RELIABILITY AND STANDARD ERROR OF MEASUREMENT: IMPLICATIONS FOR TEACHERS' DECISIONS ON STUDENTS' ACADEMIC ACHIEVEMENT

[*1]Joseph Oluwatayo Osakuade (Ph.D)
Osakuade_tayo@yahoo.co.uk
+2348067394114 (Corresponding author)
&
[2]Benjamin Babatunde Aina
ainabb687@gmail.com
07031240226; 08054355757

[1,2]Department of Guidance and Counselling
Faculty of Education,
Adekunle Ajasin University, Akungba-Akoko,
Ondo State, Nigeria

Abstract

Educational measurement serves as a basis for the quantitative description of learners' behaviours to inform decisions on the learners. Measurements in psychological and behavioural sciences are not devoid of errors and the sources of the errors are multifarious. Error is the difference between the observed score of a learner and his true score. The smaller the error of measurement, the more the observed score tends towards the true score. Measurement error can lead to wrong decision and wrong decision has great consequences on the future of learners and the society at large This paper examined the meaning of test, measurement and evaluation; qualities of a good test; concept of validity; significance of test validity; concept of reliability; significance of test reliability; relationship between reliability and validity, error of measurement; sources of error of measurement; standard error of measurement; ways of reducing measurement error; and implications of standard error of measurement to teachers' decisions on students' academic achievement. In view of the consequential effect of measurement error on teachers' decisions, teachers should strive daily to minimize the measurement error. It was recommended among others that: (i). Knowledge of the qualities of good tests and their derivations must be made essential for classroom teachers. (ii). Multiple measures of students' performance should be adopted by classroom teachers to take decisions on students' actual performance since average of multiple measures would yield the true ability of the learner than the observed score that is being contaminated with errors.

Keywords: Reliability, validity, error of measurement, teachers' decisions, academic achievement

Introduction

Education could be described as the transmission of societal norms and values from one generation to another so as to enhance perpetual development of the society. The development of any nation is determined by the kind of education given to its citizens (Orji, & Job, 2013). Education is seen in Nigeria like other developing countries of the world as a veritable tool for effecting social, economic, military and political development goals. These goals of education according to Ossai and Nwalado (2017) can only be attained with quality assessment in place most especially at the tertiary institutions. The three legs of education are: Curriculum, Instruction and Assessment.

Curriculum is the body of knowledge to be taught to the students, Instruction is the methodology employed on how to teach the body of knowledge, while Assessment is a tool used to determine the extent of the mastery of the body of knowledge by students. Assessment therefore, is inseparable from education; otherwise, the other two legs cannot stand. In the realization of the importance of assessment to education, Bandele (2006) posited that all educational endeavours stand on the three practical concepts – Test, Measurement and Evaluation. They are three-in-one compound construct which Bandele referred to as the “Educational Tripod”.

In the teaching-learning process that involves human's behaviours, testing connotes the trial of somebody to find out his/her ability, powers, knowledge, skills, achievement, abilities, and so on. The device used to carry out this trial is called a test. Test according to Kolawole (2006) is an instrument for eliciting sample of behaviour or human traits or attributes that could be cognitive, affective and psychomotor domains having measurement (score) as its end product. Test, therefore, could simply be put as the series of questions to ascertain students' knowledge or ability.

Measurement is the quantitative description of human behaviour. According to Hills in Alonge (2004), Measurement involves the assignment of numbers to attributes, objects, events or people according to rules. Measurement therefore can simply be put as the process that attempts to obtain numerical representation of the extent to which a student possesses a relevant characteristics.

From an instructional point of view, Alonge (2004) viewsevaluation as includes both quantitative and qualitative descriptions of pupils' behaviour plus value judgment concerning the desirability of that behaviour. The essence of test and measurement process in education is to generate scores on students' achievement or aptitude with the intent of making decisions on the students. Scores have been put in highest esteem in Nigeria educational system for taking wide range of decisions on students. Some of the decisions that can be taken based on the scores obtained from tests and measurement according to Alonge (2004) are:

1. Instructional decisions by classroom teachers
2. Curriculum decisions by the school authorities and Ministry of Education
3. Selection decisionsby employers or educational institutions
4. Placement or classification decisions by examination bodies like WAEC, NECO and NABTEB
5. Personal decisions made by individuals
6. Administrative decisions made by administrators

Decisions taken on students would go a long way in determining the future of such student. Decisions could mar or improve the future of students. A decision taken in error on students is as a result of measurement error.

This paper examined the qualities of a good test; concept of validity; significance of test validity; concept of reliability; significance of test reliability; relationship between reliability and validity, error of measurement; sources of error of measurement; standard error of measurement; ways of reducing measurement error; and implications of standard error of measurement to teachers' decisions on students' academic achievement.

Qualities of a Good Test

For a test to be of good quality it must possess some qualities or meeting some criteria. According to Kolawole (2006), a good test must have the following attributes:

i. the test must be valid,
ii. the test must be reliable,
iii. the test should be dependable. That is the test must be both valid and reliable,
iv. the test should be capable of discriminating between brilliant and dull students,
v. the test should be comprehensive enough to cover the body of knowledge,
vi. the test should be moderately difficult, and
vii. the test should be usable, that is easy and possible to administer and score.

Concept of Validity

A valid test must measure accurately and consistently what it is designed to measure and nothing else. Shimbery (2014) defines validity as the level of confidence which an examinee's test score could be used to infer the ability under measurement possessed by the examinees. Some of the methods that can be used to ascertain the validity of a test if not adopted according to Alonge (2004) and Gbore (2019) are: face validity, content validity, concurrent validity, predictive validity, convergent and discriminant validity, and construct validity.

Anikweze (2018) enumerated some factors that can affect the validity of a test. Some of the factors are:

1. Comprehensiveness of the test in relation to content covered
2. Appropriateness of test with reference to the standard of the testees
3. The range of ability of testees
4. The relevance of traits with reference to prognostic uses
5. Choice of items
6. Reliability of the test

Significance of Test Validity

It is very important for teachers to consider the validity of tests they give to students because of the following reasons:

i. For accurate and reliable prediction of students' future success
ii. Valid test can serve as a reliable reference points for the promotion of learners to higher class
iii. Judgments based on the results from valid tests cannot be contradicted

?

iv. For the objective grading of learners
v. To confirm if desired changes have taken place from the teachers' instruction
vi. To provide a reliable basis for the comparism of a teachers' efforts (Anikweze, 2018).

Concept of Reliability

Reliability of a test refers to its ability to measure consistently whatever it sets out to measure. By implications, the scores obtained in one administration of the test will be similar in magnitude in repeated administration of the test to some group of students in similar testing situations. There are many methods in establishing the reliability of a test. The methods according to Nkemakolam (1997) and Gbore (2019) are:

1. Test-retest method: This involves administering the same test twice to the same group after a certain time interval (of at least two weeks) has lapsed. A reliability coefficient is then calculated using Pearson Product Moment Correlation to indicate the degree of relationship the two sets of scores obtained.
2. Equivalent or parallel method: This involves administering parallel or equivalent tests to the same group of subjects and the pair observations correlated. The two tests are to be equivalent since they contain similar but not the same items. These two tests may be given at the same time or after some interval.
3. Internal consistency methods such as:
   i. Split half method: This involves administering a single test to the same group of subjects. Scoring of two halves (usually odd items versus even items) of the test is done separately for each person, and then calculating the correlation coefficient. The correlation coefficient obtained is for the half of the test. To obtain the correlation coefficient for the whole test, Spearman Brown prophesy formula $r = \frac{2r\,½}{1+r½}$ can be employed. Where r is the reliability coefficient of the half test.
   ii. Kuder-Richardson KR20/KR21 formula: Kuder-Richardson approach is perhaps the most frequently employed method for determining internal consistency of a test. The formula KR-21 can be used only if it can be assumed that the items are of equal difficulty and can be dichotomously scored. KR-20 is a measure of internal consistency reliability for measures of essay type of tests. It is similar to Cronbach's alha α except ronbach's α is also used for non-dichotomous (continuous measures). Values can range from 0.00 to 1.00. the formula for
   Kuder-Richardson KR20 $= \frac{k}{k-1}\{1 - \frac{\sum_i^k piqi}{\sigma^2 x}\}$
   Where is the sum total of the product of the proportion of the testees that answered the items correctly and wrongly. $\sigma^2 x$ is the variance of the observed total test scores.
   Kuder-Richardson KR21 $= \frac{k}{k-1}\{1 - \frac{M(K-M)}{K\,(SD^2)}\}$
   Where K = number of items on the test, M = mean of the set of test scores, and SD =

standard deviation of the set of test scores.

iii. Cronbach's alpha: This is a general form of the KR-20 formula to be used in calculating the reliability of items that are not scored right versus wrong, as in some essay tests where more than one answer is possible.

Cronbach's α is defined as: $\frac{N}{N\text{-}1}\{1 - \frac{\sum_i^N=1\sigma^2 Yi}{\sigma^2 x}\}$

Where N is the number of components (items), σ^2X is the variance of the observed total test scores, and σ^2Yi is the variance of component I

Significance of Test Reliability

It is very important that teachers should ascertain the reliability of their tests. The reasons why teachers should ascertain the reliability of their tests according to Anikweze (2018) are:

1. Test with established reliability index makes testing meaningful and dependable
2. The reliability of a test attests to the teachers' honesty in evaluation
3. Only reliable test produces test scores that can usually discriminate learners' abilities
4. It is tests with proven reliability that can perform the important function of motivating student learning
5. Teachers can depend on reliable test outcomes as basis for realizing prediction of learners' future attainments
6. Only reliable teacher-made tests can effectively prepare students formore serious external examinations.

Relationship between Reliability and Validity

Reliability is a part of validity. Reliability is a necessary but not sufficient condition for validity. Therefore:

i. For a test to be valid, it must be reliable.

ii. If a test is unreliable, it cannot be valid.

iii. A test could be reliable and still not be valid.

Error of Measurement

Error of measurement is anything that causes deviation of scores in measurement. Alonge (2004) defined error as any variable that is irrelevant to the purpose of the testing and results in inconsistencies in measurement. Measurement in psychological and behavioural sciences are not devoid of errors simply because they are involved in the use of human elements which are difficult to predict. In addition to this, the variables or constructs that are being measured are unstable. Error in measurement was initially noticed by Charles Spearman in 1901 when he propounded Classical Test Theory (CTT). In CTT, Spearman observed that the Observed score often reported by teachers is a composite function of True score and Error score. It is assumed that the Observed score (X) = True score (T) + Error score (E). Error score is the difference between the Observed score and the True score. The more the error score tends towards zero, the more reliable the test score. The more the error score is far from zero, the more the discrepancies between the Observed score and True score.

Sources of Error of Measurement

Sources of measurement error are numerous. Thorndike in Alonge (2004) classified the sources of error into three. They are:

A. Errors that are due to the test itself, such as:

i. Items in the test

ii. Ambiguity in the wording of an item

iii. Scoring procedures

B. Test administration conditions such as:

i. Side attractions around the examination hall

ii. Personality of the examiner

iii. Poor directions on the examination questions.

C. Day-to-Day changes in the examinee, such as:

i. Examinees' motivation

ii. Examination anxiety

iii. Psychological conditions of the examinees

iv. Experience with test (testwise or test sophistication)

D. Other factors, such as:

i. Length of the test

ii. Spread of scores

iii. Difficulty of the test

iv. Scores objectivity

Standard Error of Measurement (SEM)

SEM is the standard deviation of test error scores. SEM is one of the means of estimating the amount of errors in a test. SEM can be obtained from the standard deviation of the obtained scores and test reliability. The formula for calculating SEM is given thus: SEM = SD $\sqrt{(1 - r)}$

Where SEM = Standard Error of Measurement

SD = Standard deviation of the obtained score

r = Reliability of the test

With the presence of Standard deviation and reliability of a test, SEM can be obtained.

SEM is inversely related to the reliability. The more reliable the test is, the smaller the SEM the test has. The less reliable the test is, the larger the SEM the test has.

For example, in a mathematics achievement test, the reliability coefficient of the test was 0.86 and the standard deviation was 0.62, the SEM is,

$$\begin{aligned} \text{SEM} &= \text{SD}\ \sqrt{1-r} \\ &= 0.62\ \sqrt{1-0.86} \\ &= 0.62\ \sqrt{0.14} \\ &= 0.62 \text{ X } 0.375 = 0.232 \end{aligned}$$

If the estimate of the SEM is known, it is possible to construct confidence bands around the students' true score based on their obtained scores (Alonge, 2004). Confidence bands (intervals) can be described as the range of scores which can be assigned a specific probability of containing the students' true score (Alonge, 2004). The most commonly used confidence intervals, according to Alonge (2004) are:

i. 68% confidence intervals
   Obtained score ± 1.0 X SEM

ii. 95% confidence intervals
   Obtained score ± 2.0 X SEM

iii. 99% confidence intervals
   Obtained score ± 3.0 X SEM

This practice of confidence bands would afford teachers to consider a range of scores where the examinee's true score probably lies rather than just a single value (that is obtained test scores) that always contain error.

For example, if Ayo obtained a score of 40 in Mathematics Achievement Test whose SEM is 3, the true score of Ayo using 68% confidence interval would be:

Obtained score ± 1.0 X SEM
40 ± 1.0 X 3
40 + 3 - 40-3
43 – 37

Therefore, the true score of Ayo ranges between 37 to 43 using 68% confidence intervals.

In view of the inverse relationship between reliability and SEM, a more reliable test will also have smaller bands.

Considering the above score of 40 using SEM of 10 and 68% confidence bands: the true score will range between 40 ± 10 X 1 = 40 ± 10 = 50 - 30

It can be observed from the above examples that where the SEM is 3, the range of score for true score is smaller than the range of scores where the SEM is 10.

Ways of Reducing Measurement Errors

Reducing measurement errors is a *sine qua non* to quality and accurate decisions taken by teachers on their students. Measurement errors can be reduced in the following ways:

A. Increasing the reliability of measurement results. According to Ikponmwosa (1997),this can be achieved by:
   i. Reducing ambiguities in test items.

   ii. Reducing variations in scoring.
   iii. Increasing the number of items in attest.
   iv. Increasing the discrimination power of test items

B. Increasing the validity of measurement results. According to Ikponmwosa (1997),this can be achieved by:
   i. Using a well-defined sample of students drawn from the appropriate populations to reduce fluctuation errors.
   ii. Designing a different test for different groups to correct differences in ability level
   iii. Ensuring scoring reliability
   iv. Using a table of specification in test construction
   v. Ensuring consistency in test administration procedures
   vi. Promoting students' positive attitude towards examination

C. Other ways of reducing measurement error according to Tan (2005) are:
   i. Transforming scores to a predetermined scale
   ii. Weighting the scores by their reliability
   iii. Confidence interval or error of measurement could be added to students' scores

Implications of Standard Error of Measurement to Teachers' Decisions on Students' Academic Achievement

Error is very imminent in psychological and educational measurement. Errors affect the obtained scores reported by teachers on every student. Teachers are bound to commit two types of errors in measurement in the course of making decisions on students' academic achievement. The two types of errors often committed by teachers are Type 1 and Type II errors. Type I error is committed when teachers decide failure for a successful student. Type II error is committed when teachers decide success for a failing student. Type I error is more grievous than Type II error. Taking a wrong decision on students as a result of measurement error by teachers has great consequences. According to Ofem and Anagbogh (2015), the consequences of error of measurement on students' decisions by teachers are:

a. It may inhibit counsellee's chances of getting appropriate counselling
b. It may debar a credible candidatefrom getting admission in higher institution of learning
c. Qualified students may not be promoted to the next class
d. It mis-represents the real identity of individuals
e. It can give false information about item characteristics such as item difficulty, item discrimination and effectiveness of distracters.

One of the critical roles of teachers in the assessment of students is to strive in reducing error of measurement. If teachers are not striving towards reducing the sources of error of measurement or objectivity in measurement, they will continually produce assessment scores that will be far from the real ability of students. Decisions based on such spurious scores would be very inimical to the future of students and cause great calamity to the society at large. Wrong decisions could be likened to a medical practitioner sending a healthy living being to the mortuary, or leaving a corpse in the midst of living things (Joshua, 2019).

Conclusion

Measurement is an important concept in the teaching-learning process as it provides basis for quantitative description of learners' behaviour to inform decision makings by teachers. Educational measurement cannot be disentangled from error. Measurement error can lead to wrong decisions which have great consequences on students' future and the society at large. Sources of error of measurement are multifarious and multidimensional in nature. For free, fair, objective and credible decisions to be taken by teachers on learners, teachers should strive daily to reduce measurement error so that the obtained score of any learner would actually reflect his true ability**.**

Recommendations

In view of the suggested ways of reducing error of measurement and the conclusion, the following recommendations are made:

i. Knowledge of the qualities of good tests and their derivations must be made essential for classroom teachers

ii. Multiple measures of students' performance should be adopted by classroom teachers to take decisions on students' actual performance since average of multiple measures will yield the true ability of the learner than the observed score that is being contaminated with errors

iii. Test and measurement should be made a compulsory course by teacher training institutions for all prospective teachers

iv. Test and Measurement is only being offered as a specialized course in higher degrees in Nigeria, efforts should be intensified to offer the course as a specialized course in first degree since many of the teachers in elementary and secondary schools in Nigeria may not go beyond first degree.

Workshops and seminars should be organized to train in-service teachers from time to time to acquire appropriate skills for test construction, validation, administration and scoring.

References

Alonge, M.F. (2004). *Measurement and evaluation in education and psychology*. Adedayo printing (Nigeria) Limited, Iworoko, Ado-Ekiti, Ekiti State, Nigeria.

Anikweze, C.A. (2018). *Measurement and evaluation for teacher education*. Shiloh press associates, Onitsha, Anambra State, Nigeria.

Bandele, S.O. (2006). *Test, measurement and evaluation*: *The educational tripod*. 17th inaugural lecture of University of Ado-Ekiti, Ekiti State, Nigeria, held Thursday, 9th March, 2006.

Gbore, L.O. (2019). Beginning researchers and keys to valid and reliable research results. *Journal of Measurement and Evaluation in Education and Humanities.* 2(1), 1-13.

Ikponmwosa, O. (1997). *Measurement and evaluation: Principles and procedures*. United City Press. Benin city, Edo State, Nigeria.

Joshua, M.T. (2019). Battles in the classroom: Evaluation of teaching and learning to the rescue. Evaluation of teaching and learning to the rescue. 86th inaugural lecture of the University of Calabar, Calabar, Nigeria. Held 22nd May, 2019.

Kolawole, E.B. (2006). Principles of test construction and administration. Bolabay publications, Ikeja, Lagos Nigeria.

Nkemakolam, E.O. (1997). *Measurement in education, an introduction*.Benin, Nigeria: Barloz publishers.

.

Ofem, U.J. & Anagbogu, G.E. (2015). Students' variables and test/measurement anxiety among undergraduate education students in universities in Cross Rivers State, Nigeria.

Orji, K.E., & Job, M. (2013). The role of education in national development: Nigerian experience. European Scientific Journal 9(28), 312-320

Ossai, A.G., & Nwalado, E.N. (2017). Quality in higher education in Nigeria: perception oof global challenges. *Nigeria Academic Forum*, 25(1), 1596-3306.

Shimbery, G.R. (2014). FLUX (Computer program0. Moneterey , CA: CTB Macmillan/McGraw-Hill.

Tan, S. (2005). The effect of standard error of measurement to appropriateness of educational evaluation. *Manas Universitesi Sosyal Bilimier Dergisi,* 7(14), 197-201.