Valid Teacher-Made Tests: It's Implication on Students' Achievement in Business Education

By

Bertha Ese Omoruan
Institute of Education
Delta State University,
Abraka, Delta State
E-Mail: beomoruan1@hotmail.com

ABSTRACT

An attempt has been made to examine the teacher-made Test as one of the classroom practices impact on students' achievement in business education. Testing is an integral part of classroom practices that enhances students achievement; especially when valid and reliable teacher-made tests are used. It has been established that measurement instruments require systematic process of development and validation for effective classroom assessment. Test items writing is an art that must be learned. The Business Education teacher requires the knowledge and skills to develop and validate test items that can be used to monitor students' learning progress using the formative Teacher-Made Test, diagnose students' learning difficulties with the use of diagnostic Teacher-made Test and determine students' achievement of objectives set down before instruction, with the summative Teacher-made Test. A table of specifications ensures proper representation of every aspects of the trait being measured, giving the test instrument high content validity. Reliability evidence is very essential in determining the amount of measurement error expected, in order to ascertain the true variance of the observed scores accounted for by the true scores of students, which is the actual differences in the trait being measured. It is concluded that Business Education teachers be conversant with the principles of test development and validation processes or consult experts in this area for quality assessment tools to produce quality output of students and quality education.

INTRODUCTION

One of the ideal classroom practices of teachers in any institution of learning is the evaluation of students' achievement, using valid measurement instruments. In such classroom testing, teachers make use of the Teacher-made Tests. Decisions that teachers must make regarding students require certain knowledge of their aptitude, achievements and personal development (Gronlund, 1985). Such decisions have tremendous influence on students' lives, and as such, should not be causally made.

Evaluating students' achievement after instruction, using measurement instruments like the Teacher-made Tests in any subject area (Business Education for instance) plays an important role in the school programme. It is an integral part of instruction that provides information which serves as a basis for a variety of educational decisions. Although, the list of such decisions would be infeasible to be exhaustive, Gronlund (1985) identified some common questions that illustrate some of the major instructional decisions teachers generally are likely to encounter during the course of teaching. Here are a few of such questions:

i) How effective was my teaching?

Copyright © 2018 JOSTE. All Rights Reserved (www.atbuftejoste.com)

ii) How should the students be grouped for more effective learning?
iii) To what extent are students attaining the minimum essentials of the course?
iv) To what extent are students progressing beyond the minimum essentials?
v) At what point would a review be most beneficial.... etc (p.5).

An appropriate evaluation technique for answering the above questions will be the use of the Teacher-made Tests after certain stages of instruction.

A Teacher-made or classroom Test is a test instrument constructed by the classroom teacher to measure the extent of students' achievement of a certain class based on some specific objectives. There are issues of quality, scoring, grading and comparability of standards of this evaluation technique which could vary from one teacher to another. Such variation may result from lack of competence in the development, validation, administration, scoring and grading of this testing instrument. Ideally, students' performances in such testing should impact on their performances in both internally and externally conducted assessment and even out of the school system.

Observations have shown that there are flaws in teachers' classroom testing (Osadebe, 2013; Omoruan, 2015). The researchers reported that students are often examined with unvalidated Teacher-made Test items which are hurriedly set and administered to students without adhering to the procedures of test development and validation. Many of such test items fail to measure what they are supposed to measure. Meaning that, most Teacher-made Tests are not valid.

Testing of students' marketable business skills is an essential practice that enables teachers of Business Education determine the extent students have achieved the general knowledge about business art which are also needed for functioning effectively in career in other fields- like accounting, secretarial, to mention but a few (Ejafu, 2015). The use of the Teacher-made Tests for assessing students' achievement could be in the form of formative, diagnostic or summative assessment. After instruction, it is appropriate for the Business Education teacher to determine students' learning outcomes, using well valid test items based on the learning objectives. This will help to judge objectively the achievement of each student in the subject area.

The Business Education teacher requires the knowledge and ability to objectively test students in order to monitor their learning progress by using formative test instrument, identifying students with learning difficulties through diagnostic test instrument and then, be able to determine their achievement based on summative

Copyright © 2018 JOSTE. All Rights Reserved (www.atbuftejoste.com)

evaluation test technique. All these assessment the Business Education teacher will carryout using well developed and validated test instruments-which are the Teacher-made Tests.

Testing in Teaching

Teaching is an art; in whatever subject area. An art must have element of skills, practice and knowledge. The skills in teaching and the practice all form a 'body' of knowledge which is an art. In the art of teaching, teachers adapt different methods or strategies and testing techniques. Methods are the ways to understand the subject-matter; practice art goes for the 'how' aspect of teaching the subject, while the testing techniques conclude effective classroom practices. Every concept taught must be assessed. Assessment of students' achievement must support the learning of important skills in the subject area and furnish useful information to both the teacher and students. But, when invalid classroom test items are used for assessing students' achievement, such test items fail to measure what they are designed to measure. This could lead to passing wrong information and taking wrong educational decisions on students' achievement.

A test instrument is a standard set of items (questions) to be answered (Nworgu as cited in Omoruan, 2017). According to Ukwuije and Opara (2012), a test consists of series of tasks, items used to collect systematic observations, judged to be representative of the curriculum of the subject being measured. Tests are classified under different bases. For the purpose of this article, the classification based on the quality of tests or who constructs the test is considered. These are: the Teacher-made Tests and the standardized Tests. A teacher-made (classroom) Test is a test constructed by the classroom or subject teacher to measure the degree of students' achievement in specific subject areas. While a standardised Test is a test constructed by test experts; administered under a uniform set of condition based on a large normative sample (Orluwene, 2012). The researcher differentiates between standardised and Teacher-made Tests as follows:

Table 1: Differences Between standardised and Teacher-made Tests

S/N	Standardised	Teacher-Made Test
1	Constructed by experts and subject specialists	Constructed by classroom teacher.
2	Good quality is ensured and very high	Quality is unknown and lower than the standardised test
3	Items are pre-tested and selected on the basis of their effectiveness (using item analysis)	Items are not pre-tested and selection of items are not effectively done.

Copyright © 2018 JOSTE. All Rights Reserved (www.atbuftejoste.com)

4	Procedures for administration and scoring are done under uniform instructions (using test manual).	Procedures for administration and scoring are not adequately in uniform
5	Reliability of the test is high (reliability coefficient is gotten through statistical techniques).	Reliability is low and not well known
6	Interpretation of scores is done on the basis of norm	Scores interpretation is limited to school situation.
7	Test manual is provided	Test manual not provided

Source: Adapted from Orluwene, 2012:6

From the above table, the Teacher-made Test lacks proper test items development and quality (validity and reliability). Test experts develop test items following systematic procedures which the Business Education Teacher lacks. In most cases where the Teacher-made Tests are developed and used by classroom teachers, students are often examined with unreliable test, items that were hurriedly set and administered to the students.

Test Development

One of the fundamentals of test development is the writing of test items. According to Aiken as cited in Omoruan (2017), test items represent methods of obtaining information about the individual student; but the amount and kind of information vary with the nature of the tasks posed by the different types of test items. The writing of test items is an art which must be learned (Gronlund, 1985). The ability to develop high quality test items requires the knowledge and skills in the application of the principles and techniques of test development.

For proper development of test items, a table of specifications or test blue print must be drawn. A table of specifications represents the master plan which guides the writing and review of the items that make up the test instrument. It ensures high content validity (one of the qualities of a test). It is a two-way pattern or grid drawn horizontally and vertically, showing the content areas covered and the behavioural (or instructional) objectives, respectively (Iweka, 2014; Osadebe, 2003; Gronlund, 1985). According to Iweka (2014), a table of specifications provides assurance that the test will measure a representative sample of the learning outcomes on the subject-matter or topics being measured. Percentage weights are given to each content area (in terms of the duration spent during instruction) and that of the behavioural objectives, based on the proportionate emphasis given. The proportion of each behavioural objective and content area are then calculated based

Copyright © 2018 JOSTE. All Rights Reserved (www.atbuftejoste.com)

on the number of items determined by the test writer (or teacher).

When test items specifications had been clearly formulated, the writing of the test items then commences. According to Aiken as cited in Omoruan (2017), it is recommended that about 20% more items than needed be written so that an adequate number of useful items will be available (after carrying out item analysis) for the final version of the test instrument. The format which the test items writing will take, or the form of response(s) required, is decided upon by the test writer. The format could be in the form of objective types (that is, multiple-choice, matching, supply, constructed response or identification), recall, recognition or essay type.

Trial testing of Test Items/Item Analysis

Trial testing of constructed test items is a process of administering generated test items to a group of persons similar to those intended to take the final form of the measuring instrument. The general purpose of trial testing the items in a test instrument is to serve the followings:

a) To identify weak or defective items,
b) determine the difficulty level and discrimination index of each item in the test,
c) determine the appropriate time limits for the final test,
d) determine the adequacy of directions and the test format, and
e) ascertain the length of the final test (Iweka, 2014; Orluwene, 2012; Shokare, 2006; Osadebe, 2001).

The data obtained from trial testing of the items in a test instrument are used for item analysis. According to Iweka (2014) and Orluwene (2012), item analysis helps to determine items in the test instrument that are contributing to the quality of the test. The researchers stated that, this process involves computing each item difficulty and discrimination indexes and as well as that of the distracters (when multiple-choice items are used). Test items for the revised edition or that of the final form are selected from the results of item analysis. The items so selected, should be proportional to the initial master plan (the table of specifications) of the test instrument, in order to produce a test instrument with the required characteristics.

In a typical item analysis of test ability using one of the measurement theory (the classical Test theory-CTT), the test writer specifically looks for item characteristics (difficulty and discrimination indexes). In determining item characteristics in CTT, extreme groups of certain percentages (of either 25%, 27% or 33%) for both upper and lower scoring groups are used (Kline as cited in Omoruan, 2017; Osadebe, 2003). According to Osadebe (2003), the middle group of 46% (if 27% of both upper and lower scoring groups is used) is discarded;

Copyright © 2018 JOSTE. All Rights Reserved (www.atbuftejoste.com)

but assumed to follow the same trend as those in both upper and lower groups.

The difficulty index of an item in a test instrument is a measure of what percentage of the group tested, answered the item correctly. It is denoted by 'p', and defined as:

$$P = \frac{R_u + R_L}{T} \quad \ldots\ldots\ldots\ldots\ldots (1)$$

Where,

P = difficulty index,

R_u= upper percentage of students who got the item right

R_L = lower percentage of those who got the item right, and

T = Total number of students in both groups that constitute the percentage used (Ukwuije & Opara, 2012; Osadebe, 2003).

The larger the difficulty index, the easier the item, and the smaller the index, the more difficult the item (Orluwene, 2012). It can range from 0 to 1. For instance, if an item has a difficulty index (p) of 0.90, it means a very easy test item; while a p-value of 0.20 is a very difficult item. Most test writers accept a p-value of 0.5 to establish a normal distribution of item difficulty (Anastasi & urbina, 2007; Kline, 2005; Osadebe, 2003).

Discrimination index on the other hand may mean how well an item measures or discriminates in agreement with the rest of the test, or how well it predicts some external criterion (Henard, 2004). It is a measure of how performance in one test item correlates with the total test score as a whole. It tells whether an item discriminates those of the upper and lower groups of the population being measured. According to Osadebe (2003), a good test item does that. Item discrimination index is denoted by 'D' and it is defined as:

$$D = \frac{R_u - R_L}{{}^1/_2 t} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where,

D = item discrimination index,

R_u, R_L and T are as defined by equation 1. The higher the D-value, the better the item in CTT, because, such a value indicates that the item discriminates correctly (Iweka, 2014). According to Osadebe (2003), when more examinees at the lower scoring group answer an item more correctly than those at the upper scoring group, such an item is said to be negative. A range of 0.40 to 0.80 index is taken to be appropriate and should be used (Okoh as cited in Omoruan, 2017). A range of 0.30 to 0.39 should be subjected to scrutiny; while a range of 0.19 to 0.29 should be reconstructed or discarded.

Psychometric Qualities of the Teacher-Made Test Instrument

The psychometric qualities (or properties) of a good test instrument are the validity and reliability of the test. In measurement, the theory of validity and reliability requires that a measuring instrument for obtaining information or gathering data should be valid and reliable before use. Gronlund (1985), Osadebe (2013)

Copyright © 2018 JOSTE. All Rights Reserved (www.atbuftejoste.com)

and Omoruan (2017) noted that, for any measuring technique like the Teacher-made Test to give useful information, its psychometric properties must be established. This is because validity and reliability help to ensure the quality of measurement instruments, which are required for quality assessment in education. These qualities of measuring instruments are indications that the instrument is appropriate for use. Conversely, a poorly developed measuring instrument used for the assessment of students' achievement lacks psychometric qualities and as such, information gotten from such instrument will be misleading.

Validity: This refers to the degree to which a measuring instrument actually measures what it is designed to measure, and for whom it is appropriate. It should be noted that almost any information or data gathered in the process of developing and/or using a test instrument is relevant to its validity. According to Osadebe (2013), such information is relevant in the sense that, it contributes to an understanding of what the test measures. The measure of validity of any measuring instrument enables anyone to judge whether the test is right for that purpose or not. If the purpose of the test is to describe or measure students' achievement, then it should be used for that purpose only. The same applied to predicting students' success in some future activity; which should be likened to provide as accurate an estimate of future success as possible. The quality level of validity measure needed in a test instrument depends on the function the test is to serve. Hence, as there are many uses of test instruments, so also are different types of validity. Five main types of validity include: face validity, content validity, predictive validity, concurrent validity and construct validity.

Face validity of a measuring instrument is the outward look or appearance of the test items, in line with the objectives of the subject or course being measured. Subject specialists and experts in test construction are given the test items to vet and certify that the items are appropriate. The appropriate outward look of the test items implies that the test instrument has face validity. Teacher-made Test instruments with face validity are required for quality classroom assessment in Business Education.

Content validity of a Teacher-made Test is the extent to which it adequately covers a given subject area (the domain) to be measured. It is the art of testing all that the students are supposed to have learnt, regarding the domain and behavioural objectives. To measure students' achievement in a given subject or course for example, Business Education, the content areas taught have to be covered. When the

Copyright © 2018 JOSTE. All Rights Reserved (www.atbuftejoste.com)

items in the test instrument cover the content areas taught, then, the instrument is said to have content validity. Content validity can be determined by drawing a table of specifications (Ukwuije & Omoruan, 2015; Iweka, 2014; Orluwene, 2012). A table of specifications gives measuring instruments high content validity (Ukwuije & Opara, 2012). Teacher-made Tests with high content validity should be used for quality classroom practices in Business Education.

To determine individual students' behaviour in the future, predictive validity is considered. It is predictive validity when one is interested in the relationship between two measures over an extended period of time (Gronlund, 1985). For example, the scores of Reading Readiness Test (a Teacher-made Test in reading skills) might be used to predict students' future achievement in reading. This is the extent to which a test instrument is able to select individuals who are capable of doing well in their future careers (Osadebe, 2013).

It is concurrent validity when one is interested in estimating students' present status, in relationship between two measures obtained concurrently. Concurrent of validity is the extent to which two test instruments of similar domain and difficulty level are given to students to answer. If the students perform well in both tests, then, the first test has a concurrent validity with the second test (Osadebe, 2013).

Construct validity is the extent to which a measuring instrument such as the Teacher-made test reflects the existence of the construct being measured. It is concerned with psychological traits (constructs) such as: intelligence, attitude, interest, reasoning ability, to mention but a few (Iweka, 2014; Osadebe, 2013; Orluwene, 2012). When students' scores are interpreted as measures of, for example, intelligence or reasoning ability, it is impling that there is a quality that can be called intelligence or reasoning ability; which can account to some extent, for performance on the test. Hence, such interpretation in terms of some psychological quality are concerned with construct validity. Procedures such as correlation with other related measures, internal consistency evidence and factor analysis may be used to determine construct validity.

Reliability: This refers to the consistency of test scores obtained by the same persons when they are examined with the same test instrument on different occasions (Omoruan, 2017). According to Osadebe (2013), reliability is the consistency of scores over time. A test instrument like the Teaccher-made Test, is said to be reliable when it measures consistency when administered at different times to the same persons. Various statistical techniques are

Copyright © 2018 JOSTE. All Rights Reserved (www.atbuftejoste.com)

used in establishing the reliability coefficient of measurement instruments. They include: test re-test, parallel form, split-half, Kuder Richardson formula 20 and 21, Cronbach Coefficient Alpha, to mention but a few.

Test re-test: is a method of establishing reliability, it is the extent to which a measuring instrument measures the stability if scores overtime. This method of estimating the reliability coefficient of a measuring instrument indicates whether students or testees would get essentially the same scores is they took the test at different times. The instrument is administered twice on the same respondents with at least two weeks interval and then scored. The two set of scores on the different administrations are correlated using Pearson Product Moment (PPM) correlation to obtain an index of stability (Kpolovie, 2014; Iweka, 2014; Osadebe, 2013; Orluwene, 2012; Ukwuije & Opara, 2012).

Parallel form: is another method of establishing reliability, it is the degree to which two measuring instruments with similar difficulty level measure equivalence. The respondents are given the two instruments at the same time. The scores from the two instruments are then correlated using PPM to obtain a coefficient of equivalence.

Split-half: is a method of estimating reliability coefficient by administering a single measuring instrument whose scoring is based on odd and even numbers; to obtain a measure of internal consistency. Coefficient of internal consistency is an estimate of reliability that describes the extent to which the items in a test are internally consistent or homogeneous with one another. The scores so obtained are correlated using PPM to obtain a coefficient called half reliability (rh) or half internal consistency. Since the full internal consistency is required, the Spearman Brown statistical formula is then applied. Spearman Brown formula is denoted by 'rf' and defined as:

$$rf = \frac{2rh}{1+rh} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

where,

rf = full reliability, and

rh = half reliability (Omoruan, 2017; Kpolovie, 2014; Osadebe, 2013; Orluwene, 2012; Ukwuije & Opara, 2012).

Kuder Richardson formulae (KR_{20} and KR_{21}) measure internal consistency of test items. The method is also a single administration type of technique for establishing reliability of test instruments. Both formulae are denoted by r. KR_{20} determines the reliability of test items that are dichotomous. That is, test items are scored 1 or 0 for right and wrong responses respectively, based on the proportion of correct or incorrect responses to each item in a test. Proportion of individuals passing an item is denoted by 'p' and that failing the

Copyright © 2018 JOSTE. All Rights Reserved (www.atbuftejoste.com)

item by 'q' (that is, q=1-p) (Osadebe, 2013). KR_{20} is defined as:

$$r = \frac{n}{n-1}\left(1 - \frac{\sum Pq}{SD^2}\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

where,

r = internal consistency reliability,

n= number of items in the test,

pq= product of proportion of passes and failures on each item, and

SD_2 = variance of the total test scores.

KR_{21} on the other hand, is based on the mean and variance of the test scores. It assumes all the test items are of equal difficulty (or that the average difficulty level is 0.5). It is defined as:

$$r = \frac{n}{n-1}\left[1 - \frac{\bar{x}}{nSD^2}(n - \bar{x})\right] \quad \ldots\ldots\ldots\ldots\ldots\ldots (5)$$

where,

r, n and SD^2 are as defined in equation 4, and $\bar{x}$ = mean of the total test scores (Kpolovie, 2014; Osadebe, 2013; Orluwene, 2012; Ukwuije & Opara, 2012).

The difference between KR_{20} and KR_{21} is that, KR_{21} assumes all items to be of equal difficulty. That is, p is constant for all items (Ukwuije & opera, 2012). According to these researchers, KR_{21} is more useful to the classroom teacher. It does not require much computation compared with KR_{20} (Mehrens & Lehmann as cited in Ukwuije & Opara, 2012).

Cronbach Coefficient Alpha is also a single administration technique where multiple levels of response like the Likert scale type, essay test, personality tests, etc are involved. In other words, test or self-reporting inventories that the scoring of each item takes a range of values, the appropriate procedure for the determination of the internal consistency reliability is the Cronbach Coefficient Alpha (Kpolovie, 2014). This formula was developed by Cronbach (Ukwuije & Opara, 2012). It is denoted by $\propto$ (alpha); and it is defined as:

$$\propto = \frac{n}{n-1}\left(1 - \frac{\sum SD_i^2}{SD_t^2}\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots (6)$$

Where,

$\propto$ = Cronbach Coefficient Alpha reliability,

n = as defined in equation 4,

$\sum SD_i^2$ = summation of item variance, and

SD_t^2= variance of total test score (t) (Onunkwo as cited in Omoruan, 2017; Kpolovie, 2014; Osadebe, 2013; Ukwuije & Opara, 2012).

Uses of Reliability Coefficient of Teacher-Made Test Instruments

Apart from reliability being an indispensable psychometric property that measuring instruments like the Teacher-made Tests must possess, its uses can be specified in a number of ways:

(i) **Ascertaining of true variance contribution for the whole test**: reliability coefficient can be used to determine the exact extent to which testees (students) actually differ in the trait being measured (Kpolovie, 2014). In other words, reliability is used for summarising the amount of measurement error expected,

Copyright © 2018 JOSTE. All Rights Reserved (www.atbuftejoste.com)

when a particular test instrument is employed. For example, if the reliability coefficient of a test total scores in Business Education is 0.95; this will give 95% variance of the observed scores accounted for by the true scores (which is the students' actual differences in the trait being measured). The error variance will be only 0.05; which is 5%.

(2) **Determination of standard error of measurement (S_{em}) for all the Testees**: S_{em} is an index representing the amount of measurement error in a test. The degree to which a test instrument provides inaccurate readings can be estimated through standard error of measurement (S_{em}). S_{em} can be represented as the standard deviation of the error score associated with each student's obtained scores, assuming the student took the test several times. It is an index that indicates how his scores scattered across the repeated testing. The lower the S_{em}, the closer the student's obtained score is to his true score. S_{em} is defined as:

$S_{em} = s\sqrt{1 - r}$ .....(7)

Where,

S_{em} = Standard error of measurement

s = standard deviation of the test total score, and

r = reliability of the test instrument.

For example, if the reliability of a Teacher-made Test instrument is 0.85 and the standard deviation is 12.7, the S_{em} will be:

$S_{em} = 12.7\sqrt{1 - 0.85}$

$= 12.7\sqrt{0.15}$

$= 12.7\ (0.287)$

$\therefore S_{em} = 4.92$

Hence, when reliability increases, S_{em} for a given test scores decreases. Conversely, the larger the S_{em}, the less reliable is the test instrument (Kpolovie 2014; Orluwene, 2012).

(3) **Setting confidence intervals for each of the Test scores**; when the reliability coefficient and S_{em} of a test scores are known, the confidence intervals wherein each testee's true score lies at a given percentage of certainty can be computed; using the areas of the normal curve. The normal curve with its mean, mode and median at the centre, is symmetrically shaped with a total area of 1(one)- having 0.5 to the right and 0.5 to the left; where areas of normal curve between different points can be calculated (Kpolovie, 2014). According to the researcher, using the area of normal curve, percentage of certainty of 68%, 95% and 99% of test scores fall within ± 1, $\pm$ 2 and ± 3 standard deviations respectively. The formula for the confidence interval at 68% certainty is:

$C_i = X \pm (1_z \times S_{em})$.....(8)

Where,

C_i = Confidence interval of a true score,

X = a particular score obtained by a testee, and

Copyright © 2018 JOSTE. All Rights Reserved (www.atbuftejoste.com)

1_z = 1 standard deviation unit in the normal curve (Kpolovie, 2014:156).

In education, 0.05 level of confidence is used for a two-tailed test, which is 95% confidence limit (Best of Kahn, 1993:332). If a test S_{em} is 4.92, we will have: ± 2 (4.92) = ± 9.84 S_{em} for a test instrument. This can then be used to set confident limit for each student's true score.

(4) **Approximation of each Testee's True score**: Each student's true score in a test can be directly estimated by the class teacher, through the knowledge of the reliability of the test scores. Using the value of the S_{em} of ± 9.84 in (8) above, for a student whose raw score in a test in Business Education is 54, we will have; $\pm$ 9.84 + 54. That is, -9.84 + 54 and 9.84 + 54. Which is approximately 44 and 64, respectively. Here, we are 95% sure that the true score of this student lies between 44 and 64,

(5) **Computation of separate S_{em} for each student**: When the sample and population for which a test instrument is validated or standardized are very large, the actual value for S_{em} varies at different points along the range of the test. Thus, in order to make more crucial decisions based on students' true scores in a test, the reliability index of the test scores can be used for establishing each student's specific S_{em}. The formula that can be employed in finding different S_{em} for individual testee's score given as:

$$S_{emi} = \sqrt{\frac{1}{n-1} X_i (n - X_i)} \quad \text{.....(9)}$$

Where,

S_{emi} = an individual testee's S_{em},

X_i = obtained score of ith testee, and

n = number of items in the test (Lord as cited in Kpolovie, 2014). For example, if raw score of a testee (student) on a Business education test is 1 (one), out of 20 (twenty) items in the test, then, his S_{emi} will be:

$$S_{emi} = \sqrt{\frac{1}{20-1} x\ 1\ (20 - 1)}$$

$$= \sqrt{0.05\ x\ 1\ x\ 19}$$

$$= \sqrt{0.95}$$

$\therefore S_{emi} = 0.97$

Therefore, with the student's S_{emi} known to be 0.97, the confidence intervals within which the true score of his obtained score of 1 lies, will be at the specified percentage of certainty. That is, in education:

$C_i = 1 \pm 2$ x 0.97 $\equiv$ 1-2 x 0.97 and 1 + 2 x 0.97.

= -1 x 0.97 and 3 x 0.97

= -0.97 and 2.91.

The student's true score lies within -0.97 to 2.91 confidence intervals.

CONCLUSION

The need to develop valid Teacher-made Tests for the evaluation of Business Education students has been discussed. The key to sound evaluation of effective instruction is to relate as directly as possible to what is to be measured. This can only be achieved by following the principles of test development and validation. Once the principles of developing and validating

Copyright © 2018 JOSTE. All Rights Reserved (www.atbuftejoste.com)

classroom Tests are violated, interpretation of obtained results and decisions taking on them will be misleading. Hence, Business Education teachers need to be abreasted with such test development principles and validation processes or consult experts in this area for appropriate assessment techniques or tools for producing quality output of students and quality education.

REFERENCES

Anastasi, A., & Urbina, S. (2007). *Psychological testing* (seventh edition), India: Dorling Kindersley Pvt. Ltd.

Best, J. W., & Kahn, J. V. (1993). *Research in education.* USA: Allyn and Bacon.

Ejafu, U. V. (2015). Challenges of business education programme in higher education in Nigeria. *Delta Business Education Journal,* 5 (1), 213-220.

Gronlund, N. E. (1985). *Measurement and evaluation in teaching (5th ed.).* New York: Macmillan Publishing co. Inc.

Henard, D. H. (2004). Item response theory. In L. G. Grimm and P. R. Yarnold (eds). *Reading and Understanding more multivariate statistics.* Washington, D. C.: American Psychology Association.

Iweka, F. (2014). *Comprehensive guide to test construction and administration.* Omoku: Chifas Nigeria.

Kline, J. B. (2015). Psychological testing: a practical approach to design and evaluation. (online). Available: http://www.sagepub.com/upm-data/4896_kline.

Kpolovie, P. J. (2014). *Test, measurement and evaluation in education (2nd edition).* New Owerri: Spring Field publishers Ltd.

Omoruan, B. E. (2015). Computer literacy, mathematics competency and human resource development: implication for pre-service teachers. *Delta Business Education Journal,* 5 (1), 102-110.

Omoruan, B. E. (2017). Construction, Validation and application of mathematics language mastery achievement test for junior secondary school students in Delta and Edo States, Nigeria. Unpublished Doctoral Thesis, University of Port Harcourt, port Harcourt.

Orluwene, G. W. (2012). *Introduction to test* theory *and development process.* Port Harcourt: Chris-Ron Integrated Services.

Osadebe, P. U. (2001). Construction and validation of economics achievement test for senior secondary school students. Unpublished Doctoral Thesis, University of Port Harcourt, Port Harcourt.

Osadebe, P. U. (2003). Principles of test construction. A paper presented at Delta State University, Abraka.

Osadebe, P. U. (2013). Validity and reliability of evaluation techniques for quality education. *Delsu Journal of Educational Research and Development, 12* (1), 56-63.

Shokare, R. O. (2006). How to assess the cognitive domain part one. Being a paper presented at a workshop on affective, psychomotor and cognitive domain for teachers in junior secondary schools in Delta State. Organised by the Ministry of Primary and Secondary Education, Asaba, Delta State.

Ukwuije, R. P. I., & Omoruan, B. E. (2015) Utilization of factor analysis in promoting quality measures in research. *Nigeria Journal of Educational Research and evaluation, 14*(3), 33-45.

Ukwuije, R. P. I., & Opara, I. (2012). *Test and measurement for teachers.* Alakahia-Akpor: Chadik printing press.

Copyright © 2018 JOSTE. All Rights Reserved (www.atbuftejoste.com)