Inaugural Lecture series 253

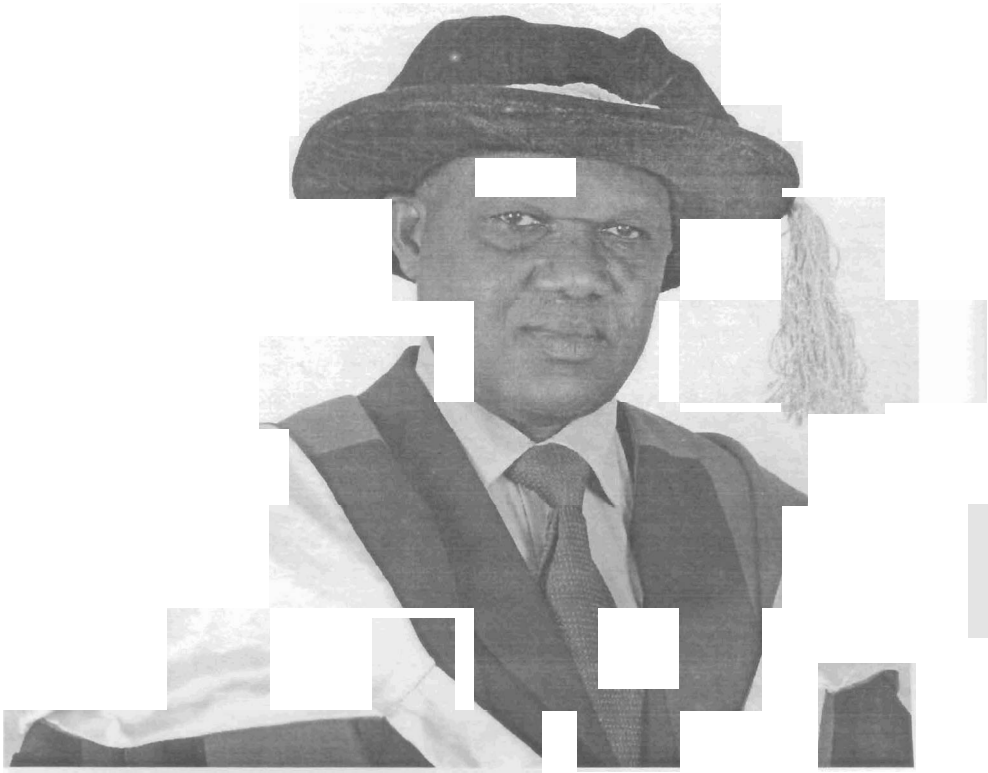
TESTS AND MEASUREMENT: A TALE BEARER OR TRUE WITNESS?

By

Eyitayo Rufus Ifedayo AFOLABI

Professor of Tests, Measurement and Evaluation

OBAFEMI AWOLOWO UNIVERSITY, ILE-IFE, NIGERIA.

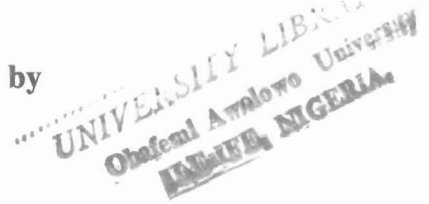


Eyitayo Rufus Ifedayo AFOLABI

Professor of Tests, Measurement and Evaluation

TESTS AND MEASUREMENT: A TALE BEARER OR TRUE WITNESS?

An Inaugural Lecture Delivered at Oduduwa Hall
Obafemi Awolowo University, Ile-Ife
On Tuesday, December 11, 2012



by

Eyitayo Rufus Ifedayo AFOLABI
Professor of Tests, Measurement and Evaluation

Inaugural Lecture Series 253

Obafemi Awolowo University Press Limited
Ile-Ife, Nigeria.

ISSN 0189-7848

Printed by
Obafemi Awolowo University Press Limited,
Ile – Ife, Nigeria

1. Introduction

Mr. Vice-Chancellor sir, and my distinguished audience, my coming into the field of Tests and Measurement was fortuitous. At the first instance, the desire to be an academic was inspired by Prof. Isaac Sodeye, who to my generation in the ECU of this university then, was a huge role model. I had been offered admission for a postgraduate programme in Mathematics of this university only to find myself in Education through the late Prof. (Mrs) Bukola Mabogunje Osibodu, and in Tests and Measurement through the first Head of my Department, Professor Olu Makinde. These events tended to support a tenet that all things work together for the good of them that love the Lord as He works out his purpose both to will and to do according to His pleasure.

Mr. Vice-Chancellor sir, it used to be said, apparently by those who find themselves at a low percentile of academic performance, that 'examination is not the best measure of a person's ability'. Perhaps, there are many people who still have this notion. There may be no need to contend with this statement, as it is everywhere evident that there is yet no viable alternative to tests and examinations in determining a person's academic ability. My mission today is therefore simple: to demonstrate that tests can validly measure, not just an individual's academic ability, but that any human characteristic, trait or attribute, can indeed be measured correctly to a significant degree of accuracy; also, to create an awareness of best test development practices, and to challenge teachers at all levels of our educational system to utilize these practices for the benefit of our students, institutions and the society. Appropriately, the title of my lecture is: "Tests and Measurement: A Tale Bearer or True Witness?"

This is the 253rd Inaugural Lecture in this university, the 11th in the Faculty of Education since its inception in 1967, the second in the Department of Educational Foundations and Counselling, and the first in the towering field of Tests and Measurement. I am humbled that the Lord has given me grace to present this lecture. I wish to

pay respect to my predecessors in the Faculty of Education who have provided a good legacy in their inaugural lectures: Professors S.O. Awokoya, A. Fajana, A. Adaralegbe, I.O. Makinde, S.A. Olatunji, J.I. Agun, T.O. Fasokun, J.A. Fawole, J.A. Akinola and E.A. Bamisaye.

In my over 30 years of service in the university as a teacher, supervisor and mentor, I have had the privilege and honour of supervising numerous postgraduate students who have obtained professional masters degree in Education (M.Ed), 56 others who have obtained Masters degree with thesis (M.A.), and 15 who have obtained doctoral degree (Ph.D) among which four are lecturers in Tests and Measurement in the Department. I am also delighted that four of these former research students are now Professors of Education. A great portion of my career development belonged to an era when a handful of academic gladiators would, at private meetings decide, using criterion other than academic merit, who should be admitted into the professorship cadre and when. That was why I spent about the same number of years to progress from the grade of Graduate Assistant to Reader, as I did from Reader to Professor. But "blessed be the Lord, who has not given us as prey to their teeth; we have escaped as a bird from the snare of the fowlers; the snare is broken, and we have escaped" (Psalm 124: 6,7). May the tribe and climes of that era diminish everywhere, everyday.

2. The Discipline of Tests and Measurement

(a) Historical Development of Testing

The earliest testing programme was in China, more than 3000 years ago, involving oral examinations to determine eligibility into public service and for promotion decisions (Dubois, 1970). They made use of test batteries (two or more tests used in conjunction) in a national multi-stage testing programme. The British government adopted a similar system of testing for its civil service

in 1855, followed by the French and German governments. The United States government established the American Civil Service Commission, which developed and administered competitive examinations for certain government jobs. This provided the impetus for the testing movement in the western world (Wiggins, 1973). The most basic concept underlying psychological and educational testing pertains to individual differences, and the publication of Charles Darwin's book, *The Origin of Species*, in 1859, was an important step toward understanding individual differences. Sir Francis Galton applied Darwin's theories to the study of human beings which he articulated in his book, *Hereditary Genius*, published in 1869, and demonstrated that individual differences exist in human sensory and motor functioning, such as reaction time, visual acuity, and physical strength (Galton,1879). Galton's work was extended by Cattell, an American psychologist, who coined the term mental test (Cattell, 1890). Thus, the first line of the development of psychological testing was based on the work of Darwin, Galton and Cattell on the measurement of individual differences.

A second major foundation of testing was based on the work of German psychophysicists Herbart, Weber, Fechner and Wundt, the fathers of experimental psychology. It was from their work that came the idea that testing, like an experiment, requires rigorous experimental control, such as administering tests under highly standardized conditions (Kaplan and Saccuzzo, 2005). Herbart developed mathematical models of the mind and used these as the basis for educational practices. Weber, Fechner Wundt, Titchner and Thurstone later built on this tradition, leading to the development of the Strong Vocational Interest Blank (SVIB).

However, the breakthrough in the development of modern tests came at the turn of the 20th century through the work of Alfred Binet, and T. Simon (who together developed the Binet–Simon Scale in 1905); with L.M. Terman who developed the Stanford–Binet Intelligence Scale in 1916, then a landmark in the testing

field. The development of Army Alpha and Army Beta human ability tests during World War 1 led to the emergence of standardized achievement tests which provided multiple–choice questions that were standardized on a large scale sample to produce norms against which the results of new testees can be compared. Personality tests, which measure presumably stable characteristics or traits that theoretically underlie behavior, also began to receive attention during this period. The earliest personality test, the Woodworth Personal Data Sheet, a structured paper–and-pencil group test, was published in 1920. Its simplistic assumption that the content of an item could be accepted at face value led to the development of projective personality tests, Rorschach Inblot Test (1921) and the Thematic Apperception Test (TAT) (1935). The two tests present ambiguous stimuli (pictures) depicting a variety of scenes and situations. While the Rorschach test required the client to explain what the inkblot might be, the TAT asked the client to make up a story about the ambiguous scene. Psychoanalytically-oriented psychologists believe that behavior is determined by unconscious processes more than by conscious ones, and that a test that asks straightforward questions is unlikely to tap the roots of an individual's personality characteristics. Projective tests assume that an individual will 'project' his or her personality into the ambigucus situation in such tests, and thus make responses that give clues to this personality. However, in a review of over 300 studies on projective tests, Lundy (1985) found low reliability and validity for projective tests such as Rorschach and Thematic Apperception Test (TAT). A comprehensive and critical review of the scientific status of projective tests (Lilienfield, Wood, & Garb, 2000) corroborate this position. Projective tests have not withstood a vigorous examination of their psychometric properties (Wood, Nezworski, Lilieafield & Garb, 2003). In 1943, the Minnesota Multiphasic Personality Inventory (MMPI) revolutionized structured personality tests, using empirical methods to determine the meaning of a test response. This era reached a greater height with the appearance of personality tests based on the statistical

procedure of factor analysis, a method of finding the minimum number of dimensions called factors, to account for a large number of variables. The introduction of the Sixteen Personality Factor Questionnaire (16PF) in the late 1940's remains an important example of a structured test developed with the aid of factor analysis. Today, factor analysis is used in the design and validation of most major tests. From the 1950's, came rapid changes in the status of testing with applications to health, industry, business, counseling, law, education, social work and schools to solve practical human problems. From the 1960s saw the emergence of a new measurement perspective, the Item Response Theory (IRT). Its seeds lie in the psychometric tradition of the Classical Test Theory (CTT) upon which testing had hitherto been based. This new theory was promoted largely by Frederic Lord (see Lord, 1952). Unlike the CTT which is based on respondent's observed score on a whole instrument, the item is the unit of focus in the IRT. Its simplest model, the Rasch or One-Parameter Logistic (IPL) model, utilizes the Item-Characteristic Curve (ICC) as building blocks, which describe the performance of an item in a test, and is unique to each item. An important attribute of IRT is that its parameters - the item and person parameters – are not test or sample dependent. Its methodology led to applications in the equating of alternate examination forms, computerized adaptive testing, item banking, and the detection of test bias among others (Lord, 1980). De-Ayala (2009) identified three reasons why IRT has not been embraced by everyday researchers. These included the large sample size required, its complicated mathematics, and the software for estimating the parameters of the model that are not readily available.

(b) Test Types

Just as there are many types of behavior, there are many types of tests. These are in two broad categories: ability and personality tests. Ability tests are concerned with capacity, potential and skill. They measure knowledge, understanding and skills in terms of speed and accuracy; and may be achievement, aptitude,

intelligence or performance tests. Achievement tests measure the knowledge gained from direct experience or teaching; aptitude tests measure an individual's capacity to benefit from future learning experience. In the school system, achievement tests are commonly used. These comprise multiple–choice, true- false, completion and matching. They are referred to as select or fixed response types. Another ability test is intelligence test. Intelligence tests, traditionally measure a person's general potential to solve problems, adapt to changing circumstances and think abstractly. **Afolabi** (1988) identified variants of multiple–choice items as one correct answer, best answer, analogy, and reverse types. Others are association, substitution, incomplete, combined response, multiple response, paired item, interpretive and answer–until–correct. There is also the free response or essay types, which include short answer, problem sets, and expanded answers. Personality tests measure patterns of behavior and thinking that prevail across time and situations and the personal characteristics that underlie and determine them.

(c) Test Use

In the school system, tests are used to :

- motivate students to study;
- determine how much students have learned;
- identify students' special difficulties and abilities;
- determine the adequacy of instructional resources;
- provide feedback on the strengths and weaknesses of teaching objectives;
- determine learning progress;
- predict students' performance;
- select learning experiences;
- determine students' vocational interest;
- ascertain curriculum efficiency;
- determine school effectiveness;
- measure teacher attitudes and competence;
- determine school needs assessment; and
- select, place, or advance students to the next class or level.

(d) Meaning of Tests and Measurement

Tests and measurement is an academic discipline in Education; however, its roots are in Psychology (where it is referred to as psychometrics) and Statistics. Mr. Vice – Chancellor sir, let me first start by explaining the meaning of the terms. A test is an instrument or systematic procedure for measuring a sample of behavior. It is a measurement device or technique used to quantify behavior (Kaplan & Soccuzzo, 2005). It may also be defined as a standard set of items which are specific stimuli, to which a person overtly responds and which can be scored. Items, that is, the specific statements, questions, tasks or problems that comprise the test, are the building blocks of tests. Responses to test items produce a measure, a numerical score.

Measurement refers to the process used in obtaining the score, and may involve test or non-test methods. A classic definition of measurement is the assigning of numerals to objects or events according to rules (or the application of rules for assigning numbers to objects or events). The numbers are not mere labels on an individual's characteristic, but may also depict the degree to which the individual possesses the particular characteristic. Through the application of mathematical techniques, the measurement process facilitates the understanding of the nature of a variable or trait. Thus, measurement may be defined as the science of assigning numerical data (discrete or continuous) to the characteristic properties of objects, events, and systems in order to precisely describe the object, event or system. Measurement makes ideas and concepts to be clearer, and knowledge and skills to be ordered and comparable. It is a trite saying that "whatever the mind can conceive can be measured", and Galileo once said "..... measure what is measurable, and what is not measurable, make measurable". This underlies the fact that characteristics or properties of people such as interests, beliefs, social relationships, attitude, honesty, hope, industry, bravery, kindness, hatred and love can be measured. We can also measure religiosity, spirituality, generosity, creativity, curiosity, motivation,

adjustment, patriotism, aggressiveness, philanthropy, depression, accountability, reputation, likeness, knowledge, aptitude, intelligence ..... The list is virtually endless.

What then is Tests and Measurement?

(i) It is a discipline concerned with developing test instruments and non-test techniques to measure the amount of epistemic, personological, and personality characteristics or traits of individuals or groups; (ii) it examines the extent to which programme or project objectives are achieved, and (iii) determines indices of prediction and improvement in measurement parameters. Its academic themes include educational, psychological and vocational testing, educational statistics, test theory, educational evaluation, and evaluation of social action programmes. Its application spans all aspects of education from curriculum studies, counseling psychology, and educational planning to physical education and economics of education. It also contributes to the understanding of clinical psychology, public health and behavioural research in general, especially in the application of statistical design, assessment and impact analysis of interventions. I wish to humbly describe Tests and Measurement as the crown prince of Education. Bandele (2006) identified seven perspectives of Tests and Measurement, a critical assessment of which suggests that the linear or traditional view of Tests, Measurement and Evaluation, is the most appealing, meaningful, logical and pragmatic. That is

Tests ⟶ Measurement ⟶ Evaluation

In this proposition, the administration of tests lead to measures or scores and these, together with quantitative and qualitative inputs from other (non – test) sources, are used in evaluative decisions.

(e) Scales of Measurement

In measurement, system of rules of assigning numbers to objects must be clearly defined. The basic feature of these types of systems is the scale of measurement. There are three important properties of measurement scales namely, magnitude, equal intervals and an absolute zero, A scale has the property of magnitude if we can say that a particular instance of the attribute represents more, less, or equal amounts of the given quantity than does another instance. By equal intervals of a scale, it means the difference between two points at any place on the scale has the same meaning as the difference between two other points that differ by the same number of scale units. Where such occurs, the relationship between the measured units and some outcome can be described by a straight line or a linear equation. An absolute zero occurs when nothing of the characteristic being measured exists.

There are four scales of measurement associated with the properties just described. They are the nominal, ordinal, interval and ratio scales. Nominal scale classifies individuals into two or more groups, the members of which differ with respect to the characteristic being scaled, without any implication of gradation or distance between the groups; dimensionality is not warranted. In the ordinal scale, individuals are ranked along the continuum of the characteristic being scaled, but without implication of distance between scale positions. The ranks are mere relative positions. The interval (or cardinal) type of scale has equal units of measurement, thus enabling the interpretation of, not only the order of scale scores, but also the distances between them. The highest level of measurement is the ratio scale, which has the properties of an interval scale together with a fixed origin or zero point. In psychological and educational measurement, the level of measurement is at best at interval level, and this delimits the statistical operations that are permissible, as well as the deductions and conclusions that can be reached. Moser and Kalton (1979) provide comprehensive discussion on the use of Thurstone, Guttman, Semantic differential, Social distance and H- scales. The

influence of the errors of central tendency, leniency, severity and halo effects, and the problem of faking are also highlighted.

3. Validity of Tests

(a) Reliability as Ingredient of Validity

Reliability is an essential requirement for validity. It describes the consistency or accuracy with which a test measures what it purports to measure. Classical test theory attributes the problem of test reliability to measurement error and the domain sampling model. In the former, error arises from the imperfection of measuring instruments; thus, the score observed for each person almost always differs from the person's true ability or characteristic. This can be represented symbolically as

$$X = T + E$$

Where,

X = observed score
T = true score, and
E = error

Errors of measurement are assumed to be random. But measurement error can be estimated using the standard deviation of errors, called standard error of measurement. In the domain sampling model, error is introduced by the use of a sample of items rather than the universe of items of the construct being measured. Thus, the greater the number of items, the higher the reliability of the test. Test reliability is usually estimated in one of three ways: test – retest (time sampling), parallel forms (item sampling) and internal consistency (item homogeneity). Test–retest reliability is concerned with the consistency of the test scores when the test is administered on different occasions, while parallel forms reliability evaluates the test across different versions of it. Internal consistency reliability examines how people perform on similar subsets of items selected from the same form of the measure. The two most popular methods for determining internal consistency are Kuder – Richardson 20 (for dichotomous items) and coefficient Alpha (for general items). Split – half method may also be used. But all the measures of internal consistency evaluate the extent to

which the different items on a test measure the same ability or trait. Where observations are used instead of psychological tests, Kappa statistic (Fleis; 1971) is used to estimate reliability. **Afolabi,** Dibu–Ojerinde, Alao and Faleye (2005) demonstrated the use of Kappa coefficient in estimating the reliability of students on teaching practice. In the study, 213 student–teachers who were supervised by three or more supervisors constituted the study sample. Twenty four lecturers supervised the student teachers in school subjects related to the academic areas in which they obtained first degrees. Table 1 presents the data obtained from supervisor ratings, which were analysed using Kappa.

Table 1: Teaching practice scores

	Number of Spervisions			
	1	2	3	4
Number	213	213	212	210
Mean	61.9	62.6	62.3	62.4
Std. Error	0.468	0.459	0.462	0.812
Range	43	39	37	48
Minimum	35	40	42	40
Maximum	78	79	79	88

From the mean ratings, a Kappa coefficient of $r = 0.545$, $p<0.05$ was obtained. This result indicated that the scores given to student - teachers during teaching practice supervision were not spurious, but comparable and showed considerable concordance.

Mr. Vice–Chancellor sir, validity is the kernel of this lecture. To inquire about the role of Tests and Measurement as tale bearer or true witness is to question the validity of tests. The meaning and import of this concept will now be elucidated. Validity is normally an assessment of the quality of an instrument or experimental design. In this lecture however, we are concerned with the validity

of test instruments. Validity is the single most important characteristic of a test. It simply means truthfulness: does the test measure what it purports to measure? Are the conclusions from test results justified by evidence? Validity is the agreement between a test score or measure and the quality it is believed to measure. It defines the meaning of tests and measures.

(b) Unified Validity

There used to be several types of validity, but in 1985, a joint committee of the American Psychological Association (APA) and the National Council on Measurement in Education (NCME) published a booklet, "Standards for Educational and Psychological Testing" (Standards) in which validity is now conceived as:

(i) specific to a particular use;
(ii) a matter of degree;
(iii) a unitary concept

In particular, the revised 1999 edition of the Standards no longer recognizes different categories of validity, but categories of evidence for validity. Just as in the Athanasian Creed of the Christian Religion, we say there are no three Gods but ONE, in the same vein, in Tests and Measurement, we say there is no longer three kinds of validity but only ONE validity, namely construct validity. All other aspects of validity are subsumed in construct validity. Messick (1988) and testing pioneer, Cronbach (1980), summarized the subsisting position on validity among a preponderance of test experts thus: "all validation is one; and in a sense, all is construct validation (p. 99).

Validity can be evidenced by examining the content of the test for sampling adequacy of the items, that is, if the test adequately represents the conceptual domain it is designed to cover. This includes the wording of the items and the appropriateness of the reading level. Inadequate content representation is regarded as *construct under representation,* while the introduction of elements unrelated to what is being measured is *construct–irrelevant variance.* Lecturers and teachers who give tests in which students who did not cover 50 percent of the course/subject content in their

preparations score 'A' (simply because such teachers selected a few of the topics covered), or those who ask questions not covered during lectures or lessons are guilty of either of these errors. Evidence of content coverage is logical rather than statistical and is often made by expert judgment. The evidence of item sampling adequacy can be corroborated in support of construct validity by correlating the test score with a well–defined criterion measure (i.e the standard against which the test is compared). When the criterion is used to forecast the power of the test, the validity evidence is predictive; when the test and the criterion score are obtained simultaneously, the validity evidence is concurrent. The relationship between a test and a criterion is expressed as a validity coefficient, which is usually not larger than 0.60. Its square is the percentage of variation in the criterion that we expect to know in advance due to our knowledge of the test scores

Campbell and Fiske (1959) distinguished between two types of evidence essential for a meaningful test: convergent and discriminant. Convergent evidence is when a measure correlates well with other tests believed to measure the same construct, while in discriminant (divergent validation) evidence, a test should have low correlations with measures of unrelated constructs. This demonstrates the uniqueness of the test. Assembling construct - related evidence for validity requires validation against many criteria. Content–related validation is an essential step in construct–related validation. Criterion–related evidence is similar to convergent and discriminant evidence.

Afolabi and Buhari (2004) investigated the validity of Holland's Self-Directed Search for use in Nigeria. Holland (1985) postulated a choice-point vocational typology theory that vocational engagements are the results of the interaction among biological constitution, social experience, maturation and aptitude. He predicted that every individual could be described as possessing personality characteristics that fit into any of the six personality types of (i) realistic, (ii) investigative, (iii) artistic, (iv) social, (v)

enterprising, and (vi) conventional. He also developed an inventory, the Self-Directed Search (SDS) which can classify people into three-letter codes representing the aggregate of the three most dominant personality types of an individual.

The SDS was administered on 615 participants, purposively selected to represent six professional groups as follows;

- Realistic: Engineers, Automobile Mechanics, Machine Operators
  Investigative: Mathematicians, Scientists, Geographers
- Artistic: Artists, Musicians
- Social: Teachers, Nurses, Counsellors
- Enterprising: Administrators, Salesmen, Personnel Managers
- Conventional: Accounting Officers

A total of 567 or 93.6% completed forms were returned. Data collected were analysed using Chi-square and Mann-Whitney tests. The results showed that 72% of the respondents were congruent, 27.4% were partly congruent and 0.6% were totally incongruent. Chi-square analysis yielded a significant value of $\chi^2 = 15.7$ ($p<0.05$). When the occupational types of the participants were compared with Holland's occupational codes of RIASEC (74.21%), 421 had their occupational types matched by the three most dominant of Holland's occupational codes. Those who did not match were 155 or 25.8%. A statistical comparison of both groups yielded U=34, which was significant. The results showed that the working class adults in the study had personality codes that were congruent (72%) with their occupation, and also showed general consistency in the profiles of the six groups. This supported the validity of the SDS as a vocational instrument and its suitability for use in Nigeria.

Concern about the quality of examinations conducted by state governments motivated Faleye and **Afolabi** (2005) to examine the (predictive) validity of the JSCE, using performance in the SSCE

as criterion. It was postulated that students who perform well in the JSCE will also perform well in the senior secondary school. Five hundred and five students from Osun State who had results in the JSCE and with intact academic records in SSI, SS2, and SSCE (WAEC) constituted the sample for the study. Examination scores of the students in six JSCE subjects, namely: English Language, Mathematics, Integrated Science, Yoruba Language, Social Studies, and Agricultural Science were matched with corresponding subjects in the senior secondary school and the SSCE. JSCE grades of A, C, P and F, and SSCE grades of Distinction, Credit, Pass and Fail were scored 3, 2, 1 and 0 points respectively. Aggregate scores were obtained for each student in all the subjects and analysed using Pearson r. Table 2 presents the results.

Table 2: Relationship between JSCE and SSS scores in SS1, SS2, and SSCE.

School	N	SS1	SS2	SSCE
A	200	0.20*	0.19	0.26*
B	100	0.24*	0.15	0.32*
C	15	0.28	0.21	0.31
D	90	0.17	0.36*	0.20
E	29	0.28	0.21	0.19
F	71	0.20	0.35*	0.44*

*significant ($p<0.05$)

The results showed that correlations were generally low and showed low predictive ability. The coefficient of determination (r^2) for schools A, B, and F which had significant correlations revealed that the variance of SSCE performance that can be accounted for by the SSCE results were 6.7%, 1.02% and 19.4% respectively.

Correlations between selected subjects in the JSCE and corresponding subjects in SS1, SS2 and SS3 were similarly low; they ranged from 0.093 to 0.333. A summary of the inter-subject performances is presented in Table 3.

Table 3: Summary of Inter-subject Performance in JSS and SSS

JSCE	N	Distinction	Credit	Pass	Fail
A	492	280(56.9)	170(34.6)	24(4.9)	179(3.5)
C	1618	489(30.3)	657(40.6)	288(17.8)	184(11.4)
P	819	65(7.9)	236(28.4)	290(34.9)	228(27.8)
F	101	12(11.9)	9(8.9)	20(19.8)	60(59.4)

The Table shows that 56.9% of the students who obtained 'A' in JSCE also obtained 'A' in SSCE, 70.8% of students who obtained C in JSCE also obtained C or better in SSCE, and 59.4% of students who obtained 'F' grade in JSCE had similar grade in the corresponding SSCE subjects. The study showed that overall performance in the JSCE of Osun State had a low capacity to predict performance in SSCE. A factor in the low validity of the JSCE could be the use of arbitrary and inflated Continuous Assessment (CA) scores as input to final results in school examination as reported by Adejumo and **Afolabi** (1990). The validity of the JSCE may be enhanced if a cluster of states (e.g. the geo-political zones in the country) prepare and conduct it as a regional examination.

A similar study by **Afolabi** and Lijoka (2005) on the validity of National Common Entrance Examination (NCEE) was conducted using 120 JSS3 students from Federal Government Colleges in three States in Southwestern Nigeria. The results showed that the aggregate scores of JSCE and NCEE had a significant correlation of $r = 0.50$, $p<0.05$, with the relationship between NCEE and JSS

results improving progressively from JSS1 to JSS3. The performance of students in the NCEE matched that in the JSS3 examination in all the selected schools. The study concluded that NCEE had high validity, and could be used to predict performance in JSS examination, especially the JSCE.

(c) Validity Evidence from Meta - Analysis

Meta – analysis refers to methods for combining and integrating results from a large number of studies. It is informed by the fact that the results of previous researches should be considered in evaluating the criterion – related validity evidence of any test or group of tests, and enables validity generalisation to be made regarding the similarity of validity coefficients across tests, jobs or settings. Adeyemo and **Afolabi** (2009) investigated the individual and overall effect sizes of 30 empirical studies conducted on the University Matriculation Examination (UME) in Nigeria using meta analysis. Empirical results reported in t, F and X^2 statistics were converted to product moment correlation following the transformation procedures by Kendal & Stuart (1967) and Rosenthal (1984). Fisher Z transformation was further applied to the r – values. The results of the study showed that the selected studies were not significantly different in terms of their probability levels ($\chi^2 = 2.68$; $p>.05$), but were significantly different in terms of their effect sizes 'r'. It was concluded that the heterogeneity of the effect sizes was a function of the sample size of the studies and that the UME demonstrated construct validity.

(d) Validity of Public Examinations

Public examinations are regional, national or state examinations, usually taken at the end of a specific course or educational level, and are external by design. By their nature, they tend to regulate educational opportunity and vocational choice (**Afolabi**, 1998). Public examinations in Nigeria include the Teachers Grade II Certificate Examination, conducted by NTI, Kaduna; Public Service Examinations, conducted by ASCON, Badagry; Junior School Certificate Examination (JSCE), conducted by the National

Examinations Council at the national level, and the Ministry of Education of each State at state level; the Senior School Certificate Examination (SSCE) conducted by WAEC and NECO; University Tertiary Matriculation Examination, (UTME) conducted by the Joint Admission and Matriculation Board (JAMB). Of these public examinations, two of the most prominent are the SSCE and the UTME. The procedures of test development and administration of the two examinations will be described, because of their standard and the contrast that they provide.

Development and Administration of the SSCE

The WAEC has been conducting public examinations in Nigeria since 1952 and included the Senior Secondary Certificate Examination (SSCE) to its portfolio in 1988. The test development procedure for this examination begins from the development of the syllabus and the provision of test specification tables in order to guide item writers. The latter specifies in detail the important topics that ought to be tested in each subject, as well as the weighting (i.e. the number of questions) to be assigned to each topic or aspects of the syllabus. Salami (1990) outlined the sequence of steps that are followed, the high points of which are listed as follows:

(i) the subject officer (an expert in a subject) commissions items from trained item writers, mostly classroom teachers, who are assigned specific topics and provided with necessary materials;

(ii) the subject officer edits the raw items and compiles them into groups for moderation;

(iii) the moderators (usually five or six classroom teachers and are experts in the subject) study the items critically and amend them where necessary to ensure clarity and standard. Items are moderated for two or more papers per subject at a time;

(iv) the subject officer and the chairman of the moderating committee select the final list of moderated items:

(v) The draft papers are then trial-tested to ensure that the items are appropriate (i.e. not too difficult or too easy) for the target population, using comparable students in randomly selected schools;

(vi) the computer analysis the trial tested items, and produces item statistics which are used to determine the items that go into the final question papers. Further improvements are normally made to the items at this stage. Any item that has been significantly changed or modified will be retrial-tested; and

(vii) the test development ends when the final draft of the paper is dispatched to the press by the subject officer, who is charged to ensure that none of the items is in any way compromised.

The administration procedure covers a wide range of activities, including pre-examination procedures, actual conduct of the examination, and post examination processes.

Pre-Examination Activities

(i) arrangement to print question papers;

(ii) arrangement in conjunction with State Ministries of Education to appoint supervisors, invigilators, and custodians;

(iii) recruitment of examiners and conduct of coordination meetings for oral and practical examinations; and

(iv) receipt, packaging and custody of question papers.

Conduct of the Examination

(i) issuance of question papers twice daily to examination centres;

(ii) conduct of the examinations in accordance with the time-table and laid down regulations;

(iii) inspection of centres while examinations are in progress; the inspectors are expected, among other things to:

- note the seating arrangement;
- check the identity of candidates with photo cards;
- investigate on the spot, cases of irregularities discovered or reported;
- find out the quality of supervision and invigilation and ascertain the security of examination materials; and

(iv) collection of scripts and relevant reports from supervisors

Post Examination Activities

(i) prompt evaluation of scripts from centres to WAEC offices;

(ii) coordination of examiners for marking, marking of scripts and checking of marking, followed by addition and transcription of marks by examiners;

(iii) retrieval of marked scripts and marksheets from examiners;

(iv) resolution of queries; e.g. missing marks, errors in names or subjects etc; and

(v) implementation of decisions of the final Award Committee on cases of irregularity.

Development and Administration of UTME

The Unified Tertiary Matriculation Examination (UTME) is conducted by the Joint Admission and Matriculation Board (JAMB). The JAMB was established in 1977, to have "the general control of the conduct of matriculation examinations for admission into all universities in Nigeria". Isemede et al (1990) illustrated the test development process for UME, which largely subsists for the UTME.

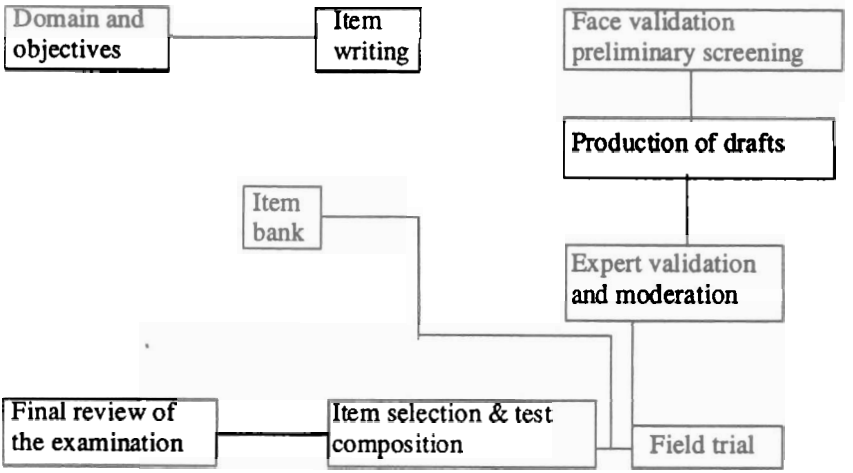


Fig. 1: Flow Chart of Test Development Process for UME.

At least, 10 seasoned subject specialists are appointed from universities and commissioned per subject to write 30 items each, except in Languages where 40 items are written by each item writer. The validation and moderation stage involves subject specialists, teachers, educationists, and experts in test construction, who review the items for content coverage, relevance and cognitive demand. In the field trial stage, the students who approximate those for whom the items are meant or a sub-sample of the target population are tested; and the item parameters from this trial are used to select the final items. Care is taken to avoid repetition of items in the previous years. The items are finally reviewed by a panel of subject officers of JAMB to ensure that all rules relating to testing are obeyed. Items are subjected again to moderation before any later use. At the production stage, the examination paper is taken through the type-setting, proof-reading, printing, packaging and delivery stages.

The UTME is a one-day examination, and it is conducted in all state capitals, and university towns, apart from some other carefully selected towns, including outside Nigeria. Each town has

a special centre and other centres with each centre accommodating a maximum of 400 candidates. The centres are housed in educational institutions from where the invigilators and centre coordinators are recruited. Supervisors who are changed annually are recruited from the universities where a university coordinator is also appointed for each year. Candidates carry identity cards containing their photographs. Usually, each invigilator has a maximum of 30 candidates assigned to him/her during the examination. In addition, five attendants are provided for each centre to ensure that invigilators remain in their respective rooms throughout the duration of the examination. The answer sheets are retrieved after the examination and then processed.

Ojerinde (2011) assessed the construct validity of the 2007 UME and determined the inter-correlations and patterns of relationships among the tests. The results showed that most of the subsets of science, social science, languages and arts groups had inter-correlations with four dominant factors.

4. Test Score Contaminants

Mr. Vice Chancellor sir, the goal of achievement tests is to obtain an examinee's best and highest level of performance. Apart from the true knowledge or proficiency that is being measured, many other factors may affect an examinee's performance in tests, especially objective tests. Thus, to evaluate test scores properly, one should not only be aware of the existence of extraneous factors, but also be able to control them; or make appropriate allowance for such factors in interpreting the result. Invalid test scores may be due to personal factors such as emotional disturbances influencing students' response to the test items or the test situation, rather than any shortcomings in the test instrument or its administration. Students may also be frightened by the test situation, and are thereby unable to respond normally. Some testees may not be motivated to put forth their best effort. These factors, called test score contaminants, may arise from the testee, the test or the testing environment.

(a) Response Changes

Response changing is generally defined with respect to objective test items. It occurs when a testee changes an option that has being previously selected as the correct or best answer, leading to the erasure or blotting out of the initial response. This change in response might be due to the recall of new information, blind guessing, a fresh perspective of the requirements of the item, cheating, anxiety or other rational factors (**Afolabi**, 1988).

When responses are changed, the changes could result in any one of (i) initial response correct, final response incorrect (C-I); initial response incorrect, final response correct (I-C), and initial response incorrect, final response incorrect (I-I). Where a testee makes more than one initial, change, and any of the responses made prior to a final incorrect responses is correct, it is regarded as (C-I). If all responses made prior to a final response are in error, they are considered as first responses, and regarded as (I-C) or (I-I). However, the latter scenarios are few to be of any consequence.

Many studies showed that it is profitable to change initial answers, yet a contrary impression persists among students and teachers. Very early studies (Matthew, 1929; Jarret, 1948; Belinky, 1972) found that the percentage of I-C changes ranged from 63 percent to 70 percent. Further, studies (Pascale, 1974; Stoffer et al, 1977) showed that students profit by changing their initial answers, and that the gains are not restricted to a few individuals but occur in the majority of respondents who changed answers. Furthermore, the more changes that are made, the larger the gain. This evidence is obtainable from the ratio of the number of answer changes made to the number of items in the test.

In a study of undergraduate students of this university, **Afolabi** (1988) found that 70 percent of the 200 students who took the MC test changed their initial responses and 84 percent changed initial responses in the TF test. Fifty percent of changes made in the MC test resulted in I-C. while 65.8 percent of the changes made in the TF test were from I-C, and of the changes that provided an incorrect final response, the first response was often correct as

incorrect. This further established the fact that response changing behavior increases students test scores, and that profit from changes is a function of the number of changes made.

Also, **Afolabi** (1990) investigated some variables in relation to response changes. These included the personality characteristics of extraversion, need achievement, risk taking, general self concept, anxiety, impulsiveness, state anxiety, academic self concept and trait anxiety. Among the variables, impulsiveness ($r = 0.360$) and extraversion ($r = 0.359$) were found to be significantly related to response changes in MC tests, while academic self concept ($r = 0.295$) and trait anxiety ($r = 0.310$) were found to be significantly related to response changes in TF test. A stepwise regression analysis showed that all the independent variables explained only 17.8 percent of the variance of response changing behavior in MC test items, while impulsiveness, extraversion and risk taking alone accounted for 14.8 percent of this variance; and had significant values of $F = 6.41$, $F = 10.93$, and $F = 2.79$ respectively ($P < .05$)

Furthermore, **Afolabi** (1992) examined response changing behavior in four objective test formats of MC, TF, matching and completion. The results showed that responses were most frequently changed in TF items and least changed in MC. The validity coefficients of each of the four test formats were consistently lower when the tests were scored without excluding changed items than when changed items were excluded. Fisher Z test failed to show significant difference when the validity coefficients were compared pairwise. However, the reliability of TF items with response changes ($r = 0.82$) and without response changes ($r = 0.63$) was significantly different ($Z = 3.55$, $p<0.05$)

(b) Testwiseness

The efficacy of objective tests may also be vitiated by testwiseness. Millman, Bishop and Ebel (1965) and Ford (1973) in their pioneering studies on testwiseness, developed a classic theoretical framework and provided an operational definition of

the construct as the ability of a test - taker to utilize the characteristics and format of a test or the testing situation to improve his/her test scores. Testwiseness is logically independent of an examinee's knowledge of the subject matter which the items are designed to measure.

Afolabi and Eso–Olawale (1999) demonstrated that students can benefit from training in testwiseness skills. One hundred and twenty students who were ascertained to be non–testwise, were randomly assigned to experimental and control groups. Both groups completed a Student Testwiseness Scale (STS) (developed by the authors) and an achievement test battery in three objective test formats. After exposure to a Testwisenes Training Package (TTP), the post–treatment score of the experimental group ($\overline{X}$ = 154.6) was found to be significantly greater than that of the control group, ($\overline{X}$ = 149.2), $t = 6.4$, $p < .05$. The study showed that students can generally be trained to be testwise and that they can have comparable gains from training in testwiseness skills notwithstanding ability or gender differences. This is perhaps because testwiseness is more a function of the test than of the testee. The study also showed that testwiseness accounted for 14.76% of the variance in test scores ($R= 0.397$, $p < 0.05$). Testwiseness principles include similar alternatives, stem cue, time using strategy, umbrella term, content information, grammatical consistency, absurd option, intent consideration and guessing cue. If trainee and practicing teachers are sufficiently familiar with these principles, it is likely to improve the quality of teacher–made tests; and if they are adequately prepared to teach the principles in schools, students will be in good stead to be more confident in examinations, and would be less likely to resort to cheating or other unwholesome behaviour during examinations.

(c) Response Set

Response set is a habit or temporary disposition that makes a person to respond to test items in a particular manner or pattern. An important type of response set is response bias, in which an

individual selects one response position significantly more often than others regardless of item content. In TF tests, test-takers have a greater tendency to choose the "true" option, and are predisposed to select the neutral category in inventory or questionnaire options that have agree–disagree continuum. In multiple–choice tests, test-takers have the tendency to favour certain positions than others. This may arise from failure to read all options, often resulting in response bias for early options Preference of test takers for first options is referred to as primary effect while preference for the last option is known as recency effect. **Afolabi** & Ibrahim (2009) reported a significant tendency of test takers to select early options. This is not unconnected with the fact that the first (and last) choices in MC test stand out than the middle two or three options.

The positions to which the correct options are assigned influence the difficulty level of MC test. Item writers tend to place the key in a middle position as much 3 or 4 times as often as at an edge position. The penultimate position has been found to be the most difficult in a 4 or 5 – MC, while items with keys in the second and third positions are of equal intermediate difficulty (**Afolabi** and Ibrahim, 2009).

Although, a preponderance of studies suggest that position preference does not constitute a significant source of invalidity in multiple–choice tests (Whiteker et al, 2002), a way to reduce the variance in test scores attributable to response bias is to randomly assign the correct choices and distracters to the options such that each position has an equal number of correct choices

(d) Guessing

In objective tests, testees guess when they have no knowledge or have partial knowledge of the correct answer to an item. Guessing can be random (blind) or informed (educated). In the former, the testee does not have any knowledge about the subject matter in the process of guessing, and the guessing is among all the options; but in informed guessing, the testee is able to eliminate one or more

wrong options prior to guessing; hence guessing, in this respect, is from the remaining options which presumably include the correct option. In objective tests, it is traditional to award one mark for each correct answer (R) such that an examinee's score(S) is represented by the Number Rights.

$$S = R \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

However, where the examinee is ignorant of the answer and guesses randomly, the expected number of right answers on Q MC items with n choices per item is

$$S = \frac{Q}{n} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

This simple scoring formula leads to upward bias in scores, particularly for students of low ability. To address it, the conventional corrected score formula is used. It is given by

$$S = R - \frac{W}{K-1} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

where W is the number of incorrect responses and k is the number of options per item. This, too, has been found to over–correct. Apart from this, its basic assumption is simplistic. Items cannot be compartmentalized into two groups of those the testee perfectly knows and those he does not know at all. There are items, that may be fairly well known, and those on which there is hazy knowledge. Sometimes, too, the formula yields negative corrected scores for low ability students. **Boyinbode** (1982) reviewed analytical and empirical approaches to correcting for guessing in TF tests, including paired item scoring, and concluded that these had limitations of usability by teachers and of misranking of students.

Since it is generally agreed among psychometricians that partial knowledge deserves partial scores, confidence scoring was developed to address the question of partial knowledge such that testees with absolute knowledge will earn maximum scores and random guessing will yield number-rights score. **Boyinbode**

(1986) applied this procedure to TF test. In addition to indicating whether the statement was true (T) or false (F), each testee was also required to circle one of the number 1, 2, 3 following each item to express the level of confidence with which the item was answered. Number 1 was to be circled when the answer to the item was made with absolute confidence; number 2 when the answer was made with little confidence (indicative of partial knowledge) and number 3 when the item was marked on the basis of random guessing. The answer sheets were then scored thrice: with all the test answers (C), with the exclusion of the answers made by random guessing (B), and with only the answers made with absolute confidence (A). Coefficient Alpha and Pearson r were used to compute the reliability and validity indices; for the latter, CGPA was used as criterion. The results are presented in Table 4

Table 4: Effects of three scoring procedures on reliability and validity

Basis of Answer	Rights	Wrongs	Omits	Reliability	Validity
Complete confidence (A)	4580	3480	1540	0.843	0.436
Complete confidence and partial knowledge (B)	5580	3600	420	0.673	0.490
Complete confidence, partial knowledge and random guessing (C)	5960	3460	180	628	462

The reliability of the TF test was highest when guessing of any form was disallowed and lowest when items were scored on the basis of random guessing. That is, the internal consistency reliability increased with a decrease in random guessing. However, scores were most valid when only the items answered by random guessing were excluded and least valid when answers were made only on the basis of absolute confidence. Furthermore, confidence scoring had no significant effect on item difficulty ($F = 608$, $p>0.05$) and discrimination ($F = 1.47$, $p>0.05$). This study showed that variance due to partial knowledge should be positively

weighted in predictor scores while random guessing may be ignored rather than negatively weighted.

Afolabi (2000) extended the application of confidence scoring to three variants of MC tests namely 3-option, 4–option and 5–option items in terms of reliability and validity. The test formats were developed over the same Mathematics course content for SS2 students, had the same stems and keys, but differed only in the number of distracters. K–R21 was used to estimate reliability while scores of the terminal mathematics test were used as criterion. Results showed that guessing had the greatest effect on 3–MC. Whereas 79.8% of the gain in total scores without absolute confidence in 5. MC was attributable to partial knowledge against 20.2 % to random guessing the corresponding proportion for 3–MC were 60.1 % and 29.9% respectively. Clearly, confidence scoring improved the reliability and validity of 4 and 5–MC, indicative of that as the number of foils increases, the value and contribution of partial knowledge also increase. However, if testees will be instructed against guessing, 3–MC appears the most suitable for classroom use, apart from its added advantage of requiring relatively fewer distracters per item.

In furtherance of minimizing the variance of test scores on account of guessing, **Afolabi** (1999) utilized another scoring procedure, a priori method of choice weighting, otherwise known as logical–choice weighting (LCW). In this method, differential weights are assigned to item options according to a priori assessments of their degree of correctness by expert judgment. The keyed choice is always assigned a weight of one. Thus, the LCW equates the number-right score plus an amount equal to the sum of the choice weights associated with the incorrect choices selected by testees. In the study, 64 part four undergraduate students of this university, participated. It examined which of Number Rights (NR) conventional (CS) and logical–choice scoring procedures would yield the most valid measure of students' cognitive ability in a 5 – MC test. The logical weights of distracters ranged from 0 to 0.8,

depending on the degree to which they are related to the key LCW had the highest validity coefficient (r = 0.744) while those of NR and CS were r = 0.411 and r = 0.535 respectively. Similarly, LCW had the highest reliability (r = 0.816) while those of CS and NR were r = 0.677 and r = 0.440 respectively. Fisher Z test revealed significant differences between pairs of the coefficients ($p<0.05$)

Afolabi and Adewolu (1999) investigated the relationship between sex, academic ability and guessing behavior. The results of the study showed that the tendency to guess increased with item option length in MC tests ($X_{3\text{-MC}} = 4.10$; $X_{4\text{-MC}} = 5.04$; $X_{5\text{-MC}} = 6.91$). The mean values were significantly different ($F = 11.70$, $p<0.05$). Female undergraduate students showed a greater propensity to guess than male students. The academic ability of students had significant effect on guessing tendency; students with higher ability probably had a greater need for achievement and hence took the greater risk

Test Administration Lapses

The principles of test administration are meant to ensure that students have a level playing ground during testing, write tests in a conducive environment and to prevent all sorts of distractions and other incidents that could adversely affect test scores. Test administration involves pre-testing, testing and post – testing phases. First, is the preparation of the test materials in the pretesting phase. Then, a suitable location for the test is identified and secured. Its suitability implies that it has adequate working space, seating facilities, ventilation and lighting. The test venue must be quiet and free from undue noise and distractions. Uncomfortable physical condition in testing has deleterious effects on test performance especially for young children (Trenthan, 1975; Anastasi, 1991). In fact Traxler and Hilket (1962) reported, in an experimental study, that students who used chairs and desks during examination had greater advantage in test scores than those who used chairs with desk arms. The pretesting process also includes advertisement of the rules and regulations guiding the examination

and arranging for adequate number of examiners and invigilators, with the responsibilities of each category duly spelt out. In all tests, especially classroom tests, students must know in advance that a test will take place. Ideally, the schedule of tests to be administered during a term or semester should be disclosed to students at the onset. Tests should not be held without prior notice either to make students "fall in line", as punishment or to take attendance. The so–called element of surprise in testing heightens anxiety and debilitates performance.

The second phase of test administration is when the test materials are directly given to the candidates at an official venue. In this regard, **Afolabi** (2005) developed a set of guidelines, which are a desiderata in the promotion of test score validity.

1. In tests conducted outside classrooms, the examiners and invigilators should arrive at the test venue early enough before the commencement of testing. This should be at least 30 minutes, in order to confirm the suitability of the test venue and or be able to make other incidental arrangements.
2. It may be necessary to post an indicator conspicuously around the test venue that a test is taking place; e.g " TESTING IS IN PROGRESS"
3. It may be helpful to arrange the testees in a special way to curb cheating and to make invigilation effective and seamless. The seats may be pre–numbered to correspond with the registration or identification number of the candidate.
4. Test materials, especially questions and answer papers should be distributed to students at about the same time, so that no student has undue advantage over another.
5. Important instruction(s) or information should be clearly and audibly pointed out before testing commences, not during testing. Where it is inevitable, such information should be communicated without distracting the candidates.

6. The starting and stopping time should be conspicuously indicated at the commencement of testing for the attention and guidance of all candidates.
7. Interruptions during testing should be avoided as these could affect the flow of candidates' thoughts, and hence their test scores. As in the case of football matches, the estimated sum of unavoidable interruptions should be added to the scheduled duration of the test.
8. Movement by invigilators during testing should be done with decorum and sensitivity. It should be done gently, quietly and gracefully, with respectable personal carriage.
9. Where a candidate appears to be misbehaving or attempts to communicate with another candidate, the invigilator should unobtrusively walk to the candidate in question and give a caution. It is not appropriate for invigilators to be shooting at candidates from a distance, thus distracting other candidates.
10. Care must be taken not to create any form of anxiety or tension during testing. The invigilator should maintain a cordial but official relationship with candidates, and must not harass or intimidate them. He should not exhibit any bossy attitude or patronizing conduct; rather, the invigilator should be friendly but firm. A long time ago, Wickes (1956) reported that the general manner and behavior of the examiner, as exemplified by smilling, nodding, etc. were found to have a significant effect on test results.
11. Invigilators should refrain from providing unnecessary information to candidates. Candidates who request for clarifications should be directed to the test instructions. Questions that candidates might ask should be anticipated and provided for. The test items should be free from error, just as the test instructions should be free from ambiguity. The idea is that no hints or academic assistance should be inadvertently given to candidates.
12. Candidates who wish to enter the test venue later than the time permitted by relevant regulations or leave the test

venue earlier than the time permissible should be politely informed what the regulations allow.

13. Examination regulations do not permit an invigilator or even an examiner to prevent a candidate, who is suspected to engage in malpractice, from writing a prescribed test; neither should any inhibition or obstacle be deliberately placed on his/her way. It is incivility for an invigilator to seize a candidate's question paper or answer booklet, let alone damaging or tearing it. Such behavior shows lack of academic maturity, presence of unstable emotionality and intolerance. Invigilators should not threaten candidates in any way or exhibit terrorist tendencies. All that is necessary is to record every significant event that occurs during testing, including suspected acts of cheating or examination malpractice. This may be the completion of appropriate forms by both the invigilator and the candidate(s) suspected of infringements. If, on the contrary, a candidate is prevented from completing a test because of alleged misdemeanor, and if after due investigation, the candidate is exonerated, then will the invigilator not be liable to a charge of injustice and highhandedness? How will the candidate's loss be adequately restored?

14. If the test has subtests that have to be collected during testing (e.g Section 1, Part A or Paper 1), adequate instructions on this should be given at the beginning, and the collection should be done without disturbing the serenity of the testing environment. Furthermore, while it is helpful to alert candidates that testing time will soon be over, this should be done rather infrequently and without creating unwarranted nervousness. The time remaining, for example, can be indicated on the board at regular time intervals.

15. When the time allowed for testing is over, test materials should be collected immediately and in an orderly manner, counted, checked and properly packed in prescribed envelopes. All test materials must be accounted for. There

are candidates who are marked present for a test but who will not submit the answer booklet/sheets for lack of confidence or who simply want to play pranks and come for the test score later on. Proper documentation of examination materials will take care of such pranks. Also, once the time allowed for testing is over, it is not right for an invigilator, using his own initiative, to give time than is allowed in the test instruction (because candidates are yet to finish) or to reduce the time allowed (because almost everyone has finished).

16. It is inappropriate for invigilators, during testing, to
    - engage in private conversation;
    - write a letter;
    - grade a script;
    - read a novel;
    - complete a report;
    - leave the candidates unattended

Invigilation involves more than physical presence in the examination room. The invigilator should demonstrate commitment, discipline, skill, and presence of mind on duty. He/she should not just sit down somewhere, carried away in thought, imagination, or with any other preoccupation. Rather, he should be alert, move quietly but purposefully round the testing room, turning unpredictably at intervals, with eyes spanning the whole room or a large portion of it at any time. Invigilators should be conscious of candidates' non–verbal (body) language during testing and be familiar with students' methods of cheating. Invigilation has its own methodology. Therefore, invigilators need proper training including techniques for the establishment of rapport, methods of relieving candidates' anxiety, and of detecting cheating.

(f) Examination Malpractice

Examination malpractice is a form of academic dishonesty. It is an act of deception, fraud, trickering, imposture or imposition. It is academic cheating, using unethical and immoral ways to obtain

undeserved scores, and gain undue advantage over others. "Examination malpractice is a deliberate act of wrongdoing contrary to official examination rules with the intention of placing a candidate at an unfair advantage" (World Bank, 2001). It can also be described as irregular, illegal and unethical behavior exhibited before, during or after a test or examination, which infringes on the operational rules and regulations of such tests or examinations. Any activity of a student or group of students with the purpose of giving any of them higher grades than they would be likely to receive on the basis of their own achievements is cheating (Ebel, 1979).

The implications of examination malpractice have been articulated by **Afolabi** (2012). According to him, examination malpractice suppresses the creative power in students' mind, and makes them loose the desire for mastery and excellence. The moral compass that students need to guide personal conduct in both school and society can be thrown off-course. If imbibed during the teenage years, cheating can become a lifelong habit. It will create gaps in knowledge and skills that can adversely impact later success when the foundation of knowledge necessary to understand processes in higher-level courses has not been acquired; and this can lead to frustration and criminal tendencies. The incidence of examination malpractice is a threat to the reliability and validity of examinations as the practitioners might obtain unearned and undeserved marks or grades, leading to a mismatch between ability or competence and position or ranking. Examination malpractice undermines the credibility of examinations and the integrity of the results or certificates awarded. Ultimately, its influence extends to the workplace, national institutions, public life and the professions. It can influence admission to higher institutions or appointment to jobs, benefiting dishonest students at the expense of their honest peers. Beneficiaries will be unable to actualize the promise of the examination results or certificates they hold, engendering system failure, loss and wastage. Socially, persons who cheat to pass and who are unable to progress or secure admission into higher

institutions or appointments in the workplace will be a disappointment to their parents, and might constitute a danger to social order and harmony. The future of the country in terms of manpower development and planning will also be jeopardized. If the educational system breaks down under the burden of examination malpractice, the full implications and consequences will be hard to imagine. The problem of examination malpractice, therefore, cannot be ignored.

Prevalence

Cheating by students is rampant across the continents. Surveys among higher education students in China reported that 83 percent of the 900 students that were surveyed cheated (The Epoch Times, 2006). Widespread cheating among middle school, high school, and college students also has been reported in Australia, England, India, Japan, Korea, Spain and Scotland (Callaham 2004). In the US, a nationwide survey of 36,000 high school students found that 60 percent admitted to cheating on tests and assignments (Josephson Institute of Ethics, 2006). Students who cheat were not only those characterized by marginal abilities causing them to do so in order to keep pace with more intelligent classmates; even among students categorized as high achievers in the US, 80 percent acknowledged cheating on teacher-made and state tests (Lathrop and Foss, 2005).

Davis and Ludvigson (1995) also reported that between 40 and 60 percent of undergraduate students admitted to having cheated in at least one examination. Students from all age groups and achievement levels participate in cheating, however, female students tend to cheat more than male students, and non-honours students tend to cheat more than honours students (Rittman, 1996, Strom and Strom, 2006). In Nigeria, the first publicly reported case of examination malpractice occurred in 1914 when there was a leakage of question papers in the Senior Cambridge local examination (Alutu and Aluede, 2006). Ever since, cases of irregularities have been a regular phenomenon.

Forms of Malpractice

The major forms of malpractice are impersonation or contract (writing a test on behalf of someone else for favour or bribe), bringing foreign materials, substituting worked scripts and collusion (sharing of information or copying in the examination room). Others are swapping of scripts, use of covert notes or crib sheets, leakage (obtaining the questions or answers to an examination ahead of time), and copying (reproducing another candidates' work with or without permission). Lateness to the examination hall, leaving the hall during the examination without permission, absconding with answer scripts and physical or verbal assault on invigilators or supervisors similarly constitute malpractice offences. There are also organized or mass cheating involving assistance from invigilators, supervisors or security agents and mercenary students who hide within the vicinity of the examination hall. The sophistication of examination malpractice has led to the use of cell phones or even advanced electronic devices like electronic pens which can store data.

Ojerinde (2004) identified pre-examination malpractice at the Senior School Certificate Examination (SSCE) conducted by NECO to include registration of non-school candidates, registration of candidates too many for available physical facilities, and registration to allow for impersonation. Some teachers are involved in inappropriate monitoring of examinations. This includes leaving the examination room or permitting candidates to do so during an examination without surveillance, both of which increase the chances of malpractice. Johnson (2003) was of the view that no student should be out of a teacher's sight while taking a test. In fact, some teachers or invigilators indulge in reading private materials or discussing with colleagues during examinations. There are others that stay aloof at one edge of the examination room, thus having the view of only a portion of the test takers.

In an interview, the Minister of Education in Nigeria, Prof. Rukayyatu Ahmed Rufai, commenting on the 2010 NECO examination, said inter alia, that government was also concerned about reports of teachers compromising the integrity of internal examinations, and sometimes aiding or even abetting in the forgery of certificate for pupils Similarly, the JAMB Registrar/Chief Executive, Prof. Dibu Ojerinde, in an interview with the Punch on the conduct of the 2010 UTME, stated that "candidates in some examination centres, beat up invigilators or supervisors (intimidation) to enable them perpetuate all forms of malpractice". He continued, "there were also reported cases of collusion between invigilators to allow malpractice; some parents even brought their children into the examination hall, collected information and pass it on to their children during the examination". Thus, although students are the main culprits and beneficiary of examination malpractice, they have collaborators in all manner of people ranging from item writers, computer operators, printers, custodians, examination bodies or agencies, teachers, principals, supervisors, invigilators, parents, guardians and even security personnel hired to secure examination halls!

Reasons for Cheating

But why do students cheat? Decisions about examination malpractice by students are clearly influenced by societal and school norms, as well as the attitudes of teachers, and most importantly, friends. Social approval from parents, teachers or friends is, perhaps the most important factor in cheating. Students with a high need for social approval and are certain of receiving same are likely to cheat more often because of their concern about negative evaluation if they fail. Theories on deviance have indicated that the threat of being caught and punished might act as a deterrent, yet research suggests that many students feel that they are not likely to get caught or that they will be left off the hook even if they are caught (McCabe and Trevino, 1993, 1997). Students belief that they will not get into trouble tends to embolden them. They are aware that examination malpractice seldom

produces punishment and therefore, poses a low risk. Lathrop and Foss (2005) reported that 95 percent of students who acknowledged they had cheated said that they were never caught and even considered themselves as morally responsible individuals.

When lecturers suspect or detect malpractice incident on an individual, on discretional basis the lecturer is more likely to be lenient when he is aware that reporting the student that is engaged in malpractice may lead to the student's suspension or expulsion. Many are also unwilling to go through the bureaucratic processes of panel investigation. Some teachers worry that they may erroneously accuse a student of cheating and then have to suffer unpalatable and dreadful consequences, especially in countries where parents can institute lawsuits when their children are accused of cheating, or in developing countries where the alleged student can instigate gang or criminal attack against the teacher.

It is not only teachers that are reluctant to report erring students; students themselves turn the other eye when their peers cheat during tests. A student said: "if you report, you will make yourself look worse in the eyes of other students than the person who cheated". Another student said: "everybody starts looking at you like, why did you have to go and tell? Why don't you mind your business". It appears cheating does not really weigh heavily on the conscience of most students!

Other reasons for students' involvement in examination malpractice include:

* Desire for high grades
* Increasing emphasis on test scores and grades
* Stiff competition for admission, scholarship or job opportunities
* Pressure for grades and certificates from parents, relations and the school

* Others are cheating and getting grades easily with much less effort
* Some students who are high performers aid cheating because they want to help their friends
* Lack of confidence in own ability
* Laziness and lack of planning leading to inadequate preparation
* Students also cheat when they feel that a particular course is of no use to their career or programme
* Large and crowded class or examination hall with insufficient invigilators
* When students receive little or no teaching, they are vulnerable and feel justified to cheat.
* Declining ethical standards
* Desperation; the more extreme the desperation, the more ambitious and serious the attempt to cheat is likely to be

In a study, **Afolabi** (2010) examined the ability of Didactic Ethical Therapy (DET) to influence the attitude of students towards examination malpractice; this was based on the premise that attitudes affect behavior. An instrument, Students' Attitude Toward Examination Malpractice (SATEM) was developed to elicit responses on core dimensions of examination malpractice. The SATEM was administered on an initial group of 300 students, from which 60 students, who scored highly on the construct (indicative of very favourable attitude towards examination malpractice), were purposively selected to constitute the study sample. They were randomly assigned to experimental and control groups. The experimental group was exposed to an interactive strategy, DET, which explored the various aspects of examination malpractice, its meaning, scope, predisposing factors, methods, consequences and implications. The control group was exposed to a placebo. After the re-administration of SATEM to both groups, the results showed that the experimental group had a lower mean score ($\overline{X}$ = 17.1) after than before the experiment ($\overline{X}$ = 25.4). The

corresponding scores for the control group were $\overline{X}$ = 23.6 and $\overline{X}$ = 24.9 respectively. This revealed a significant change in the attitude of the experimental group towards examination malpractice ($t = 2.71$, $p<0.05$). This study shows a window of opportunity for school counselors and administration to address the problem of examination malpractice through systematic involvement in ethical dialogue with students.

(g) Scoring

No matter how well prepared a test is, if it is poorly scored, no reliable and valid inferences may be made from it. In fact, in the case of objective tests, it is the consistent scores that it yields irrespective of how many people do the scoring, that make it objective. In essay tests, scoring the answers has been its greatest alleged weakness. How do we ensure that improper grading practices do not ruin the beauty of a well prepared essay question? Two factors are important; the scorer and the scoring method employed. Unlike in the scoring of objective items, the scorer of essay questions must have adequate knowledge of the subject matter. Since it is an unrestricted response format, he should be able to know where the answer is merely tangential and where it is totally off the mark. It is possible for a student to come up with points or arguments that are somewhat different from those of his peers, but which are nonetheless relevant, and even superior. So, if the scorer is not very competent, he may dismiss such response as wrong, simply because it does not tally exactly with the pattern of responses of other students or with the model answer that the scorer has provided.

The second issue is the scoring method Mehrens and Lehmann (1978) suggested that in grading essay tests, one must (i) use appropriate methods to minimize biases, (ii) pay attention only to the significant and relevant aspects of the answer, (iii) be careful not to let personal idiosyncrasies affect the grading, and (iv) apply uniform standards to all the papers. Of these suggestions, by far, the most crucial is the uniformity of scoring standards, as this is

the greatest determinant of test reliability. Without uniformity, validity of students' performance will not be meaningful. But, how is uniformity of scoring of essay answers to be achieved?

There are two conventional methods of scoring essay answers, the analytical method and the global method. The method to be used depends on the type of essay item, the number of scripts to be graded, the time available and the purpose of the test. Usually the analytical method is recommended for the restricted – response items and when the number of essays is not very large.

In this method, a model answer is prepared and divided into specific points and component parts by the scorer to reflect the points expected to be seen in the answers provided by the students. The points are generally based on the content domain covered during instruction. As Coffman (1971) remarked, such procedure does not allow extraneous considerations such as hand writing to affect the assignment of scores.

The process of preparing the model answer (before test administration) enables the teacher to have a more concrete and realistic feel of the demands of the questions, especially in respect of wording, time limit, difficulty and or complexity. Further, by accounting for every point that the student makes with the point method, each answer is compared with the ideal answer in the scoring key, and a given number of credit points is assigned to the answer in terms of assigning appropriate weights depending on its relevance, appropriateness and adequacy. The analytical method tends to reduce the influence of subjective impression and reduces inter or scorer variability; thus, reliability is enhanced. However, it is very laborious and time consuming.

Global scoring

Global scoring, also called holistic or rating method, like the analytic method, involves preparing an ideal or model answer; but this is not divided into specific points and component parts. The

answer papers are then perused, and from these, those that correspond to particular standards are identified and selected, and are used as anchor points in assigning other papers to the identified categories of standard. As the answers are read, each paper is placed in one of a number of piles. These piles represent degrees of quality and determine the weight assigned to each answer. The crux of this method is to select papers that serve as anchor points that vary in quality, and to train readers to go rapidly through a response, and give a global impression of the quality of the response. In rating the responses, any number of rating categories may be established, from two to ten.

Hopkins (1998) provided a suitable guideline for the use of anchor points in scoring essay tests. These include

1. before scoring, compare the marking scheme with some actual responses
2. score one question at a time for all papers;
3. ensure uniformity in scoring by adhering strictly to the scoring criteria, this will reduce halo – horn effect.
4. responses should be scored anonymously; that is papers should be identified by numbers rather than by names;
5. criteria for scoring should focus on the substance of the subject (or course), not on mechanics of language; and
6. responses to any one question should be scored without undue interruption

Award of Zero Scores

The award of zero score to a testee as a result of incorrect response to objective items is defensible, but not so in essay items, requiring free response and many dimensions to an item. **Afolabi** (1998) argued that the absence of a true zero in educational measurement makes the interpretation of a zero score ontologically inconsistent. If the marking that gives rise to a zero score derives from the criteria of relevance and appropriateness, the award of a zero score to responses that fall within a continuum of permutations of these criteria would be non-discriminating and non-inconclusive. It may

also boil down to the interrogative quality of the questions asked. The award of zero aggregate score to students in essay tests, therefore, casts doubts on the validity, not only of the items but also on the scoring system.

5. Validation of Locally Developed Tests

There are hundreds of test instruments that measure various traits around the world; yet there is always the need to be conscious not only of the validity of those instruments for use in a particular socio–cultural setting, but also the issues of usability (accessibility, availability, timeliness, cost, administration and scoring). Cognizance must also be taken of our social, economic, cultural and education milieu in terms of the nature and wording of foreign instruments. Furthermore, there is the need for capacity building with a view to developing testing skills and entrepreneurship among younger scholars. In this regard, we have developed, mostly with our students, a couple of psychological tests based on constructs related or tangential to Education. These are apart from foreign test instruments that have been adapted for use in Nigeria. Ten of the psychological tests that have been locally developed following psychometric best practices, and the collaborating students are:

(i) Counsellor Effectiveness Scale
Shaba, A.A. (2000)

(ii) Patient Satisfaction Scale
Afolabi, M.O. (2005)

(iii) Teacher Self-Efficacy Scale
Faleye, B.A. (2006)

(iv) Nurses Effectiveness Scale
Jemilugba, M.O. (2006)

(v) Mathematics Teachers' Competence Scale
Akinrogunde, F.I. (2009)

(vi) Teachers' Need Assessment Scale
Agboola, O.O. (2010)

(vii) Bankers' Self–Efficacy Scale
Odeyemi, A.A. (2011)

(viii) Teachers' Job Satisfaction Scale
Edogboh, K.L. (2011)

(ix) Science Teaching Competence Scale
Ajisegiri, R.O. (2012)

(x) Test-Taking Motivation Scale
Oyebamiji, R.O. (2012)

Necessary supportive information will be provided to some of these tests in order to upgrade them to commercial status for national and international use by researchers, employers, institutions, and practitioners.

Afolabi and Oyebamiji (2012) demonstrated the process of developing and validating one of the instruments, the Test-Taking Motivation Scale.

Definitions

(a) Motivation is an internally generated or externally stimulated move to do something. It gives rise to a drive or energy that compels an individual to engage in tasks or activities that are directed towards achieving set goals and objectives and is a fundamental aspect of learning.

(b). Test-taking motivation is the willingness to engage in working on test items and to exert effort and endurance in doing so. It is the extent to which testees give their best efforts to the test, with the goal of being able to adequately represent what they know and can complete in the test.

Sample

A sample of 600 students was selected, which spanned across 20 senior secondary schools from five local government areas of Osun State, using multi-stage sampling technique.

Generation of Items

A review of the literature identified seven dimensions of test-taking motivation, namely: peer influence; teachers, parents and society influence; test stakes; mastery of subject; performance expectancies; test/course characteristics; and testing environment.

First, a pool of 50 items was generated from the literature based on the aforementioned themes. Following expert judgment and review, three items were considered inappropriate and therefore deleted. The items were distributed over the dimensions earlier identified, ranging from 4 to 9 items and administered on the study sample, with a Likert–type response format. Using SPSS, the appropriateness of the items for measuring test – taking motivation was determined using the following rules:

1. Items with mean values less than the scale mean (MLSM) of 3.871 were deleted;

2. Items with low item-total correlation (LITC) that were below item-total correction (ITC) mean of 0.143 were deleted;

3. Items with Cronbach alpha-if-item deleted higher than the scale's Cronbach alpha of 0.867 were deleted; and

4. Subscale reliability analysis was conducted; items with low (less than 0.3 coefficients) and negative corrected item-total correlation coefficients were also deleted.

   In addition, to the item mean and standard deviation statistics, the item-total correlation of the scale was computed in order to make item reduction decisions. The

evaluation of the corrected item-total correlation and Cronbach's alpha if-item-deleted is presented in Table 5.

Table 5: Item-Total Statistics of TTMS

Item No	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Total Corre-lation	Squared Multiple Corre-lation	Cronbach's Alpha if Item Deleted
1.	178	428	0.373	0.337	0.864
2.	178	436	0.197	0.303	0.867
3.	178	431	0.295	0.218	0.865
4.	179	453	-0.085	0.166	0.874
5.	177	427	0.497	0.330	0.862
6.	177	426	0.551	0.406	0.861
7.	177	424	0.503	0.385	0.861
8.	178	427	0.390	0.282	0.863
9.	177	427	0.528	0.409	0.862
10.	178	429	0.410	0.309	0.863
11.	177	426	0.448	0.400	0.862
12.	177	427	0.483	0.330	0.862
13.	178	425	0.477	0.464	0.862
14.	177	425	0.506	0.377	0.861
15.	178	426	0.450	0.444	0.862

16.	177	425	0.512	0.446	0.861
17.	177	426	0.576	0.452	0.861
18.	177	426	0.533	0.470	0.861
19.	178	442	0.085	0.259	0.870
20.	177	425	0.530	0.429	0.870
21.	178	450	-0.034	0.321	0.872
22.	177	426	0.493	0.483	0.862
23.	178	447	0.003	0.338	0.872
24.	179	472	-0.449	0.388	0.878
25.	177	426	0.539	0.401	0.861
26.	177	426	0.539	0.401	0.861
27.	177	431	0.389	0.328	0.863
28.	177	425	0.500	0.425	0.861
29.	178	440	0.123	0.411	0.869
30.	178	430	0.363	0.320	0.864
31.	178	442	0.093	0.395	0.870
32.	177	430	0.426	0.314	0.863
33.	177	429	0.527	0.445	0.863
34.	177	428	0.494	0.370	0.862
35.	179	461	-0.232	0.294	0.875

36.	177	428	0.488	0.381	0.862
37.	178	436	0.230	0.255	0.866
38.	177	430	0.413	0.304	0.862
39.	178	430	0.365	0.354	0.864
40.	177	427	0.466	0.364	0.862
41.	177	426	0.507	0.407	0.862
42.	177	427	0.497	0.345	0.862
43.	178	434	0.260	0.295	0.866
44.	178	428	0.354	0.295	0.864
45.	177	427	0.399	0.269	0.863
46.	178	440	0.111	0.273	0.869
47.	177	429	0.514	0.426	0.862

Table 5 shows the item-total correlation of the items. Item 24 had the highest Cronbach-alpha-if item deleted value of 0.878, followed by item 35 with 0.875, and item 4, with 0.874. These values depict the relatedness of the items to the whole scale. The items identified by these criteria were items 4, 19, 21, 23, 24, 29, 31, 35, and 46. Thirteen additional items, apart from the nine items earlier identified, were deleted after subscale reliabilities were conducted. Thus, the final scale had items, which were considered to have relatively superior psychometric qualities.

Construct Validity

The construct validity of the scale was determined using two methods. The first was Kaiser or eigenvalues greater-than-one

criterion (KI), and the second method was the scree test, which involves an examination of a plot of the eigenvalues for breaks or discontinuities. The data was subjected to KMO test of sampling adequacy, which yielded a value of 0.930, indicative that the items were suitable for factor analysis. From the initial eigenvalues, five factors of test-taking motivation emerged, which accounted for 50.4% of the total scale variance on the TTMS. Table 6 presents the results.

Table 6: Eigenvalue and Total Variance on the TTMS

Components	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	7.684	30.735	30.735
2	1.434	5.737	36.472
3	1.336	5.344	41.816
4	1.103	4.413	46.230
5	1.033	4.131	50.360
6	0.943	3.771	
7	0.898	3.594	
8	0.861	3.444	
9	0.821	3.284	
10	0.807	3.230	
11	0.727	2.907	
12	0.682	2.727	

Components	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
13	0.670	2.681	
14	0.647	2.589	
15	0.618	2.470	
16	0.571	2.283	
17	0.549	2.197	
18	0.549	2.197	
19	0.509	2.034	
20	0.474	1.897	
21	0.465	1.859	
22	0.445	1.780	
23	0.425	1.701	
24	0.417	1.668	
25	0.353	1.413	

Furthermore, the scree plot of the 25-item scale was also determined

Reliability Components

The stability of the scale was determined by administering the instrument to the study sample twice with an interval of three weeks. Thereafter, the test and retest scores of the scale were

correlated. The result shows a high correlation coefficient of 0.730 between the first and second administration of the scale, which was significant at $p<0.05$. this implies that the reponses of the sample were stable over time.

In addition, Cronbach Alpha, Spearman-Brown Coefficient, and Guttman Split-half Coefficient were conducted on both the first scale (47-item) and final scale (25-item) items

Table 6: Internal Consistency Reliability Coefficients of the TTMS

Scale Items	Cronbach Alpha	Spearman Brown	Guttman Coefficient
N=47	0.867	0.820	0.820
N=25	0.900	0.866	0.862

Table 6 shows that Cronbach Alpha, Spearman Brown and Guttman Coefficients for the initial 47 – item scale were 0.867, 0.820 and 0.820 respectively; the corresponding values for the 25-item final scale were 0.900, 0.866 and 0.862. The results showed that the reliability of the final 25-item TTMS was consistently greater than that of the initial 47-item scale for each of the three reliability measures. Table 7 shows the final scale items of the TTMS

Table 7 Test-Taking motivation Scale (Final Version)

Old S/No	**New S/No**	**To what extent do you feel encouraged**	**Response options**
6	1	By your parents' expectation of your good performances in tests?	1 2 3 4 5
7	2	When your teacher praises you for doing well in your class exercises and tests?	1 2 3 4 5

8	3	When your siblings are very good at providing the correct answers to test questions?	1 2 3 4 5
9	4	When you come across people that are successful in life through good performances in tests?	1 2 3 4 5
13	5	When your teacher gives you problems as homework/assignment?	1 2 3 4 5
14	6	By the way the society generally regards persons who excel in tests?	1 2 3 4 5
15	7	When your teacher gives you problems to solve in the classroom?	1 2 3 4 5
16	8	By the need to devote more time on your studies than any other thing?	1 2 3 4 5
17	9	To do your best in whatever you are doing?	1 2 3 4 5
18	10	To be the best in whatever you are doing?	1 2 3 4 5
20	11	To meet up with the highest grade in tests?	1 2 3 4 5
22	12	To do your best in SSCE tests?	1 2 3 4 5
25	13	When academic talks are organized in your school on the importance of performing well	1 2 3 4 5

		in tests?	
26	14	By the effort you need to provide the correct answers to the questions?	1 2 3 4 5
28	15	When you notice you are making progress in your studies?	1 2 3 4 5
32	16	By the result you expect from your tests?	1 2 3 4 5
33	17	By the need to get on well in life?	1 2 3 4 5
34	18	When you are informed that your performances in tests is a very important prerequisite you need to master to study your dream course?	1 2 3 4 5
36	19	By the many prospective disciplines/courses which success in your subjects will make you eligible to take in higher institutions?	1 2 3 4 5
39	20	By the length of time given to write your tests?	1 2 3 4 5
42	21	When a scholarship award is attached to good performances in your tests?	1 2 3 4 5
43	22	By the way invigilators manage your tests?	1 2 3 4 5

44	23	When learning equipments are adequately provided for learning and during testing?	1 2 3 4 5
45	24	By the provision of good classroom condition (well ventilated, good sitting arrangement, lighting e.t.c) for your tests?	1 2 3 4 5

6. Conclusion

Mr. Vice Chancellor sir, and my distinguished audience, we have observed that tests measure core traits of overt and covert human behaviour. They fulfill important needs in the decision-making process permeating all facets of human endeavour. Since errors inherent in testing can be estimated and evaluated, they can be controlled and reduced to the barest minimum. Tests and Measurement, in its pristine form is evidence - based and scientifically driven. Just as the decision of a judge in the temple of justice may be based on the balance of evidence in a case, the psychometric parameters of a test provide the fundamental ingredients for determining its worth and usability. Obtaining data in validity studies is like gathering evidence for a court trial. For tests and the measure they provide to be a true witness, the evidence must be plausible, persuasive and credible. Evidence for validity comes from establishing significant association between each item of the test and the whole test (item-test), among all the items (inter-item) and between the test and other variables (by providing convergent or discriminate evidence). An extension of this evidence is the ability of the test to predict behaviour. Tests are capable of predicting which engaged couples will have marital satisfaction and which will get divorced. Tests can tell how well an applicant will do in a job (like whether a Graduate Assistant will likely become a Professor or stagnated along the line). Tests can

show which applicants for postgraduate studies we should select based on the likelihood to complete a particular Masters or Doctoral Programme. Then some may ask; are tests clairvoyant? No, tests are not; not at the slightest.

It is clear (as Mathematicians wont to say) that Tests and Measurement, properly applied, provides credible, not perfect, evidence of the attributes or characteristics of persons, objects, events or systems, and of behavioural tendencies of individuals. It cannot therefore be said to be a tale bearer. It is in the hands of untrained persons that Tests and Measurement may wear that garment, presenting pre-conceived, distorted and biased evidence as observed scores. Indeed, Tests and Measurement is a true witness which can be called upon to provide evidence in respect of decisions to be taken on any matter involving human behaviour. A wholesome caveat, however, is that decisions involving human behavior, especially high stake decisions should be premised on multiple evidence. This is in consonance with an ancient admonition that "...in the mouth of two or three witnesses, a matter shall be established" (Matt. 18:16). Consequently, we urge students, parents, administrators and employers of labour to accept the unassailable role and contribution of Tests and Measurement as a purveyor of truth. It is thus incumbent on teachers, lecturers, examination bodies and agencies of government responsible for developing, administering and using tests and other assessment measures to embrace and utilize best practices in testing and continually take steps on improvement.

Suggestions

This community can leverage on the insight and services of Tests and Measurement in many ways. We will highlight only three of these, that impact on students and lecturers.

First, Continuous Assessment (CA), as intended, should at least provide students with opportunities to demonstrate their abilities and skills at regular intervals in courses offered during each semester. This provides the motivation for students to study for

mastery throughout the semester if it is properly implemented, and its guidance-oriented function of letting students know where they stand, including their strengths and weaknesses in their courses, and what more they need to do. Conducting CA a few weeks or days to examinations appears to be cosmetic, merely to fulfill statutory requirement, not to monitor students' learning progress or provide feedback to lecturers on what they can do differently to enhance students' understanding and performance. Moreover, releasing CA results close to examinations (if at all) is even less effective. Sometimes, CA scores are kept on the chest of lecturers, only to be used for discriminatory purposes after examination scores are known; sometimes to upgrade the final scores of, perhaps, undeserving candidates. There may therefore be the need to adjust the timing of CA on the academic calendar to reflect its continuous nature, and to encourage lecturers to submit CA scores to Heads of Departments and publish same before the commencement of semester examinations.

Second, lecturers routinely make use of tests to determine students' knowledge and performance. It is part of normal academic duties. But test construction requires the use of appropriate psychometric procedures and techniques for test design, standardization and administration. Notwithstanding that as a result of ICT, the administration and scoring of test items may be technology driven, the construction and development of the items remain a professional enterprise. Towards this end, the university may wish to organize regular workshops on educational assessment, especially educational testing, which academic staff of all cadres should be encouraged to attend. This could also be made part of the orientation that all new academic staff should have before assignment of duties in their departments, in line with best practices in reputable corporate organizations.

Third, the assessment of academic staff, hitherto, reflects only the views of individual lecturers, their colleagues and different levels of the management. A significant other is missing, that is the students. Students are largely the direct recipients of the services of

lecturers, and key stakeholders in the system. Their views should count in the assessment of the teaching component of lecturers' job. Appropriate test assessment instrument that can serve this purpose can be developed and which will be objective, dependable and fair. This will contribute to the validity of the assessment scores of lecturers on teaching.

Final Remarks

I wish to acknowledge persons who have made significant contributions to my education and career advancement. My parents, Mr. Samuel Afolabi Boyinbode (now late) and Chief (Mrs) Janet Eyemowa Boyinbode, who without any formal education, made great sacrifices to ensure that all their children (Olatunde, Ifedayo, Solape, Oladeji and Abimbola) had university education; my other mother and aunt and her caring husband Pa J.O. and Mrs. B.A. Ogunsuyi; they are always there for me through thick and thin. I am profoundly grateful to Professor Olu Makinde (my first Head of Department) and Prof (Mrs.) B.M. Osibodu (now late) (my first degree supervisor) through whom I came into the field of Tests and Measurement; they were also instrumental to my becoming an academic staff. Prof Dayo Adejumo (Uncle D) (now late) was my academic mentor and confidant. Prof. Dibu Ojerinde, an icon and cedar in Tests and Measurement, was my M.A. supervisor; he taught and showed me the rudiments of testing. Professors Wole Falayajo, Joseph Obemeata (my Ph.D supervisor), Pai Obanya and E.A. Yoloye widened the scope of my knowledge of Measurement and Evaluation at the International Centre for Educational Evaluation, University of Ibadan where I obtained my Ph.D. in 1988. I thank Prof. D.O. Owuamanam for his labour of love; Prof A.A. Olowu for believing in me; and my professional colleague and friend, Prof. S.O. Bandele for his optimism when my desire to move on was at a low ebb. I thank my Bishops, especially Rt. Revd. E.B. Gbonigi, who gave me the first opportunity to travel abroad as Mission Partner to the Church of England while on sabbatical leave at the University of Newcastle-upon-Thyne; and my other spiritual fathers for their prayerful

support. The Very Revd. J. O. Olugasa (then my Provost) and the Rt. Revd. M. O. Ipinmoye (my Bishop) were always asking me at a time: when will you become a Professor? The answer is now joyfully before them, more so at this inauguration. Thank you sirs. Dr (Mrs) Kemi Adamolekun (now late) wiped my tears and did so much to get me afloat the muddy, debilitating politics of our Faculty until she resigned; Prof J.T. Ogundari laboured without success to extricate me from the nest of the then lords. I also thank Prof. S. A. Adeyanju for his encouragement and active support at all times. I appreciate all my former postgraduate students, some of whom endured persecution for associating with me when others were mere pretenders. Dr. Dele Faleye and Dr. Demola Odeyemi were particularly outstanding in their support and encouragement. I appreciate my siblings and their spouses for their moral and prayerful support: Engr. (Elder) E. O. and Mrs V. O. Boyinbode; Engr. J. S. and Dr. (Mrs) M. O. Jemilugba; Engr. (Pastor) and Dr. (Mrs) K. O. Boyinbode; and Pastor B. and Mrs M. A. Owojori. I appreciate my children: Toosin, Tolu and Seyi for their emotional support even when they could not comprehend what was going on. I express my deep appreciation to my delightful and delectable wife, my queen, Margaret Olubunmi Afolabi (Ph.D); she gave her all for the course of my well-being and achievement, far beyond my imagination.

Surely, with God all things are possible; it is He who raises the poor from the dust and lifts the needy from the ash heap. He seats them with princes, with the princes of their people. He has made everything beautiful in His time. Certainly, it is beautiful now! Everyone can test and measure it.

Thank you all for your attention.

References

Adejumo, J.A. **Afolabi**, E.R.I (1990). Assessing educational attainment in Junior Secondary School: from policy to practice. Nigerian Educational Forum, 12(1), 133-143.

Adeyemo, E.O. & **Afolabi**, E.R.I (2009). Estimation of effect size in a meta analysis of series of validity studies on matriculation examinations in Nigeria. **International Journal of Educational Research and Administration,** 6(4), 61 – 68.

Afolabi, E.R.I (1994). Towards an accurate assessment of students' cognitive capacity: An evaluation of three scoring procedures in objective test, **Ondo State University Journal of Education,** 1(1), 41 – 49.

Afolabi, E.R.I & Ibrahim, A. (2009). **Effects of positional response bias and correct response location on the difficulty level of multiple – choice questions.** Unpublished Manuscript.

Afolabi, E.R.I & Lijoka, J.O. (2005). The relationship between Common Entrance Examination and the Junior Secondary Certificate Examination in Osun State. **African Journal of Education Studies,** 1, 117 – 119.

Afolabi, E.R.I. (1990). Effects of test format, self–concept and anxiety on response changing behavior. **Journal of Education and Society,** 2 (1), 39 – 46.

Afolabi, E.R.I. (1991). Effects of response changing behavior on the reliability and validity of four objective test formats. **Journal of Issues in Social Science** 1(3), 55 – 60.

Afolabi, E.R.I. (1995). Locus of prediction and scholastic self–rating: Any difference in predictive validity! **Ife Journal of Educational Studies,** 3(1), 16 – 28.

Afolabi, E.R.I. (1998). The paradox of zero in educational measurement. **Nigerian Journal of Social and Educational Research,** 1(1), 71 - 74.

Afolabi, E.R.I. (1998). Validity of Public Examinations: the Environment and Sustainable National Development. In S. S. Obidi; E.R.I. **Afolabi** and M. A. Adelabu (Eds.), **Books of Readings on Education, Environment and Sustainable National Development.** Ibadan: Cardinal Crest.

Afolabi, E.R.I. (1999). Academic self-estimate among secondary school students: Implications for educational guidance. **Nigerian Journal of Social and Educational Research,** 1(2), 68 - 73.

Afolabi, E.R.I. (1999). Effect of testwiseness training on performance in objective tests. **African Journal of Educational Research,** 5 (2), 91 – 98.

Afolabi, E.R.I. (1999). Six honest men for continuous assessment: Evaluating the equating of achievement scores in Nigerian secondary schools. **Ife Journal of Behavioural Research,** 1(2), 7 - 15.

Afolabi, E.R.I. (2000). The changing of initial answers in true–false tests; discrepancy in sex, age and ability groups, **Ife Journal of Theory and Research in Education.** 5(1), 1 - 8.

Afolabi, E.R.I. (2005). **Principles of Test Administration.** Paper Delivered at the Staff Training Workshop, Osun State Polytechnic, Iree.

Afolabi, E.R.I. (2010). Effects of Didactic Ethical Therapy on the Attitude of Secondary School Students to Examination Malpractice. Unpublished Manuscript, Obafemi Awolowo University, Ile-Ife.

Afolabi, E.R.I. (2010) *Effects of Didactic* Ethical Therapy on the Attitude of Secondary School Students to Examination Malpractice. Unpublished Manuscript, Obafemi Awolowo University, Il-Ife.

Afolabi, ERI. (2011). The Problems and Challenges of Curbing Examination Malpractice in the Nigeria Educational System. In K.A. Alao; E.R.I **Afolabi**, & B.A. Faleye (eds.), *Educational Assessment and National Development,* Department of Educational Foundations and Counselling, Obafemi Awolowo University.

Afolabi, E.R.I. (2011). The problems and challenges of curbing examination malpractice in the Nigerian educational system. In K.A. Alao, E.R.I. **Afolabi** & B.A. (Eds.), **Educational Assessment and National Development.** Ile-Ife: Department of Educational Foundations and Counselling, Obafemi Awolowo University.

Afolabi, E.R.I. & Adewolu, B.A. (1999). Effects of item option length on guessing behavior in multiple–choice tests. **Ife Journal of Behavioural Research,** 1(4), 92 – 98.

Afolabi, E.R.I.; Dibu–Ojerinde, O.O., Alao, K.A. & Faleye, B.A. (2005). Inter–rater reliability of Obafemi Awolowo University teaching Practice scores. Yearbook of International Council of Education for Teaching, Illinois (USA), 12 – 15.

Afolabi; E.R.I. (1992). Effects of test format, self-concept and anxiety on answer changing behavior. **Journal of Education and Society** 2(1), 39 – 46.

Afolabi; E.R.I. & Buhari, J.O. (2004). The predictive validity of Holland's self – directed search to vocational guidance in Nigeria. **The Africa Symposium,** 4(2),

Attalli, Y. & Bar–Hillel, M. (2001). **Seek whence: Answer sequences and their consequences in key–balanced multiple-choice tests.** Unpublished Manuscript.

Boyinbode, I.R. (1986). Effects of confidence level on some psychometric properties of true – false test answers, **Nigerian Journal of Educational Psychology.** 1(1), 97 – 106.

Campbell, D.T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matix. **Psychological Bulletin,** 56, 81-105.

Cattell, J.M. (1890). Mental tests and measurements. **Mind,** 15, 373 – 380.

Cronbach, L.J. (1980). Validity on parole: How can we go straight? **New Direction for Testing and Measurement,** 5, 99-108.

De-Ayala, R.J. (2009). **The theory and practice of item response theory.** New York: Guilford Press.

DuBois, P.H. (1970). **A history of psychological testing.** Boston: Allyn & Bacon

Faleye, B.A & **Afolabi**, E.R.I. (2006). Continuous assessment practices in Osun State secondary schools: From policy to practice **International Journal of Learning,** 12(12), 11-16.

Faleye, B.A. & **Afolabi**, E.R.I (2005). Predictive validity of the Osun State Junior Secondary Certificate Examination. **Electronic Journal of Research in Educational Psychology,** Issue 5, Vol.3(1), 131 - 144.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. **Psychological Bulletin,** 76, 378 – 382.

Ford, W.A. (1973). **Everything you wanted to know about testwiseness.** Princeton, New Jersey: Educational Testing Services

Galton, F. (1879). Psychometric experiments. **Brain,** 2, 149 – 162.

Holland, J.L. (1985). Making vocational choice: A theory of vocational personalities and work environments. New Jersey Prentice – Hall.

Kaplan, R.M. & Saccuzzo, D.P. (2005). Psychological testing, Principles, Applications and Issues (6^{th} ed.) Belmont, C.A: Thomson Wadsworth.

Kendal, M.G. and Stuart, A. (1967). The advanced theory of statistics. London: Griffin.

Lilienfield, S.O., Wood, J.M. & Garb, H.N. (2000). The scientific status of projective techniques. **Psychological Science in the Public Interest,** 1, 27 – 66.

Lord, F.M. (1952). A theory of test scores. **Psychometric Monograph,** No. 7

Lord, F.M. (1980). **Applications of item response theory to practical testing problems.** Hillsdale, NJ: Erlbaum.

Lumsdan, J. (1978). Tests are perfectly reliable. **British Journal of Mathematical and Statistical Psychology,** 31, 19-26.

Lundy, A.C. (1985). The reliability of the Thematic Apperception Test. **Journal of Personality Assessment,** 49, 141 – 145.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Broom (Eds.), **Test Validity.** Hillsdale, NJ: Erlbaum.

Micall, A. (1920). A new kind of school examination. **Journal of Educational Research,** 6, 33 – 40.

Millman, J.; Bishop, C.H & Ebel, R. (1965). An analysis of testwiseness. **Educational and Psychological Measurement,** 25, 701 – 726.

Moser, C. & Kalton, G. (1979). **Survey methods in social investigation.** London: Heinemann.

Murphy, K.R. & Daridshofer, A. (2005). **Psychological testing.** New Jessey: Prentice Hall.

Ojerinde, A. & **Afolabi,** E.R.I. (1993). "Tests and measurement" in A. Uba; O. Makinde; D. Adejumo and A. Aladejana (Eds.). **Essentials of Educational Foundations and Counselling.** Ibadan: Claverianum Press.

Ojerinde, D. (2004). Examination Malpractice in our educational system: The NECO experience. **Faculty of Education Lecture Series,** Obafemi Awolowo University, Ile-Ife.

Ojerinde, D. (2011). **Public examinations in Nigeria.** Lagos: Melrose Books.

Rosenthal, R. (1984). **Meta – analytic procedures for social research.** New York: Sage.

Terman, L.M. (1916). **The measurement of intelligence.** Boston: Houghton Mifflin.

Whitaker, K.; Bradley, R.; Navine, A. and Hoghes, E. (2002). Positional response set among high school students in multiple–choice examinations. **Journal of Educational Measurement,** 7, 161-163.

Wickes, T.A. Jr. (1956). Examiner influence in a testing situation, **Journal of Consulting Psychology,** 20, 23 – 26.

Wiggins, J.S. (1973). **Personality Assessment.** Reading, MA: Addison–Wesley.

Wood, J.M; Nezworski, M.T.; Lilienfield, S.O. & Garb, H.N. (2003). **What's wrong with the Rorschach? Science confronts the controversial inkblot test.** San Francisco: Jossey Bass.

Woodworth, R.S. (1920). **Personal data sheet.** Chicago: Stoelting.