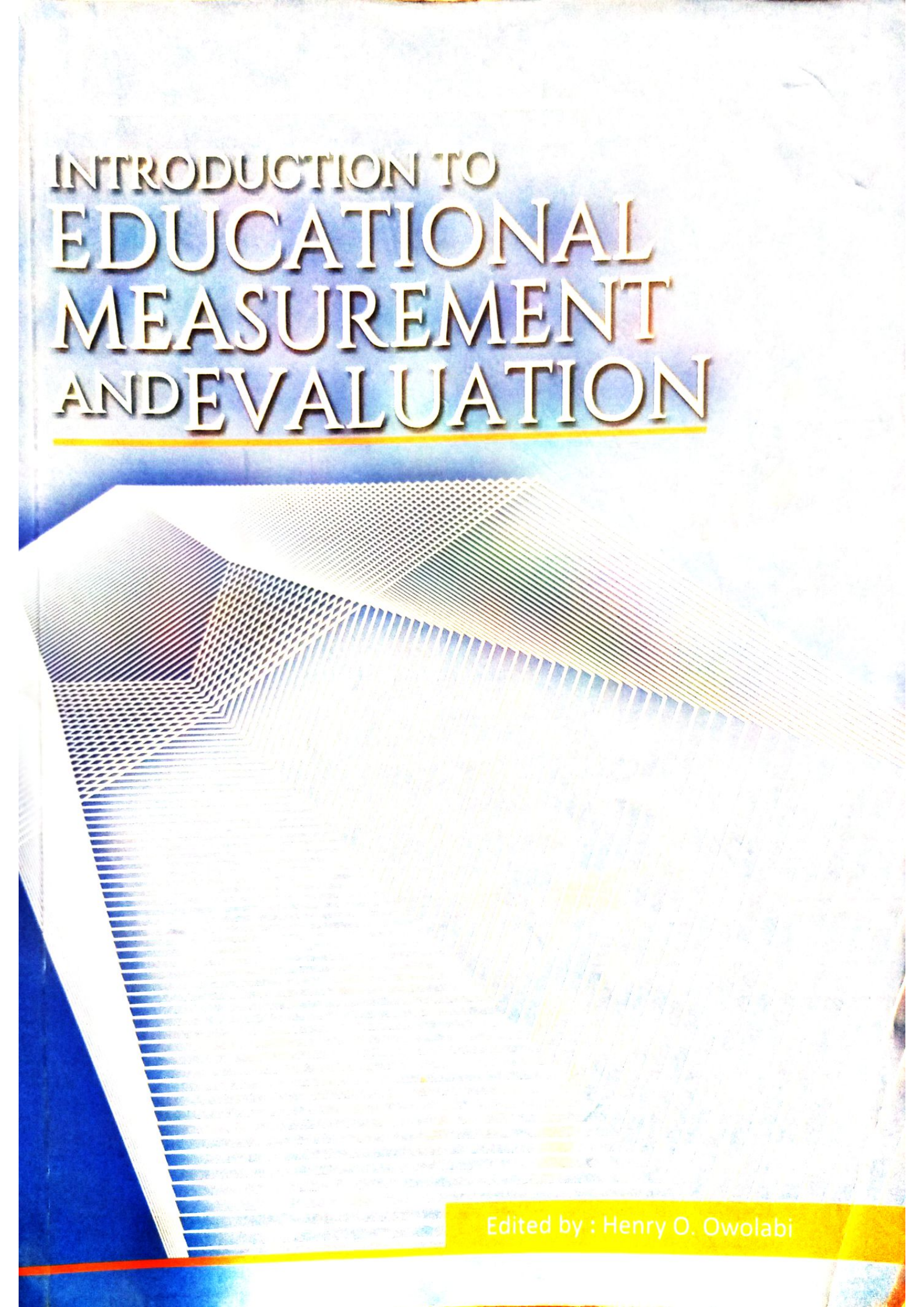


INTRODUCTION TO EDUCATIONAL MEASUREMENT AND EVALUATION



Edited by : Henry O. Owolabi

Copyright ©

Department of Social Sciences Education,
University of Ilorin.

January, 2019

All right reserved

No part of this publication may be reproduced or transmitted in any form or by means of electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without permission in writing from the publisher

ISBN: 978-978-972-067-5

Printed in Nigeria by
Wils & Bankleys Publisher (Nig) Ltd
15, Police Road, Ilorin

TABLE OF CONTENTS

Basic Concepts in Educational Evaluation

Owolabi, Henry O. 1

Purpose of Evaluation in Education

Daramola, Dorcas S. 8

The Questionnaire

Bamidele, Esther. O. 11

Observation Technique

Effrim, Paul. K. 22

Checklist

Olutola, Adekunle. T. 28

Interview

Mahmud, Jamiu. A. 34

Rating Scale

Adewuni, David. A. 38

Projective Techniques

Abdusalam-Nuhu, Rasheedat. 41

Inventories

Sheu Adaramaja .Lukman. 46

Test Development Techniques

Ogunjimi, Mayowa. O. 49

Concept of Test Validity

Ogunjimi, Mayowa. O. 70

Test Reliability

Bamidele, Esther. O. & Oladele, Jumoke. I. 75

Test Administration and Interpretation

Cobbinah, Andrews. 92

Basic Statistical Concepts

Abdusalam-Nuhu, Rasheedat. 100

Descriptive Statistics in Education

Jimoh, Mohammed. I. & Akinboboye, Joseph. T. 104

Measures of Variability/dispersion

Daramola, Dorcas. S. 120

Correlation Statistics

Jimoh, Mohammed. I. 133

TEST RELIABILITY

BAMIDELE, E. O. & OLADELE, J. I.

Introduction

One of the challenges in educational measurement is the weakness of an instrument to obtain the similar scores from two or more administrations. This is the problem of consistency of scores. Educational measurement must yield accurate and consistent results in order to serve the purpose for which it is constructed. A test has little value if the score it yields today is quite different from the score it would yield another time when the test is administered under similar condition. Measuring instruments will yield similar results when the systematic error of measurement such instruments are minimal. Hence, a test is reliable when the test measures consistently whatever it does measure.

Concept of Reliability

Reliability is the consistency with which a measuring instrument measures what it is designed to measure. It means the extent to which the results obtained from such instrument can be relied upon as the true score. It is the degree to which a particular test yields similar scores over two or more administrations or the degree to which two similar tests yield about the same scores. Hopkins (1998) defined reliability as a synonym for consistency and replicability over time, instrument and person(s). As such, reliability is the extent to which two tests measure identical concepts at an identical level of difficulty. The meaning given to the concept of reliability is similar across authors (Thorndike & Hagen, 1977; Guilford & Fruchter, 1986; Nwana, 1982; Hassan, 1986; & Okpala, Onocha & Oyediji, 1993). These definitions points to reliability as the extent to which an assessment result is dependable, stable and consistent. For instance, how far can the same or similar results are obtained if a group of persons' heights are measured in one day and repeated the following week(s)? Reliability of measures is in degrees, i.e. it can either be low, moderate or high. It is obtained through correlation statistics which yields a coefficient of reliability. This coefficient varies from zero which indicate very low or no reliability to unity showing absolute reliability. For a test, a reliability of not less than 0.60 may be regarded as satisfactory (Abiri, 2007).

Perfect reliability in behavioural sciences may not be possible to attain because most of the measurements are not as precise as those of the physical sciences such as mass, weight, height, etc. Measurement of human traits in psychology and education may changes over time since human beings are dynamic. It is possible for a teacher to give a test of mathematics achievement to a group of students and after two weeks repeat the same test but obtain different scores for each student from the score obtained in the first administration. This is possible because changes in the health, motivation, experience, disposition and mood of testees as well as fatigue may bring about changes in their scores, however reliability must be estimated. Theoretically, scores in measurement referred to as observed score are made up of the actual or "true score" and the "error score". Classical test theory assumes that the variance of observed scores (S^2_{obs}) for a group of examinees is equal to the variance of their true score (S^2_{true}) plus the variance of the systematic error of measurement (S^2_{error}).

$$S^2_{obs} = S^2_{true} + S^2_{error}$$

Mathematically, reliability is the ratio of the true variance to observed variance, or the

proportion of the observed variance accounted for by true variance. Thus, the error component is involved due to measurement error.

Approaches for Estimating Test Reliability

There are different approaches to obtaining test reliability. They are:

1. Test-retest Reliability
2. Equivalent Forms Reliability
3. Internal Consistency Reliability
4. Scorer Agreement Reliability

These approaches are aimed at ascertaining the measures of stability over time, measures of equivalence, measure of construct or internal consistency, and measure of rater consistency respectively. These methods can further be depicted in Figure 1.

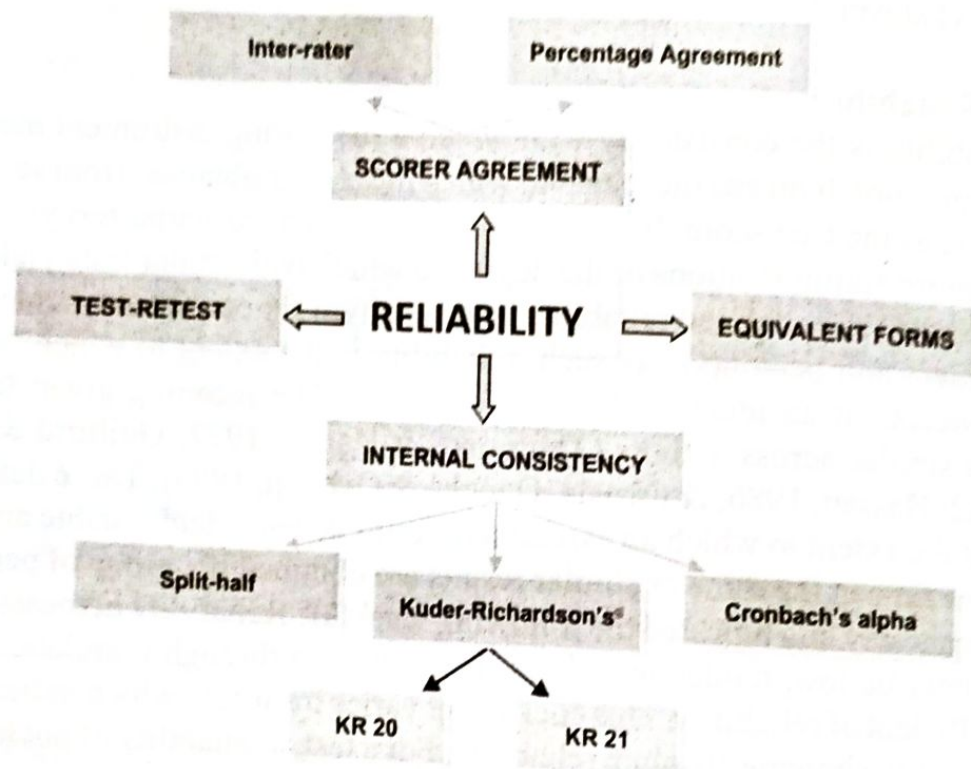


Figure 1: Reliability Conceptual Frame

While the formula behind these concepts would be discussed respectively, in reality students need not always go through these manual calculations due to the availability of computer software available for statistical computations such as IBM Statistical Package for Social Sciences (SPSS). As such, how to obtain reliability statistics with SPSS would also be explained below where applicable with the goal of exposing users of this book to current trends in reliability statistical computation.

TEST RE-TEST RELIABILITY: this is a measure of stability and the evidence is obtained over a period of time. Test re-test reliability is about the most often used method of reliability. It is a reliability procedure which reliability is computed after an instrument has been administered on a group of respondents twice, within a given time lag mostly two to three weeks. This method aims at finding out whether the same set of testees took a test on two different times, they would

record approximately the same scores with the assumption that there is no additional learning between test administrations. Therefore, the two sets of scores so obtained from repeated conduct of the test will thereafter be correlated using the Pearson's correlation. In reporting test re-test reliability, the time interval should be clearly stated. Test re-test reliability emphasizes stability over time. The extent to which the reliability co-efficient is high may point to the reliability of the instrument, but may also be attributable to other extraneous factors like change in condition of administration and maturation as they have seen the items during the first administration etc. With this type of reliability estimate is possible to determine how confidently to generalize from the score a person earns at one time to what he would earn if the test had been given at a different time.

The formula for obtaining coefficients of stability is given below:

Coefficients of Stability

$$R_{A_1A_2} = \frac{\sum A_1A_2}{N^1} - \frac{\sum A_1}{N^1} \cdot \frac{\sum A_2}{N^2}$$

Where: A_1A_2 = Test 1 and 2

N^1 = Number of items in test 1

N^2 = Number of items in test 2

EQUIVALENT FORMS RELIABILITY: this is also referred to as alternative or parallel forms reliability which is a measure of equivalence of similar tests in difficulty, scope, etc. Two sets of similar tests administered at intervals of time as against administering one single test twice are called parallel tests. As such, they are assumed to satisfy both item-by-item equivalence and stability over forms. The concern here is that if a different but similar test was used, would the testees record the same or nearly the same scores individually? The correlation co-efficient obtained by computing the relationship between the scores on the two parallel tests gives the reliability coefficient. The decision on time interval of administration could be shorter than the case of test re-tests to alleviate the problem of test wiseness. Thus, these two test qualities (i.e item difficulty and discrimination levels) are expected to be present in similar or close proportion in the two tests that are to be used as parallel forms of the same test to ensure equivalence. The procedure to be employed in determining the reliability of a test using the parallel forms method is correlation using the Pearson's correlation similar to that of test re -test method.

In constructing equivalent tests, care must be taken to ensure that the two measures are equivalent in statistical sense with equal means, variances and item inter-correlations. Hence, equality of contents should be put into consideration, the table of specification of the two tests must follow the same pattern or form, items should show the same difficulty level and discriminating power and format and administrative instruction must be the same for both tests. Furthermore, if one wishes to use the test results for long range prediction, then the use of a coefficient of stability is appropriate, else it could lead to failure in long term prediction. This would make the reliability estimate to reflect any trait change as error so that the confidence in any prediction would be appropriately tempered by a lower reliability coefficient. If the purpose of giving a test is not for long range prediction but rather, for the purpose of making interference about the knowledge an individual has in a subject area, then the interest would primarily be in a coefficient of equivalence.

At times researchers are concerned with both long-range production inferences to a

domain of knowledge. This is actually more likely to occur than content for just stability. For example, the measurements of constructs (unobservable phenomena) such as intelligence, creativity, aggressiveness or musical interest are probably not depending on a specific set of questions. In that case a coefficient of equivalence and stability could be obtained by giving one form of the test and, after a period of time, administering other form and correlating the results. This procedure allows for both changes in scores due to trait instability and item specificity to be considered. This estimate of reliability is thus usually lower than other procedures.

The formulas for obtaining coefficients of equivalence are given below:
Coefficients of Equivalent

$$R_{AB} = \frac{N \sum AB - \sum A \sum B}{\sqrt{[N \sum A^2 - (\sum A)^2][N \sum B^2 - (\sum B)^2]}}$$

Where: AB = Test 1 and 2
N = Number of items in test

Hypothetical Example 1:

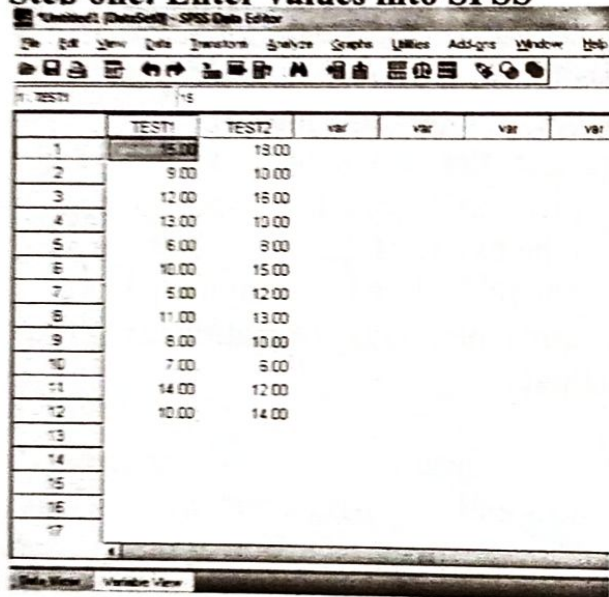
The following scores were obtained by administering two sets of equivalent tests administered at an interval of two weeks as shown on table 1:

Table 1: Student's scores obtained from administering equivalent tests

Test 1 (Form 1)	15	9	12	13	6	10	5	11	8	7	14	10
Test 2 (Form 2)	18	10	16	10	8	15	12	13	10	6	12	14

The following steps can be followed to obtain r for determining coefficient of equivalence.

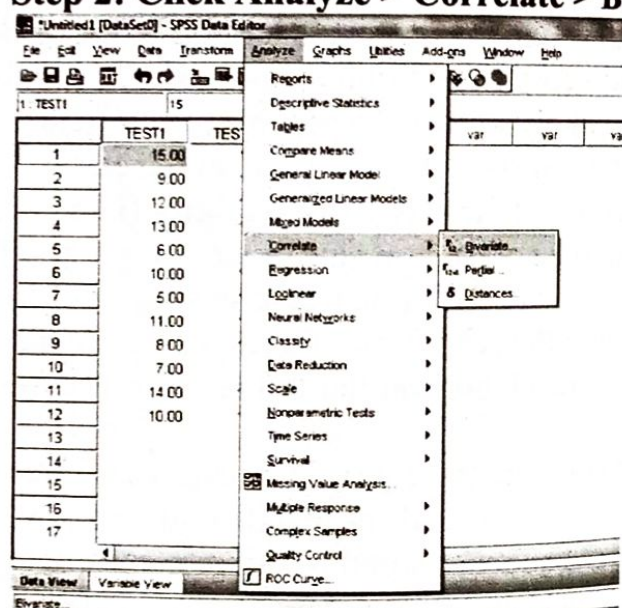
Step one: Enter values into SPSS



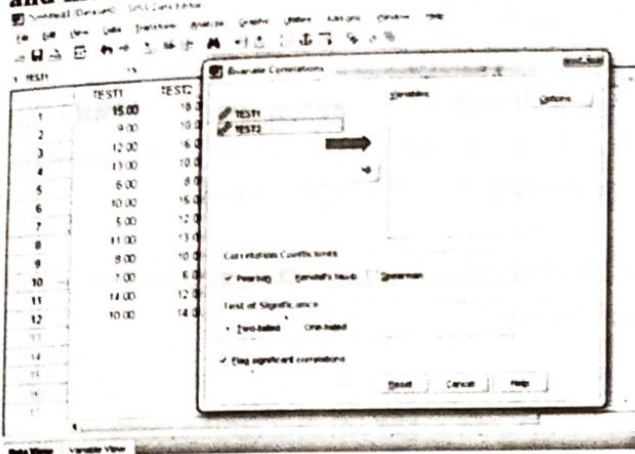
The screenshot shows the SPSS Data Editor with two columns, TEST1 and TEST2, and 17 rows of data. The data is as follows:

	TEST1	TEST2
1	15.00	18.00
2	9.00	10.00
3	12.00	16.00
4	13.00	10.00
5	6.00	8.00
6	10.00	15.00
7	5.00	12.00
8	11.00	13.00
9	8.00	10.00
10	7.00	6.00
11	14.00	12.00
12	10.00	14.00
13		
14		
15		
16		
17		

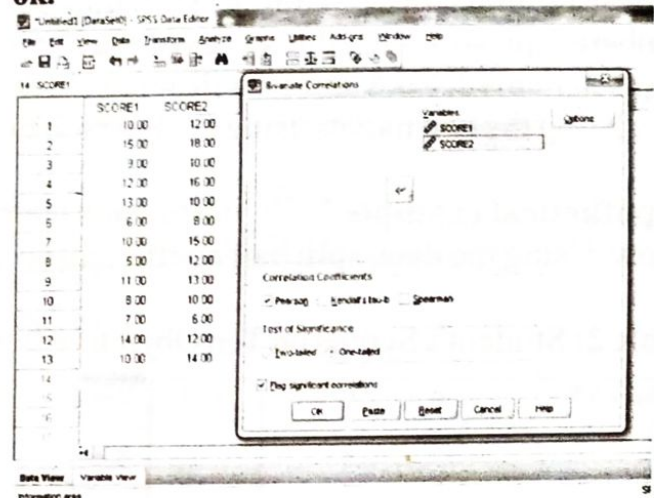
Step 2: Click Analyze > Correlate > Bivariate.



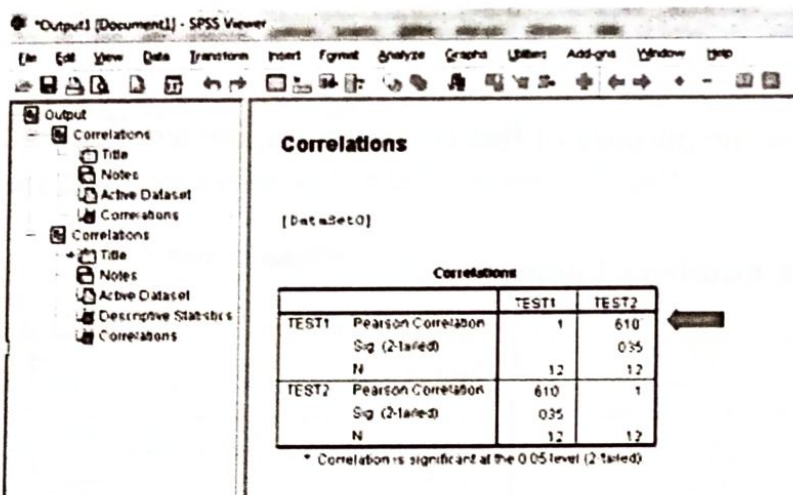
Step 3: Select the variables Test 1 and Test 2 and move them to the Variables box.



Step 4: In the "Correlation Coefficient" Area, Click Pearson and in the "Test of Significance" area, select two-tailed and click ok.



Step 5: View result in IBM SPSS output window show



Step 6: Report value for Pearson's r. As shown in the above display, the value for r is 0.61

Step 7: Interpret result: an r of +0.61 means that both tests obtained by administering two set of similar tests at the same time is reliable.

Worthy of note is that the value obtained for r from the manual or computer programme computational method is usually within the same range if not the same.

INTERNAL CONSISTENCY RELIABILITY: this is a measure of Construct Consistency also referred to as inter-stem consistency. Internal consistency is typically a measure based on the correlations between different items on the same test (or the same subscale on a larger test) which measures whether several items used to measure the same general construct produce similar scores. Methods for determining the internal consistency of test includes the use of Split-half, Kuder-Richardson's KR20 and KR21, and Coefficient of alpha (α) procedures. Generally, all the internal consistency methods are inappropriate for speeded tests.

Method 1: Split-half Reliability

Split-half is one of the techniques of calculating the internal consistency of an instrument. is a process of splitting a test into two halves such that a test of ten items has the scores of items 1,2,3,4 and 5 as the first half and the scores of items 6,7,8,9 and 10 as the second half of the test or the scores of the even numbered questions correlated with their scores in odd numbered questions e.g the scores of the items 1,3,5,7, and 9 will serve as the first test while the scores in items 2,4,6,8 and 10 will serve as second test. The reliability coefficient is obtained by correlating the two halves using the Pearson product moment correlation procedure.

Hypothetical example 2: Eight students were given 10 objective items. Their scores are given below. Using the data, split half coefficient is computed below:

Table 2: Student's Scores on Ten Objective Items

Cand S/N	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
1	x	✓	x	x	x	✓	✓	x	✓	x
2	✓	x	x	x	✓	x	✓	x	✓	✓
3	x	x	✓	✓	x	✓	x	✓	x	x
4	✓	✓	x	✓	x	✓	x	✓	x	x
5	x	✓	x	✓	x	✓	✓	✓	x	✓
6	✓	x	✓	✓	x	x	✓	✓	x	✓
7	x	✓	✓	x	✓	x	x	✓	x	✓
8	✓	x	✓	x	x	x	✓	x	✓	x

SOLUTION

Step 1: Split the text into two halves. For the purpose of this computation, the test was split on the basis of even and odd numbers

Table 3: First test made up of all even numbered items

Cand. S/N	Item 2	Item 4	Item 6	Item 8	Item 10	Total Over 5
1	✓	x	✓	x	x	2
2	x	x	x	x	✓	1
3	x	✓	✓	x	✓	3
4	✓	✓	✓	✓	x	4
5	✓	✓	✓	✓	x	4
6	x	✓	x	✓	✓	3
7	✓	x	x	✓	✓	3
8	x	x	x	x	x	0

Table 4: Second test made up of all odd numbered items

Cand S N	Item 1	Item 3	Item 5	Item 7	Item 9	Total Over 5
1	x	x	x	✓	✓	2
2	✓	x	✓	✓	✓	4
3	x	✓	x	x	✓	2
4	✓	x	x	x	x	1
5	x	x	x	x	x	0
6	✓	✓	x	✓	x	3
7	x	✓	✓	x	x	2
8	✓	✓	x	✓	✓	4

The formula for obtaining $r^{1/2}_{1/2}$ (reliability of half of the test) is:

$$r^{1/2}_{1/2} = \frac{N \sum OE - \sum E \sum O}{\sqrt{[N \sum O^2 - (\sum O)^2][N \sum E^2 - (\sum E)^2]}}$$

The variables needed as displayed in the formula are E, O, OE, and O^2 where:

E = total number of items answered correctly on one half of the test even numbers and

O = total number of items answered correctly on other half of the test odd numbers by each testee

This can easily be accomplished using a computational table as shown below:

Step 2: Computation value for $r_{11/12}$

Table 5: Computation Table for $r_{11/12}$

S/N	E	O	OE	E ²	O ²
1	2	2	4	4	4
2	1	4	4	1	16
3	3	2	6	9	4
4	4	1	4	16	1
5	4	0	0	16	0
6	3	3	9	9	9
7	3	2	6	9	4
8	0	4	0	0	16
	$\sum E = 20$	$\sum O = 18$	$\sum OE = 33$	$\sum E^2 = 64$	$\sum O^2 = 54$

Step 3: Having obtained the values of E, O, OE, and O^2 , the next step is to substitute the values into the formula. Therefore,

$$\begin{aligned}
 r^{1/2}_{1/2} &= \frac{8(33) - (20)(18)}{\sqrt{[8(54) - (18)^2][8(64) - (20)^2]}} \\
 &= \frac{264 - 360}{\sqrt{(432 - 324)(512 - 400)}} \\
 &= \frac{-96}{\sqrt{(108)(112)}} = \frac{-96}{109.98} = -0.873
 \end{aligned}$$

At this junction, readers should note that the reliability coefficient so computed is in respect of one half of the test. The formula for correcting this coefficient to full length of test as given by Spearman-Brown's formula as presented below

$$r_{xx} = \frac{2r_{1/2, 1/2}}{1+r_{1/2, 1/2}}$$

where r_{xx} = estimated reliability of the whole test

$r_{1/2, 1/2}$ = reliability of the half-test

Next is to substitute the value for the reliability of half of the test as shown below:

$$r_{xx} = \frac{2r_{1/2, 1/2}}{1+r_{1/2, 1/2}} = \frac{2(0.873)}{1+0.873} = \frac{1.746}{1.873} = 0.932$$

Interpretation: There is a high relationship within the test. This indicates that the items are internally consistent

Method 2: Kuder-Richardson Reliability

Kuder-Richardson KR 20 and KR 21 are two reliability estimates of an instrument's internal consistency such as test. KR 20 is meant to compute internal consistency of dichotomously scored items with a range of difficulty such as multiple-choice, short answer and fill in the blank objective test while KR 21 is a modified version of KR 20 for Essay test.

$$KR_{21} = \frac{K}{K-1} \left[1 - \frac{M(K-M)}{KS^2} \right], \quad KR_{20} = \frac{K}{K-1} \left[1 - \frac{\sum p q}{S^2} \right]$$

Where: K is the number of items in the test.

M = the mean (arithmetic average) of the test scores.

p = proportion of those who get the item correctly.

q = proportion of those who get the item incorrectly ($q = 1 - p$).

p, q = variance of a single item scored dichotomously.

S^2_x = variance of the total test.

Hypothetical Example 3: the table below resulted from an item by item analysis of the performance of twelve students on a ten-item standardized objective achievement test.

Table 6: Students Scores on a Ten-item Standardized Objective Achievement Test

Students	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
5	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
6	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
7	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
8	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
9	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
10	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
11	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
12	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Questions: Estimate the Kuder-Richardson reliability of the test and interpret result?

Note: KR20 formula will be used being a dichotomously scored objective test.

SOLUTION USING THE FORMULA:

STEP 1: Arrange the scores of the twelve students in descending order of magnitude and provide a frequency table for the scores.

Table 7: Student's scores in descending order of magnitude

S/N	Score	Position	S/N on Merit	Score
1	9	3 rd	12	10
2	2	10 th	5	10
3	4	8 th	1	9
4	3	9 th	6	8
5	10	1.5 th	7	7
6	8	4 th	8	6
7	7	5 th	9	5
8	6	6 th	3	4
9	5	7 th	4	3
10	0	12 th	2	2
11	1	11 th	11	1
12	10	15 th	10	0

STEP 2: Find the mean, variance and standard deviation.

$$\text{Mean} = \bar{X} = \frac{10+10+9+8+7+6+5+4+3+2+1+0}{12} = \frac{65}{12} = 5.42$$

$$\text{Variance} = \sigma^2 = \frac{\sum X^2}{N} - \left[\frac{\sum x}{N} \right]^2 = \frac{485}{12} - \left[\frac{65}{12} \right]^2 = 40.417 - 29.34 = 11.077$$

$$\text{Standard deviation} = \sigma = \sqrt{11.077} = 3.33$$

STEP 3: Substitute the values into the formula to compute KR 20

Table 8: Students Scores

Items	$p/_{12}$	$q/_{12}$	$pXq/_{144}$
1	8	4	32
2	9	3	27
3	8	4	32
4	6	6	36
5	5	7	35
6	5	7	35
7	6	6	36
8	5	7	35
9	7	5	35
10	6	6	36
	65		339/144 2.354
	$\sum x$		$\sum p q$

$$KR20 = \frac{K}{K-1} \left[1 - \frac{\sum Pq}{\sigma^2} \right] = \frac{10}{9} \left[1 - \frac{\left[\frac{339}{11.08} \right]}{11.08} \right]$$

$$KR20 = \frac{10}{9} \left[1 - \frac{2.354}{11.08} \right] = \frac{10}{9} [1 - 0.212]$$

$$= (0.788) = 0.876$$

The test from the obtained result is interpreted as highly reliable.

On SPSS procedure for KR20, note that KR20 is a special case of Cronbach's Alpha in which the items are binary variables (usually scored as 0 or 1). The RELIABILITY procedure can calculate Alpha and therefore can calculate KR20 for binary item variables as would be explained in the subsequent section. For KR21 which is based on the (usually unrealistic) assumption that all items in the scale have equal difficulty levels (i.e. equal proportions of 'successes' on the items), there is no direct procedure in SPSS. However, KR21 could be calculated in SPSS by using the AGGREGATE procedure to save KR21 formula components to an active file, where they could be input into a COMPUTE command to calculate KR21. This is clearly beyond the level of the readers of this text. Though both techniques have their uniqueness, KR20 is generally considered to be a better reliability estimate than KR21.

Method 3: Cronbach's or Coefficient Alpha (α)

This formula was developed by Cronbach in 1951 and is suitable for items not scored dichotomously like essay tests and in Likert response styled instrument such as: Strongly Agree (SA), Agree (A), Disagree (D) and Strongly Disagree (SD). As such, Cronbach's alpha is a test used to determine if multiple-question with Likert scale format surveys are reliable. These questions measure latent variables which are hidden or unobservable variables such as: a person's conscientiousness, neurosis or openness. These are very difficult to measure in real life. Cronbach's alpha will tell you if the test you have designed is accurately measuring the variable of interest usually referred to as educational constructs. The formula for calculating Cronbach's Alpha (α)

$$\text{Cronbach Alpha } (\alpha) = \frac{k}{k-1} \left[1 - \frac{\sum \sigma_i^2}{\sigma^2} \right]$$

Where: k is number of items in the instrument.

$\sum \sigma_i^2$ = variance of each of the test item.

Hypothetical Example 4: a researcher devised a 5-point Likert nine-item questionnaire to measure how safe people feel at work at an industrial complex. In order to understand whether the questions in this questionnaire all reliably measure the same latent variable (feeling of safety); the following is the manual calculation of Cronbach's alpha.

Table 9: Respondent questionnaire Results

S/N	Items 1	Items 2	Items 3	Items 4	Items 5	Items 6	Items 7	Items 8	Items 9
1	2	5	1	2	5	1	2	3	3
2	3	3	3	3	3	3	3	3	3
3	5	4	5	5	4	5	5	3	2
4	1	4	1	1	4	1	1	3	1
5	3	3	5	3	3	5	3	1	4
6	5	5	3	5	5	3	5	2	1
7	1	1	4	1	1	4	1	4	5
8	5	5	2	5	5	2	5	4	3
9	3	2	3	3	2	3	3	4	2
10	4	3	5	4	3	5	4	4	1
11	4	5	1	4	5	1	4	5	4
12	3	1	2	3	1	2	3	3	1
13	5	3	3	5	3	3	5	2	3
14	1	5	5	1	5	5	1	1	1
15	5	1	1	5	1	1	5	1	3
Σ	50	50	44	50	50	44	50	43	35
Mean	3.33	3.33	2.93	3.33	3.33	2.93	3.33	2.87	2.33
σ_i^2	1.85	2.1	2.33	1.85	2.1	2.33	1.85	1.4	1.67

SOLUTION

Recall that the formula for calculating Cronbach's alpha is: $(\alpha) = \alpha = \frac{k}{k-1} \left[1 - \frac{\sum \sigma_i^2}{\sigma_x^2} \right]$

Where: k = number of items in the instrument

σ_i^2 = variance of each of the item across respondents

σ_x^2 = variance associated with the observed total scores

STEP 1: Calculate the variance for each of the item across respondents as shown below:

$$\sigma^2 \text{ of } (2,3,5,1,3,5,1,5,3,4,4,3,5,1,5) = 1.85$$

$$\sigma^2 \text{ of } (1,3,5,1,5,3,4,2,3,5,1,2,3,5,1) = 2.33$$

$$\sigma^2 \text{ of } (5,3,4,4,3,5,1,5,2,3,5,1,3,5,1) = 2.1$$

$$\sigma^2 \text{ of } (2,3,5,1,3,5,1,5,3,4,4,3,5,1,5) = 1.85$$

$$\sigma^2 \text{ of } (3,2,1,4,1,5,3,2,1,4,1,3,1,3,1) = 1.67$$

$$\sigma^2 \text{ of } (5,3,4,4,3,5,1,5,2,3,5,1,3,5,1) = 2.1$$

$$\sigma^2 \text{ of } (2,3,5,1,3,5,1,5,3,4,4,3,5,1,5) = 1.85$$

$$\sigma^2 \text{ of } (1,3,5,1,5,3,4,2,3,5,1,2,3,5,1) = 2.33$$

$$\sigma^2 \text{ of } (3,3,3,1,2,4,4,4,4,5,3,2,1,1,3) = 1.4$$

STEP 2: Obtain the summation of the variance for all the items

$$\Sigma \sigma_i^2 = 1.85 + 2.1 + 2.33 + 1.85 + 2.1 + 2.33 + 1.85 + 1.4 + 1.67 = 17.48$$

STEP 3: Calculate the variance of the observed scores (x) total across respondents.

Therefore, the variance of $(24,26,37,18,28,40,20,35,24,37,28,20,29,27,23) = 43.44$

$$\sigma_x^2 = 41.95$$

STEP 4: Substitute the values into the formula

$$\alpha = \frac{k}{k-1} \left[1 - \frac{\sum r^2}{n} \right]$$

$$\alpha = \frac{9}{9-1} \left[1 - \frac{17.48}{43.44} \right]$$

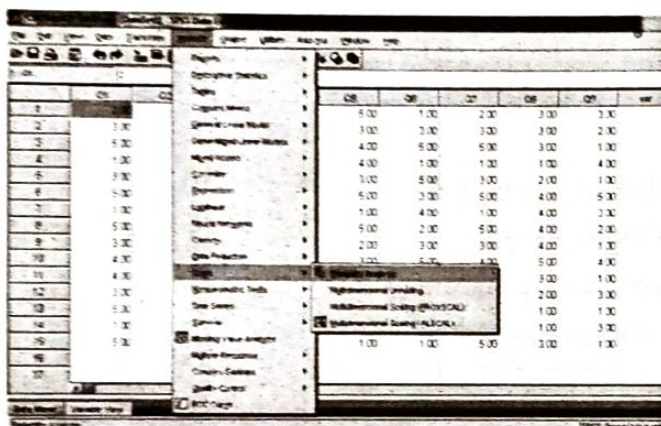
$$\alpha = - \left[1 - 0.41 \right]$$

$$\alpha = 1.125 \left[1 - 0.4 \right] = 1.125 \times 0.6 = 0.68. \text{ Therefore; Cronbach's alpha } (\alpha) = 0.68$$

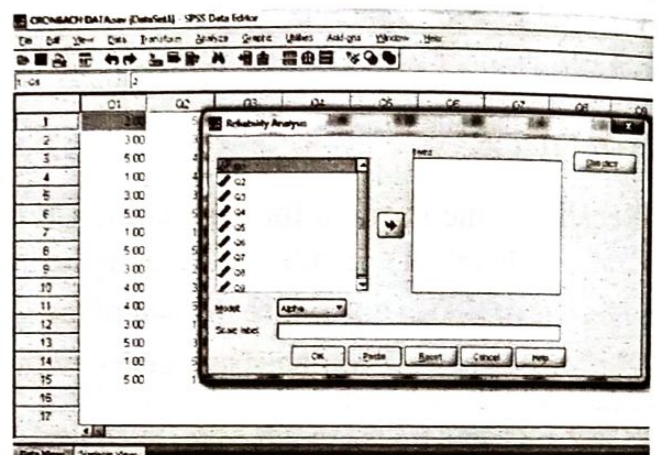
Cronbach's alpha Test Procedure in SPSS Statistics

Using the same hypothetical example 5, Cronbach's alpha can be calculated using SPSS going through the following steps.

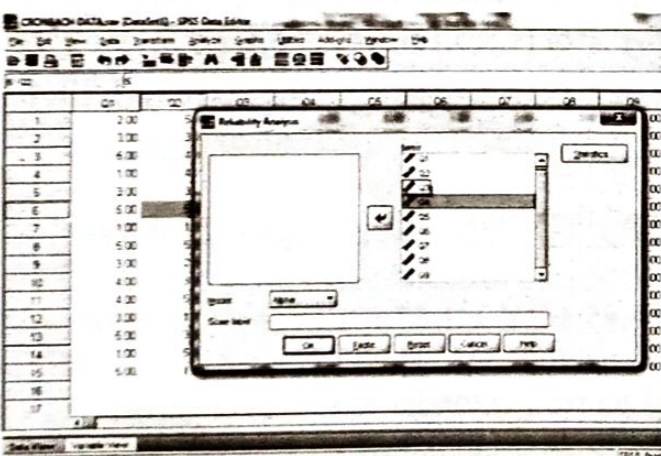
STEP 1: Click Analyze, then 'Scale' and '**Reliability Analysis**' on the top menu, as shown in the screen shot below:



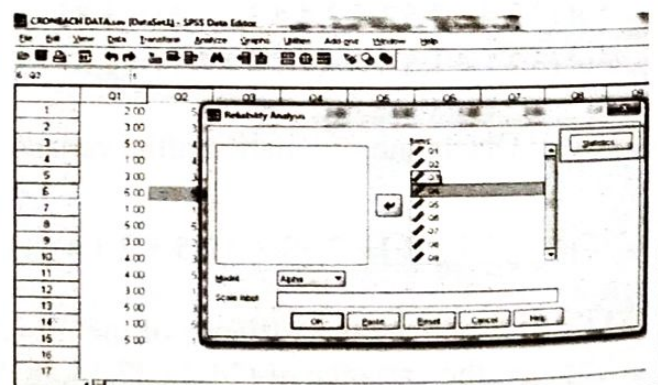
STEP 2: the Reliability Analysis dialogue box will appear as shown below:



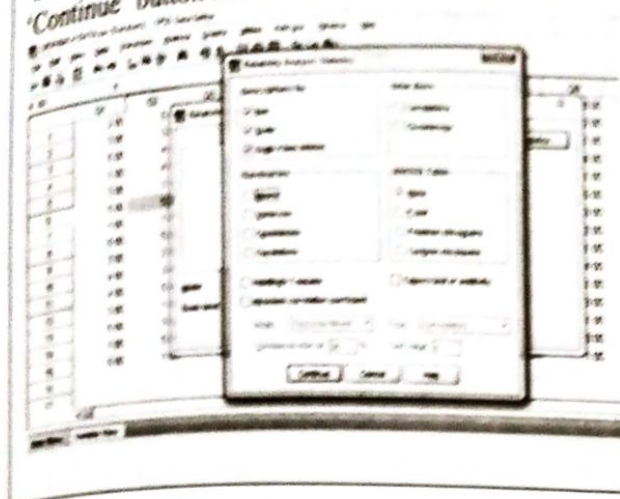
STEP 3: Select all the items and click on the arrow to move them to the box on the right for the analysis



STEP 4: Select all the items and click on the arrow to move them to the box on the right for the analysis. Alpha is the default selection on Model then click the 'Statistics' button for other options



STEP 5: In the statistics option box, check the 'Item', 'Scale' and 'Scale if item is deleted' options and the 'Continue' button and then 'OK' to produce result.



STEP 6: SPSS output for cronbach's alpha is produced

SPSS OUTPUT

Case Processing Summary

Case Processing Summary	N	%
Total	15	100.0
Excluded from analysis	0	0.0
Total	15	100.0

Reliability Statistics

Reliability Statistics	Cronbach's Alpha	N of Items
	.633	9

Item Statistics

Item Statistics	Item	Mean	Std. Deviation	N
1	1	2.0000	.80000	15
2	2	2.0000	.80000	15
3	3	2.0000	.80000	15
4	4	2.0000	.80000	15
5	5	2.0000	.80000	15
6	6	2.0000	.80000	15
7	7	2.0000	.80000	15
8	8	2.0000	.80000	15
9	9	2.0000	.80000	15

Item-Total Statistics

Item-Total Statistics	Item	Total Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item Total Correlation	Squared Multiple Correlation	Partial Correlation
1	1	24.0000	24.0000	.000	.000	.000
2	2	24.0000	24.0000	.000	.000	.000
3	3	24.0000	24.0000	.000	.000	.000
4	4	24.0000	24.0000	.000	.000	.000
5	5	24.0000	24.0000	.000	.000	.000
6	6	24.0000	24.0000	.000	.000	.000
7	7	24.0000	24.0000	.000	.000	.000
8	8	24.0000	24.0000	.000	.000	.000
9	9	24.0000	24.0000	.000	.000	.000

Scale Statistics

Scale Statistics	Mean	Variance	Std. Deviation	N of Items
1	24.0000	24.0000	4.89897	9

STEP 7: Interpret results

CASE PROCESSING SUMMARY shows how many participants or cases were valid and ensure that none was missing or how many if missing

ITEM STATISTICS show the mean and standard deviation and how many times each item included in the analysis was answered. This ensures that each question was answered for all participants.

RELIABILITY STATISTICS this value shows how well does the measure capture the 'feeling of safety which is the concept' the researcher is trying to measure. From the result, Cronbach's alpha as 0.63 and the goal is for alpha to be close to one (1) as much as possible.

SCALE STATISTICS shows the variance which explain the range between the items and how they were selected on the scale

TOTAL ITEM STATISTICS shows how Cronbach's alpha will fare if particular items are deleted shown on the last column. Recall that the aim of conducting the analysis is to ensure that Cronbach's alpha is close to 1 as much as possible. The column shows the resulting alpha if a particular item is deleted. With Cronbach's alpha at 0.63 approximately 0.6, the result show that alpha is increased to 0.684 approximately 0.7 if item 15 is deleted.

The foregoing shows that calculating using SPSS statistical package opens the student to understanding Cronbach's alpha better with item, scale, and total item statistics all in one analysis. On this premise, students are advised to acquaint themselves to the use of the package.

Interpretation:

Alpha is measured on the same scale as Pearson's r (correlation coefficient) which typically varies between 0 and 1. Although a negative value is possible which indicates a scale in which some items measure the opposite of what other items measure? The closer the alpha is to 1.00, the greater the internal consistency of the items in the instrument being assessed.

CORER AGREEMENT RELIABILITY (measures of scorer or rater consistency)

Evidence of behavioral changes as a result of students' learning can be assessed using practical assessment techniques. Such assessments are firmly rooted in observation of behavioral changes as a result of learning. Though, tests have been largely used in measuring learning achievements; however, Kautz, Heckman, Diris, Weel, and Borghans (2014) explained that Ralph Tyler, one of the two scholars credited with developing achievement tests, having recognized the limitations achievement tests, suggested using measures of behavior such as performance, participation in student activities, and other behavioral observational techniques by teachers to complement achievement tests when assessing students in schools. As such, the extent to which learning objectives have been achieved can be ascertained using practical methods; having proved to be a relevant source of information about professional performance. Such assessments are often based on scores or judgment by individuals who are trained to evaluate the performance or product. The goal of using this reliability technique is to ascertain the measures of error as a result of an individual scorer. According to Siegle (2002), the consistency between ratings can be calculated using inter-rater reliability. Some methods for ascertaining scorer agreement reliability discussed in this section are inter-rater agreement, percentage agreement and Inter Class Correlation (ICC).

Interrater Agreement: This is the relationship existing between the scores submitted by three or more judges on a context or activity. In this situation, two judges can evaluate a group of student and the correlation between their ratings is calculated with an acceptable correlation coefficient (r) of 0.90. The coefficient obtain is interpreted the same way bi-variate correlation coefficients are interpreted.

Percentage Agreement: this is a numerical estimate/measure of the degree of agreement between two raters or judges. Therefore, the two judges can evaluate a group of products and a percentage for the number of times they agree is calculated with an acceptable percentage (%) of 80.

Hypothetical Example 5: Five students on teaching practice were rated by lecturers allotted to them on two occasions as displayed on table 10. Calculate the percentage agreement of both raters.

Table 10: Students' Ratings

S/N	Rater 1	Rater 2	Agreement Ratio
1	60	60	1
2	40	55	0
3	55	55	1
4	65	65	1
5	70	65	0

SOLUTION:

Step 1: Count the number of ratings that are in agreement indicated as 1 and those that are not indicated as 0.

Step 2: Count the number of ratings in agreement

Step 3: Divide the total by the number in agreement to get a fraction using the formula: $\frac{AR}{n}$

Where: AR = Agreed Rating and N = total number of students rated = $\frac{3}{5}$

Step 3: Convert the fraction to a percentage

Therefore, the ratio of rating agreement = $\frac{3}{5} \times 100 = 60\%$

Step 4: Interpret result: with the acceptable benchmark of 80%, the inter-rater agreement is low.

Worthy of note, calculating percent agreements for more than a handful of raters can quickly become cumbersome. Therefore, a major flaw with this type of inter-rater reliability is that it doesn't take chance agreement into account and so overestimate the level of agreement.

Factors Affecting Test Reliability

The following factors are capable of affecting reliability estimates:

1. Test Characteristics

Vocabulary and sentence structure, order of items, clarity of the stem and the foils, and appropriateness of test format relative to the level of knowledge being measured will all affect test reliability. When test items are poorly constructed, directions are unclear or inadequate and when test forms that are suitable for measuring factual knowledge are used to measure higher order cognitive skills, then the reliability of tests is lowered. Items that do not discriminate effectively among students (e.g. very easy, very difficult, no clear answer and no correct answer at all items) will lower reliability.

2. Test length

The number of questions should be increased to be able to get a better sampling of student behavior. Adequate time must be given to questions, insufficient time, that may prevent many students from completing the items or surplus time that may aid cheating because the test score of a student that cheated will not be an accurate representation of his/her level of knowledge. The test administrator should not use his/her own initiative, to give more time than is allowed in the instruction because candidates have almost finished or to reduce the time because almost everybody has finished.

3. Test Administrator

The administrator should ensure adequate preparation needed for the conduct of the examination such as lightening, sitting arrangement, availability of examination materials, time, venue and conducive environment. The test administrator should demonstrate commitment, discipline skill and presence of mind at his task, he/she should be familiar with student methods of cheating, he/she should not be engaged in private conversation during testing, reading novel and should be physically present.

4. Group homogeneity

A low correlation in group comparison due to homogeneity is likely to occur when the group on which the correlation has been obtained has already been selected on the basis of the test (or other measure that correlates with the test score). For example, if we want to correlate cumulative graduation GPA with success in masters' degree, and we select graduate students with a wide range of GPAs, the correlation with success in the masters' degree will be higher than when we restrict the predictor variable to graduate students with only first class degrees.

5. Item difficulty

Item difficulty is simply the percentage of students taking the test who answered the item correctly. The larger the percentage getting an item right, the easier the item. The higher the difficulty index, the easier the item is understood to be (Wood 1960). This has a profound effect on both the variability of test scores and the precision with which test

scores discriminate among different groups of examinees (Thorndike et al, 1991) When all of the test items are extremely difficult, the great majority of the test scores will be very low. When all items are extremely easy, most test scores will be extremely high. In either case, the test scores will show very little variability. The test administrator should construct questions that their range of difficulties will allow the candidate to demonstrate their competence, not less than 50% should be able to answer the question to some extent.

6. **Objectivity**

A test should normally be developed to serve a purpose and the results used for that purpose. No test will serve a universal use. It is therefore wrong when teachers simply handpick tests prepared by others for particular purposes and then adopt these tests for use in the classrooms, even when the content domain on which the test is based are not identical. A selection test that typically focuses on whether the candidate possesses the knowledge and skills needed to be admitted to a course calls for the use of aptitude tests. But the result cannot be used to identify students who should represent the institution in a competition.

7. **Marking and scoring**

There should be uniformity in marking and scoring standards because a test that is well prepared but poorly scored can never be reliable. The score of papers should be based on realistic and appropriate standards one paper should be scored at a time for all papers with a marking guide to avoid variation in scoring from paper to paper.

References

- American Educational Research Association, American Psychology Association and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*, Washington, DC: AERA.
- Cronbach, L. J. (1951). Coefficient alpha and the International Structure of tests. *Psychometrika*, 16, 297-334.
- Guilford, J.P. & Fruchter, B. (1986). *Fundamental Statistics in Psychology and Education* (6th Ed.) New York: McGraw-Hill Book Co.
- Hassan, T. (1998). *Introduction to Test and Measurement*. Ibadan: Alpha Publishers.
- Hopkins, K.D. & Stanley, J.C. (1998). *Educational and Psychological Measurement and Evaluation* (8th ed). New Jersey: Prentice-Hall, Inc.
- Kuder, G.F, and Richardson, M.W. (1937). The theory of estimation of test reliability, *Psychometrika*, 2, 151-160.
- Mehrens W.A. and Lehmann (1978) *Measurement and Evaluation in Education and Psychology* (2nd ed). New York: Holt, Rinehart and Winston..
- Nwana, O.C. (1982). *Educational Measurement for Teachers*. Lagos: Thomas Nelson (Nig.) Ltd.
- Ojerinde, D. (2004). Examination malpractice in our educational system: The NECO experience, 2004 faculty of Education Lecture delivered at the Obafemi Awolowo University on 23rd February.

- Okpala, P.N.; Onocha, C.O. & Oyedeji, O.A. (1993). *Measurement and Evaluation in Education*. Ibadan: Stirling-Horden Publishers (Nig.) Ltd.
- Osterlind, G.J. (1983). *The item bias*. Newbury Park: Sage publication.
- Thorndike, R.L. & Hagen, E.P. (1977). *Measurement and Evaluation in Psychology and Education*. New York: John Wiley & Son.
- Wood, D. A (1996). *Test Construction: Development and Interpretation of Achievement tests*. Columbus, OH: Charles E. Merrill Books, Inc.