*Prestige Journal of Counselling Psychology, Vol. 3, No. 2, September 2020* ISSN: 2651-5687 (Print)
*A publication of the Association for the Promotion of Innovation in Education (APIE)* ISSN: 2651-5709 (Online)

# *Assessing the Technical and Practical Qualities of a Good Test as a Measuring Instrument*

[1]**Bassey A. Bassey, Ph.D**
*babssey67@gmail.com*

[1]**Eme O. I. Amanso**
*emeamanso@yahoo.com*
[1]*Department of Educational Foundations*
*University of Calabar, Calabar*
*Cross River State, Nigeria*

## Abstract
*A good test, as a measuring instrument, has both technical and practical characteristics, qualities or criteria. The primary purpose of this paper is to identify and discuss the qualities or characteristics of a good measuring instrument (test) for schools. Technical characteristics discussed include test items, standardization, objectivity, validity, reliability, discrimination, and norms; while the practical qualities include usability, acceptability, adequacy, purpose, economy, meaningfulness of test scores, and comparability. Related concepts such as measurement, assessment and evaluation are also defined in this paper. The authors posit that all these characteristics and more are interdependent. That is, they are mutually causal and have direct bearing on each other and are not exhaustive. This paper concludes that for a test to be a good measuring instrument, among others, it must measure (validity) what it is supposed to measure, accurately and consistently (reliability); it must be fair to the examinees (objectivity) and be comprehensive enough to serve its purpose (adequacy); it must be easily utilized (usability); and able to pick out, the bright, average and dull students separately (discrimination); it should not result in objections (acceptability); and should be interpreted in terms of a common base that has natural or accepted meanings (comparability).*
**Keywords:** technical, practical, properties, measuring, instrument, reliability

## Introduction
Test is a valuable instrument in the hands of teachers for measuring students' learning outcome. A test is a formal, systematic, usually paper-and-pencil procedure used to gather information about learners' behaviour. It is the most often used instrument for the appraisal, assessment, and evaluation of cognitive outcomes of learning (Falayajo, 2016). Tests are only one of the many types of assessment information teachers deal with; thus, tests are a subcategory in the general domain of assessment approaches

(Airasian, 2004). Other evidence-gathering strategies that also fit within the general domain of assessment are observations, interviews, projects, questionnaires, and so on. Test is an objective approach for measuring the behavioural pattern of a sample or group of students (Anastasi & Urbina, 1997). Cronbach (1960) sees a test as a systemic procedure for comparing behaviour of two or more persons. Test is also seen as an instrument for systematic measure of a sample of behaviour; or a systematic procedure for observing a person's behaviour and describing it with the aid of numerical scale or category system (Joshua, 1998; Brown, 1983). From these definitions, it is evident that some people see test only as an instrument, while some see it as both an instrument and a procedure for using or applying the instrument (Joshua, 2005).

Tests, therefore, are instruments or measurement devices that provide the teacher with information about learners, as well as the instruction. In addition to instructional functions, tests provide useful information for guidance and counselling, along with administrative and programme evaluation decisions.

When constructing or selecting tests and other evaluation instruments, the most important question is: how well will the interpreted scores be appropriate, consistent, or meaningful and useful for the intended application of the results? Tests and other evaluation instruments serve a variety of uses in the school. For example, tests of achievement might be used for selection, placement, diagnosis, or certification of mastery; aptitudes tests might be used for predicting success in future learning activities or occupations; appraisal of personal social development might be used to understand better pupils' learning problems or to evaluate the effects of a school programme. Regardless of the type of instrument used or how the results are to be used, however, all measuring instruments and indeed measurements should possess certain technical and practical criteria, qualities or characteristics. The most essential of these qualities, in the words of Gronlund and Linn (1990), are validity, reliability, and usability.

Any measuring instrument must fulfil specific conditions, criteria, qualities, or characteristics. Criterion is a standard by which something may be judged or decided (Hornby, 2014). Quality implies the standard of how good something is; a distinctive feature of something (Hornby, 2014). While characteristic is a quality typical of a person or something (Hornby, 2014).

Whenever we must construct/develop a test or select a test out of many available tests, we have to ensure that the test fulfils the highest standards (Bassey, 2019). Certain criteria, qualities, or characteristics would culminate into a test or evaluation instrument fulfilling these high standards. A test or evaluation technique is judged for its appropriateness, adequacy, efficiency, consistency, and suitability of a measuring

device based on commonly accepted characteristics. A critical assessment of these characteristics, both technical and practical, which include validity, reliability, objectivity, adequacy, usability, discrimination, and so on, is the concern of this paper. But before this, it is pertinent to discuss briefly some concepts that are related to test and testing. Test has already been defined elsewhere in this article.

## Measurement, Assessment, and Evaluation Defined

**Measurement** is the process of quantifying or assigning a number to performance. It is a process of obtaining a numerical description of the degree to which an individual or object possesses a particular attribute or characteristic. It is a process of describing people's behaviours, events, objects, things, and expressing them quantitatively, that is, in numbers or scores (Joshua, 2005). The most common example of measurement in the classroom occurs when a teacher scores a test or a quiz. Scoring produces a numerical description of performance, for instance, Nukak got 18 out of 25 items correct on Civic Education test; Asuquo got a score of 65% on his Mathematics test; Vivian's score on the Chemistry test was 59 items correct. Other common measurements are made of say, students' height and weight. In all these examples, a numerical score is used to represent the individual's performance, height or weight. Therefore, the outcome of any measurement process is number or score. It should be noted that man has been depending on the outcome of the measurement process to take some decisions. In a classroom situation, measurement uses test as an instrument to obtain these numerical descriptions (scores). Measurement answers the question: "how much"? "How many"?

**Assessment** is the process of collecting, synthesizing, and interpreting information to aid in decision-making. Assessment, as a concept, in a school setting, includes the full range of information teachers gather in their classrooms: information that helps them understand their students, monitor their instruction, and establish a viable classroom culture. It also includes the variety of ways teachers gather, synthesize, and interpret that information for decision-making. In the words of Joshua (2005), assessment is a global process of synthesizing information about individuals so as to describe, understand, and perhaps help them better. It is a process which involves the collection of meaningful information to understand and help people (students) cope with problem. Assessment process is not restricted to the use of test only; it uses other formal and non-formal measurement procedures of gathering information, such as observations, interviews, and projects. Assessment, therefore, is a general term that includes all the ways teachers gather information in their classrooms. Once assessment information is collected, it is used to make judgement about students, instruction, or classroom climate.

**Evaluation** involves making judgements about the quality of students' performance or a possible course of action. When the assessment information collected has been synthesized and thought about, the teacher is in a position to judge the quality of a student's performance or which classroom course of action is best (Airasian, 2004). Joshua (2005) sees evaluation, in generic sense, as a systematic process of judging the worth, desirability, effectiveness, adequacy of something according to definite criteria and purposes. It includes obtaining information (qualitative or quantitative) for use in judging the worth of a programme, product, procedure, course of study, curriculum, or objective, or the potential utility of alternative approaches designed to attain specific objectives. Thus, evaluation is a process, and involves obtaining, generating, or providing useful information or data (e.g. test scores), and taking decisions or making value judgements on these information, data or scores. However, those decisions or judgements so made are based on pre-determined criteria or standards (Joshua, 2005; Gronlund, 1985; Wentling, 1980).

A critical look at these definitions of evaluation indicates that a properly conducted evaluation should involve three fundamental processes, namely:

i) Prescription of specific values or goals to be achieved;

ii) Securing specific evidence regarding the existence, quantity and quality of a condition or process; and

iii) Making a judgement in the light of available evidence concerning the extent to which the desired values or goals have been attained (Joshua, 2005).

Accordingly, therefore, a properly conducted evaluation should lead to the improvement of whatever or whoever is being evaluated.

Test, therefore, is the instrument in the hand of the teacher for measuring students' performance, like tape in the hand of a tailor; and the process of using that instrument to gather quantitative data (numbers or scores) constitutes measurement process; the collection, synthesis, and interpretation of the data (information) to aid the teacher in decision making constitute assessment; while the value judgements made on the results/data/scores of measurement/assessment constitutes evaluation. Measurement answers the question: how much? while evaluation answers the question: how well? or how good?

It is imperative, therefore, that for the right decision-making or correct value judgement about the students, the data/information/scores gotten from the measurement process must be authentic and accurate. To obtain these authentic and accurate information/data/scores about students, the measuring instrument (test) must possess appropriate technical and practical qualities or characteristics, assessment of which, is the focus of this article.

**Qualities of a Good Test as a Measuring Instrument**

The qualities or characteristics of a good test can be classified into two categories, namely:

A. Technical qualities, and
B. Practical qualities (Sidhu, 2007).

**A. Technical qualities of a good test as a measuring instrument**

Technical qualities of a good test include the following: test items, objectivity, validity, reliability, discrimination, standardization, and norms (Sidhu, 2007).

1. **Test items:** Items of good quality in a test are the first requisite of a good test. A test item is bad if it shows a low item-test correlation. An item on which bright learners surpass the poor onces is judged as being good, while the one which shows no difference in respect of bright and dull learners or in which the dull group is more successful than the bright group is a bad test item. The test item fails to discriminate between bright and dull learners on account of anyone of these reasons:
    i. It is so easy that everyone passes it or it is so hard that everyone fails it.
    ii. It is ambiguous or confusing.
    iii. It measures something different from what the test measures (Sidhu, 2007).

Good test items automatically satisfy various criteria of a good test namely; purpose, acceptability, adequacy, usability, standardization, objectivity, validity, reliability, discrimination, and so on, which are both technical and practical characteristics of a good test.

2. **Objectivity**: The objectivity of a test can be viewed from two aspects, namely objectivity of items and objectivity of scoring (Sidhu, 2007). Objectivity of test items can be adjudged if such test items maintain the same meaning from person to person. For a test item to be objective, there should be no difference between the examiner's and examinee's interpretation of the item; if the examinee takes an item in different sense, its objectivity will be considerably reduced. Words like perhaps, always, never, nevertheless, could harm the objectivity of an item. Objectivity of scoring implies the uniformity of scores in the hands of different markers. The personal judgement of the marker should not affect the scores. Variation in his mood and feelings, his attitudes and prejudices should have no bearing on the scores being awarded by him. Essay type tests generally suffer from objectivity of scoring (Sidhu, 2007; Bassey, 2019).

Generally, one of the core factors that must be considered during the development of any test is objectivity. Objectivity is a prerequisite to the reliability of a test; as subjective judgements are considered unreliable. The objectivity of a test is ensured to the extent that its items can readily be scored as right or wrong. The objective test items are so worded that only one answer satisfies the requirements of the statement

or question. Objectivity of a test may be expressed using a coefficient of correlation. The correlation coefficient obtained between scores assigned to a group of papers by the same examiner at two different times is sometimes called the coefficient of objectivity.

3. **Validity**: The validity of a test simply refers to the extent to which such a test measures what it was intended to measure. One of the most important characteristics of a good test is validity (Gronlund, 1985). The absence of validity implies that the inferences and conclusions made from the test results will be either erroneous, misleading or both. The validity of a test is situation specific, and teachers must try and understand this. That is, a test may be valid for one specific purpose or situation and may not be valid for others. No test is valid for all purposes. A test designed to measure what a student has learnt in mathematics for example, should measure his achievement in that course and nothing else. If the test is so constructed that an intelligent learner can determine the correct answer without knowing the subject matter, the test measures general intelligence rather than achievement in mathematics (Sidhu, 2007). It follows therefore, that there are different types of validity, depending on the purposes for which tests are used. The types or kinds of validity include face validity, content validity, construct validity, criterion-related validity, concurrent validity, predictive validity, congruent validity, statistical validity, and differential validity. Detailed discussion on each of the kinds of validity listed here is outside the focus of this article (Joshua, 2005; Sidhu, 2007).

4. **Reliability**: Reliability of measurement is consistency. In essence, it means consistency in measuring whatever the instrument is measuring. A test can be considered reliable based on the extent to which repeated measurements obtained using it are able to yield consistent or similar results for an individual or groups. Reliability deals with the question: Do we get the same score (or approximately the same score) when we measure that person with this instrument (test) more than once? It may be that the test is not measuring what was intended, but it is measuring in a consistent manner. Next to validity, reliability is a very important characteristic of evaluation results since it provides the consistency that makes validity possible and indicates how much confidence is placed in the results (Joshua, 2005; Bassey, 2019).

Reliability is necessary but not enough condition for validity; a test that produces totally inconsistent results cannot possibly provide valid information about the performance being measured. On the other hand, highly consistent test results may be measuring the wrong thing or may be used in inappropriate ways. Thus, low reliability can be expected to restrict the degree of validity obtained from the results of an instrument, but a high reliability may not necessarily guarantee that a satisfactory

degree of validity will be present. Simply, put, reliability merely provides the consistency that makes validity possible.

Correlation coefficient is a statistic which shows the degree of relationship between two sets of scores obtained from a group of individuals; for example, the relationship between students' performance in mathematics and physics from a test. There are several methods of estimating reliability coefficients. They include:

i) Test-retest method that yields measures of stability
ii) Equivalent forms method that yields measures of equivalence
iii) Split half method that yields measures of internal consistency
iv) Kuder-Richardson method that yields measures of internal consistency
v) Cronbach Alpha method that yields measures of internal consistency.

Detailed discussion on the above methods is outside the scope of this article (Joshua, 2005; Sidhu, 2007).

5. **Discrimination**: Discriminating power of a test shows how well an item discriminates between the bright and the dull students. It indicates whether the item is measuring the same ability as the test measures. It is a measure of correlation between the item and the total test score. Like the coefficient of correlation, the discrimination index ranges between -1.00 to + 1.00. A high positive discrimination index is desirable. A discrimination index closer to zero indicates that the bright and the dull students both have done equally well on the item. It means that the item is not discriminating between the high scoring and the low scoring groups. A negative discrimination index shows the dull students scored higher than the bright ones on the item. The standard for the discrimination index is that it should be as high as possible; usually, an index of .30 and above is acceptable. Ebel and David (1991) have given the standards as guidelines:

| **Discrimination index** | **Item Evaluation** |
|---|---|
| .40 and above | Very good item |
| .30 to .39 | Reasonably good but subject to improvement |
| .20 to .29 | Marginal items, needing improvement |
| Below .19 | Poor items, to be rejected |

6. **Standardization**: Standardization is a special application of the need for controlled conditions in all scientific observations (Sidhu, 2007). Testing conditions must be the same for all individuals in order to make the scores obtained by different persons comparable. Standardization ensures uniformity of testing conditions through specified and fixed procedure, apparatus and scoring. The process of standardization

requires that exact materials are employed; time limit, instructions, demonstrations and every other detail of the testing situation are made equally available to all students.

The process of standardization according to Sidhu (2007) is carried out through many steps, namely: specifying the objectives and outcomes of learning, preparing suitable items in relation to objectives, outcomes of learning, contents, subtopics, weightage and importance; selecting representative sample, administration of the test on this sample, preparing a scoring key, working out reliability, validity, and norms. A test without standardization automatically loses most of the qualities and characteristics of a good test.

7. **Norms**: Every good test must accompany the requisite tables of norms. Norms are levels of performance in a test attained by well-defined groups of examinees, such as age norms, grade norms, percentiles, standard scores, and quotients. A raw score provides only a numerical summary of a pupil's performance. Norms provide adequate interpretation to raw scores (Gronlund, 1985; Sidhu, 2007).

**B. Practical qualities of a good test as a measuring instrument**

In selecting test and other evaluation instruments, practical considerations cannot be neglected. Practical characteristics of a good test as a measuring instrument include: Usability, Acceptability, Adequacy, Purpose, Economy, Meaningfulness of test score, and Comparability (Sidhu, 2007).

1. **Usability**: No matter how valid and reliable a test might appear, if it is not usable to solve desired problems or solve intended constructs, then such a test is as good as not having it designed in the first place (Bassey, 2019). A test that can be handled adequately by the regular classroom teacher without much of special briefing is better than a test requiring specially trained administrators. The usability of a test depends upon several factors, such as: ease of administration, ease of scoring, ease of interpretation and application, time required for administration, cost of testing (Sidhu, 2007). A good test should contain clear and complete instructions so that all the examinees read them and follow them equally well. The results of the test should be obtainable in a simple, rapid, and routine manner. The success or failure of a testing programme is determined by the ease and accuracy of interpretation. A good test is one which can easily be used and its results easily interpreted by an average teacher (Joshua, 2005).

2. **Acceptability**: A good test should be acceptable to the learners for whom it is intended. It should also be acceptable to teachers, parents, and other members of the society (Sidhu, 2007). A too easy or a too difficult test will not be acceptable to any concerned group. Its acceptability increases as people obtain desirable results from it year after year. The results obtained from a test should be satisfactory to almost

everyone viewing the output obtained from it. Acceptable tests are those that comprehensively cover the contents or courses for which it is designed, and nothing more (Sidhu, 2007).

3. **Adequacy**: We cannot assume that a comprehensive test can measure all the elements of knowledge and skills that a learner must acquire in completing a course. Although, it is expected that tests items should widely represent all the possible outcomes expected of the learners, the sampled items should give the scores as representatives of the pupil's achievement for the entire covered area. However, if the test is too short or too long, it goes against the criterion of adequacy. The test should be adequate from all angles of contents, age, grade, local emphases, expected learning outcomes, objectives, and other related factors (Sidhu, 2007; Bassey, 2019).

4. **Purpose**: A test may possess all the important characteristics of a good test and yet it may be of no value for use in a situation. Unless tests are selected or constructed for definite purposes and used in an intelligent manner to achieve the desired results, they are of little value and may even be harmful (Sidhu, 2007). A test can fulfil only the purpose for which it has been standardized. If the test is constructed by the teacher, its utility depends largely upon whether the results will serve the desired purposes of the test.

5. **Economy**: Another important characteristic of a good test as a measuring instrument is economy. This may be economy of time and/or economy in cost. On the economy of time, tests requiring long time are not acceptable to students, parents, as well as markers. Other things being equal, shorter tests should be preferred to longer tests. At the same time, too short a test would be lacking in its reliability and validity. Using separate answer sheet could represent economy of time. On the economy in cost, a test should be within the financial resources of the tester. Fortunately, there seems to be no relation between the cost of a test and its quality, for even a limited budget can be enough to have a well-constructed test (Sidhu, 2007; Bassey, 2019). Standardized and re-usable tests with separate answer booklets/sheets are always economical than the test which are usable only once. The testing can also be made economical if a teacher decides to assess learners in groups as opposed to individual testing.

6. **Meaningfulness of test score**: Generally, a single score is obtained from a test which is likely to be more meaningful than the several different scores. The single score becomes meaningful in view of the specific purpose of the test. In a battery of tests, it has to be specified what the overall score conveys, what scores on separate sub-tests convey, or what the various combinations of scores convey to the tester (Sidhu, 2007).

7. **Comparability**: The last but not the least characteristics of a good test for consideration here is comparability. A test possesses comparability when scores resulting from its administration can be interpreted in terms of a common base that has natural or accepted meanings. As suggested by Sidhu (2007), there could be two ways by which comparability of results of standardized test is established:

   a. Availability of parallel forms of the test, and
   b. Availability of adequate norms.

Through norms, the scores of an individual can be compared with the age and grade norms. Through parallel forms, the individuals or groups can be compared from class to class, school to school, and year to year.

**Conclusion/recommendations**

This paper has presented a somewhat comprehensive list of characteristics or qualities of a good test. The paper posits that just like metre rules, tapes, balance, barometers, thermometers, speedometers, tests as a measuring instrument must possess specific characteristics that would make them (tests) serve in schools the respective purposes for which they were constructed. A good test for schools must actually measure what it is supposed to measure (validity); measure accurately and consistently (reliability); must be fair to the examinees (objectivity); must be comprehensive enough to serve its purpose (adequacy); must be easy to utilize (usability); should pick out the bright, average and dull students separately (discrimination); should not result into objections (acceptability); and should be interpreted in terms of a common base that has natural or accepted meanings (comparability). All these characteristics and more are interdependent. That is, they are mutually causal and have direct bearing on each other and are not exhaustive.

The purpose of testing, measurement and/or evaluation to the teacher, primarily, is that of providing more comprehensive, systematic and objective evidence on which to base instructional or educational decisions. The teacher depends on some data or information to be able to take correct and informed decision on the students/children that will help transform them educationally, socially, morally, and so on.

Besides the teacher, each of the stakeholders, such as the parents, the counsellor, the school administrator, ministry of education official, the researcher, and the priest, among others, has his/her set of informed decisions to make on behalf of the children. Each of these significant persons requires some form of data or information to take proper and informed decisions that could help transform the young ones generally. Certainly, good tests for appropriate measurement and evaluation provide these data and information for educational, social and moral transformation of the youth to face societal challenges.

It is recommended that nothing should be taken as final about these characteristics or qualities. A keen examiner and inquisitive/curious evaluator will always find that there is need and scope for improving even the best among available tests for growth and development of students and schools as these remain the valid and dependable panacea for educational, social and moral transformation of the youth to face societal challenges in all ramifications.

**References**

Airasian, P. W. (2004). *Classroom assessment*. New York: McGraw-Hill.

Anastasi, A. & Urbina, S. (1997). *Psychological testing* (7th ed.) London: Prentice-Hall International.

Bassey, B. A. (2019). Good tests for schools: A panacea for basic education teachers. In J. E. Tabotndip, K. Achuonye & T. E. Agboghoroma (Eds.), *Basic Education in Nigeria: Matters Arising* (pp. 34-48). Onitsha: West and Solomon Corporate Ideals Ltd.

Brown, F. G. (1983). *Principles of educational and psychological testing* (3rd ed.). New York: Holt, Rinehart and Winston.

Cronbach, L. J. (1960). *Essentials of psychological testing.* New York: Harper and Brothers.

Ebel, R. I. & David, A. F. (1991). *Essentials of educational* measurement. New Delhi: Prentice-Hall.

Falayajo, A. (2016). Methods of evaluation. In J. G. Adewale (Ed.), *Emerging Trends in Educational Management, Assessment, and Evaluation in Africa.* Yaounde: Educational Assessment & Research Network in Africa (EARNiA).

Gronlund, N. E. (1985). *Measurement and evaluation in teaching* (2nd ed.). New York: Macmillan Publishers.

Gronlund, N. E. & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th ed.). New York: Macmillan Publishers.

Hornby, A. S. (2014). *Oxford advanced learners' dictionary*. Oxford: Oxford University Press.

Joshua, M. T. (1998). Role of test measurement and evaluation in counseling. *Nigeria Education Journal, 2*(1), 55-63.

Joshua, M. T. (2005). *Fundamentals of test and measurement in education.* Calabar: University of Calabar Press.

Sidhu, K. S. (2007). *New approaches to measurement and evaluation.* New Delhi: Sterling Publishers.

Wentling, T. L. (1980). *Evaluating occupational education and training (*2nd ed.). Urbana, Illinois: Griffon Press.