CHAPTER ONE

THE NATURE OF MEASUREMENT AND EVALUATION

Introduction

The terms of measurement and evaluation are often used interchangeably with little regard for their meanings but they are technically never the same. Ebel (1972) defines measurement as a process of assigning numbers to the individual members of a set of objects for the purpose of indicating differences among them in the degree to which they possess the characteristics being measured. Sax (1972) also perceived measurement as the assigning of numbers to attributes or characteristic of persons or events according to explicit rules or principles. Measurement therefore refers to a systematic process of assigning numbers or symbols to observations that confirm the true attribute of what is being measured and answers the question ‘how much?’

Measurement does not involve qualitative descriptions or value judgements as such it is relatively objective. The objectivity of measurement, however, depends on the accuracy and the reliability of the instrument used in measuring. Measurement involves physical, concrete, abstract, and mental figures. In physical measurement, there is direct measurement; for instance measuring the dimensions of a chalkboard with a ruler to indicate its size or dimensions.

Learning is the end product of educational endeavour. It is difficult to measure the outcome of learning (i.e. behavioural change) since it is intangible or qualitative. As a result the degree of learning in an individual is measured differently, and that is what is referred to as educational measurement.

There is also a uniform graduation unit of physical measurement but in educational or mental measurement, the units are not equal in graduation. For instance, it can be said that a 10cm stick is equal to a 10cm on a ruler but it cannot be said, for instance, that student ‘A’ obtained 80% in a test and student ‘B’ 40% so student ‘B’ put in only half the effort of student ‘A’ or that student B knows half of what student ‘A’ knows. Mental measurement thus comparatively falls short of logical accuracy and consistency compared to physical measurement.

Steps in Measurement

Thomdike and Hagab (1977) indicate that measurement of any attribute entails three principal steps:

- Identifying and defining characteristics that are to be measured. This demands precision to guide against straying. In education, identifying and defining characters that are to be measured is to measure intelligence.
- Determining a set of operations by which the characteristics may be manifested and observed. That is the means by which the behaviour being measured can be demonstrated or exhibited. Tests (intelligence, aptitude, achievement, etc.) are used to make testees demonstrate their abilities.
- Establishing a set of procedures for translating the observation into quantitative statement. This involves assigning numerals to the attribute to indicate the extent and

the degree of behaviour exhibited. In Education, scoring scheme is drawn to give weights to the attributes.

Definition of Evaluation

Evaluation goes beyond measurement, ‘how much’ to concern itself with the question, ‘what value?’ It seeks to answer an important question for testers and testees – ‘What progress am I making?’ Evaluation therefore presupposes a definition of goals to be reached i.e. objectives that have been set forth. In Education, we evaluate to find out whether we are reaching the goal of our teaching.

By analysing the method and results, we are able to find ways of improving them. Evaluation is therefore, not an extra chore imposed on instruction but rather an integral part of what a good teacher does to make teaching more effective. Evaluation is not just a testing programme, for tests are but one of the many different techniques (observation, check lists, questionnaires, interviews, etc.) that may contribute to the total evaluation of a programme.

Evaluation is a continuous inspection of all available information concerning student’s educational programmes and the teaching – learning process to ascertain the degree of change in students and for making valid judgement about effectiveness of the programme. Cronbach (1982) defines evaluation as the collection and use of information to make decisions about educational programmes. For Payne (1975), evaluation is the process by which quantitative and qualitative data are processed to arrive at value and worth of effectiveness. The objective of education is to make judgement about the quality or worth of an educational activity. Evaluation is therefore is not a culminating activity but has a primary purpose of seeking questions about a programmer. For example questions such as:

i. Are the instructor’s objectives achievable?
ii. Are they worthwhile?
iii. Are the methods effective?
iv. Is the instructor actually changing students’ behaviour in the desired direction?

Continued evaluation is therefore very essential in educational programmes. Evaluation is more comprehensive than measurement for it deals with both qualitative and quantitative characteristics of events of attributes. It is also judgemental in nature and deals with the worth, the goodness or badness of a performance or decision. Evaluation is therefore relatively subjective.

Process of Evaluation

Evaluation follows three steps:

- It involves collection of information
- Analysis of information collected
- Using the information collected

Evaluation is usually confused with assessment but it is more comprehensive than assessment. While evaluation bothers on the collection of data and making of decisions on an educational programme concerning the learner (testee), the teacher (tester) and the programme as a whole, assessment on the other hand concerns itself with collecting data and making decisions on the performance of the testee (learner).

Gray (1975) defines assessment as an attempt to measure the pupil not as a whole, but some particular ability, knowledge, skill or attitude which he may or may not possess. Assessment is therefore limited in scope than evaluation.

Purpose of Evaluation

Evaluation is carried out for various purposes some of which are the following:

- One of the most important purposes of evaluation is to adapt instruction to the differing needs of individual pupils. Evaluation techniques help teachers to identify pupils needing specialized work and the kind of specialization required. Without evaluation techniques teachers may over-estimate or under-estimate the extent to which they should differentiate their treatment of pupils. Evaluation leads to better-directed and more effective methods of carrying out educational activities.
- Another use of evaluation is educational guidance. Evaluation provides information on how much aptitude pupils possess for scholastic work in which he is most likely to succeed.
- Furthermore, evaluation provides a basis for long-range counselling, placement and follow-up work as well as assistance in dealing with immediate problems of pupils.
- In personal guidance, evaluation is used to identify the most troublesome educational, vocational, social and emotional problems which pupils face.
- In addition to purposes pointed directly towards pupil needs, pupil evaluation helps in the overall appraisal of the total school programme by revealing specific strengths and weakness in an educational programme.
- Evaluation provides a basis upon which to compare one school's programme with another. It makes possible a study of a programme between different dates, school standards, school norms and the nature of needs in curriculum improvement.
- Pupil reports to parents and school patrons may also be used as a basis for the improvement of public relations and the mobilization of public opinion.

Differences between Measurement and Evaluation

	Measurement	**Evaluation**
1	Measurement is an old concept	Evaluation is a new concept
2	Measurement is a simple word	Evaluation is a technical term
3	The scope of measurement is narrow	The scope of evaluation is wider
4	In measurement only quantitative progress of the pupils can be explored	In evaluation pupil's qualitative progress and behavioural changes are tested

5	In measurement, the content, skill and achievement of the ability are not tested on the basis of some objectives but the result of the testing is expressed in numerals, scores, average and percentage	In evaluation, the learning experiences are provided to the pupils in accordance with predetermined teaching objectives are tested
6	In measurement, the qualities are measured as separate units.	The qualities are measured in the evaluation as a whole
7	Measurement means only those techniques which are used to test a particular ability of the pupil.	Evaluation is the process by which the previous effects and hence caused behavioural changes are tested
8	In measurement, personality test, intelligence test and achievement test etc. are included	In evaluation, various techniques like observation, hierarchy, criteria, interest and tendencies measurement etc. are used for testing the behavioural changes
9	By measurement, the interests, attitudes tendencies, ideals and behaviours cannot be tested	Evaluation is that process by which the interests, attitudes, tendencies, mental abilities, ideals, behaviours and social adjustment etc. of pupils are tested
10	Measurement aims at measurement only	The evaluation aims at the modification of education system by bringing a change in the behaviour

THE IMPORTANCE OF TESTING IN EDUCATION

Testing represents an attempt to provide objective data that can be used with subjective impressions to make better more defensible decisions. Tests are indispensable tools in educational enterprise, for without test there can be no evaluation and without evaluation there can be no feedback to facilitate learning. We test to provide objective information which we combine with our subjective common sense impressions to make better educational decisions. Measurement data enter into decisions at all levels of education, from those made by the individual classroom teacher to those made by the state or society.

Thomdike and Hagan (1977) categorized the decisions made from test under: instructional, grading, diagnostic, selection, placement, counselling and guidance, programme or curriculum and administrative policy. Tests are sometimes characterized as a “necessary evil” in education. Almost every student approach a test with apprehension, and those who do less well than they had expected can easily find some basis for regarding examinations as unfair. Cheating in examination is reported often enough to cause some shadows of disrepute over test; instructors too, sometimes dislike assuming the role of examiners. Most of them prefer to be helpful rather than critical.

There is also something inconsiderate about probing the minds of other human beings and passing judgement on their shortcomings. Unfortunately, there is no effective substitute for tests or examinations in most classrooms. "To teach without testing is unthinkable….. the evaluation process enables those involved to get their bearing, to know in which direction they are going" (Joint Commission of the American Association of School Administrators, 1962).

Anxiety, unfairness, dishonesty, humiliation and presumptuousness are associated with tests. The process of examining and evaluating cannot be dispensed with if education is to proceed effectively. It should also be emphasized that those who regard test as "evil" that must be tolerated usually do not mean to imply that good education is possible without any assessment of student achievement whatsoever. What they suggest is that a good teacher working with a reasonable class size has no need for tests in order to make sufficiently accurate judgement of students' achievement. They may also suggest that tests, which they have seen or perhaps they have been used to leave so much to be desired that a teacher is better off without the kind of "help" such tests are likely to give him/her.

The major function of test is to measure student achievement and thus to contribute to the evaluation of his/her educational progress and attainment. To say as some critics of testing have said that what a student knows and can do is more important than his score on a test or his grades in a course implies quite incorrectly in most cases. Again, to say that testing solely to measure achievement has no educational value also implies quite incorrectly in most cases. Tests facilitate decision making and decisions can be made concerning the learner, the teacher, guidance and administration.

Let us examine the decisions that can be made on each of these through testing:

Learner

1. Tests motivate and direct students learning. The experience of almost all students and teachers support the view that students tend to study better when they expect an examination than when they do not.
2. Tests provide feedback to the student to reveal his strengths and weakness and therefore facilitate guidance and counselling. Tests ensure good learning habits through its feedback and guidance and counselling to the student.

Teacher

1. Tests help teachers to give more valid and reliable grades to students as the grades are intended to summarise concisely a comprehensive evaluation of the student's achievement.
2. Tests facilitate instructional directions – the process of constructing them if it is approached carefully, a test may cause teachers to think about instructional goals and help in linking pedagogy with educational objectives.
3. Tests provide feedback to teachers and help them to modify instructional methods.

4. Tests equip teachers with the knowledge of the strengths and weakness of pupils (diagnostic) and therefore encourage remediation and individualized attention.
5. Tests provide teachers with information on the entry behaviour of pupils and help them to determine the standing level of instructional programmes.

Guidance Decisions

Guidance decisions refer to vocational choices, educational and personal problems of the student. Usually, tests scores are used as the basis for providing guidance and counselling services on placement for students. For example, tests scores provide:

- Constructive feedback (information) to both the teacher and the student on the strengths and weaknesses of the student to facilitate guidance and counselling;
- Information on the students ability, interest and capability, which can be gleaned from the results of a text provide an index for vocational and educational choices (guidance)
- Information of the student's performances in a test at times helps the teacher and parents to counsel the student.

Administrative Decisions (Social Functions)

Testing and test scores are used for selection, classification, placement and certification.

- **Selection:** Test scores (grades) are used for sorting students into programmes or occupations.
- **Classification:** test results (scores) are used to categorize (assign) students into programmes and vocations.
- **Placement:** test results are used for educational and vocational placement.
- Tests are used for curriculum development by providing curriculum developers with the level of students' competence through test scores.
- Test results are used for certification denoting the type of skills or knowledge of the individual.

Overall, tests are needed in order to provide information about the achievement of groups of learners without which it is difficult to see how rational educational decisions can be made. While for some purposes, teachers' assessments of their own students are both appropriate and sufficient; this may not be true for other cases. Even without considering the possibility of bias, we have to recognize the need for a common yardstick, which tests provide, and if we care about testing and its effects on teaching and learning, the other conclusion to be drawn from recognition of the poor quality of so much testing is that we should do everything that we can make to contribute to the improvement of testing by:

- Helping teachers to write better tests items, and
- Putting pressure on others, including professional testers and examining boards to improve their tests.

Disadvantages of Testing

As important as test is in education, sometimes, it is regarded as a necessary evil. Some of the reasons are:

1. Tests create fear and anxiety in students and interfere with learning. The apprehension in students has the potential to destroy self-confidence and kill the desire to learn.
2. Tests are regarded as an invasion of privacy since they lead to the closure of the values, beliefs, interests and the ability of the testee.
3. Tests lead to labelling of students into groups like bad, good, average, slow, and intelligent. This does not mean that they are fixed but such labelling go a long way to affect the personality development of students.
4. Tests tend to encourage unhealthy competition among students and breed selfishness and retards group cohesion.
5. Tests penalize bright and creative students because of its nature of conformity. Tests deny the creative person a significant opportunity to demonstrate his creativity and favours shrewd candidates over the one who has something to say.
6. Test results reveal only a superficial aspect of the individual as the system of assessment is not holistic.
7. Test also encourages "teach test" or the narrow curriculum where some teachers teach only examination syllabus.

ASSESSMENT

Definition of assessment

Assessment refers to the wide varieties of methods or tools that educators use to evaluate, measure and document the academic readiness, learning progress, skill acquisition or educational needs of the students.

In other words, educational assessment is seen as a systematic process of documenting and using empirical data on the knowledge, skills, attitude and beliefs to refine programs and improve students learning.

Assessment stands at the heart of effective school system. Without adequate system of assessment, selection will be difficult to legitimize, certification will carry a wide range of meanings, monitoring of the school performance will be difficult and diagnostic and remediation of learning problems will be haphazard. Gray (1984) in Aku (2018) defines assessment as an attempt to measure some particular ability, knowledge, skills or attribute of a pupil.

Approaches to assessment vary because the range of options in choosing models of assessment is very wide. They include individual and group testing; written, oral and practical task; open and closed book conditions, self, school-based or external assessment, continuous / formative and summative assessment. Assessment is more than measurement as it deals with qualitative data like evaluation. It is however limited in scope compared to evaluation for whilst evaluation concerns the teacher/tester, the learner/testee and the programme as a whole, assessment deals only with the learner or the testee.

Fundamental Principles of Assessment

Assessment is an integrated process for determining the nature and extent of student learning and development. This process will be most effective when the following principles are taken into consideration:

- Clearly specifying, what is to be assessed: the effectiveness of assessment depends as much on a careful description of what to assess as it does on the technical qualities of the assessment procedures used. Thus, specification of the characteristics to be measured should precede the selection or development of assessment procedure. When assessing student learning the intended learning goals should be clearly specified before selecting the assessment procedures to use.
- An assessment procedure should be selected because of its relevance to the characteristics or performance to be measured. It is to be noted that assessment procedures are frequently selected on the basis of their objectivity, accuracy or convenience.
- Comprehensive assessment requires a variety of procedures. It is important to note that no single type of instrument or procedure can assess the vast array of learning and development outcomes emphasized in a school programme. Multi-choice and short-answer tests of achievement are useful for measuring knowledge, understanding and application outcomes, but essay tests and other written projects that require students to formulate problems, accumulate information through library research or collect data (for example, through experimental observations and interviews) are needed to measure certain skills in formulating and solving problems.
- Proper use of assessment procedures requires an awareness of their limitations. Assessment procedures range from very highly developed measuring instruments, for example, achievement tests, to rather crude assessment devices, for example self-report technique. It is important to note even the best educational and psychological measuring instruments yield results that are subject to various types of measurement error.

Functions of Assessment

Assessment performs various roles that can be broadly classified into facilitating and inhibiting. Assessment becomes facilitative when it motivates learning, reinforces learning goals and access to good life. On the inhibitory and preventive side, assessment has been argued to eliminate the learner from the process and enjoyment of learning. For instance, failure reduces self-esteem in the learner.

The role/functions of assessment are classified under six main headings:

- Diagnostic
- Evaluative
- Guidance
- Prediction
- Selection and
- Grading.

A more extensive classification can be made from above classification even though there may be some overlapping between categories. Generally, assessment performs the following functions:

1. Certification and qualification
2. Selection and social control
3. Clear recording and reporting of attainment
4. Prediction
5. Measurement of individual differences (psychometrics)
6. Student-pupil motivation (whether teaching-learning structures are competitive, co-operative or individualistic).
7. Monitoring students' progress and providing effective feedback to students
8. Diagnosing and remediation of individual difficulties
9. Guidance
10. Curriculum evaluation
11. Provision of feedback on teaching and organization effectiveness
12. Teacher motivation and teacher appraisal
13. Provision of evidence for accountability and distribution of resources
14. Curriculum control and
15. Maintaining or raising of standards.

Differences between Assessment and Evaluation

BASIS FOR COMPARISON	ASSESSMENT	EVALUATION
Meaning	Assessment is a process of collecting, reviewing and using data, for the purpose of improvement in the current performance.	Evaluation is described as an act of passing judgement on the basis of set of standards.
Nature	Diagnostic	Judgemental
What it does?	Provides feedback on performance and areas of improvement.	Determines the extent to which objectives are achieved.
Purpose	Formative	Summative
Orientation	Process Oriented	Product Oriented
Feedback	Based on observation and positive & negative points.	Based on the level of quality as per set standard.
Relationship between parties	Reflective	Prescriptive
Criteria	Set by both the parties jointly.	Set by the evaluator.
Measurement Standards	Absolute	Comparative

Review Questions

1. Differentiate measurement from evaluation using five points only in a tabular.
2. Why is measurement and evaluation necessary in education?
3. List and explain the three steps/processes of evaluation.
4. What is test?

4a.List and explain five types of test.

5. Outline five importance of test to each of the following:
    - Teacher
    - Student
    - Curriculum planners
    - Guidance and Counselling
    - Administrator

5b.Test is regarded as a necessary evil, justify.

CHAPTER TWO

CONTINUOUS ASSESSMENT

Definition of Continuous Assessment

Continuous Assessment is the system in which the quality of student’s work is judged by various pieces of work during a course and not by one final examination.

Kajola (2010) defined Continuous Assessment as a period examination of the students at different stages of learning for feedback purposes.

Trend of Usage

Evaluation can be **summative** (i.e. terminal or one-shot) or **formative.** It is from formative evaluation that Continuous Assessment (CA) is derived. The distinctive feature of Continuous Assessment is the frequency of assessment by which the final grade of a student is the aggregate of his/her performance in a course. There is a significant international trend towards continuous assessment as many countries with a variety of political ideologies have introduced CA to operate in parallel with external examinations in their system of education. Continuous Assessment is in operation in several countries including Tanzania, Papua New Guinea, Nigeria, Seychelles, Sri Lanka and Ghana. Continuous Assessment was introduced in Tanzania in 1974 with the passing of the Musoma Resolution to get rid of the “ambush” type of examination and to reduce the emphasis placed on written examination (TANU, 1974, quoted in Nall (1987).

In Tanzania, continuous assessment score contributes 50 percent of total weighting in students’ final result, Njabill (1987) argues that the main purpose of having continuous assessment scheme as an integral component of assessment procedures in the Tanzanian education system is to minimize the elements of risk associated with a single examination and to give a valid indication of student achievement, because it is felt that no student who works conscientiously should fail.

Another country with a long standing commitment to Continuous Assessment is Papua New Guinea, which has used CA since independence in 1975 and until 1981. In Papua New Guinea, like Tanzania, CA and external examination marks have each carried 50 percent of the weighting for students' final results. Nigeria adopted the CA in 1977 and Seychelles in 1987. In Nigeria and Seychelles, CA takes the form of on-going observation, folios of work, oral or written tests or home work. The national policy on Education in Nigeria (1981 Revised) and the National Youth Service (NYS, 1987) in Seychelles both indicated that the CA should consider the student's achievement in both cognitive and psychomotor domains.

In Ghana, the trimester system was abolished and the semester introduced in tertiary institutions. The adoption of the semester system brought in its wake the concept of Continuous Assessment. Continuous Assessment was accepted and implemented in1987. In Ghana, the weighting for external assessment is sixty percent (60%) and that of is forty percent (40%). Continuous assessment is now practised from the basic level of education to the university level in Ghana. In all the countries considered, the introduction of CA has had two major effects. First, assessment leading to final result has been spread over a period of time. Second, a substantial element of that assessment has become school-based. Again, although there are differences of emphasis in the various countries, the reasons for introducing CA in most cases fell within the following broad alms, which are interrelated.

Reasons for the Introduction of Continuous Assessment (CA)

A. **To enhance the validity of Assessment:** it is argued that a one-off formal examination is not good test of pupils' achievement. For example course work allows candidates who do not perform well under examination conditions to demonstrate their ability in a more relaxed atmosphere. Course work can also be used to assess those skills that cannot be measured or assessed in written examination (Mkndawire, 1984). Although in some cases, continuous assessment may consist merely of a series of written tests; it is a general aim of CA to assess and report a wider range of students' achievement. Thus, CA may include a wide variety of styles like project, interviews, questionnaires and teacher observation. Continuous assessment is also intended to cover a much wider range of skills that may span cognitive, affective and psycho-motor domains and in the case of the cognitive domains emphasis order skills. It is felt that the validity of students' scores is increased by gathering assessment over a substantial period of time and by maximizing the range of educational objectives that are assessed.

B. **To integrate Curriculum, Pedagogy and Assessment**

Changes in what is assessed are likely to be associated with changes in what is valued, and the concept of assessment linked (if not assessment – led) curriculum development leads to emphasis on relevant education. Certainly CA can be argued to reduce undesirable backwash effect of external examinations. The introduction of CA may also be related to concern about the quality of education provision. A key feature of CA in all the countries considered is the responsibility of teachers for Continuous Assessment of their own pupils and their involvement in both the planning and implementation of CA. The introduction of CA thus provides considerable opportunities for in-service education and training (INSET) of teachers, which enhances their professional efficiency. Another key feature of the CA system is feed-back of assessment data about individual students and about curricular effectiveness, which is associated with the clarity of the objectives to be assessed. Thus, CA encourages a focus on curriculum objectives, instructional procedures and the system of assessment.

C. **To serve a broader range of assessment functions and in particular to emphasize formative functions**

The shift of emphasis away from summative evaluation to formative evaluation appears to be of great importance at any rate within the world if education itself as it facilitates a holistic assessment of the individual. Nevertheless, it will be a mistake to conclude that assessments are no longer designed to discriminate between candidates.

Continuous assessment has both formative and summative aims. The aims of CA can be designed to discriminate between candidates:

- To know the performance achieved by the students in various fields of learning in which they are involved.
- To appreciate particular knowledge and skills acquired by the students individually or in groups.
- To identify the strengths and weaknesses of the teaching/learning process.
- To generate an information device for guidance and counselling.
- To give to the students feedback about their attainments vis-a –vis different learning targets.
- To provide information for consideration of students' vocational and occupational guidance and decision making.
- To give the teacher greater involvement in the overall assessment of his/her pupils.
- To provide a more valid assessment of the students' overall ability and performance.
- To enable teachers to be more flexible and innovative in their instruction.
- To reduce examination malpractice.

Characteristics of Continuous Assessment

Continuous assessment has characteristics that can be classified as comprehensive, formative, cumulative, systematic, diagnostic and guidance oriented.

1. **Comprehensive:** this comprehensive nature of CA lies in the extent of coverage and the holistic nature of assessment. Continuous assessment takes into consideration the totality of the individual (personally) and the assessment procedures cover the cognitive, affective and the psychomotor domains. Thus the learner's interest, ability, capability and skills are all evaluated. Furthermore, CA uses varied evaluation procedures like observation, standardized and teacher-made tests, projects, class assignments, interviews and rating scales.
2. **Formative:** this involves the collection of data on the student on regular bases; the effective analysis of the results and the breaking down into smaller units of instructional materials or into manageable units to make learning meaningful. The formative nature, i.e. the regular collection of data and the sequential presentation of instructional materials facilitate evaluation of the teaching learning process (transfer of learning).

3. **Cumulative:** assessment of students is not based on one-shot examination (summative) but the aggregate of all attainments through the period of programme. Thus, the total (final) grade of a student is determined by the marks obtained in class assignments, contribution in class, projects, class tests, mid-semester examinations and end-of-term examinations.
4. **Systematic:** it is well planned and designed in an orderly manner and done at short predetermined intervals. It is not episodic. Again, the procedure indicates explicitly what is to be measured, the instrument to be used and the type of trait or performance to be assessed.
5. **Diagnostic:** it provides reliable information about the learner and facilitates the identifications of strengths and weaknesses, individual difficulties and attention and ensures remediation of problems.
6. **Guidance-Oriented**: the formative and holistic of assessment provides information (feedback) to the teacher and the learner which helps the learner to discover and develop his potentialities. The learner therefore, knows his strengths and weaknesses which facilitates educational and vocational guidance and eventually leads to occupational/vocational congruence.

Merits of Continuous Assessment

1. Continuous assessment reduces fear and anxiety in students as the fear of failure in examination is reduced by the cumulative nature of assessment in this case.
2. Continuous assessment reduces examination malpractice since anxiety and fear that compel students to resort to foul means of passing examination associated with one-short examination is reduced by the continuous (cumulative) nature of assessment.
3. It discourages teaching to syllabus (narrow-curriculum). The involvement of the class teacher in the assessment which covers a wide range ensures the inclusion of relevant materials in the instruction programme that helps the total development of the learner.
4. It provides information to the class teacher on the strength and weaknesses of an educational programme for the necessary correction.
5. It helps in the development of an integrated personality as the assessment procedures touch on the cognitive, affective and the psychomotor domains.
6. Continuous assessment motivates learning by the short and predetermined nature of assessment which keeps students active and alert and therefore, reduces laxity and procrastination.
7. The diagnostic nature of continuous assessment through feedback promotes effective and healthy learning habits.

Weaknesses/Challenges of Continuous Assessment

Continuous assessment is not without its problems. Countries considering the introduction or operation of CA should weigh up the pros and the cons. The problems are both technical and practical, and some are more easily solved than others. The major problem areas of CA are:

i. Inadequate conceptualization
ii. Doubtful validity
iii. Inadequate structural and administrative support

Specific problems that may affect the implementation of CA schemes include:

I. Teachers may lack experience and expertise in CA. In particular, the quality of many classroom tests may be low, tending to negate gains in the validity of assessment made possible by the introduction of CA. possible solutions are the provision of adequate INSET support and/or the construction of item banks.

II. Teacher work load may be substantially increased by CA as teachers spend most of their time constructing test items, scoring and recording of grades. There is evidence from England (Pennychucick and Murphy, 1988) that teachers are prepared to make the necessary effort if they perceive the benefits to themselves and to their pupils of an innovatory assessment system. Nevertheless, schemes should be designed to take account of pressure on teachers (for example, in avoiding excessive demands of record keeping and reporting). Again, INSET can help.

III. Administration of CA within the school may not be straight forward. For instance, considering what to do when pupils are absent from CA test, or when a pupil transfers from one school to another as well as how to deal with normal aggregation and weighting of marks. Possible solutions are the establishments of an assessment committee within each school and clear instructions and documentation from the national body responsible for CA to all satisfactory administration.

a) There are several possible sources of unreliability in school-based assessment. They include administrative mistakes, teacher or assessor bids, conscious or unconscious (e.g. the "Halo" effect). The solution to this is constant vigilance to minimise these factors, but it should be remembered that there may be unreliability in external examinations too. Reliability can be increased by assessing on multiple occasions.

b) There is also the problem of comparability between classes within schemes and between schools. Methods of ensuring comparability usually involve some form of accreditation and or moderation. Moderation could be statistical, through visitation and by consensus. In the context of most developing countries, statistical moderation is probably the cheapest and easiest to apply but there is a danger if external examination is used as the reference test because of the backwash effect of external examinations will continue to dominate the school curriculum.

IV. There may be overload on pupils undertaking projects in several subjects simultaneously. Also, pupils from relatively wealthy backgrounds may be at an advantage as they may have greater access to resources and help their parents. Project work should be planned and given to pupils in a manageable manner as much as possible.

V. Inadequate materials and equipment pose problems to teachers. Stationary, computers and other accessories that may be necessary for effective continuous assessment may be lacking. Governments and the Ministry of Education should ensure the provision and supply of equipment to facilitate effective assessment in schools.

VI. Colleges of Education and universities engaged in pre-service teacher preparation should make CA an integral part of the curriculum by making it examinable to help would-be teachers acquire the expertise before entering the school environment.

CHAPTER THREE

GOALS AND LEARNING TARGETS OF INSTRUCTION

Introduction

Definition of terms:

Goal: a general aim or purpose.

Educational goals: they are human activities that contribute to the functioning of a society and acquired through learning. Educational goals are usually stated in broad terms and they give direction and purpose to the overall planning and execution of an educational activity. For example, the educational goal at the Basic level of Education includes literacy, numeracy and good citizenship. Educational goals are derived from societal needs. Such broad goals when established are organized into subject matter areas for study in the school system such as Mathematics, English, History, Science, etc.

Outcome: an outcome is what occurs as a result of an experience.

Educational outcomes: they are the product of learning experiences. In other words, they are the end result of learning. Some educational outcomes are knowledge, understanding, application of knowledge and understanding to situations, thinking skills, general skills attitude etc. Outcomes are also broad and we can break them down to specifics the knowledge on facts, knowledge of concepts, knowledge of terms etc. Because educational goals and outcomes are c broad, they cannot be achieved in a single instructional setting. For specific instructional settings we need to reduce broad educational settings we need to reduce broad educational goals and outcomes into specific objectives.

What is an Objective?

An Objective can be defined as an intended behavioural change that a learner is expected to exhibit after undergoing a learning experience. Tuckman (1976) perceives an objective as an intended and measured. Similarly, Kubiszyn and Borich (1987) commenting on instructional objectives indicate that they should be stated in clear and concise manner to indicate the skills which students will be expected to perform after a unit of instruction. They added that instructional objectives should include the level of proficiency to be demonstrated. Instructional objectives should be stated in observable, behavioural terms. A complete instructional objective should include an observable behaviour (action verb specifying the learning outcomes), any special conditions under which the behaviour most be displayed, and a performance level considered sufficient to demonstrate mastery. Stating instructional objective in measurable terms is advocated by behaviourists like Bloom (1956) and Tyler (1949). Some of the action verbs used in stating instructional objectives are; describe, recite, solve, draw, label, state, identify, classify, complete, name and explain.

Objectives, goals and aims are used inter-changeably but they are technically not the same,

Goals are said to be broad and long term;

Aims are broad and medium term and

Objectives are specific and short term. Objectives help teachers to evaluate the appropriateness of an instructional programme and test items.

The Importance of Learning Objectives for Classroom Assessment

i. Learning objectives make the planning for assessment procedure easier through the knowledge of specific outcomes.

ii. The selection, designing and construction of assessment instruments depends on knowing which specific should be assessed.

iii. Evaluating an existing assessment instrument becomes easier when specific outcomes are known.

iv. Learning objectives help to judge the content and relevance of an assessment procedure. Specific learning outcomes provide information for judgement.

Taxonomy of Educational Objectives

Taxonomy is a system of classification in which particular entities are arranged in accordance with clear guidelines and principles. In other words, taxonomy is an ordered classification system which has hierarchical schemes for classifying learning objectives into various levels of complexity. Instructional objectives have been categorized into three by matching them with test items.

They are:

Cognitive Domain,

Affective Domain and

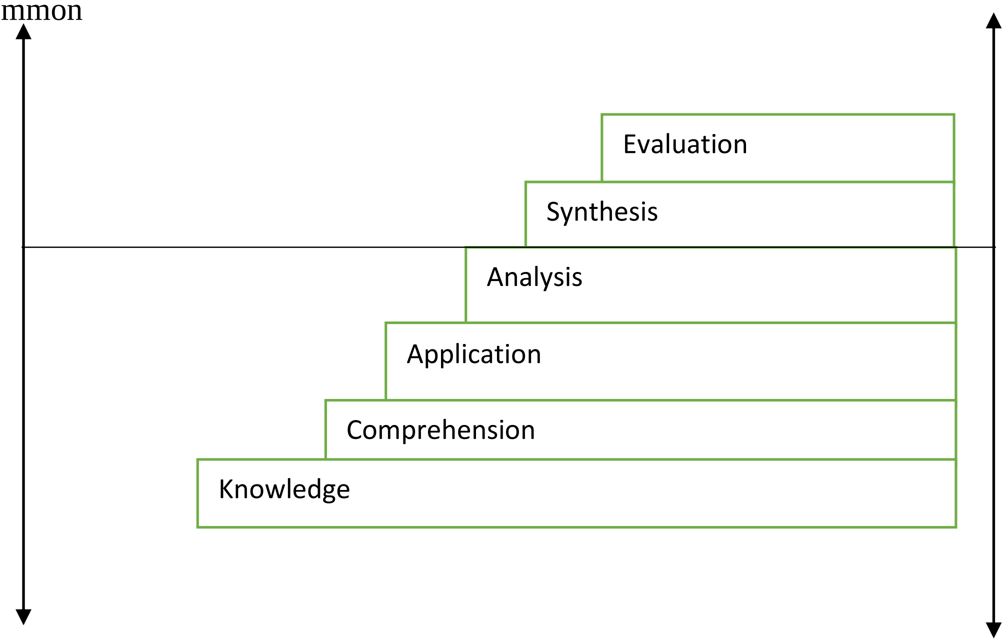
Psychomotor Domain

This method of categorising objectives into complex cognitive was derived by Bloom et al (1956). This is taxonomy of educational objectives for the cognitive domain and delineates six levels to the evaluations level which is the most complex. This is illustrated in the diagram below:

Taxonomy of Educational Objectives: the Cognitive Domain

Highest Level

Least common



Lowest level Most

Common

The three main domains of educational objectives are:

i. Cognitive (understanding, knowledge, facts)
ii. Affective (attitude, values, beliefs)
iii. Psychomotor (motor skills)

The affective domain borders on objectives that describe interest, values and attitudes. This also has five categories arranged hierarchically on levels of involvement-receiving, responding, valuing, organization and characterisation. The psychomotor domain also has varying categories outlined by educational psychologists. These are manipulative or motif skills, non-verbal communications and coordinated bodily activities.

The cognitive domain deals very much with educational objectives and test instruction. The cognitive domain is arranged in hierarchical manner based on the intellectual capabilities of the learner.

Knowledge: knowledge is defined as the remembering of previously learned material. It may involve the recall of materials from specific facts to complete theories. Knowledge represents the lowest level of learning outcomes in the cognitive domain. Objectives at the knowledge level require students to remember. Test items ask the students to recall or recognise facts, terminology, problem-solving strategies and rules. Some of the illustrative verbs used at the knowledge level are; define, describe, identify, list, label, state, outline etc. Questions that can be asked could be: How much is…..Who is…. When was…..How did… e.g. what is the capital of Ghana?

Comprehension: comprehension is defined as the ability to grasp the meaning of material. This may be shown by translating materials from one form to another or by interpreting material or summarizing and estimating future trends. Comprehension goes beyond the simple remembering of materials or facts. Objectives at this level require some understanding. Test items require the student to translate, restate what has been read, to see connections or relationships among parts of communication, interpret or generate conclusions or consequences from the information. Illustrative verbs that can be used at the comprehension level are: illustrate, classify, estimate, distinguish, explain, rewrite, summarise. Paraphrase, etc. Questions set at this level could be: Paraphrase in your own words, Give an example of….How are these ideas similar to….? Etc.

Application: application refers to the ability to use learned material in new and concrete situations. This may include the adoption of such things as rules, methods, concepts, principles, laws, theories, etc. Learning outcomes in this area require a higher level of understanding than those under comprehension. Objectives at this level require the students to use previously acquired information in a setting other than that in which it was learned. Application differs from comprehension in that application questions present the problem in a different and often applied context. Illustrative verbs that can be used at the application level are: compare, discover, modify, differentiate, apply, compute, change, etc. Questions asked at this level could be: Apply the formula to the following problem…..What would happen if….? etc.

Analysis: refers to the ability to break down material into its component parts so that its organizational structure may be understood. This may include the identification of the parts, analysis and recognition of the organizational principles involved. Learning outcomes here represent a higher intellectual level than comprehension and application because they require an understanding of both the content and the structural form of the material. Objectives written at this level require the student to identify logical errors or to differentiate among facts, opinions, assumptions, hypotheses or conclusions. Some of the illustrative verbs that can be used at the analysis level are: break down, discriminate, distinguish, point out, analyse devise, etc. Questions set at the analysis are: List the basic assumptions….., separate the major and minor themes…., Distinguish between theory and facts, etc.

Synthesis: synthesis refers to the ability to put parts together to form a new whole. This may involve the production of a unique communication (theme or speech) a plan of operations (research proposal), or a set if abstract relations (scheme for classifying information). Learning outcomes in this area stress creative behaviours with major emphasis on the formulation of new patterns of structures. Objectives written at this level require the student to judge in unique way logical and factual information. Some of the illustrative verbs used stating specific learning outcomes are: categorise, combine, compile, explain, rearrange, reconstruct, relate, synthesize, etc. Questions that can be asked at the synthesis level include: Write an essay proposing a new solution to the problem of….., Describe the three major theories and show how they may be combined in enhancing learning.

Evaluation: evaluation is concerned with the ability to judge the value of material (statement, novel, poem, research report, etc.) for a given purpose. The judgements are to be based on definite criteria. These may be internal criteria (organization) or external criteria (relevance to the purpose) and the student may determine the criteria or be given them. Learning outcomes in this area are highest in the cognitive hierarchy because they contain elements of all the other categories, plus value judgements based on clearly defined criteria. Instructional objectives written at this level require the student to form judgement about the value or worth of methods, ideas, people or products that have a specific purpose. Instructional objectives, as stated earlier are matched with test items and objectives in test construction (test blue print). Each level is an extension of all previous levels. Illustrative verbs for stating specific learning outcomes at the evaluation level are: appraise, assess, compare, contrast, discriminate, justify, explain, summarise, etc. Some of the questions that can be asked at the evaluation level include: Evaluate the recent decisions by the electoral commission regarding electoral procedures to be adopted in the impending elections, Write a careful critique of theory X and theory Y and justify your conclusions, etc.

The Affective Domain

The affective domain describes our feelings, likes and dislikes, our experiences as well as the resulting behaviours (reactions). It is demonstrated by behaviours indicating attitudes of awareness, interest, attention, concern, responsibility, ability to listen and respond in interactions with others and ability to demonstrate those attitudinal characteristics or value which are appropriate to the test situation and the field of study. Krathwohl, Bloom and Masia

(1964) proposed a five level taxonomy of objectives in the affective domain. The taxonomy was developed around organised levels of commitment. The levels are:

1. Receiving
2. Responding
3. Valuing
4. Organising
5. Characterising by value

Receiving: this refers to willingness to accept or to attend to a particular phenomena or stimuli (classroom activities, textbooks, assignments, etc.) from the teaching standpoint; receiving is concerned with getting, holding and directing the students' attention. Receiving has been divided into three subcategories: awareness (being conscious of something), willingness to receive (being willing to tolerate a given stimulus) and controlled or selected attention. Learning outcomes in this area range from the simple awareness that a thing exists to selective attention on the part of the learner.

Responding: this refers to active participation on the part of the student. The student is sufficiently motivated not to just to be attentive but is actively attentive. It indicates the desire that the student has become sufficiently involved in or committed to the lesson or subject. Learning outcomes involve obedience or compliance or willingness.

Valuing: here the student sees worth or value in the activity. An important element of valuing is that the student is motivated not by the teacher to comply or obey by his underlying value guiding the behaviour. Learning outcomes are concerned with behaviour that is consistent to make the value clearly identifiable.

Organising: this entails bringing together complex values or possible disparate values or resolving conflicts and beginning to build an internally consistent value system. The emphasis is on comparing, relating and synthesizing values. A typical example would be recognising the need for balance between freedom and responsible behaviour.

Characterising: this implies a pervasive, consistent and predictable behaviour. Instructional objectives are concerned with the students' general patterns of adjustment (personal, social, emotional). A typical example would be how he/she cooperates in groups.

The Psychomotor Domain

This refers to educational outcomes that focus on motor skills and perceptual processes. It includes physical movements, coordination, and use of motor skills. According to Seel and Glasgow (1990) Harrow's taxonomy of the psychomotor domain is organised according to the degree of coordination including involuntary responses as well as learned capabilities. Simple reflexes are at the lowest level of the taxonomy while complex neuromuscular coordination makes up the highest levels.

The six main categories of objectives of Harrow's taxonomy are:

1. Reflex movements: actions elicited without learning in response to some stimuli (flexion, stretch, postural adjustments).

2. Basic fundamental movement – inherent movement patterns that are formed from a combination of reflex movements and are the basis of complex skilled movements (walking, running, pushing, gripping, grasping, etc.).
3. Perceptual abilities – interpretation of stimuli from various modalities providing information for an individual to make adjustments to his/her environment (coordinated movements such as jumping a rope and catching).
4. Physical activities – this requires endurance, strength and vigour (all activities which requires strenuous effort for long periods of time, long distance running; muscular exertion).
5. Skilled movements – this refers to communication through bodily movement ranging from facial expression through sophisticated choreographies (body postures, gestures, etc.).

Review Questions

1. Define educational goal.

1b. Using three points only, explain the relationship between educational goals and outcome.

2. Educational objectives are important to classroom assessment; support the assertion using five points only.

2b. Assessment can be wrongly conducted without proper cognizance of the educational goals, justify.

3. Write an elaborate note on the taxonomy of educational objectives.

3b. List and explain the three domains as pointed out by Benjamin (195

CHAPTER FOUR

THE STAGES IN CLASSROOM TEST CONSTRUCTION

Introduction

Testing plays an important role in education. It is as important as teaching and learning. The use of test at all levels of our educational system that is, from the nursery stage to the university; necessitate the need to take a critical look at tests and how they are constructed, administered and interpreted. According to Etsey (2001) the principal stages involved in classroom testing are:

- Constructing the test
- Administering the test
- Scoring the test
- Analysing the test result

Constructing the test

Test construction like any other purposeful activity needs to be adequately planned and executed. There are eight steps to follow in the construction of a good classroom test. These are referred to as principles of test construction. These include:

Defining the purpose of the test – the basic question to ask is "why am I testing?" Several purposes are served by classroom tests and the teacher has to be clear on the purpose of the test. Test items must be related to teacher's classroom instructional objectives. This forms part of the planning stage so the teacher has to answer other questions like why is the test being given at this time in the course? Who will take the test? Have the testees been informed? How will the scores be used?

Determining the item format to use – the choice of format must be appropriate for testing particular topics and objectives. Here the teacher needs to list the objectives of the subject matter for which the test is being constructed and the main topics covered or to be covered. The test items could be essay, objective or performance type. It is important at times to use more than one format for a single test. Mwehrens and Lehmann (2001) have suggested eight factors to consider in the choice of appropriate format. These include:

- The purpose of the test
- The time available to prepare and score the test
- The number of students to be tested
- The skill to be tested
- The difficulty desired
- Physical facilities that are available (like reproduction materials)
- Age of the pupils
- Test constructor or teacher's skills in writing the different type of items

Preparing a test Blue print or Table of Specification

Just like a blue print used by a builder to guide building construction, the test blue print is used by a teacher to guide in test construction. It ensures that the teacher does not overlook details considered essential to a good test. Specifically, it ensures that a test will sample whether learning has taken place across the range of content areas covered in class and cognitive processes considered important. Here the teacher has to determine what topics or units the test will cover as well as what knowledge, skills and attitudes to measure. This he can do by asking himself/herself questions like: what is it that I wish to measure?

Below are examples of a text blue print for a unit of instruction.

Example 1

Table of Specification for a fifty item test in Geography

Topics	Knowledge of terms	Understanding of principles	Application of Principles	Interpretation of charts	Total
Drainage	2	3	2	3	10
Climate	3	4	3	4	14
Relief	4	3	2	5	14
Vegetation	3	3	3	3	12
Total	12	13	10	15	50

Example 2

A teacher is set to construct an objective question of 120 items in Agricultural Science.

Topics	Understanding	Application	Analysis	Total
Mixed Cropping	10	10	10	30
Mixed Farming	10	20	20	50
Yam & Plantain	5	10	15	30
Forestry	5	5	0	10
Total	30	45	45	120

Advantages of the Test Blue Print

The test blue print is important for a number of reasons. Firstly, the test blue print helps one to plan adequately to set items to cover all the topics treated as well as the behaviours. That is to say, plunging into item writing without the specification table is likely to produce a test which may be lopsided. Secondly, the procedure facilitates meaningful weighting of the items in each cell of the table in accordance with the importance attached to them. Thirdly, the blue print ensures content validity of the test. Content validity in this sense means the items adequately sample the universe content. This is achieved through the selection and writing of appropriate items in both behavioural and content areas.

Writing the individual items

This is the phase at which specific items are written in accordance with the table of test specification or blue print. Whichever test items are being constructed should follow the basic principles laid down for them. For convenience the original draft of items should exceed the number of items intended for the test. The rationale behind should exceed the number of items intended for the test. The rationale behind this is that after eliminating unsuitable items enough number of items could be left for the final test. The following principles must be considered when writing the individual items:

- Keep the table of specification before you and refer to it as you write test items
- Items must match instructional objectives
- Formulate well-defined items that are not vague and ambiguous and free from grammatical and spelling errors
- Avoid needlessly complex sentences
- Write the test items simply and clearly
- Prepare more items than you will actually need
- The task to be performed and the type of answers required should be clearly defined
- Include questions of varying difficulty
- Avoid textbook or stereotyped language.

Reviewing the items

In reviewing the items one has to check on whether each item measures the specific learning outcome and subject-matter content it is supposed to measure. A check is also made on any ambiguity of the items and whether the items are free from irrelevant clues and each item is edited for its representativeness and clarity. Bad items are removed or eliminated.

Preparing the Scoring Key or the Marking Scheme

Having constructed a test that is both valid and reliable, it is necessary to produce a marking or scoring scheme that will enable the tester to evaluate the responses as fairly and accurately as possible. Frith and Macintosh (2001) recommend the use of the following checklist for preparing a marking scheme.

- Are suggested answers appropriate to the questions?
- Are suggested answers technically and/or numerically correct?
- Does the scheme embraces every point required by the question and allocates marks for each point?
- Are the marks allocated strictly according to knowledge and abilities which the questions require the testees to demonstrate?
- Is there adequate provision for alternative answers?
- Are marks commensurate to degree of difficulty of questions and time needed to answer them?
- Is time allowance appropriate for work required?
- Is scheme sufficiently broken down to allow marking to be as objective as possible?
- Is the totalling of marks correct?

Writing Directions

This entails writing clear, concise and specific directions or instructions. Directions must include number of items to respond to, mode of responding, amount of time available, credit for orderly presentation of material and mode of identification of respondent.

Evaluating the Test

A test should be evaluated for its worth before administration. The main criteria in this direction are validity, practicality and efficiency. In considering validity, the test constructor finds out whether the items are measuring what they are supposed to measure. He should ask the question: Are the items representative of the content and the behaviours they are intended to measure?

Clarity refers to how the items are stated and phrased taking cognisance of the ability and level of the testees.

Practicality on the other hand is concerned with the necessary materials and the time allotted to the test.

Administering the Test

Test administration is as important as its construction. According to Kubiszyn and Borich (1987) the following principles must be observed in administering test:

1. Candidates must be made aware of rules and regulations governing the conduct of test. Penalties for malpractice such as cheating should be clearly spelt out.

2. The sitting arrangement must allow enough space so that candidates may not copy each others' work.
3. Adequate ventilation and lighting is expected in the lighting room
4. Candidates should start the test promptly and stop on time.
5. Announcement must be made about the time at regular intervals.
6. Invigilators are expected to stand at a point where they could view all students.
7. They should once in a while move among the students to check malpractices
8. Such movements should not disturb the students
9. Invigilators must be vigilant
10. Threatening behaviours should be avoided by the invigilators. Speeches like, if you don't write fast you will fail are threatening. Students should be made to feel at ease.
11. The testing environment should be free from distractions.
12. Noise should be kept at a very low level if it cannot be eliminated or removed
13. Interruptions within and outside the classroom should be reduced.
14. Expect and prepare for emergency.

Review Questions

1. List and explain the stages of classroom test construction
2. What is a test blue print?
3. Why is it necessary to prepare a scoring guide?
4. Outline five out of the principles that should be observed when administering test.
5. Explain the psychometric properties of test instrument

CHAPTER FIVE

THE MAJOR TYPES OF TEST FORMAT

Standardized Test

A **standardized test** is a test that is administered and scored in a consistent, or "standard", manner. Standardized tests are designed in such a way that the questions, conditions for administering, scoring procedures, and interpretations are consistent and are administered and scored in a predetermined, standard manner.

Any test in which the same test is given in the same manner to all test takers is a standardized test. Standardized tests need not be high-stakes tests, time-limited tests, or multiple-choice tests. The opposite of a standardized test is a non-standardized test. Non-standardized testing gives significantly different tests to different test takers, or gives the same test under significantly different conditions (e.g., one group is permitted far less time to complete the test than the next group), or evaluates them differently (e.g., the same answer is counted right for one student, but wrong for another student).

Standardized tests are perceived as being more fair than non-standardized tests. The consistency also permits more reliable comparison of outcomes across all test takers.

Design and scoring

Some standardized testing uses multiple-choice tests, which are relatively inexpensive to score, but any form of assessment can be used.

Standardized testing can be composed of multiple-choice questions, true-false questions, essay questions, authentic assessments, or nearly any other form of assessment. Multiple-choice and true-false items are often chosen because they can be given and scored inexpensively and quickly by scoring special answer sheets by computer or via computer-adaptive testing. Some standardized tests have short-answer or essay writing components that are assigned a score by independent evaluators who use rubrics (rules or guidelines) and benchmark papers (examples of papers for each possible score) to determine the grade to be given to a response. Most assessments, however, are not scored by people; people are used to score items that are not able to be scored easily by computer (i.e., essays). For example, the Graduate Record Exam is a computer-adaptive assessment that requires no scoring by people (except for the writing portion).

Scoring issues

Human scoring is often variable, which is why computer scoring is preferred when feasible. For example, some believe that poorly paid employees will score tests badly. Agreement between scorers can vary between 60 to 85 percent, depending on the test and the scoring session. Sometimes states pay to have two or more scorers read each paper; if their scores do not agree, then the paper is passed to additional scorers.

Open-ended components of tests are often only a small proportion of the test. Most commonly, a major test includes both human-scored and computer-scored sections. These major tests do not measure the student's overall ability in learning.

Scoring

There are two types of standardized test score interpretations: a norm-referenced score interpretation or a criterion-referenced score interpretation.

- **Norm-referenced score interpretations** compare test-takers to a sample of peers. The goal is to rank students as being better or worse than other students. Norm-referenced test score interpretations are associated with traditional education. Students who perform better than others pass the test, and students who perform worse than others fail the test.
- **Criterion-referenced score interpretations** compare test-takers to a criterion (a formal definition of content), regardless of the scores of other examinees. These may also be described as standards-based assessments, as they are aligned with the standards-based education reform movement. Criterion-referenced score

> interpretations are concerned solely with whether or not this particular student's answer is correct and complete. Under criterion-referenced systems, it is possible for all students to pass the test, or for all students to fail the test.

Either of these systems can be used in standardized testing. What is important to standardized testing is whether all students are asked equivalent questions, under equivalent circumstances, and graded equally. In a standardized test, if a given answer is correct for one student, it is correct for all students. Graders do not accept an answer as good enough for one student but reject the same answer as inadequate for another student.

Standards

The considerations of validity and reliability typically are viewed as essential elements for determining the quality of any standardized test. However, professional and practitioner associations frequently have placed these concerns within broader contexts when developing standards and making overall judgments about the quality of any standardized test as a whole within a given context.

Evaluation standards

In the field of evaluation, and in particular educational evaluation, the Joint Committee on Standards for Educational Evaluation has published three sets of standards for evaluations. The Personnel Evaluation Standards was published in 1988, The Program Evaluation Standards (2nd edition) was published in 1994, and The Student Evaluation Standards was published in 2003.

Each publication presents and elaborates a set of standards for use in a variety of educational settings. The standards provide guidelines for designing, implementing, assessing and improving the identified form of evaluation. Each of the standards has been placed in one of four fundamental categories to promote educational evaluations that are proper, useful, feasible, and accurate. In these sets of standards, validity and reliability considerations are covered under the accuracy topic. For example, the student accuracy standards help ensure that student evaluations will provide sound, accurate, and credible information about student learning and performance.

Testing standards

In the field of psychometrics, the Standards for Educational and Psychological Testing place standards about validity and reliability, along with errors of measurement and issues related to the accommodation of individuals with disabilities. The third and final major topic covers standards related to testing applications, credentialing, plus testing in program evaluation and public policy.

Advantages

One of the main advantages of standardized testing is that the results can be empirically documented; therefore, the test scores can be shown to have a relative degree of validity and reliability, as well as results which are generalizable and replicable. This is often contrasted with grades on a school transcript, which are assigned by individual teachers. It may be difficult to account for differences in educational culture across schools, difficulty of a given teacher's curriculum, differences in teaching style, and techniques and biases that affect grading. This makes standardized tests useful for admissions purposes in higher education, where a school is trying to compare students from across the nation or across the world.

Another advantage is aggregation. A well designed standardized test provides an assessment of an individual's mastery of a domain of knowledge or skill which at some level of aggregation will provide useful information. That is, while individual assessments may not be accurate enough for practical purposes, the mean scores of classes, schools, branches of a company, or other groups may well provide useful information because of the reduction of error accomplished by increasing the sample size.

Standardized tests, which by definition give all test-takers the same test under the same (or reasonably equal) conditions, are also perceived as being more fair than assessments that use different questions or different conditions for students according to their race, socioeconomic status, or other considerations.

Effects

Standardized tests are used in every school around the United States in nearly every grade level. These tests are referred to as high stakes testing and come with many names such as Iowa Tests of Basic Skills, ACT, and SAT; however they all serve the same purpose. All of the testing given in this manner is used to judge the performance of the nations' students and determine their proficiency amongst their peers. Teachers are also measured based on students' results on standardized tests. If a student is found to be less than average it is said to reflect on the teacher and his/her abilities. It is with these perceptions that the United States puts its students in danger. The other problem with SAT, and ACT, is that the tests don't test people who are talented in other domains such as, art, athletics, creative writing and many others.

Testing in schools is used in a wide variety of ways: placing children into learning groups, ranking schools amongst others in the region, state, and nation, and creating a visual for where the United States as a whole is heading. What surprises many is standardized testing may also be a way schools determine merit pay for teachers. Teachers in all grade levels are encouraged to shape their classroom around the upcoming test in hopes that their students outperform others. The effects of this kind of teaching are not beneficial to anyone, except potentially the teacher whose students do well. In the article "Standardized Testing and Its

Victims" author Alfie Kohn states, “Schools across the country are cutting back or even eliminating programs in the arts, recess for young children, electives for high schoolers, class meetings (and other activities intended to promote social and moral learning), discussions about current events (since that material will not appear on the test), the use of literature in the early grades (if the tests are focused narrowly on decoding skills), and entire subject areas such as science (if the tests only cover language arts and math)” (Kohn 1).

Cutting back on real classroom learning is taking its toll on teachers who were genuinely interested in reaching out to the youth and helping them grow. “Many educators are leaving the field because of what is being done to schools in the name of ‘accountability’ and ‘tougher standards’” (Kohn 1). Teachers are becoming displeased with the field and the ones who genuinely care about student growth are abdicating their roles as educators simply because it has become a twisted version of what it used to be. Prospective educators are now second guessing their choice of careers due to the pressure that will be put upon them to produce the high test scores that matter the most to their potential employers.

With all the stress teachers and administrators are under it would be unreasonable to think it does not rub off on the students as well. Some schools go as far as putting up a visual aid to show where their students fall compared to their classmates. This allows the students to see which of their classmates are proficient, which can be embarrassing for students who fall below the given line. Teachers have many chances to attain their merit pay; a student may only have one chance to pass a test allowing them to move to the next grade level. A single test can determine the outcome of a student’s entire educational career, not doing well can be a detriment to their self-esteem. A fourth grader does not need to feel devalued because of a test, they are still developing at an unsteady pace and expecting them all to fall into a neat category of proficiency is simply not acceptable. “Virtually all specialists condemn the practice of giving standardized tests to children younger than 8 or 9 years old” (Kohn 1).

Students feel the pressure put upon them in a completely different way than an adult would. When asked if students feel the pressure to achieve higher scores on standardized tests educator, Ashley Grossman, states, “I don't think they fully understand it. They feel intimidated and stressed around test time but some of them are like that with any test” (Grossman). Children cannot feel pressured constantly without it having a negative impact on their emotional and potentially physical state. Stress impacts children much the same as it can an adult, sometimes more severe. According to author Josh Ska, “Symptoms of too much stress are usually very evident in children, although they might be mistaken for being rebellious or difficult. A child who frequently blows up over nothing may be having problems at school or at home which are causing chronic stress. Another possible sign of stress is jumpiness or nervousness and poor concentration, which may affect schoolwork. Children who are stressed out may also stop eating or get sick more frequently. The constant adrenaline rush can keep them awake at night and you might notice that your child seems to have insomnia, although she is exhausted. Stomachaches are a common complaint among children suffering from this problem, as are bowel problems and headaches” (Ska 1).

Machines scoring tests do not lessen the bias of testing in any way. For the multiple choice problems it is a simple right or wrong; however, computers have been used to score essay portions as well. Criterion is the name of grading software to determine the proficiency of a student's writing abilities. The University of California was considering using this software to determine if students were eligible to skip a writing course, which the instructor was opposed to. In order to prove his point the instructor, Andy Jones, took a letter of recommendation he had written to score it. Author Alain Jehlen notes, "[He] replaced the student's name with a few words from a Criterion writing prompt, and substituted 'chimpanzee' for every 'the.' Criterion loved the result, calling it 'cogent' and 'well-articulated'" (Jehlen 3). If changing a single word and creating a nonsensical paper was scored so well, then one can only imagine what kinds of writing samples this machine was letting through and calling "wonderful."

Claims have been brought against standardized tests in court due to bias. The legality of a test is based on seven factors: disparate impact (unjustified adverse impact on members of a protected class), validation studies (tests must be validated), state interest, notice and implementation (due process), judicial deference (deferring to a professional educator), remediation and retakes (the amount of remediation offered and the number of retakes), and if the test is homemade. With all these thinpgs taken into account there are still several cases where a test was found to be biased and was ruled unreliable by the judicial system.

Debra P. v. Turlington is a case documented where a standardized test was challenged on the basis of racial bias. The SSAT II was claimed to be unconstitutional in the way it was able to deny the students who failed the test high school diplomas. The students in question were provided inadequate notice of the graduation requirements and not given adequate time to prepare themselves for the test. Shelly Mack notes in her research, "The court found that the SSAT II had a clear disproportionate impact on African American students, and noted that Florida intended to discriminate against African American children between 1967-1971 (when the current graduating class was in school under the dual system)" (Mack 2). The state admitted to knowingly discriminating against these children so they would not graduate.

Crump v. Gilmer Independent School District is another case in which graduation was hanging on a single test. Three students had all failed the Texas Assessment of Academic Skills Examination (TAAS). Two of these students had successfully completed all other graduation requirements, while the third student had not. TAAS had only become a graduation requirement in 1991, two years before this case was presented in court, making the argument the students had insufficient time to prepare for the test, as per Debra P. v. Turlington stated that there be at least four to six years of preparation time from its announcement before a new process could be implemented. The two students who had successfully completed all other graduation requirements were granted their diplomas, while the third student was not. It was deemed that the third student's denial was constitutional because there seemed to be no effort on his part from an academic standpoint.

Despite the biases of standardized testing the question remains of if these tests even show actual learning or learning potential of a student. The answer seems to be a resounding “no” from all sources. The number of guesses that are marked correct do not indicate the student has mastered the skill in question; more often than not they had a one in four chance of being correct. Wrong answers are measured correctly, as the student clearly did not know the material, but the correct answers are not indicative of knowledge. A correct answer can point to two other possibilities than mastery of skill; “A correct answer can be achieved using memorization without any profound understanding of the underlying content or conceptual structure of the problem posed” (“Standardized Test” 2) or simply a blind guess resulting in a positive outcome.

However there are positive aspects to standardized tests; specifically for young children. The purpose of standardized tests for young children is to identify developmental delays and to evaluate a young child’s development. The standardized tests used for young children are screening tests, diagnostic tests, language tests, and achievement tests. A screening test is used in order to detect an indication of a developmental problem—it identifies if a problem needs to be investigated further. A diagnostic test is done if a child has already been screen tests and indicates further evaluation. Diagnostic tests are designed to assess developmental problems related to learning disabilities. A language test is often administered to students who are considered at-risk. Language tests determine if a student would benefit from a language enrichment program. The achievement test was designed for children in the Head Start program and was introduced by the George W. Bush administration (Wortham, 2008). Overall standardized tests are not solely used to assess young children but is a great way to detect developmental problems in young children.

One proponent of standardized testing is the No Child Left Behind Act (NCLB). This bill supports standards-based education reform, “the belief that setting high standards and establishing measurable goals can improve individual outcomes in education” ("No Child Left Behind Act” 1). NCLB is what set the testing frenzy of the United States in motion. The national government felt it had to step in and take over where state governments had been failing. All the act seemed to do in reality was set up a system of incentives for educators if test results improved. “The system of incentives and penalties sets up a strong motivation for schools, districts, and states to manipulate test results. For example, schools have been shown to employ ‘creative reclassification’ of drop-outs (to reduce unfavorable statistics)” (“No Child Left Behind Act” 3).

NCLB has encouraged the “teach to the test” method more and more schools have put into place, which leads to students not properly interpreting the test materials despite having been trained for them. Teachers are taught to anticipate what will be on the test and teach the students only that material, leading to students having vague, if any, understanding of any other concepts they may need. “Many teachers who practice ‘teaching to the test’ actually misinterpret the educational outcomes the tests are designed to measure. On two state tests (New York State and Michigan) and the National Assessment of Educational Progress (NAEP) almost two-thirds of eighth graders missed math word problems that required an application of the Pythagorean theorem to calculate the distance between two points" ("No Child Left Behind Act" 3).

Standardized testing is a detriment to students, affecting them psychologically, emotionally, and intellectually. Their self-esteem is lowered when they do not receive scores they may be aiming for, or when they do not do as well as their classmates. Students are put under undue stress to outperform, simply because teachers are put under stress to make sure their students do well. Important programs are slowly being taken from schools in order to focus on "teaching to the test." Students should be learning the social and moral skills that come with being in particular extracurricular groups or elective classes along with their basic subjects, but with classrooms being test oriented some of the most important real world skills are being taken away from them. This sends up a very real red flag for the future about the kinds of people that will be running the United States. They may be goal oriented, but being people oriented is just as important quality to have.Part of the blame falls to the educators, administrators, and states for not speaking out, but most of the blame lies with the government for increasing standards in a way that is unhealthy.

Public policy

Standardized testing is used as a public policy strategy to establish stronger accountability measures for public education. While the National Assessment of Education Progress (NAEP) has served as an educational barometer for some thirty years by administering standardized tests on a regular basis to random schools throughout the United States, efforts over the last decade at the state and federal levels have mandated annual standardized test administration for all public schools across the country.

The idea behind the standardized testing policy movement is that testing is the first step to improving schools, teaching practice, and educational methods through data collection. Proponents argue that the data generated by the standardized tests act like a 'report card' for the community, demonstrating how well local schools are performing. Critics of the movement, however, point to various discrepancies that result from current state standardized testing practices, including problems with test validity and reliability and false correlations (see Simpson's paradox).

Critics charge that standardized tests became a mandatory curriculum placed into schools without public debate and without any accountability measures of its own. Many feel this ignores basic democratic principles in that control of schools' curricula is removed from local school boards, which are the nominal curricular authority in the U.S. While some maintain that it would be preferable to simply introduce mandatory national curricula, others feel that state mandated standardized testing should stop altogether in order that schools can focus their efforts on instructing their students as they see fit.

Critics also charge that standardized tests encourage "teaching to the test" at the expense of creativity and in-depth coverage of subjects not on the test. Multiple choice tests are criticized for failing to assess skills such as writing. Furthermore, student's success is being tracked to a teacher's relative performance, making teacher advancement contingent upon a teacher's success with a student's academic performance. Ethical and economical questions arise for teachers when faced with clearly underperforming or underskilled students and a standardized test.

Disadvantages and criticism

Standardized tests are useful tools for assessing student achievement, and can be used to focus instruction on desired outcomes, such as reading and math skills. However, critics feel that overuse and misuse of these tests harms teaching and learning by narrowing the curriculum. According to the group FairTest, when standardized tests are the primary factor in accountability, schools use the tests to narrowly define curriculum and focus instruction. FairTest says that negative consequences of test misuse include narrowing the curriculum, teaching to the test, pushing students out of school, driving teachers out of the profession, and undermining student engagement and school climate. Critics say that "teaching to the test" disfavors higher-order learning. While it is possible to use a standardized test without letting its contents determine curriculum and instruction, frequently, what is not tested is not taught, and how the subject is tested often becomes a model for how to teach the subject.

Uncritical use of standardized test scores to evaluate teacher and school performance is inappropriate, because the students' scores are influenced by three things: what students learn in school, what students learn outside of school, and the students' innate intelligence. The school only has control over one of these three factors. Value-added modeling has been proposed to cope with this criticism by statistically controlling for innate ability and out-of-school contextual factors. In a value-added system of interpreting test scores, analysts estimate an expected score for each student, based on factors such as the student's own previous test scores, primary language, or socioeconomic status. The difference between the student's expected score and actual score is presumed to be due primarily to the teacher's efforts.

Supporters of standardized testing respond that these are not reasons to abandon standardized testing in favor of either non-standardized testing or of no assessment at all, but rather criticisms of poorly designed testing regimes. They argue that testing does and should focus educational resources on the most important aspects of education — imparting a pre-defined set of knowledge and skills — and that other aspects are either less important, or should be added to the testing scheme.

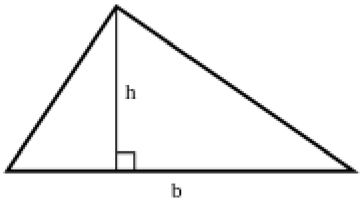
In her book, Now You See It, Cathy Davidson criticizes standardized tests. She describes our youth as "assembly line kids on an assembly line model," meaning the use of standardized test as a part of a one-size-fits-all educational model. She also criticizes the narrowness of skills being tested and labeling children without these skills as failures or as students with disabilities. Widespread and organized cheating has been a growing culture in today's reformation of schools.

Education theorist Bill Ayers has commented on the limitations of the standardized test, writing that "Standardized tests can't measure initiative, creativity, imagination, conceptual thinking, curiosity, effort, irony, judgment, commitment, nuance, good will, ethical reflection, or a host of other valuable dispositions and attributes. What they can measure and count are isolated skills, specific facts and function, content knowledge, the least interesting and least significant aspects of learning."

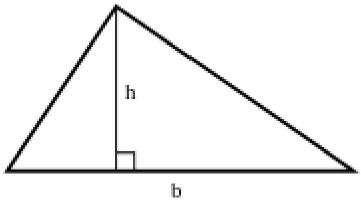
Another disadvantage to Standardized Tests is the cost. It has been reported that the United States spends about 1.7 billion dollars annually on these tests.

Scoring information loss

A test question might require a student to calculate the area of a triangle. Compare the information provided in these two answers.



Area = 7.5 cm^2



Base = 5 cm; Height = 3 cm
Area = $^1/_2$(Base × Height)
= $^1/_2$(5 cm × 3 cm)
= 7.5 cm^2

The first shows scoring information loss. The teacher knows whether the student got the right answer, but does not know how the student arrived at the answer. If the answer is wrong, the teacher does not know whether the student was guessing, made a simple error, or fundamentally misunderstands the subject.

When tests are scored right-wrong, an important assumption has been made about learning. The number of right answers or the sum of item scores (where partial credit is given) is assumed to be the appropriate and sufficient measure of current performance status. In addition, a secondary assumption is made that there is no meaningful information in the wrong answers.

In the first place, a correct answer can be achieved using memorization without any profound understanding of the underlying content or conceptual structure of the problem posed. Second, when more than one step for solution is required, there are often a variety of approaches to answering that will lead to a correct result. The fact that the answer is correct does not indicate which of the several possible procedures were used. When the student supplies the answer (or shows the work) this information is readily available from the original documents.

Second, if the wrong answers were blind guesses, there would be no information to be found among these answers. On the other hand, if wrong answers reflect interpretation departures from the expected one, these answers should show an ordered relationship to whatever the overall test is measuring. This departure should be dependent upon the level of psycholinguistic maturity of the student choosing or giving the answer in the vernacular in which the test is written.

In this second case it should be possible to extract this order from the responses to the test items. Such extraction processes, the Rasch model for instance, are standard practice for item development among professionals. However, because the wrong answers are discarded during the scoring process, attempts to interpret these answers for the information they might contain is seldom undertaken.

Third, although topic-based subtest scores are sometimes provided, the more common practice is to report the total score or a rescaled version of it. This rescaling is intended to compare these scores to a standard of some sort. This further collapse of the test results systematically removes all the information about which particular items were missed.

Thus, scoring a test right–wrong loses 1) how students achieved their correct answers, 2) what led them astray towards unacceptable answers and 3) where within the body of the test this departure from expectation occurred.

This commentary suggests that the current scoring procedure conceals the dynamics of the test-taking process and obscures the capabilities of the students being assessed. Current scoring practice oversimplifies these data in the initial scoring step. The result of this procedural error is to obscure of the diagnostic information that could help teachers serve their students better. It further prevents those who are diligently preparing these tests from being able to observe the information that would otherwise have alerted them to the presence of this error.

A solution to this problem, known as Response Spectrum Evaluation (RSE), is currently being developed that appears to be capable of recovering all three of these forms of information loss, while still providing a numerical scale to establish current performance status and to track performance change.

This RSE approach provides an interpretation of the thinking processes behind every answer (both the right and the wrong ones) that tells teachers how they were thinking for every answer they provide. Among other findings, this chapter reports that the recoverable information explains between two and three times more of the test variability than considering only the right answers. This massive loss of information can be explained by the fact that the "wrong" answers are removed from the test information being collected during the scoring process and is no longer available to reveal the procedural error inherent in right-wrong scoring. The procedure bypasses the limitations produced by the linear dependencies inherent in test data.

Testing bias occurs when a test systematically favors one group over another, even though both groups are equal on the trait the test measures. Critics allege that test makers and facilitators tend to represent a middle class, white background. Critics claim that standardized testing match the values, habits, and language of the test makers. However, being that most tests come from a white, middle-class background, it is important to note that the highest scoring groups are not people of that background, but rather tend to come from Asian populations.

Not all tests are well-written, for example, containing multiple-choice questions with ambiguous answers, or poor coverage of the desired curriculum. Some standardized tests include essay questions, and some have criticized the effectiveness of the grading methods. Recently, partial computerized grading of essays has been introduced for some tests, which is even more controversial.

Educational decisions

Test scores are in some cases used as a sole, mandatory, or primary criterion for admissions or certification. For example, some U.S. states require high school graduation examinations. Adequate scores on these exit exams are required for high school graduation. The General Educational Development test is often used as an alternative to a high school diploma.

Other applications include tracking (deciding whether a student should be enrolled in the "fast" or "slow" version of a course) and awarding scholarships. In the United States, many colleges and universities automatically translate scores on Advanced Placement tests into college credit, satisfaction of graduation requirements, or placement in more advanced courses. Generalized tests such as the SAT or GRE are more often used as one measure among several, when making admissions decisions. Some public institutions have cutoff scores for the SAT, GPA, or class rank, for creating classes of applicants to automatically accept or reject.

Heavy reliance on standardized tests for decision-making is often controversial, for the reasons noted above. Critics often propose emphasizing cumulative or even non-numerical measures, such as classroom grades or brief individual assessments (written in prose) from teachers. Supporters argue that test scores provide a clear-cut, objective standard that minimizes the potential for political influence or favoritism.

The National Academy of Sciences recommends that major educational decisions not be based solely on a test score. The use of minimum cut-scores for entrance or graduation does not imply a single standard, since test scores are nearly always combined with other minimal criteria such as number of credits, prerequisite courses, attendance, etc. Test scores are often perceived as the "sole criteria" simply because they are the most difficult, or the fulfillment of other criteria is automatically assumed. One exception to this rule is the GED, which has allowed many people to have their skills recognized even though they did not meet traditional criteria.

Aptitude Test

An **aptitude** is a component of a competency to do a certain kind of work at a certain level, which can also be considered "talent". Aptitudes may be physical or mental. Aptitude is not developed knowledge, understanding, learned or acquired abilities (skills) or attitude. The innate nature of aptitude is in contrast to achievement, which represents knowledge or ability that is gained through learning.

Intelligence

Aptitude and intelligence quotient are related, and in some ways opposite views of human mental ability. Whereas intelligence quotient sees intelligence as being a single measurable characteristic affecting all mental ability, aptitude refers to one of many different characteristics which can be independent of each other, such as aptitude for military flight, air traffic control, or computer programming. This is more similar to the theory of multiple intelligences.

Concerning a single measurable characteristic affecting all mental ability, analysis of any group of intelligence test scores will nearly always show them to be highly correlated. The U.S. Department of Labor's General Learning Ability, for instance, is determined by combining Verbal, Numerical and Spatial aptitude subtests. In a given person some are low and others high. In the context of an aptitude test the "high" and "low" scores are usually not far apart, because all ability test scores tend to be correlated. Aptitude is better applied intra-individually to determine what tasks a given individual is more skilled at performing. Inter-individual aptitude differences are typically not very significant due to IQ differences. Of course this assumes individuals have not already been pre-screened for aptitude through some other process such as SAT scores, GRE scores, or finishing medical school.

Combined aptitude and knowledge tests

Tests that assess learned skills or knowledge are frequently called achievement tests. However, certain tests can assess both types of constructs. An example that leans both ways is the Armed Services Vocational Aptitude Battery (**ASVAB**), which is given to recruits entering the armed forces of the United States. Another is the SAT, which is designed as a test of aptitude for college in the United States, but has achievement elements. For example, it tests mathematical reasoning, which depends both on innate mathematical ability and education received in mathematics.

Aptitude tests can typically be grouped according to the type of cognitive ability they measure:

1. **Fluid intelligence**: the ability to think and reason abstractly, effectively solve problems and think strategically. It's more commonly known as 'street smarts' or the ability to 'quickly think on your feet'. An example of what employers can learn from your fluid intelligence is your suitability for the role for which you are applying
2. **Crystallised intelligence**: the ability to learn from past experiences and to apply this learning to work-related situations. Work situations that require crystallised intelligence include producing and analysing written reports, comprehending work instructions, using numbers as a tool to make effective decisions, etc.

Achievement Test

An **achievement test** is a test of developed skill or knowledge. The most common type of achievement test is a standardized test developed to measure skills and knowledge learned in a given grade level, usually through planned instruction, such as training or classroom instruction. Achievement tests are often contrasted with tests that measure aptitude, a more general and stable cognitive trait.

Achievement test scores are often used in an educational system to determine what level of instruction for which a student is prepared. High achievement scores usually indicate a mastery of grade-level material, and the readiness for advanced instruction. Low achievement scores can indicate the need for remediation or repeating a course grade.

Under No Child Left Behind, achievement tests have taken on an additional role of assessing proficiency of students. Proficiency is defined as the amount of grade-appropriate knowledge and skills a student has acquired up to the point of testing. Better teaching practices are expected to increase the amount learned in a school year, and therefore to increase achievement scores, and yield more "proficient" students than before.

When writing achievement test items, writers usually begin with a list of content standards (either written by content specialists or based on state-created content standards) which specify exactly what students are expected to learn in a given school year. The goal of item writers is to create test items that measure the most important skills and knowledge attained in a given grade-level. The number and type of test items written is determined by the grade-level content standards. Content validity is determined by the representativeness of the items included on the final test.

Essay Type of Test

Classroom teachers construct and use a number of tests either to determine the achievement of their students motivates or encourage them to learn, identify their strengths and weaknesses, and prompt them to develop good study habits and so on. Most classroom test could be classified under two main types of test. These are the essay type and the objective type tests.

Kubiszyn and Borich (1984) defined an essay test as one for which the student supplies rather than selects the correct answer and demands that the student composes a response often extensive to the question from which no single response or pattern of responses can be cited as correct to the exclusion of all other answers

Types of Essay Tests

Essay test items are usually classified into two groups; restricted response item and extended response item.

Restricted Response Items (Controlled Response)

The restricted response essay item tends to limit the content, form or the number of words for testees. The limitations in terms of content and form of response are generally indicated in the statement of the item or question. The item therefore, tends to be specific and task related. Example;

1. State and discuss four functions of a wholesaler
2. Give an account of an interesting scene you have watched. Your essay should be written between 300-350 words.

The Open or Extended Response Item

This allows the testee to determine the length and complexity of the response. The student is therefore free in this case to identify and select any information which he thinks are pertinent to the item. Organise and interpret the ideas into a coherent and logical sequence to the best of his ability in response to the item. The freedom makes it possible for the student or the testee to demonstrate his competence in particular areas such as selection, organisation and integration of ideas. Extended response essay is most useful at the synthesis or evaluation levels in cognitive taxonomy. Example;

1. Discuss the advantages of essay type tests
2. Write an essay on the Economic Community of West African States (ECOWAS)

Suggestion for writing and using Essay test

Most teachers construct essay test without due regard to the principles of essay construction. When essay test is constructed without due regard, it becomes too open minded, disputable and unclear. The following guidelines are recommended for essay writing:

I. Have clearly in mind what mental processes you want the testee (student) to use before starting to write the question. For instance, if you want the testee to analyse, judge, or think critically include mental processes that involve analysis, judgement or critical thinking.

II. Write the questions in such a way that the task is clearly and unambiguously defined for the student/testee. This can be achieved by explaining in the overall instructions preceding the test items, the task and in the test items themselves.

III. Restrict the use of essay type test to those learning outcomes that are applicable. They should measure comprehension, application, analysis, synthesis and evaluation.

IV. Avoid using optional items. Optional items create the following problems

- Decreases test validity
- Decreases the basis for comparison among students
- Gifted students may be penalised because they may be challenged by complex and difficult questions

V. Establish reasonable time and/or page limits for each essay item to help the students complete the entire test and to give indications of the level of detail you have in mind for each item.

VI. Indicate the breadth and scope of the essay clearly to ensure precision and specificity of responses.

VII. Prepare a scoring scheme while preparing the test items. This will:

- Prevent under and overestimation of time needed in responding to the test items.
- Provide an estimation of the framework within which the student must operate.
- Provide an estimation of the length and complexity of the question
- Help in tailoring the scheme to be probable answers of the testees

VIII. Pitch the length and complexity of the test items to the level of achievement of the testees.

IX. Ask questions which are determining in the sense that experts could agree that the answer is better than another to reduce subjectivity and biases on the part of examiners.

When to use Essay Test

I. When the group to be tested (class size) is small. The time-consuming nature of essay test especially the scoring procedure makes it in appropriate for large classes.

II. When the instructor wishes to encourage and reward the development of students' skill in writing, critical thinking, originality, and the ability to organise and integrate ideas.

III. When an instructor is more interested in exploring students' attitude than in measuring their achievement.

IV. When an instructor is more confident of his proficiency as a critical reader than as an imaginative writer of good objective test item. Test scoring is very controversial even in the hands of specialists and becomes, dangerous and unreliable in the hands of classroom teachers and amateurish in scoring.

V. When time available for test preparation is shorter than the time available for test grading i.e. test preparation with the development of test blue print is time demanding and consuming and it demands adequate and sufficient time which should not be done hastily.

VI. When test security is a consideration. If one is afraid that test items will be passed on to future students and consequently decides not to re-use a test, it is better to use essay test. In general, a good essay test makes less time to construct than a good objective test.

Scoring of Essay Tests

Essay type test tends to be difficult to score. Coffman (1971) states that an essay answer may be given an 'A' by one scored and a 'B' by another scorer or the same essay and may be graded 'A' on one occasion but B or C on another occasion by the same scorer. Similarly, Chase (1978) citing Ashbum reports that the passing or failing of 40 percent of students depends on not what they know or do not know, but on who reads the papers and that passing or failing of about 10 percent depends on when the papers are read. There is therefore, inter-rater and intra-rater variability. Raters tend to assign different grades to the same paper on different occasions.

Various factors account for the unreliability in scoring essays. Factors like language usage, hand writing, sex, length of essay and the number of students' scripts are likely to affect essay test.

Qualgrain (1992) indicates that teacher characteristics like fatigue, illness, mood and restlessness influence grading of essay test. To minimize the unreliability of essay grading, the following are recommended:

1. **Good essay items:** poorly written questions are one source of scorer unreliability. Questions that are long and do not specify response-length is an important source of unreliability.
2. **Use of several restricted items:** rather than a single extended-range item, writing good items and using restricted range essays rather than extended range essays help improve scoring reliability.
3. **Use of predetermined scoring scheme:** this point is an important one. All too often, essays are graded without the scorer specifying in advance what he/she is looking for in a good answer. In scoring an essay, one makes an evaluation and in making an evaluation, criteria are very necessary. If a teacher does not determine and specify the relevant criteria beforehand, the reliability of scoring will be greatly reduced. Two main scoring methods are generally used by scorers. These are the analytic/point/key method; and global/quality/holistic method.

Analytic/Point/Key Method: this involves the essential elements in the form of a model answer which indicates the elements and points awarded for each element. According to Qualgrain (1992), in Ghana teachers prefer the analytic method to the global method for the following reasons:

- It ensures uniform scoring criteria
- Fairness in scoring is maintained
- Readers are less swayed by irrelevant factors, verbosity and unnecessary digression
- Script could be scored by another person using the same marking scheme
- Scoring point award to students can be explained easily; and
- It establishes evaluative criteria ahead of time.

The weaknesses in analytic method include the following:

- It is time consuming
- Assessment is based on point and not the integrated whole
- It does not reward good presentation and quality.

Holistic/Global/Quality Method: with this rating method, the teacher generally is more interested in the overall quality of the answer than specific points. The holistic/rating method is done by simply scoring papers into piles, usually five, if the letter grades are given. After scoring, the answers in each pile are reread and attempt is made to ensure that all the 'A' papers of comparable quality do not include 'B' and 'C' and so forth. This step is very important since the problem of changing the criteria above is always present in rating answers. The rating method helps the likelihood, for example, that an A paper may get sorted into the C pile because it was graded while the teacher was maintaining “strict” criteria. The rating method is certainly an improvement over simply reading each answer and assigning a grade based on some nebulous, undefined criteria and rationale.

Suggestions for improving Essay Scoring

1. Use the scoring scheme consistently. Do not favour one student over another or get stricter or more relaxed over time.
2. Remove or cover the names on the papers before scoring. This will help the ratter score the paper on its merit, rather than an overall impression of the student.
3. Score each student answer to the same question before going on the next answer. This avoids a student score having influence on another student score and it helps in maintaining the scoring criteria.
4. Keep scores for precious items hidden when scoring subsequent items to avoid straying from the scoring criteria.
5. Evaluate the papers before returning them. They help detect discrepant rating for correction.
6. Mark essays when you are emotionally stable. Physically and mentally alert to avoid dumping.
7. Provide feedback information to students to help in diagnosing their problems to facilitate learning.
8. Score particular setting at a sitting to help maintenance of scoring scheme.
9. Break when fatigue sets in
10. Before marking, shuffle the papers to ensure equal probability for students because the position of a script in a pile influences scoring.

Advantages of Essay items

1. Essay items are relatively easy to construct. Time spent in constructing essay items is comparatively shorter than objective test.
2. Essay test helps in assessing complex learning outcomes as it helps students to organize information constructively, analyse and synthesize information (high-level cognitive skills)
3. Essay test skills are essential in academic discipline, if developing communicative skill is an instructional objective.
4. Guessing is reduced by essay test since no questions are provided.
5. It allows student greater freedom in expressing themselves and therefore encourages critical thinking.
6. It encourages global learning by its holistic and integrated nature

Disadvantages of Essay Tests

1. Essay test encourage bluffing. By its length and score, essay test encourage verbosity and digression.
2. Essay test are difficult to score. By their nature there is a degree of high subjectivity in the hands of an inefficient examiner

3. It is tedious and time consuming in both writing and scoring. It is tedious to wade through pages and pages of students hand writing and students also spend a lot of time in writing.
4. Essay test suffer from limited sampling. The items are inadequate due to the needed time to respond to them.
5. Scorers of essay test are unreliable as it is difficult to maintain a common set of criteria for all the students

OBJECTIVE TYPE TEST

What is an objective test?

An objective test is a test for which correct responses are provided and students/testees are requested to select the right response from the number of responses. The items are called objective because they can be scored more objectively than any other type of them used to measure students' performance.

Types of objective tests

There are two major types of objective tests. These are the selection type and the supply type.

Selection type

The selection type consists of the multiple-choice type, true and false type and matching type.

Supply type

The supply type has variations as sentence completion, fill-in-the-blanks and short answer.

Objective test items are popular with classroom teachers for several reasons:

First, teachers can use them to measure many types of learning from verbal information to the use of rules and principles. Second, a wide range of content can be covered and measured because the items can sample the scope covered. Third, objective tests are easier to administer, score, and analyse than other types of tests. Fourth, they can often be adapted for use with computers and be scored and analysed by machine. Fifthly, fewer scoring are likely to be made, yielding scores generally more reliable than those from other types of tests.

Suggestions for constructing supply test item

- To enhance **scoring**, answers should be placed in the right hand margin
- The degree of response or precision should be expressed in an explicit and lucid manner
- Use limited blank spaces
- Avoid ambiguous questions
- Provide adequate space for the answer to be supplied
- Leave blank space either in the middle or at the end in to facilitate easy response
- Avoid "lifting"

Advantages of supply test

i. It is used for testing knowledge on definitions and terminologies
ii. It facilitates computational skills in sciences- mathematics, etc.
iii. It is more discriminative than multi-choice and True-false test
iv. It reduces copying and guessing
v. It allows students to exercise their ability in thinking thoroughly for the answer
vi. It facilitates vocabulary and concept development in students.

Disadvantages of supply test

i. Scoring is cumbersome, tedious and subjective;
ii. It does not encourage analytic thinking as they usually require symbols or phrases
iii. It encourages rote learning as some of the answer are factual
iv. Items may be ambiguous if written by incompetent testers

Selection type

This involves choices or response from which students are allowed to select the probable response for the items. The selection type includes:

a) True-False
b) Matching test
c) Multiple choice

True-False Items: true-false items are popular probabilities because they are quick and easy to write or at least they seem to be. In true-false item test, the student is made to ascertain whether the statement of proposition is true or false. The student is made to underline, circle or tick the right answer.

Suggestions for writing True-False Items

I. The statement should be clear and lucid i.e. the statement should be definitely true or definitely false without additional qualifications.
II. Uses relatively short statement and eliminate extraneous material.
III. Keep true and false statement at approximately the same length
IV. Avoid using double – negative statement. Avoid verbal clues, specific determiners and complex sentences
V. Avoid broad general statements that are usually true and false without further qualifications
VI. Arrange test items so that there is no discernible pattern of true/false without further qualifications
VII. Avoid terms denoting indefinite degree. For example, never, only.
VIII. Avoid taking statements directly from the text books and presenting them out of context
IX. Approximately, half (50%) of the total number of items should be false because it is easy to construct statements that are true and the tendency to have more true statements
X. Arrange the items such that the correct responses do not form a discernible pattern.

Advantages of True-False Test item

I. They provide simple, fundamental and direct test of students' knowledge
II. It is easy to score quickly and objectively

III. True and false items are quite efficient. The number of independently scrabble responses tend to be higher than multiple-choice test

IV. Most testers find the task of writing true-false items simple and less time consuming.

V. It can be adequate to most question areas

VI. It is good at lower classes for children who are not good at reading

VII. Test students' ability to recall information. Students can be asked to determine whether a definition, rule, or a principle is stated correctly.

Disadvantages of True-False items

I. They are suspected with good reason of being particularly susceptible to chance error resulting from guessing

II. They are less reliable than multiple-choice test of equal length due to chance errors

III. They are usually judged to be trivial

IV. Most true-false items are based directly from textbooks and there is the danger that they might encourage and reward sheer verbal memory

V. True-false questions may lack background information or qualifications to enable even an expert to judge with assurance whether they are True/False

VI. They do not provide explicit alternative in relation to which the relative truth or falsity of the item can be judged.

MULTIPLE-CHOICE TEST (MCQS)

Multiple-choice is one of the selected response test item formats and the most popular and the most frequently used of the selected response formats. It is a type of objective test in which the respondent/testee is to select from among the alternatives (options or responses) the one that best completes the item. The incorrect options are called foils or distracters.

Good multiple-choice items are the most time-consuming kind of objective items to write. There are two types of multiple-choice tests. These are the **correct answer type** and **best answer type**. In the 'correct-answer type' there is only one correct answer, all other alternatives are wrong. The distinguishing characteristics of this variety is that one of the responses must be unambiguously correct and the other responses unambiguously incorrect. Example:

I. What will be the effect on the standard deviation of a set of 100 test scores if five points were added to each score?
   a) The standard deviation would increase by five points
   b) The standard deviation would increase by an amount equal to the square root of 0.05
   c) There would be no effect; the standard deviation would remain the same.
   d) There is no way of predicting the effect in advance of actual calculations

II. Which of the following criteria in evaluating essay test item involves checking?

   Do the items relate to instructional objectives?

   a) Clarity
   b) Fairness
   c) Practicality
   d) Validity
   e) Reliability

III. In the development of a new test, the extent to which performance on the new test predicts performance on a similar but well known test is referred to as:

   a) Concurrent validity
   b) Face validity
   c) Content validity
   d) Construct validity

In the **Best-answer type,** the candidate is presented with answers of varying degree of acceptability (or a number of possible reasons some of which are better than the others, or several possible procedures some of which are more desirable than others) and is expected to select the best answer e.g.

1. Learning objectives are important for classroom assessment because they:
   a. Communicate to students the performance they are expected to learn
   b. Give direction to the selection and construction of assessment instruments
   c. Help evaluate existing assessment instruments when specific outcomes are unknown
   d. Provide information for classifying construct accuracy of assessment procedures and teaching.
2. Which general assessment principle underlies the use of continuous assessment in Ghanaian schools?
   a. Clearly specifying what is to be learned
   b. Ensuring reliability of the scores
   c. Relevance of characteristics of what is being assessed
   d. Ensuring the use of a variety of assessment procedures
3. A physics lecturer observed her students during a science laboratory session to determine how effectively students can carry out experiments. The physics lecturer
   a. Tested her student's performance
   b. Measured her student's performance
   c. Assessed her student's performance
   d. Evaluated her student's performance

The multiple-choice test may also have the **multiple response type** which consists of a stem followed by several true/false statements or words. The respondent is to select which statement or statements that complete the stem.

Example:

Which of the following signs and symptoms is/are common to malaria?

1. Fever
2. Slow pulse rate
3. Vomiting
   a. 1 only
   b. 2 only
   c. 1 and 3
   d. 2 and 3
   e. 1, 2 and 3

Alkali metals have the following characteristic properties

1. Their emission spectra contain a closely spaced doublet
2. They form basic oxides
3. They have a relatively high low ionization energy
4. They have a relatively high electron affinity
   a. 1, 2 and 3 only
   b. 1 and 3 only
   c. 2 and 4 only
   d. 4 only

Guidelines for constructing Multiple-Choice tests

I. The stem should contain the central issue of the item, and should be concise, clear to read and understand.

II. Options should be plausible. Distracters must be plausibly attractive to the uninformed.

III. All options for a given item should be homogenous in content, form and grammatical structure

IV. Avoid the repetition of words in the options

V. Avoid specific determiners, which are clues to the correct option.

VI. Vary the placement of the correct options

VII. Avoid items measuring options; one option should clearly be correct or best

VIII. The responses must be parallel in form i.e. one word, about same sentence, length, etc. in an alphabetical or sequential order itemized vertically, and not horizontally

IX. Avoid overlapping options; each option must be distinct.

X. Avoid using "all of the above" as an option, but use "none of the above" sparingly, if at all. "None of the above" should be used only when an item is of the "**correct answer**" type and Not the "**best-answer**" type

Advantages of Multiple-Choice test items

i. Multiple-choice questions have considerable versatility in measuring objectives from knowledge to evaluation

ii. A substantial amount of course material can be sampled in a relatively short time.

iii. Scoring is highly objective acquiring only a count of the number of correct response

iv. The degree of discrimination among the correct options is high and this allows the student to select the best alternative and avoid the absolute judgement found in True/False test items.

v. The multiple options reduce the effect of guessing

Disadvantages

i. Multiple-choice questions can be time-consuming to construct

ii. Multiple-choice questions can at times have more than one defensible correct answer

iii. The error of guessing is only reduced by the discriminatory element but not eliminated

iv. To an extent promotes rote learning

Matching Test items

Matching-test items consist of premises from which the student selects the **response** to match each item

Suggestions for writing matching test items

i. Keep both list of discriminations and the list of options fairly short and homogeneous
ii. Make sure that all the options are plausible distracters for each description to ensure homogeneity of list
iii. The list of descriptions should contain the longer phrases or statement while the options should consist of short phrase words or symbols
iv. Each description in the list should be numbered
v. Include more options than descriptions
vi. In the directions, specify the basis for matching and whether options can be used more than once

Please, find below an example of a matching test:

Match the following names with their countries

A. George W. Bush	I. Nigerian
B. Tony Blair	II. Ghana
C. J. A. Kuffour	III. United States of America
D. Musiveni	IV. Great Britain
E. Obasanjo	V. Uganda

Advantages

i. Matching questions are usually simple to construct and score
ii. Matching items are ideally suited to measure associations between facts
iii. Matching questions can be more efficient than multiple-choice questions because they avoid repetition of options in measuring associations
iv. Matching tests are amenable to machine scoring
v. Matching questions reduce the effect of guessing
vi. They require little reading time
vii. They require students to integrate their knowledge by matching the items in the columns

Disadvantages

i. Matching questions sometimes tend to ask students trivial information
ii. Matching test sometimes emphasize memorization
iii. If the items are not well arranged matching-test may encourage serial memorization.
iv. The size of matching test may be limited by the size of commercial answer which will reduce the options

Comparison of Essay and Objective test items

	Essay Test		**Objective test**
I.	Requires the student to plan and organise his answers		Requires the student to choose among several alternatives
II.	Consists of relatively few or more general questions which call for extended answers		Consists of many rather specific questions which require brief answers.
III.	Demands a lot of thinking and writing from students		Less time is spent on thinking and writing due to guessing and the options
IV.	It allows the student the freedom to express him/herself		It affords freedom for a test constructor but limits the freedom of the student
V.	There is subjectivity in grading		It measures a minute aspect of the individual (Low level ability)
VI.	It is easy to construct but difficult to score		It is relatively tedious and difficult to prepare but easy to score.
X	The task and objective for judging the student are not clearly stated.		The task and objective for judging the student are clearly stated.
VIII.	Permits and occasionally encourages bluffing		Permits and encourages guessing
IX.	Distribution of scores obtained can be controlled to a considerable degree by the grader.		Distribution of numerical scores obtained is determined almost entirely by test
X.	Examiners task and scoring criteria are seldom made.		Examiner’s task and scoring procedures are usually stated.
XI.	Item analysis is tedious and cumbersome.		Item analysis is simple and clear

``````````````````````````````**INSTRUMENT, VALIDITY, RELIABILITY**

## *Part I: The Instrument*

*Instrument* is the generic term that researchers use for a measurement device (survey, test, 0that the *instrument is the device* and *instrumentation is the course of action* (the process of developing, testing, and using the device).

Instruments fall into two broad categories, researcher-completed and subject-completed, distinguished by those instruments that researchers administer versus those that are completed by participants. Researchers chose which type of instrument, or instruments, to use based on the research question. Examples are listed below:
``````````````````````````````

Researcher-completed Instruments	**Subject-completed Instruments**
Rating scales	Questionnaires
Interview schedules/guides	Self-checklists
Tally sheets	Attitude scales
Flowcharts	Personality inventories
Performance checklists	Achievement/aptitude tests
Time-and-motion logs	Projective devices
Observation forms	Sociometric devices

Usability

Usability refers to the ease with which an instrument can be administered, interpreted by the participant, and scored/interpreted by the researcher. Example usability problems include:

1. Students are asked to rate a lesson immediately after class, but there are only a few minutes before the next class begins (problem with administration).
2. Students are asked to keep self-checklists of their after school activities, but the directions are complicated and the item descriptions confusing (problem with interpretation).
3. Teachers are asked about their attitudes regarding school policy, but some questions are worded poorly which results in low completion rates (problem with scoring/interpretation).

Validity and reliability concerns (discussed below) will help alleviate usability issues. For now, we can identify five usability considerations:

1. How long will it take to administer?
2. Are the directions clear?
3. How easy is it to score?
4. Do equivalent forms exist?
5. Have any problems been reported by others who used it?

It is best to use an existing instrument, one that has been developed and tested numerous times.

Validity

Validity is the extent to which an instrument measures what it is supposed to measure and performs as it is designed to perform. It is rare, if nearly impossible, that an instrument be 100% valid, so validity is generally measured in degrees. As a process, validation involves collecting and analyzing data to assess the accuracy of an instrument. There are numerous statistical tests and measures to assess the validity of quantitative instruments, which generally involves pilot testing.

Face validity *referred to value judgment of the instrument as it occurs to the sight. Looking at the instrument to determine weather is worthy to carry out the anticipated task.*

External validity is the extent to which the results of a study can be *generalized* from a sample to a population. Establishing eternal validity for an instrument, then, follows directly from sampling. Recall that a sample should be an accurate representation of a population, because the total population may not be available. An instrument that is externally valid helps obtain population generalisability, or the degree to which a sample represents the population.

Content validity refers to the appropriateness of the content of an instrument. In other words, do the measures (questions, observation logs, etc.) accurately assess what you want to know? This is particularly important with achievement tests. Consider that a test developer wants to maximize the validity of a unit test for 7th grade mathematics. This would involve taking representative questions from each of the sections of the unit and evaluating them against the desired outcomes.

Reliability

Reliability can be thought of as consistency. Does the instrument consistently measure what it is intended to measure? It is not possible to calculate reliability; however, there are four general estimators that you may encounter in reading research:

1. *Inter-Rater/Observer Reliability*: The degree to which different raters/observers give consistent answers or estimates.
2. *Test-Retest Reliability*: The consistency of a measure evaluated over time.
3. *Parallel-Forms Reliability:* The reliability of two tests constructed the same way, from the same content.
4. *Internal Consistency Reliability:* The consistency of results across items, often measured with Cronbach’s Alpha.

Relating Reliability and Validity

Reliability is directly related to the validity of the measure. There are several important principles. First, a test can be considered reliable, but not valid. Consider the SAT, used as a predictor of success in college. It is a reliable test (high scores relate to high GPA), though only a moderately valid indicator of success (due to the lack of structured environment – class attendance, parent-regulated study, and sleeping habits – each holistically related to success).

Second, validity is more important than reliability. Using the above example, college admissions may consider the SAT a reliable test, but not necessarily a valid measure of other quantities colleges seek, such as leadership capability, altruism, and civic involvement. The combination of these aspects, alongside the SAT, is a more valid measure of the applicant’s potential for graduation, later social involvement, and generosity (alumni giving) toward the alma mater.

Finally, the most useful instrument is both valid and reliable. Proponents of the SAT argue that it is both. It is a moderately reliable predictor of future success and a moderately valid measure of a student’s knowledge in Mathematics, Critical Reading, and Writing.

Validity and Reliability in Qualitative Research

Thus far, we have discussed Instrumentation as related to mostly quantitative measurement. Establishing validity and reliability in qualitative research can be less precise, though participant/member checks, peer evaluation (another researcher checks the researcher’s inferences based on the instrument (Denzin & Lincoln, 2005), and multiple methods are 3 constructivist viewpoint that reality is unique to the individual, and cannot be generalized. These researchers argue for a different standard for judging research quality.