpha = 2\left[\frac{Sx^2}{Sx^2} \frac{-[Sy1^2 + Sy2^2]}{Sx^2}\right]$$

$$\alpha = 2\left[1 - \frac{[Sy1^2 + Sy2^2]}{Sx^2}\right]$$

It estimate the error variance and the formula is parallel to Rulon (1935)

Kuder – Richardson (Measures of Internal Consistency)

Kuder and Richerson (1937) developed several formulae for obtaining reliability estimate using scores from one test administration. The correlation co-efficient computed with the scores is a measure of internal consistency. A basic assumption of the method is that items in the test are homogeneous and therefore posses inter-item consistency. Kuder-Richardson formulae 20 and 21 are widely applicable in research and evaluation.

The formulae are:

K – R 20: $\mathbf{r_{xx}} = \frac{Ns}{N-1}\left[\frac{S^2 - \sum PQ}{S^2}\right]$

$$\mathbf{r_{xx}} = \frac{n}{N-1}\left[1 - \bar{x}\left[\frac{n - \bar{x}}{ns^2\, t}\right]\right]$$

Where: r_{xx} = reliability Coefficient

n = number of item

p = proportion of people who got the item right

q = Proportion of people who got the item wrong

pq = Variance of a single item scored dichotomously

$\sum$ = Summation sign indicating that pq is summed for all items

St^2=Variance of the test

$\bar{x}$ =Mean of the total Test.

The calculation of Reliability Using KR_{20}

European Journal of Statistics and Probability, 12 (1) 54-78, 2024
Print ISSN: 2055-0154(Print),
Online ISSN 2055-0162(Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Formula KP_{20} =

N_s = number of test takers = 50

N = number of items = 6

S^2 = Variance = 2.8

Example: 1

Item	**Numbers of Test taken responding correctly**	**P**	**q = 1-P**	**Pq**
1	12	$\frac{12}{50}$	0.76	0.18
2	41	$\frac{41}{50}$	0.18	0.15
3	18	$\frac{18}{50}$	0.64	0.23
4	29	$\frac{29}{50}$	0.42	0.24
5	30	$\frac{30}{50}$	0.40	0.24
6	47	$\frac{47}{50}$	0.06	0.06

$\sum pq$ = 1.10s

$$\mathbf{S^2 = \frac{\sum x^2 - \left[\frac{\sum x^2}{Ns}\right]}{Ns-1}}$$

S^2= 2.8 $S^2 - \sum pq$ = 2.8 – 1.1 = 1.7

$$\mathbf{\frac{S^2 - \sum pq}{S^2} = \frac{1.7}{2.8}}$$

= 0. 607

$$N = 6\frac{N}{N-1} = \frac{6}{6-1}$$

= 1.2

KR-20 = (1.2)(0.67) = 0.73

KR-20 is applicable when test items are scored dichotomously, that is, simply right or wrong, as O or 1. It also deals with test items that produce scores with multiple values (e.g. 0, 1 or 2).

Example II on K20

Consider these data for a five item test administered to Six Students. Each item could receive a score of either 1 or 0.

Student	**Item 1**	**Item 2**	**Item 3**	**Item 4**	**Item 5**	**TOTAL SCORE**
A1		0	1	1	1	4
B1		1	1	1	1	5
C1		0	1	0	0	2
D0		0	0	1	0	1
E1		1	1	1	1	5
F1		1	0	1	1	4
P_1 =	0.8333	0.5	0.6667	0.8333	0.6667	SD^2 = 2.25
P_2=	0.1667	0.5	0.3333	0.1667	0.3333	
P_1 x q_1 =	0.1389	0.25	0.2222	0.1389	0.2222	

$\sum P_1q_1$ = **0.1389 + 0.25 + 0.2222 + 0.1389 + 0.2222**

= **0.972**

$KR_{20} = \frac{5}{4}\left[\frac{2.25-0.972}{2.5}\right]$

$KR_{20} = 1.25\left(\frac{1.278}{2.25}\right)$

KR_{20} = **1.25 (0.568)**

KR_{20} = **0.71**

Coefficient Alpha (Measures of Internal Consistency)

Cronbach (1951) argued that Cronbach Alpha is more general form of KR-20 that, deals with test items that produce scores with multiple values (e.g. 0, 1 or 2).For items that require more than two response like in interest inventories personality test (agree undecided disagree) Weight could be given to the responses say, 3, 2 , 1 or +1, 0, -1, 1 or any other system of Weight. In case of essay test, the students' score on each question could take a range of Values. Coefficient alpha is sensitive

to measurement error due to content sampling and is also a measure of item heterogeneity. It can be applied to tests with items that are scored dichotomously or that have multiple values. The formula for calculating coefficient alpha is:

Coefficient alpha = $\left(\frac{K}{K-1}\right)\left(1-\frac{\sum SDi^2}{SD^2t}\right)$ **or** $\left(\frac{K}{K-1}\right)\left(1-\frac{Vi}{Vt}\right)$

Where: **k** = number of items

SDi^2 = Variance of individual items.

S = Variances of total test first scores-Co-efficient alpha is more broadly applicable. It has become the preferred statistic for estimating internal consistency (Keith & Reynolds, 1990).

Example: iv

Consider these data for a five item test that was administrated to six students. Each item could receive a score ranging from 1 to 5

	Item 1	**Item 2**	**Item 3**	**Item 4**	**Item 5**	**Total Score**
Student 1	4	3	4	5	5	21
Student 2	3	3	2	3	3	14
Student 3	2	3	2	2	1	10
Student 4	4	4	5	3	4	20
Student5	2	3	4	2	3	14
Student 6	2	2	2	1	3	10
S	0.8056	0.3333	1.4722	1.555%	1.4722	S=18.81

∑S0.8056 + 0.333 + 1.4722 + 1.5556 + 1.4722 = 5.63889

Coefficient alpha = $(\frac{5}{4})$ **(1 -** $\frac{5.63889}{18.81}$ **)**

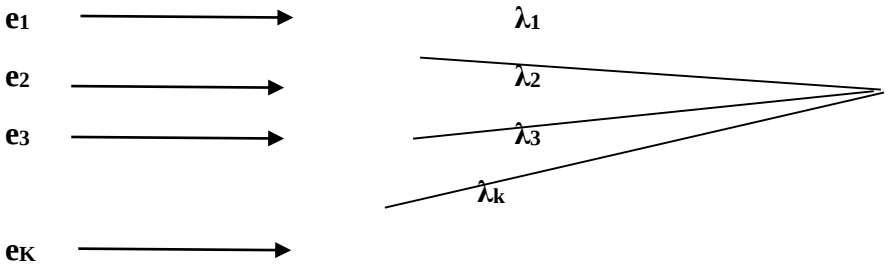
= **1.25 [1 – 0.29978]**

= **1.25 [0.70]**

= **0.875**

Cronbach's Alpha

A = Tau – equivalent Measurement Model



A tau – equivalent measurement model is a special case of a congeneric measurement model, hereby assuming all factors loadings to be the same i.e. $\lambda = \lambda_1 = \lambda_2 = \lambda_3 = \ldots = \lambda_k$

Uzoagulu(1998) posited that, the following should be noted:

- The Cronbach alpha (α) estimates correlation of an instrument with an alternative form which is composed of the same number of items.
- It can be applied for polychotomous items.
- It is a generalization of the Kuder-Richardson coefficient.
- The alpha depends on the number of items in the test or instrument as well as on the average inter-item correlation.
- As the average correlation among items increases, the alpha value also increases.
- As the number of items in a test increases, the value of alpha also increases.

Gronlund (1976) stated that, the longer the test, the higher the reliability. That is to say, the internal consistency appreciates with increase in the number of items in the test. For example, in a 130-item test instrument captioned: Introductory Technology Achievement Test (ITAT), the internal consistencies of the subtests were found to be as shown below:

Internal Consistencies of ITAT and its Subtests

S/N	Subtests	Cronbach Alpha Coefficients	Number of Items in the Test
1	Ceramics	0.224	6
2	Plastic and Rubber	0.386	6
3	Food Technology	0.390	7
4	Technical Drawing	0.392	15
5	Building Works	0.440	22

6	Wood works	0.483	24
7	Electricity/Electronics	0.520	25
8	Metal Works	0.597	35

Standardized Cronbach's alpha

The Standardized Cronbach's alpha can be defined as:

α Standardized=

Where k is item and the mean of the k (k – 1)/2 non – redundant correlation Coefficient (i.e., the mean of an upper triangular, or lower triangular correlation matrix)

Cronbach's alpha Internal Consistency

$0.9 \leq \alpha$	Excellent
$0.8 \leq \alpha < 0.9$	Good
$0.7 \leq \alpha < 0.8$	Acceptable
$0.6 \leq \alpha < 0.7$	Questionable
$0.5 \leq \alpha < 0.6$	Poor
$\alpha < 0.5$	Unacceptable

To determine the inter-item Correlation Coefficient, the Cronbach alpha can be used with the formula:

$$\boldsymbol{\alpha} = \frac{k\bar{r}}{1+\overline{r(k-1)}}$$

Where k = number of items in the test

$\bar{r}$ = The mean of inter-item Correlation.

The can be obtained by summing all the coefficient values of the correlation between all the items. For example, the sub-tests of the Introductory Technology Achievement Test (ITAT) had a Correlation Matrix as shown in the table below:

European Journal of Statistics and Probability, 12 (1) 54-78, 2024
Print ISSN: 2055-0154(Print),
Online ISSN 2055-0162(Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Correlation al Matrix of the ITAT and its Subtests

Content Items	MW	EE	WW	BD	TD	FT	PR	CR
Metal Work (MW)	1.000							
Electricity/Electronic (EE)	- 0.123	1.000						
Wood work (WW)	0.126	0.545	1.000					
Building (BD)	- 0.472	0.001	0.008	1.000				
Technical Drawing (TD)	0.881	- 0.096	0.258	- 0.414	1.000			
Food Technology (FT)	- 0.589	0.305	0.348	0.578	-0.546	1.000		
Plastic and Rubber (PR)	- 0.377	0.213	- 0.013	0.347	0.511	- 0.505	- 1.000	
Ceramic (CR)	0.218	0.359	0.587	- 0.181	0.242	-0.011	- 0.591	1.000

The mean inter-item correlation r can be obtained by summing up all these Correlations and dividing the sum by 28 (The number of correlations) coefficients the r for this coefficients matrix can be computed as follows:

$\bar{r}$ = - 0.123 + 0.123 +0.126 + (-0.472) + 0.881 + (-0.589) + (-0.337) + 0.218 + 0.545 + (-0.001) + (-0.096) + 0.305 + 0.263 + 0.359 + 0.008 + 0.258 + 0.348 + (-0.013) + 0.587 + (-0.414) + 0.578 + 0.347 + (-0.181) + (-0.546) + 0.511 + 0.242 + (-0.505) + (- 0.011) + (-0.593)

= 2.664

This depicts the mean inter-item correlation for the ITAT which indicates a low and positive inter-item correlation coefficient. The inter-item correlation can be found using the formula:

α Standardized

$$\alpha = \frac{k\bar{r}}{1+\bar{r}(k-1)}$$

k = 130 $\bar{r}$ = 0.0951

$$= \frac{130 \times 0.0951}{1+0.0951\,(130-1)}$$

α = 0.91. This is an excellent Reliability Coefficient

European Journal of Statistics and Probability, 12 (1) 54-78, 2024
Print ISSN: 2055-0154(Print),
Online ISSN 2055-0162(Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Hoyt's reliability (Measures of Internal Consistency)

This is the statistical procedure used when analyzing variability in a set of data is called an analysis of variance (ANOVA). The purpose of this procedure is to learn the degree to which the Variation of the individual item scores, also called the total Variance, consists of differences among student average per-item scores and differences among item averages. A third source of variation is an interaction between students and test items, something considered to be "error" which relates to inconsistency in ranking. The tests of significance associated with ANOVA are not employed here. Only the partitioning of total variance is of interest. The calculation works as follows. Variability in data can be expressed in terms of various mean squares, each represented by symbol of the form MS. The subscript identifies the data used in the calculation, as described subsequently. Each of these mean squares is found from a related sum of squares represented by the symbol SS. To calculate any mean square, first find the corresponding SS by adding the squared deviations from the grand mean of each data value of the given type as reflected in the subscript, and then divide this sum of squares by the respective degree of freedom, df which represents the number of free choices within a certain constraint. When the MS_s is known, Hoyt's reliability is when the MS are known Hoyt's reliability is calculated as follows:

$$\textbf{Reliability} = \frac{\textbf{MSst} - \textbf{MSe}}{\textbf{MSst}}$$

Where; $\mathbf{MS_{st}}$ = is a measure of the amount of variation to student average- peer item scores

$\mathbf{MS_e}$ = is the amount of Variation associated with error a measure of the amount of interaction between students and items with ranking of student s differing from item to item. A reliability of 1 is perfect and is approached when MS_e is much smaller than MS_{st}.

The seven – step calculation of Hoyt's reliability are:

Calculate SS_t: The total sum of squares based on each individual item score.

Calculate SS_{st}, the sum of square attributable to student achievement, by in essence replacing each of the individual item scores in each row with the corresponding student average-per-item score.

Calculate the sum of squares attributable to differences between items, SS_i by in essence replacing each score in a column by the corresponding item average given the row.

Calculate SS_t, the amount of variation associated with error. The total variation is made up of three components, namely, student variation item – score variation and error variation that is;

European Journal of Statistics and Probability, 12 (1) 54-78, 2024
Print ISSN: 2055-0154(Print),
Online ISSN 2055-0162(Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

SS_t = $SS_{st} + SS_i +$

while the respective MS_s are found using the formula

$$\mathbf{MS_?} = \frac{SS}{df}$$

$$\mathbf{MS_e} = \frac{SSe}{Dfe}$$

$$\textbf{Reliability } \mathbf{r_{xx}} = \frac{\mathbf{MSst - MSe}}{\mathbf{MSst}}$$

Example: Student score on each item

Student	**Item**			**Student score (Average per item)**
	1	2	3	
A	1	2	3	2.00
B	0	1	2	1.00
C	1	2	3	2.00
D	0	1	2	1.00
Item average	0.5	1.5	2.5	1.50 grand average

Calculate the reliability from above table

Compute SS_t

SS_t = **(**$1 - 1.5)^2 + (2 - 1.5)^2 + (3 - 1.5)^2 + (0 - 1.5)^2 +$

$(1 - 1.5)^2 + (2 - 1.5)^2 + (3 - 1.5)^2 + (0 - 1.5)^2 +$

$(1 - 1.5)^2 + (2 - 1.5)^2 = 11$ which is total variance in the data.

Compute SS_{st} = $(2 - 1.5)^2$ x 3 scores/row + **(**$1 - 1.5)^2$ x 3 +

$(2 - 1.5)^2$ x 3 + **(**$1 - 1.5)^2$ x 3

= 3 which is the Variation associated with Student achievement

= MS_{st} = 1 i.e. MS_{st} = 3/3 = 1

Compute SS_i = $(1.0 – 1.5)^2$ X 4 score/column

+ $(1.0 – 1.5)^2$ SK 4 + $(2.5 – 1.5)^2$

= 6 which is the variation associated with item difficulty

SS_e = 11 – 3 - 6 = 2

MS_e = i.e. $df_e = df_{st}$ x df

I.e. df_{st} x df_i = 3 x 2 = 6

MS_e = = 0.33

Reliability = = = 0.67.

Scorer (judge) reliability (Measures of internal consistency)

The concern for scorers reliability is less for most objective questions (items). But for essay tests which are subjective threatens reliability. It is necessary to determine how much error is due to the person scoring. Two or more examiners are used to mark the same set of papers. For all the examiners the resulting scores are correlated, the resulting correlation coefficient is the scorers reliability (Agwubike & Momoh, 1995).

Inter - Rater Reliability

Inter-rater reliability assesses the degree of agreement between two or more raters in their appraisals. It is a numerical estimate/measure if the degree of agreement among raters. It provides adequate levels ensure accuracy and consistency in the assessment. While inadequate levels indicate scale inadequacy and need for additional rater training. The basic model for calculating inter-rater reliability is percentage agreement in the two-rater model. The test, this is referred to as inter rater reliability. If the Scoring of an assessment relies on subjective judgment, it is important to evaluate the degree of agreement when a different individual score was given, this is referred to as inter-rater reliability.

European Journal of Statistics and Probability, 12 (1) 54-78, 2024
Print ISSN: 2055-0154(Print),
Online ISSN 2055-0162(Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Example: Table 8

STUDENTS	RATER 1	RATER 2	AGREEMENT
1	2	2	1
2	3	4	0
3	2	2	1
4	3	3	1
5	3	4	0

(1) Calculate the number/rater of ratings that are in agreement.

(2) Calculate the total number of ratings.

(3) Convert the Fraction to a percentage.

Table 9

STUDENTS	RATER 1	RATER 2	RATER 3
1	2	2	1
2	3	4	0
3	2	2	1
4	3	3	1
5	3	4	0

Percent Agreement = 60% **3/5**

Benchmarking inter-rater reliability percentage

Rules-of-Thumb for Present Agreement

Number of Ratings	High Agreement	Minimal Agreement	Qualification
4 or Fewer Categories	90%	75%	No ratings more than one level apart
5-7 Categories	75%		Approximately 90% of ratings identical or adjacent

Percentage Agreement = 60%, what does this mean? Since 60% is lower than minimal benchmark, inter-rater reliability is unacceptable.

Problems with the percentage agreement statistic are:

Unintuitive and more difficult to hand calculate with multiple raters

Absolute agreement is an unforgiving standard A common solution is to count adjacent rating as being in agreement and this can result in meaningless reliability estimate.

Does not take chance agreement into account-over-estimating the inter-rater reliability estimate.

Cohen's Kappa Statistic (k)

Is a measure of the agreement between two raters, where agreement due to chance is factored out (Smeeton, 1985). Kappa Statistic (k) measures inter-rater agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation, as k takes into account the possibility of the agreement occurring by chance. Robert & Marco (2011) argued that some researchers suggested that, it is conceptually simpler to evaluate disagreement between items.

$$\mathbf{k} = \frac{Po-Pe}{1-Pe} = 1 - \frac{Po-Pe}{1-Pe}$$

Where $\overline{\overline{P_o}}$ = the relative observed agreement among raters (identical to accuracy).

$\mathbf{P_e}$ = the hypothetical probability of chance agreement using the observed data to calculate the probabilities of each observer randomly seeing each category.

K = 1 Implies the raters are incomplete agreement

And if

K = 0 There is no agreement among the raters other than what would be expected by chance (as given by P_e).

It is possible for statistic to be negative, which implies that there is no effective agreement between the raters or the agreement is worse than random.

Cohen (1960) and Simi & Wright (2005)

In terms of symbols this, is: $\mathbf{K} = \frac{Po-Pe}{1-Pe}$

Where P_o is the proportion of observed agreement and P_e is the proportion of agreement expected by chance. The data for paired ratings on a 2x2 contingency table is given as:

		Rater B		Total
		Yes	**No**	
Rater A	**Yes**	a 22	b 2	g 24
	No	c 4	d 11	g 15
Total		F 26	f 13	n 39

$\mathbf{P_o} = \frac{a+d}{a+b+c+d} = \frac{22+11}{22+2+4+11} = \mathbf{0.8462}$

$\mathbf{P_e} = \mathbf{P_{yes}} + \mathbf{P_{no}} = \mathbf{0.5385}$ **where**

$\mathbf{k} = \frac{Po-Pe}{1-Pe}$

$\mathbf{P_{yes}} = \frac{a+d}{a+b+c+d} \cdot \frac{a+c}{a+b+c+d} = \frac{22+2}{22+2+4+11} \times \frac{22+4}{22+2+4+11}$

$= \mathbf{0.42}$

$\mathbf{P_{no}} = \frac{c+d}{a+b+c+d} = \frac{b+d}{a+b+c+d}$

$= \frac{4+11}{22+2+4+11} \times \frac{2+11}{22+2+4+11}$

$= \mathbf{0.13}$

$\mathbf{P_e} = \mathbf{P_{yes}} + \mathbf{P_{no}} = 0.42 + 0.13$

$= 0.55$ where

$\mathbf{K} = \frac{Po-Pe}{1-Pe} = \frac{0.85-0.55}{1-0.55}$

$= 0.67$

The range of possible value of kappa is -1 to 1.Though; it usually falls between o and 1. Unity represents perfect agreement, indicating that the raters agree in there classification of every case. Zero indicates agreement no better than that expected by chance as if the raters had simply "guessed" every rating. A negative kappa would indicate agreement worse than that expected by chance (Bartko & Carpenter, 1976).

Using the Reliability coefficient as a Test-Quality Indicator

Using reliability measurement has significant test – design implications as well as ethical consideration. The reliability of a test is improved by restricting it to item for which student performance on the item is strongly related to student total score. That is to say, test developer tends to prefer items for which a positive relationship exists between how well students do on the

item and how well they do on the total test. This relationship is a major Criterion for including an item in a standardized test design ranking students. The higher the relationship between item score and total test score for the item on a test, the greater the reliability of the test.

Test – items for which test-takers' scores' correlate poorly with their total test scores may be replaced. On such a low –relationship items, students who are typically unsuccessful tend to do better and students who are typically successful experience difficulty. This result is often anathema to standardized – test designers and is perhaps unfair to students.

Note however that, test items that consistently separate low achievers from high achievers are preferred. Items that fail to distinguish among students in this way are replaced by items that do. Therefore, one cannot detect improvements in the uniformity of student achievement, much as one would be unable to ever see most students become i.e. "above average". The question of reliability is expanded to respond to such questions as "with what consistency does test categories students with respect to mastery in a certain domain?" or what probable range of scores would this student get on other tests like this, one from the same content domain? Such a view of reliability goes beyond comparing one student with another. Its use will become more frequent with the advent of standards and tests designed to measure student achievement relative to these standards (Crocker & James, 1986; Nitko, 1983; Tranb, 1994).

Variables that affect Reliability Estimates

Variables associated with the test that can be manipulated or controlled to enhance score reliability with the examinees or testing conditions will be considered under the following as posited by (David, 1988).

Test length: Scores from a longer test are apt to be more reliable than the scores from a shorter one. This is true because the longer test is likely to yield a greater spread of scores "(David 1988).

Test Content: Tests that measure the achievement of a somewhat homogeneous set of topics are likely yield more reliable scores than test that measure a potpourri of somewhat unrelated ideas. Each of the methods of internal analysis described above for estimating reliability is an index of item homogeneity an indication of the extent to which all the items in the test measure a single domain of content (David, 1988; Abe 2005).

Item difficulty: All the items in a test need to be in the moderate range of difficulty neither too hard nor too easy for the group, to help identify differences in achievement among students (David 1988; Wilson, Downing & Ebel; 1977).

Item discrimination: Items that discriminate properties are answered correctly by most of the students. Who earn high scores on the test and are missed by most of those who earn low test scores. Items that discriminate properly help to accumulate high scores for those who have learned and keep low achievers from obtaining high scores on the test. Highly discriminating items help to distinguish between examinees of different achievement levels and consequently, they contribute substantially to the score reliability. That is to say, the test with the highest average item discrimination index likely to yield scores of highest reliability (David, 1988).

Group heterogeneity: The reliability estimate will be high for a group that is heterogeneous with respect to achievement of the test content than it will be if the group is homogeneous when a group is very homogeneous it is more difficult to achieve a spread of scores and to detect the small differences that actually exist. When inter individual differences are greater, as in a more heterogeneous group, the rank ordering of individual is likely to be replicated more easily on a retest.

Student Motivation: if students are motivated to do their best on a test, their scores are not apt to represent their actual achievement levels very well. But when the consequences of scoring high or low are important to examinees, the scores are likely to be more accurate. Indifferences, lack of motivation or under enthusiasm for whatever or reasons can depress test scores just as much as anxiety or over enthusiasm may.

Student testwiseness: When the amount of test-taking experience and levels of testwiseness vary considerably within a group, such background and skills may cause scores to be less reliable than they otherwise would be when all examinees in the group are experienced and sophisticated test takers or when all are relatively naive about test taking, such homogeneity probably will not lead to much random measurement error. The rank order of scores is likely to be influenced only when there is obvious variability in testwiseness within the group. Students who answer an item correctly because of their testwiseness rather than their achievement of content, cause the item to discriminate improperly that is poor item discrimination contributes to lowered reliability estimates.

Time Limits: It is customary for classroom achievement tests to be administered with generous time limits so that nearly all, if not all students can finish. However, when time becomes a factor when test can be regarded as speeded, the result is a reliability coefficient that somewhat misrepresents score accuracy. The reliability estimate obtained under speeded conditions by the method of internal analysis is artificially high, an artifact of the method itself.

Security Precautions: Occurrences of cheating by student during a test contribute random errors to the test scores (Bandele & Abe, 2017). Some students are able to provide correct answer for questions to which they actually do not know the answers, coping of answers, use of cribs or cheat sheets, and the passing of information give unfair advantage to some and cause their scores to be higher than they would be an retesting. The passing of information from class to class when the same test be given to different classes at different times also reduces overall score reliability (Abe, 2005; Cracker & Algein, 1986; Ebel & Frishine, 1989).

CONCLUSION

This paper explains and demonstrates the procedures that are commonly used to determine the reliability coefficient of a test in such a way that a person who has modest mathematical or statistical skills can carry out the same analysis on a classroom test or examination in order to ascertain the quality of the items given in a particular test. Therefore, the issues arise from these approaches or methods to assessing test quality were presented while the quick way to estimate

reliability for classroom examination was addressed with illustration examples all through the methods of reliability estimates.

Recommendations

Every professional teacher who has gone through NCE, Degree in Education or Postgraduate Diploma in Education (PGDE or PDE) should atoned or abreast to the procedural ways of estimating reliability coefficient of the test conducted in the classroom, this will serve as a quality indicator in the teachers' made test.

Specialists in the field of Tests, Measurement and Evaluation should be allowed to handle the course at NCE, Degree or Post graduate levels of our educational strata this will minimize the tendency of mystifying the course for the students.

References

Abe, T.O & Alonge, M.F (2010) Statistical Moderation of Continuous Assessment and Terminal Examination Scores of Junior Secondary Schools in Ekiti State, Nigeria. *International Journal of Research in Counseling and Sports Sciences IJORECS,* 1(1), 6 -18.

Abe, T.O (2006) Statistical Moderation of Internal Assessment Scores in Senior Secondary School in Ekiti State, Nigeria. Unpublished Ph.D. Thesis, University of Ado-Ekiti, Ekiti State Nigeria.

Abe, T.O & Gbore L.O (2006) Kurtosis as a Means of Comparing Internal Assessment Scores in School Subjects. *Journal of Educational Thoughts,* 5(1), 74-82.

Abe, T.O (2005) Reliability Estimate of Instrument's Internal Consistency, *Ikere Journal of the Science Teacher,* 2(2), 8 – 15.

Abe, T.O (1995) Scaling Moderation of Internal Assessment Scores in Mathematics in Selected Secondary Schools in Owo Local Government Area of Ondo State, Unpublished M.Ed Thesis, Ondo State University, now Ekiti State University, Ado ekiti.

Agwuibike, E.O & Momoh, F.O (1995). *An Introduction to Tests, Measurement and Evaluation in Education and Psychology,* Benin, Jodah Publication 1st Ed, pp 54 – 68.

Alonge, M.F (1989). *Measurement and Evaluation in Education and Psychology, 1st edition* Ado-Ekiti, Adabayo Publishing and Printing Press (Nigeria) Ltd.

Alonge, M.F (2004). *Measurement and Evaluation in Education and Psychology,* 4th edition, Ado Ekiti Adabayo Publishing and Printing Press (Nigeria) Ltd.

Bandele, S.O & Abe, T.O (2017) Ensuring Fairness and Reliability Of School – Based Assessment Scores through Moderation. In Tests, Measurement, Assessment, Evaluation and Research in Education. *A book of Reading in honour of Professor M.F Alonge,* Published by Institute of Education, Faculty of Education, Ekiti State University, Ado-Ekiti, pp 20-46.

Bartko, J.S & Carpenter W.T (1976) On the Methods and Theory of Reliability. *J.New, Ment. Dis.* 163, 307-317.

Brennan R.L & Prediger D.J (1981) Coefficient λ: Some Uses, Misuses and alternative *Educational and Psychological Measurement* 41:687-699. doi:10.1177100131644810410030 7.

Brown, F.G (1970) Principles of Education and Psychological Testing, *the Dryden Press Inc.* 97-156.

Cecil, R.R, Ronald, B.L & Victor W. (2011) *Measurement and Assessment in Education 2nd Ed*. PHI learning Pivoted, Limited, New Delhi-110001, pp 90 -122.

Cohen, J.J (1960). A Co-efficient of agreement for nominal scales. *Educational, Psychology and Measurement* 20:37-46.

Croker Linda & James Algina (1986) *Introduction to classical and Modern Test Theory,* New York: Holt Rinehart & Winston.

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

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

Davidshofer, C.D, & Murphey K.R (2005) *Psychological Testing: Principles and Applications* (6th), Upper Saddle River N.J: Pearson/ Prentice Hall.

Flanagan, J.C (1937): A proposal Procedure for increasing the efficiency of Objective tests. *Journal of Psychology*, 8(2), 28 – 81.

Gronlund, N. E (1976). *Measurement and Evaluation in Teaching* (3rd Ed): New York: MacMillan publishing Co_Inc.

Keith, T.Z & Reynolds, C.R (1990) Measurement and design issues in Child assessment Research. In C.R. Reynolds & R.W kampheus (Eds). *Handbook of Psychological and educational assessment of Childern: Intelligence and achievement.* (pp. 29 – 62). New York: Guilford Press.

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

Nitko A.J (1983) *Educational Tests and Measurement: An Introduction New and measurement.* An Jovanovich.

Ojerinde, Dibu (1986) *Tests and Measurement in Education.* University press Limited Ibandan..

Onocha, C.D & Okpala, P.N (1995) *Tools for Educational Research.* Stirling – Horden Publisher (Nigeria) Ltd, Jatta-Uzairue.

Pontius R. & Miltones M. (2011). Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment *International Journal of Romote Sensing* 32, 4407 – 4429.

Rulon, P.J (1939) A Simplified Procedure for determining the reliability of a test of split-halves. *Harvard Educational review,* 9, 99 – 103.

Sim, J & Wright, C.C (2005). The Kappa Statistic in Reliability Studies: Use interpretation and Sample Size Requirement. *Physical Therapy*, 85, 257-268.

Spiegel, M.R (1970) *Schaum's Outline Series. Theory and Problems of Statistics* MC Graw Hill Book Company, New York.

Traub, R.F (1994). *Reliability for the Social Sciences. Theory and Application: Measurement Methods in the Social Socials.* Thousand Daks, Calf:sage Publications.

Trochim, W.M (2006). *The Research Method Knowlegde Base*, 2nd Edition. Internet www.page, at URL:http://www.socialresearchmethods.net.

William, M.K (2006) *Types of Reliability: The Research Methods knowledge* Base: Last Revised; 20 October 2006.

Ugwuja, J.O (2004) *Basic Inferential Statistics for Higher Education* High Class Quality Integrated Press Nsukka.

Uzoagulu A.E (1998) *Practical Guide to writing Research Project Reports in Tertiary Institutions* Enugu. John Jacob's Classic Publishers.