# NEW TRENDS IN EDUCATIONAL ASSESSMENT AND INSTRUMENTATION

**First Edition**

**Omachi, Daniel, Ph.D California,**
Educational Measurement, Evaluation and Research
Department of Science and Computer Education
Godfrey Okoye University, Enugu

First Edition, Published in 2021 by
Angusco Nigeria Enterprise
No. 37 Edinburgh Road,
Ogui New Layout, Enugu State
Nigeria

## DEDICATION

This book is dedicated to all my lovely students

## ACKNOWLEDGEMENT

I wish to acknowledge first, all my lovely students in all universities and colleges whose encouragement informs this edition of the book. I also acknowledge my dear Provost of Peaceland College of Education, Rev. Fr. Prof. Leonard Ilechukwu for his love for education. Thank you sir for all your encouragement, supports and advice. A very big thank you to other management and staff members of Peaceland College of Education, Enugu. Many thanks to the management and staff of Elyland College of Education, Ankpa for their encouragement and support. A special thanks to Mr. Arome Alfa, Mr. Raluchukwu Okwudili, Mr. Nwadike Bornaventure, Mr. James Monday and others.

May I thank in a very special way my academic giants and worriors whose shoulder I stand. Prof. Romey Okoye of Nnamdi Azikiwe University, Prof. Ngozi Agu, Prof. Kan Nwankwo, Prof. Esomuonu, and Prof. Sunday Abonyi for their wealth of knowledge and what they have made out of the author. God bless you all.

I appreciate greatly the understanding of my dear wife Mrs. Grace Omachi Daniel, my children Elyon and Stainless, my Dad Dr. Adams Baba Daniel and all my siblings. I appreciate all your efforts and understanding that led to the success of this work.

To all the people worthy of this acknowledgement but not mentioned, I tell you, you are the best and I appreciate your prayers, encouragement and supports. God bless you all.

## FOREWORD

The purpose of this book is to provide the reader the major theoretical and practical considerations that inform the development, use and interpretation of continuous assessment and instrumentation in education. The book addresses three major areas of educational concern expressed by university trained teachers or those who guaranteed from colleges of education. These concerns are (1) the lack of their firm grounding in many aspects of continous assessment (2) instrumentation and (3) basic statistics used in instrumentation.

It is against these backgrounds that I am pleased to write the foreword to this invaluable book on the new trends in continous assessment. This book addresses the three areas exceptionally well and at the classroom level. This is because the presentation of concepts in the book is clear, simplified and logical; and these are done in such a way as not to be too technical for the teacher.

The organization of the text, an essential feature of any textbook, is appropriate and well thought out. Indeed each of the 8 chapters adequately covers each of the subject matters components. The content of each of the chapter flows into the other chapter, smoothly and coherently. In totality, the 8 chapters adequately cover the essential subject matter of educational assessment and instrumentation.

The substance and style of presentation of the content are quiet commendable. Illustrative examples that are teachers friendly were used in this textbook. I have no doubt in my mind that the readers will find this text very useful especially as an invaluable guide in the process of knowing how to develop educational assessment and instrumentation. I therefore recommend this text strongly to colleges of education and university students and their teachers to demystify the mysteries of continous assessment and instrumentation in education and other social fields.

**Dr. Obi Leonard**
Department of Educational Foundation,
Federal University, OyiEkiti

**TABLE OF CONTENT**

**Chapter 7**

Reliability of Instrument

- Measurement of stability
- Measurement of equivalence
- Measurement of internal consistency
- Scorer reilability

**Chapter 8**

Basic Statistics in Continous Assessment

- Presentation and organization of data
- Frequency distribution
- Graphs of frequency distribution
- Measurement of central tendencies

## CHAPTER 1

**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.

**Table 1: 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 |

## CHAPTER 12

## 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).

**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).

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.

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.

**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.

**Weaknesses/Problems 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

## CHAPTER 18

## 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 text blue print for a unit of instruction.

***Example 1***

Table 2. 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***

**Table 3**

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.

## CHAPTER
## INSTRUMENTATION

Instruments are measurement tools (for example, questionnaires or scales) designed to obtain data on a topic of interest from research subjects**.**It helps the assessor or researcher to obtain, measure, and analyze quantitative data from subjects area of interest.

The need for an educationist to be acqinted with ideas of construction, choice, administration and interpretation of instrument cannot be over-emphasized. One may have an intetion of measuring an elephant but if a wrong instrument is selected or constructed, the instrument might end up measuring coackroach.

The efficacy of the instrument in education is dependent on the variables mentioned above. This chapter shall therefore dymistify and explore to a reasonable extent on ideas of instrumentation.
In education, there are many instrument that can be used to measure different characteristics or trait, such as: questionnaire, observation, interview, rating scales, inventories etc.

**Characteristics of a good instrument**

- Valid and reliable ;
- Based on a conceptual framework, or the researcher's understanding of how the particular variables in the study connect with each other ;
- Must gather data suitable for and relevant to the test/research topic ;
- Able to test hypothesis and/or answer proposed research questions under investigation;
- Free of bias and appropriate for the context, culture, and diversity of the study site ;
- Contains clear and definite instructions to use the instrument ;
- The langauge must be simple and clear enough.

**Types of Data Collection Instruments**
**Questionnaire**

Questionnaire is one of the best instrument if one need to collect data from a large number of people. They contain multiple choice questions, attitude scales, closed questions and open-ended questions. This data collection instrument is flexible as there is no rush or pressure on respondents to provide immediate answers. Respondents can take their time to think about the questions and then provide answers to them at their most convenient time. This ensures that the answers provided are not influenced by time rush or experiences from a bad day the respondent may be having. Again, questionnaires can be administered in different forms by post, email attachments, administered in conferences or posted on Internet sites. Researchers may even decide to administer the questionnaire in person. This method has an advantage to those people that have difficulty reading and writing. In this case, the participant orally answers each question on the questionnaire as the researcher notes down the responses. Since questionnaires do not require names, Participants are more comfortable to state their views or feelings privately without worrying about what other people might think of them or the possible reaction of the researcher. One major drawback in using questionnaires which may result in the researchers drawing false conclusions from their study is that they usually have a fairly low response rate; while some may not answer the questions completely, others may give no response at all. Again, some people may give socially acceptable answers. Respondents are however encouraged to answer all questions as honestly as possible.

**Interview**

This type of data collection instrument can be described as an oral questionnaire. Interviews are usually done in a face to face meeting. They can also be conducted via phone conversations, or through video chats, during which the interviewer takes notes with a pen and paper or a tape recorder. The interviews are conducted either formally, informally or even semi-formally. In an informal interview, the interviewer in this case allows the respondents to speak freely on a

particular topic. While in a formal interview the interviewer seeks answers to particular questions that are thus presented to the interviewees. Here, a list of structured questions centered around the subject matter is prepared by the researcher prior to the interview.

**Experiments**

This type of data collection instrument is used in pure and applied sciences research. Experiments are carried out in laboratories by researchers. The experiments are strictly centered on the research topic for the sole purpose of meeting the research objectives. If the experiments are carried out properly, its results are viable and error free. However, one limitation with this method is that; it is quite expensive to carry out science experiments and if the researcher is not careful in the laboratory and does not protect himself properly with laboratory gears, when chemicals spill, they may cause damage to the researcher.

**Participant and Non-Participant Observation**

Observation as a method of collecting data is popular in behavioral and social sciences. This method involves observing and recording individual behaviors. Individual behaviors may be observed under these categories; what people do, why they do them, the roles they have, relationships that connect these 'activities' and features of the situation in which they find themselves. In participant observation studies, the researcher becomes part of the group to be observed. He has to fit in the group and gain the trust of its members. But at the same time, he needs to be careful enough to be detached in a way that he is able to carry out the observation. Non-participant observation is the direct opposite of what happens in participant observation. A good advantage of non-participant observation is that the result is more inclined to be viable and free from bias as the researcher is not part of the group being observed and thus has no attachments to the group. But it has the problem of inaccuracy and delayed result. The observation carried out could be continuous or over a set period of time (1 hour daily for

3weeks) or randomly for shorter periods of time (for 60 seconds every so often). These two types of observation methods are informative, flexible and cheap to be carried out. However, special skills are required to access behavioral observations in research.

**Categorization of Instrument**

**Ability Tests/Scales**

Ability tests are tests designed to assess competence in an activity or occupation based on one's skill, capacity, means or other special qualifications. The term ability test is more generally used as "measures of a cognitive behaviour". Anastasi (1997) rightly notes that "any cognitive tests, regardless of what it has been called traditionally, provides a sample of what the individual knows at the time he or she is tested and measures the level of development attained in one or more abilities". The clumping of "aptitude tests" and “achievement tests” as ability test became very -necessary in view of the current misuse of test results by researchers. It must be appreciated that test errors abound in correlating achievement scores with aptitude scores in so far as no two performance indicators correlate perfectly. To reduce errors of over-prediction or under-prediction it is necessary to consider and measure both aptitude and achievement as ability.

Ability tests may include individual tests, tests for a special population and group tests. Although researchers are encouraged to develop these types of tests when necessary, there are however well developed and standardized tests which researchers can adapt in their study provided they meet all necessary conditions for adaptation of instruments.

Standardized ability tests, which can be adapted by researchers, are:

1. The Stanford-Binet Intelligent Scales
2. The Wechster Intelligent Scales
3. The Kaufman Assessment Battery
4. Detriot Test of Learning Aptitudes

5. Elliot's Differential Ability Scales
6. McCarthy Scales
7. Piagetian Scales
8. Differential Aptitude Tests (DAT)
9. Multidimensional Aptitude Battery (MAB).

The problem with adapted instrument is that in many cases they do not match with the present background under which the study is conducted. An important issue, which all researchers must bear in mind, is that no two research conditions are purely identical in all respects. Because our research is of the social science type we should not assume that we could achieve maximum controls.

**Personality Tests/Scales**

Personality Tests according to Anastasi and Urbina (1997) are "Instruments for the measurement of emotional, motivational, interpersonal, and attitudinal characteristics, as distinguished from abilities". Behavioural research scientists have viewed the issue of personality test with a lot of seriousness. The reason is that it has to do with human traits, which change, not only over time but with varying circumstances. Trait measures, therefore, are subjected to detailed scrutiny before drawing any conclusion based on data collected with it.

Based on the Anastasian classification, personality tests/scales are categorized as Self Report, Personality Inventories, Attitude and Interest Measures, Projective Tests and Situational Tests. Generally, most of these instruments measure the following traits: emotional disposition, depression, mania, paranoia, hysteria, masculinity/feminity, psychotenia, schizophrenia, social Introversion, hypochondriasis, psychopathic deviate, anxiety, alcohol/drug dependence, stress disorder, delusion, aggression, avoidant dysthymia, etc.

There are also standardized personality tests/scales, which researchers readily adapt. They include:

1. The Minnesota Personality Inventories (MPI)
2. Multi-Stage Personality Inventories
3. Millon Clinical Muitiaxial Inventory (MCMI)
4. Edward Personality Preference Schedule (EPPS)
5. The Strong Interest Inventory (Sll)
6. Jackson Vocational Interest Survey (JVIS)
7. The Rorschach Inkblot
8. The Holtzman Inkblot
9. General Thematic Apperception Test (TAT)
10. Word Association Tests (WAT)
11. Early Memory Procedures (EMP)
12. Draw -a- Person Test (D-A-T)
13. The Semantic Differential Scale
14. Role Construct Repertory Test

Although there are a number of standardized tests or scales, which researchers can adapt, it is very much advisable that researchers take the necessary pains in developing their own research instruments. The reason is that adapted instruments are often stereotyped irrespective of whatever adaptation precautions the researchers may have taken during the adaptation processes. For instance, most of these -instruments were developed in an entirely differing physical, social, anthropological and psychological environment. As such, the extent to which they can fit into the new research environment is always uncertain. This is to say that its validity in a new situation is obviously in doubt.We realize the fact that researchers find it difficult to generate entirely instruments for some given research studies. This is not necessarily because the procedures are unattainable, but obviously due to lack of direction and necessary guidance. This text has in the subsequent units provided, from a practical perspective, the major instrumentation approaches in behavioural research.

**Considerations in Choice of Instruments**

A number of factors must be borne in mind before deciding on instrument for data collection in a given research. They are:

- The purposes of the study/test
- The nature of the population/test
- The design of the study/test
- The expected tool for data analysis

## CHAPTER 13

## TAXONOMIES OF EDUCATIONAL OBJECTIVES AND TEST DEVELOPMENT

The idea of creating a taxonomy of educational objectives was conceived by Benjamin Bloom in the 1950s, the assistant director of the University of Chicago's Board of Examinations. Bloom sought to reduce the extensive labor of test development by exchanging test items among universities and other higher institutions.In this unit we will take a look at the practical applications of the taxonomies of educational objectives in test development. While we make few references to the theoretical procedures, the major focus here is the practical aspect, which automatically guides you, should you find yourself in a situation that you have to generate your own test items to suit your specific purpose. As we already know, taxonomies were classified into three main domains: the affective, the psychomotor, and the cognitive.

Let us present our illustrations under the following three sections:

a) Considerations in development of tests for assessing affective behaviours
b) Considerations in development of tests for assessing psychomotor behaviours
c) Considerations in development of cognitive tests

**Development of Tests for Assessing Affective Behaviours**

Benjamin Bloom and his colleagues in 1956 developed a classification system or taxonomy now popularly called the Bloom's Taxonomy. The taxonomy classifies educational objectives into three principal domains: the cognitive, the affective and the psychomotor domains. While the cognitive domain includes the characteristics that deal with the recall or recognition of knowledge and development of intellectual skills, the affective domain includes objectives related to emotions, feelings and attitudes. The psychomotor domain on the other hand deals with the objectives related to muscular or motor skills or manipulation of

materials and objects and neuro-muscular coordination (Mehrens and Lehmann, 1991; Bloom et al 1956).

Emphasizing the indispensability of assessment in the affective domain, Mehrens and Lehmann (1991:200) wrote that "because the affective disposition of the student has direct relevance to his ability to learn, his interest in learning and attitudes toward the value of education, educators in general and classroom teachers in particular should know something about affective measurement especially attitudes".

In many occasions we have made comments such as these:

a) John will make a very intelligent scholar if he puts interest in his studies
b) Nkechi has very high numerical skill but lacks interest in mathematics
c) George would have been a wonderful scientist if he had put interest in science.

In all these statements, it is inherent that affective behaviour is in control of all cognitive processes. Most learning difficulties originate from individual's affective responses - I *am a girl, I know I cannot study engineering.* People generally develop problems in mastering a particular task because of their inner driving force, which propels them to achieve or fail in a given task. These forces reside within the affective domain. As such more emphasis should be given to the affective domain in educational evaluation. We need to measure it; we also need to build a positive affective behaviour in learners and generally in people striving to succeed in various enterprises.

Krathwohl (1964) in his *Taxonomy of Education Objectives, Handbook II: The* Affective *Domain* identified five objectives related to emotional responses to tasks, which he most appropriately called the objectives of the affective domain.

According to Santrock (2004) each of the five objectives requires the individual to show some degree of commitment or emotional intensity. The five objectives of the affective domain are:

- Receiving
- Responding
- Valuing
- Organizing, and
- Value characterizing

**Receiving**

This entails individuals' self-awareness of the immediate environment. The individual at the receiving stage begins to recognize that he is in a new environment, which offers a challenge or insight that may lead to something. Assuming students are on a field trip to a typical rainforest, some may be seen taking their time observing and noting striking feature of vegetation. At least they will realize that they are in a new environment and that there is something to learn there. Receiving also involves attentiveness. Take for example a situation where a guest speaker visited a school to give a talk on science arid society. The major affective objective is for students to listen carefully knowing very well that there is always something to get from the interaction. This is also an aspect of receiving. It is a demonstration of willingness and acceptance.

**Responding**

In this objective individuals/students exhibit motivation to learn and display new behaviours as a result of experience and the interaction. In the case of the field trip, the students may respond by trying to name tree of the habitat they were observing. In case of the guest speaker on science and society, students may respond by asking questions on science based vocations and basic requirements for such vocations.

**Valuing**

In valuing students become involved in or committed to some experiences. As an objective within the affective domain students may begin to develop great values for systematics. In the case of the guest speaker on science and society, students may begin to develop values for subjects that lead to major careers in science and technology.

**Organizing**

This objective involves students integrating new values into already existing sets of values and giving it proper priority (Santrock, 2004; Krathwohl et al 1964). An activity here could be students developing personal and collective herbarium. By building on their values in systematics they proceed into developing herbarium in the school and at home and also forming environmental clubs with the desire to protect the plant species. In the case of the guest speaker in sciences the objective might be for students to form or join science clubs within and outside the school.

**Value Characterizing**

This is the last objective within the affective domain. The gradual developments in the other four objectives are crystallized here. It concerns students' actions with respect to the preceding developments. Students are seen acting in accordance with the values and are firmly committed to it. Throughout the individual's stay in the school he/she will be seen devoting his time to the herbarium and may even extend by involving others in developing the habit. As for the case of the guest speaker, students may increasingly value science and act in that direction at all time.

For your personal guidance, some actions verbs for writing objectives in the affective domain are presented in Table 4:

Action verbs for writing objectives in the affective domain

| Objective Category | Action Verbs |
|---|---|
| Receiving | Accept, differentiate, listen, separate, select, share, agree |
| Responding | Approve, applaud, comply, follow, discuss, volunteer, practice, spend time with paraphrase |
| Valuing | Argue, debate, deny, help, support, protest, participate, subsidize, praise |
| Organizing | Discuss, compare, balance, define, abstract, formulate, theorize, organize |
| Value characterizing | Change, avoid, complete, manage, resolve, revise, resist, require |

Adopted Abonyi (2004)

**Procedures for Assessing Objectives in the Affective Domain**

In schools, assessment program demands that all behaviour domains (the cognitive, the affective and the psychomotor) be regularly assessed on the basis of which evaluation judgments are made on the learners. The common practice in all schools is that both the affective and psychomotor domains are neglected. The current argument is that researchers and teacher lack the requisite skills for assessment in the affective domain. Generally, teachers are not expected to employ standardized personality scales in the assessment of affective behaviour in the classroom rather those basic affective attributes that are already enshrined in students report booklets should be taken into consideration. Let us take a look at a sample of students' affective checklist and discuss how teachers could include the affective behavior in the assessment of students.

Table 5 : Rater's guide for affective behaviors

| SN | Items | Rating Options | | | |
|---|---|---|---|---|---|
| | | E | G | F | P |
| 1. | Attendance to class | | | | |
| 2. | Attentiveness during lesson | | | | |
| 3. | Observance and sensitivity | | | | |
| 4. | Questioning ability | | | | |
| 5. | initiative- and responsiveness | | | | |
| 6. | Carrying out assignments | | | | |
| 7. | Sense of commitment | | | | |
| 8. | Innovativeness | | | | |
| 9. | Spirit of integration | | | | |
| 10. | Organizational ability | | | | |
| 11. | Spirit of cooperation | | | | |
| 12. | Perseverance | | | | |
| 13. | Insistence on completion of work | | | | |

*Key to the rating options*
*E* - Excellent (4 points) G - Good (3 - points)
F - Fair (2 - points) P - Poor (1 - point)
A close look at this checklist will reveal a split of the five objectives of the affective domain into thirteen. Let us take them one after the other.

**Assessing the Objective "Receiving"**
The rating scale above reveals a list of behaviours that measure the five objectives. As you can see, the first three behaviours lend themselves to receiving which is the first objective within the affective domain. Attendance to class is the first indication that the candidate is interested, followed by attentiveness and observance/sensitivity. As a classroom teacher/assessor it is your duty to monitor attendance to class and assign marks to it as a part of the CA score. In the same vein you must monitor students' attentiveness during lesson and also allocate score to it. Finally, you must take a critical look at how observant and

sensitive the students are to situations arising in the classroom and during field trips. Scores should also be assigned to them as appropriate. This type of assessment goes beyond the classroom. In a non-formal educational or social setup the affective behaviour could be ascertained using this approach. Note that this example is hypothetical because there are cases of personality assessment that has nothing to do with *attendance to class.* What is important here is that you grasp the concept of receiving as willingness and its demonstrations. The same thing is applicable to all other categories.

**Assessing the Objective "Responding"**

Going by our description of responding as objective within the affective domain, easily recall that items 4 and 5 measure exactly this behaviour. These items are questioning ability, and initiative/responsiveness. As an experienced teacher/assessor, you should take note on daily basis students/individuals level of participation in lesson or activities in question. It is indicated in their questioning, strength, initiative and responsiveness to the situation in question. Please note that what is being measured is the level of participation and not correctness of response. As you rate these behaviours on daily basis you will be in a position to assign definite scores to the behaviours using the rating guide prepared for or as provided in the result sheet as the case may be.

**Assessing the Objective "Valuing"**

With our experience in personality assessment as researchers/teachers and following the neat categorizations in Krathwohl's (1964) taxonomy, we can easily realize that items 6 and 7 of the affective domain in the raters guide in Table 2.2 measure the objective "valuing". Valuing has to do with getting involved and committed. As a trained teacher should know that carrying out assignment implies getting involved. We must therefore rate students based on the extent to which they carry out assignments and show a sense of commitment. Care should be exercised here also. What we are rating is not the precision or

correctness in the work done but the commitment exhibited and the exhibition of interest in carrying out the assignment. The precision and correctness in the assignment should be left for the cognitive and psychomotor domains.

**Assessing the Objective "Organizing"**

As we already have discussed, organizing involves integrating new values into already existing ones. This is all about innovation and integration. How have yum M intents introduced exciting innovations into issues at hand and how have they in integrating ideas and practices in very striking innovative ways that even you as a teacher never ceased marveling. We also know that it takes good organizational skills to integrate ideas and practices while introducing an innovation. As a good teacher you must rate the students in this capacity and incorporate the scores in the assessment schedule.

**Assessing the Objective "Value Characterizing"**

In most classroom settings value characterizing are displayed through spirits of cooperation and perseverance. The teacher can give a group project and paired tasks and observe within group and inter-group interaction/cooperation among the students and rate them as appropriate. For difficult projects/tasks it is easy to assess perseverance. While some will insist on completing the task, others may give up. The teacher/assessor must be sensitive to all these behaviours and rate them accordingly.

**Important Note**

It is necessary to note that instruments for assessing the affective behaviours are generally of the rating types. The assessor or instrument developer generally determines scales of the instrument. Whichever instrument is developed to assess affective behaviours must recognize the objectives within the affective domain, in a normal school setting the outcome of the assessments are presented in result booklets. In such

a situation average ratings for the identified categories are taken into consideration. This could be illustrated using conventional report sheets as below.

Table 6: Sample Result Booklet indicating guides for ratings of the effective behavior

| **Post Primary School Management Board Senior Secondary Termly Report** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Name of School: ……<br>Name of Student:……<br>Admission Number:……<br>Session: …… | | | | | | Key to Grades<br>5 – Excellent<br>4 – Good<br>3 – Fair<br>2 – Poor<br>1 – Very Poor | 1st Rating | 2nd Ratings | 3rd Ratings | Total |
| Key to Grades<br>A (distinction) 70% and above<br>C (Credit) 55 – 69%<br>P (Pass) 40 – 54%<br>F (Fail) Below 40% | **CA** | **End of Term** | **Total Score** | **Class Average** | | **AFFECTIVE** | | | | |
| Core Subjects | | | | | | Receiving | | | | |
| English Lang. | | | | | | | | | | |
| Igbo | | | | | | | | | | |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Mathematics (Gen). | | | | | | Responding | | | | |
| Physics | | | | | | | | | | |
| Biology | | | | | | Valuing | | | | |
| Chemistry | | | | | | Organizing | | | | |
| | | | | | | | | | | |
| | | | | | | Value | | | | |
| | | | | | | Characterizing | | | | |
| **ELECTIVES** | | | | | | **PSYCHOMOTOR** | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |

Form Master/Mistress Comment: …….

Name:……..

Signature:…….

Principals' Comments:……...

**Development of Tests For Assessing Psychomotor Behaviours**

As Mehrens and Lehmann (1991:35) noted "psychomotor domain includes objectives related to muscular or motor skills, manipulation of materials and objects and neuromuscular co-ordination". In the real sense, psychomotor skills are functions of motor & perceptual, spatial or mechanical aptitudes. The behaviours are however governed by cognitive and the affective dispositions. It therefore takes a very careful observation to isolate the skills from those of the cognitive and affective. The actual demonstration of psychomotor learning is by physical skills: coordination, dexterity, manipulation, strength, speed; actions which demonstrate the fine motor skills such as use of precision

instruments or tools, or actions which evidence gross motor skills such as the use of the body in dance or athletic performance.

Harrow (1972 and Simpson (1972) made separate attempts to develop taxonomies of the psychomotor domain. Both the Harrow and the Simpson taxonomies are acceptable to psychologists and evaluators in particular.

**Harrows Taxonomy**

This taxonomy developed by Harrow in 1972 is most popularly used for assessing behaviours that particularly has to do with physical body movement. Teachers of physical education and dance have used it extensively in primary schools. The categories identified in Harrows taxonomy are:

- Reflex movement
- Basic Fundamental movement
- Perceptual Abilities
- Physical Abilities
- Skilled Movements
- Non-discursive communication

**Reflex Movement**

The reflex movements are actions elicited without learning in response to some stimuli. Examples include: flexion, extension, stretch, postural adjustments. This type of movement sometimes occurs spontaneously.

**Basic Fundamental Movement:**

These types of movements according to Harrow (1972) are "inherent movement patterns which are formed by combining of reflex movements and are the basis for complex skilled movements". Examples are: walking, running, pushing, twisting, gripping, grasping, manipulating etc.

**Perceptual Abilities**
This, according to Harrow is the "interpretation of various stimuli that enable one to make adjustments to the environment:. Visual, auditory, kinaesthetic or tactile discrimination suggests cognitive as well as psychomotor behaviors. Examples include: coordinated movements such as jumping rope, punting or catching.

**Physical Abilities**
Behaviors in this psychomotor domain require endurance, strength, vigour and agility which produce a sound, efficiently functioning body (Harrow, 1972). Harrow gave the examples as all activities, which require strenuous effort for long periods of time; muscular exertion; a quick, wide range of motion at the hip joints; and quick, precise movements.

**Skilled Movements**
According to Harrows (1972) category, skilled movements are the result of the acquisition of a degree of efficiency when performing a complex task. Examples are: all skilled activities obvious in sports, recreation and dance.

**Non-Discursive Communication**
This according to Harrow (1972) is the communication through bodily movements ranging from facial expressions through sophisticated choreographic. Examples include: body postures, gestures and facial expr4essions efficiently executed in skilled dance movement and choreographic.

**Procedures for Assessing Objectives in the Harrow's Psychomotor Domain**
It is important to note that both the cognitive and psychomotor behaviours are aspects of ability and all instruments/tests for assessing cognitive and psychomotor behaviours are categorized as ability tests.

In physical education, these behaviours are content based and as such the tests are drawn from specified contents of a curriculum, just like in the cognitive tests, tests of psychomotor are either achievement or aptitude test. The only difference is in the multiplicity of instruments in the assessment of psychomotor behaviours. Let us take a look at the tools for the assessment of the various objectives of the psychomotor domains.

**Reflex movement**

Both physical and psychological tools are employed in assessing this objective. It depends to a large extent on the specific behaviour in questions. Many reflexes are assessed with medical kids, others are assessed through simple observation and rating. Researchers, especially those in human kinetics are advised to employ instruments that strictly address the behaviours in question. In most cases organic function tests are employed in assessing reflex movements. Such tools include kits for determining pulse rate, pulse pressure, standing and sitting blood pressures etc.

**Basic Fundamental movement**

Determination of "inherent movement patterns which are formed by combining of reflex movements are usually measured using standard kits. Manuometer is usually employed in assessing strength of the grip (finger flexors). In the same vein Tensiometer is employed in measuring the pulling force of a cable.

**Perceptual Abilities**

The assessment of perceptual motor is hinged on the premise that the efficiency of the higher thought processes is a direct function of the basic motor abilities upon which they are based. Mathews (1983) argued that for a child's higher thought processes to function at their best his/her neuromuscular development must be adequate. The essence of this category in the psychomotor domain is to find out those with

retarded motor development. The teacher in developing a scorecard must ensure that the test is sensitive to a sharp assessment of the ways in which a given task is accomplished e.g. in walking a beam the assessor must focus on whether the task is performed in "easy, relaxed and coordinated movement" or is he "stiff, fearful and unrelaxed"? (Mathews 1983:197). The emphasis here is that the items of the scorecard for perceptual ability must focus strictly on perceptual motor coordination whether it is in dance, field events or any other psychomotor task.

**Physical Abilities**

In this category we experience a combination of behaviour that requires a number of measurements. Measures of physical fitness/ability comprise the assessment of muscular performance, organic functions and a combination of the two. Tests in this category involves those that are skillfully designed to m wide range of muscular activities involving the larger muscles of the body, such as running, pull-ups, squat jumps and broad jumps. It also includes organic function tests involving similar measures as seen in basic reflex category and also a combination where effects of a specific exercise are recorded in terms of pulse rate, pulse pressure and blood pressure usually both before and after exercise. A Scorecard is usually provided when rating physical abilities.

**Skilled Movements**

Skilled movement is usually assessed using skill tests. In developing skill test, the test developer is concerned primarily with a combination of the most essential skills required for a particular psychomotor activity e.g. baseball, football etc. In developing a test for measuring skilled movement in football field the test developers should focus specifically on such skills as dribbling skills, speedy movement will ball, exact short and long passing, "chesting" and heading skills. In the rating schedules the test developer must first and foremost list all the

skills involved in the activities he intends to test and ensure that these skills are included in the scorecard.

**Non-discursive communication**

Bodily movements ranging from facial expressions through sophisticated choreographies are usually assessed in the field of human kinetics. The test developers must ensure a comprehensive listing of non-discursive behaviours inherent in specific psychomotor tasks and ensure their inclusion in a scorecard developed for the purpose of assessing the behaviour. Through simple observations such behaviours are rated using a scorecard specifically designed for the particular psychomotor behaviour.

**Development of Tests for Assessing Cognitive Behaviours**

Bloom published the Handbook of the cognitive domain in 1956 and identified objectives namely: knowledge, comprehension, application, analysis, synthesis and evaluation. The Bloom taxonomy has been a major guide in instrumentation and determination of validity of cognitive tests. Let then look at the objectives.

**Knowledge**

Mehrens and Lehman defined knowledge simply as remembering of previously learnt materials. This includes knowledge of specifics (e.g. terminology and specific facts), knowledge of ways and means will of dealing with specifics (e.g. knowledge of conventions, trends & sequences, and categories, criteria and knowledge of methodology), knowledge of universals and abstraction in a field (e.g. knowledge of principles, generalization, theories and structures). In generating items for this particular objective the test developer must ensure that the item measures only knowledge as it concerns recall of previously learnt materials.

**Comprehension**
This has to do with the ability to understand, to know, to recognize, realize or comprehend the meaning of materials. Comprehension skill entails ability to translate, interpret and extrapolate. In generating items to measure comprehension the item developer must ensure that the items measure comprehension only. Table *2A* provides definite guide to ensure that the items are tied to the specific objective in question.

**Application**
This has to do with the ability to cross-fertilize knowledge. It refers to the ability to apply what has been learnt in diverse or new areas. Items of this objective must measure the ability of the learner to apply previous knowledge in new areas or tasks.

**Analysis**
Analysis as an objective in the cognitive domain refers to the ability to "break materials down into specific parts so that the overall organizational structure may be comprehended" (Mehrens and Lehman, 1991: 32). It involves a thorough examination and scrutiny of specific materials in such a way that constituent parts are understood and can be isolated. This includes analysis of elements, analysis of relationships, and analysis of organizational principles. Items in this objective should focus on the ability of the learner to isolate constituent parts through clear analysis, distinctions, classifications, discriminations, categorizing, deductions, comparisons etc.

**Synthesis**
Synthesis here refers to creation, amalgamation, forming a whole or blending. It deals with ability of the learners to put learnt materials or parts together to form a whole. This objective includes production of a unique communication, a plan or proposed set of operations, and Derivation of a set of abstractions.

**Evaluation**

This involves value judgement. It has to do with the ability to judge the worth of a material for a specified purpose. This could be judgement in terms of internal evidence (e.g. accuracy/accuracies, consistency/consistencies, fallacies, reliability, flaws, errors, precision, and exactness) or judgement in terms of external criteria (e.g. ends, means, efficiency, economy/economies, utility, alternatives, causes of action, standard, theories, generalizations). This is the most complex of the six objectives and is better assessed for higher level -learners.

In generating items care should be taken to ensure appropriateness of tenses and use of infinitives to ensure that objectives are not misdirected. A guide is provided below.

**Table 7 : Taxonomies classification**

| **Key Words** | | |
|---|---|---|
| **Taxonomy classification** | **Examples of infinitives** | **Examples of direct objects** |
| Knowledge | To define, to distinguish, to acquire, to identify, to recall, to recognize | **Knowledge of terminology:** Vocabulary, terms, terminology, meaning(s), definitions, referents, elements<br>**Knowledge of Specifics:** Facts, factual information, (sources), (names), (dates), (events), (persons), (places), (time periods), properties, examples, phenomena.<br>**Knowledge of Convention:** Forms, conventions, uses, usage, rules, ways, devices, symbols, representations, style(s), format(s). |

| | | |
|---|---|---|
| | | **Knowledge of Trends:** Actions, processes, movement(s), development(s), trend(s), sequence(s), cause(s), relationship(s), forces, influences. **Knowledge of Classification and categories:** Area(s), type(s), feature(s), class(es), set(s), division(s), arrangement(s), classification(s), category/categories. **Knowledge of methodology:** methods, techniques, approaches, uses, procedures, treatments. **Knowledge of principles:** principles(s), generalization(s), proposition(s), fundamentals, laws, principal elements, implication(s). Knowledge of theories and structures: theories, bases, interrelations, structure(s), organization(s), formulation(s). |
| Comprehension | To translate, to transform, to give in words, to illustrate, to prepare, to read, to represent, to change, to | **Translation:** meaning(s), definitions, abstractions, representations, words, phrases. **Interpretation:** relevancies, relationships, essentials, aspects, new view(s), |

| | | |
|---|---|---|
| | rephrases, to restate, to interpret, to reorder, to rearrange, to differentiates, to distinguish, to make, to draw, to explain, to demonstrate, to estimate, to infer, to conclude, to predict, to interpolate, to extrapolate, to fill in | qualifications conclusions, methods, theories, abstractions. **Extrapolation:** consequences, implications, conclusions, factors, ramifications, meanings, corollaries, effects, probabilities |
| **Application** | To apply, to generalize, to relate, to choose, to develop, to organize, to use, to employ, to transfer, to restructure, to classify | Principles, laws, conclusions, effects, methods, theories, abstractions, situations, generalizations, processes, phenomena, procedures |
| **Analysis** | To analyze, to distinguish, to detect, to classify, to discriminate, to recognize, to categorize, to deduce, to contrast to | **Analysis of elements:** elements, hypothesis/hypotheses, conclusions, assumptions, statement (of facts), statement (of intents), arguments, particulars. **Analysis of relationships:** |

| | | |
|---|---|---|
| | compare, to distinguish, to deduce | relationships, interrelationships, relevance/relevancies, themes, evidence fallacies, arguments, cause-effect(s), consistency/consistencies, parts, ideas, assumptions. Analysis of organizational Principles: form(s), pattern(s), purpose(s), point(s) of view, techniques, bias(es), structure(s), theme(s), arrangement(s), organization(s). |
| **Synthesis** | To write, to tell, to relate, to produce, to constitute, to transmit, to originate, to modify, to document, to propose, to plan, to design, to specify, to derive, to develop, to combine, to synthesize, to classify, to deduce, to formulate. | **Production of a unique communication:** structure(s), pattern(s), product(s), performance(s), design(s), work(s), communication(s), effort(s), specifics, composition(s). **Production of a plan or proposed set of operations:** plans, objectives, specification(s), schematic(s), operations, way(s), solution(s), means. **Derivation of a set of abstractions:** phenomena, taxonomies, concept(s), scheme(s), theories, relationships, abstractions, generalizations, hypothesis/hypotheses, |

| | | |
|---|---|---|
| | | perceptions, ways, discoverie4s |
| **Evaluation** | To judge, to argue, to validate, to assess, to decide, to consider, to compare, to contrast, to standardize, to appraise | **Judgement in terms of internal evidence:** accuracy/accuracies, consistency/consistencies, fallacies, reliability, flaws, errors, precision, exactness **Judgment in terms of external criteria:** ends, means, efficiency, economy/economies, utility, alternatives, causes of action, standard, theories, generalizations. |

*Adapted from Mehrens & Lehman (1991)*

Now that we have seen the objectives of the cognitive domain, it is necessary that we ar4e guided by the specification of what constitutes each objective when developing tests.

Let us take a look at the blueprint of a newly developed Geometry Achievement test by Ezike in 2008.

Table 8: Test blueprint for a Newly Developed Geography Achievement Test (GAT)

| **Content** | **Know. 8%** | **Comp. 4%** | **App. 30%** | **Ana. 20%** | **Synt. 26%** | **Eva. 12%** | **Tot.** |
|---|---|---|---|---|---|---|---|
| Erosion 2% | - | - | 1 | - | - | - | 1 |
| Soil 2% | - | - | - | 1 | - | - | 1 |
| Climate 10% | 1 | - | 2 | 1 | 1 | - | 5 |
| Map 2% | - | - | 1 | - | - | - | 1 |
| Mountain 16% | 1 | - | 2 | 1 | 2 | 1 | 7 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Rock 14% | - | - | 2 | 2 | 2 | 1 | 7 |
| Weather 20% | - | - | 2 | 2 | 3 | 1 | 8 |
| Dune 22% | 1 | 1 | 1 | 2 | 2 | 2 | 9 |
| Longitude/Latitude 2% | 1 | - | - | - | - | - | 1 |
| Gas 6% | - | - | 1 | 1 | - | 1 | 3 |
| River 4% | - | - | 1 | - | 1 | - | 2 |
| **Total 100%** | **4** | **1** | **`13** | **10** | **11** | **6** | **45** |

*Adapted from Abonyi (2020)*

As you can see the contents are listed vertically while the objectives are listed horizontally. For both the contents and objectives percentage coverage are clearly specified to guide item generation.

**Notes:** The assessment of validity of a test depends on item representativeness of the objectives in line with a guiding test blueprint. Let it be known that a test blueprint is not a prerequisite but a prelude to content validity of tests. We will take a firm look at this in the subsequent units.

# CHAPTER 15

## VALIDATION OF INSTRUMENT

In both pure scientific and behavioral research, instruments are devised to measure what the researcher intends to measure. The major concern of the researcher is the extent to which that instrument measured what is designed to measure. According to Anastasi and Urbina (1997:113) "the validity of a test concerns what the test measures and how well it does so". Test scores, as we know, are used to draw inferences. The essence of validation is to provide some evidence on the basis of which such inferences can be substantiated. Based on this premise, Mehrens and Lehmann (1991) rightly conceptualized validity as "the extent to which certain inferences can be made accurately from and certain actions should be based on - test scores or other measurement". Mehren and his colleague drew their definition of validity from Messiek's (1989) conceptualization of the term validity'. Messiek (1989) while making a contribution in Lin's Educational Measurement notes that "Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inference and actions based on test scores or other modes of assessment". Validity, therefore, is the "worthiness of dependence or trust".

In physical sciences, the issue of validity of instrument is not at much controversy. For example when a ruler is used to measure the height of an object, or when a well calibrated flask is used in assessing the volume of a given substance, there will not be much doubts about the validity of the measure. But in behavioural research, a researcher may wish to assess interest of a group of people or other psychological constructs such as authoritarianism or neuroticism where there is no definite attribute that "points unmistakably" to such constructs. What behavioural researchers normally do is to devise indirect means to

assess such constructs even though the validity of such an assessment may be questionable.

It must be appreciated, however, that no test can be said to have a high or low validity in isolation of the particular use for which it is designed. A test that is valid in assessing neuroticism may not be valid in assessing students' reading skills, in the same vein, a test that may be valid for grouping students in terms of comprehension of passages may have low validity in grouping students with respect to motor skills. What is pertinent here is that the legitimate use of any test must be determined by the specific function for which the test has been validated.

Validity of research instruments can be determined through a number of procedures. In fact, there are about four types of validity. They are:

- Content Validity
- Construct Validity
- Criterion Related Validity
- Face Validity

**Content Validity**

Content validity, according to Mehrens and Lehmann (1991) is "related to how adequately the content of - and responses to - the test samples the domain about which inferences are to be made". Content validation is strictly the extent to which items of a test has achieved representativeness of the content from where the measuring instrument was of validation is primarily judgemental in so far as it is not possible to draw random samples of items from a universe of content. This is very much true considering the fact that "such universe exist only theoretically".

In the light of these evidence, it is pertinent, therefore, to agree with the fact that content validity cannot be expressed as coefficient but through a logical procedure that is determined, by commonsense. Nunnally

(1972) is of the opinion that the one way to ensure validity is to clearly outline the goals to be implemented in a course of instruction and then to compose examinations relating to that outline. To ensure this, adequate provision must be made by test compilers at the planning stage of the test. This is achieved by making a table of specification which provides the operational blue print which guides the test builders and ensures a sampling adequacy of the content or item representativeness. As Anastasi and Urbina (1997) rightly pointed out "these specifications should show the content areas or topics to be covered, the instructional objectives or processes to be tested, and the relative importance of individual topics and processes.

**Table 9:** A table of specification for an achievement test designed to assess the ability of students in specified units in biology is represented

| S/N | **Syllabus Section (Topics)** | **Bloom's Categories** | | | | | | **TOTAL** |
|---|---|---|---|---|---|---|---|---|
| | | Know l. 40% | Compr . 25% | Appl . 5% | Analysis 10% | Syth 5% | Eval. 5% | |
| 1. | The cell and its environment | 8 | 5 | 3 | 2 | 1 | 1 | 20 |
| 2. | Digestion of food | 12 | 8 | 4 | 3 | 2 | 1 | 30 |
| 3 | Transport in plants and animals | 24 | 15 | 9 | 6 | 3 | 3 | 60 |
| 4. | Excretion | 12 | 7 | 5 | 3 | 1 | 2 | 30 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 5. | Ecology | 24 | 15 | 9 | 6 | 3 | 3 | 60 |
| **Total** | | 80 | 50 | 30 | 20 | 10 | 10 | 200 |

It must be appreciated that a complete table of specification should cover all the six major categories in the cognitive domain as identified by Bloom et al (1956). For Beginners, however, the table of specification may exclude the higher order categories since they are not expected to acquire such skills at that stage of their academic development. Cognitive domain refers to the domain which deals with the "recall or recognition of knowledge and the development of intellectual abilities and skills" (Bloom 1956:7).

Unfortunately many researchers and test developers usually find it very difficult to prepare a table of specification. Before you can prepare a table of specification you will first of all state the intellectual objectives which your test should cover. When you have stated the intellectual objectives, then you decide on the percentage of the total item that should be allotted to each-objective. You can state the objectives as follows with the following percentage of the total scores say for a test in Agricultural Science.
Knowledge - 30%
Comprehension = 35%
Application- = 20%;
Analysis = 8%
Synthesis = 5%
Evaluation =*2%*

Having stated the objectives and assigned percentage to each of them, then you proceed to list content area you wish to cover and the percentage of the total item each of the content should have. We can list them as follows:
Meaning of Agriculture = 15%

| | | |
|---|---|---|
| Farm Tools | = | 10% |
| System of Cropping | = | 20% |
| Animal Husbandry | = | 25% |
| Soil | = | 30% |

The next step is to arrange them in tables as follows:

**Example of table of specification showing % and appropriation figure**

**Table 10:** A teacher is set to construct a 100 items objective questions below is the table of specification and the procedures.

| | Knowledge 15% | Comprehension 20% | Application 25% | Analysis 10% | Syntax 18% | Evaluation 12% | Total 100% |
|---|---|---|---|---|---|---|---|
| Measurement | A | B | C | D | E | F | |
| | 3.75 | 5.0 | 6.25 | 2.5 | 4.5 | 3 | |
| 25% | 4 | 5 | 6 | 2 | 5 | 3 | 25 |
| Continuous | G | H | I | J | K | L | |
| Assessment | 2.25 | 3.0 | 3.75 | 1.5 | 2.7 | 1.8 | 15 |
| 15% | 2 | 3 | 4 | 1 | 3 | 2 | |
| Standardize | M | N | O | P | Q | R | |
| Test | 5.25 | 7.0 | 8.75 | 3.5 | 6.3 | 4.2 | |
| 35% | 5 | 7 | 9 | 4 | 6 | 4 | 35 |
| Evaluation | S | T | U | V | W | X | |
| | 3.75 | 5.0 | 6.25 | 2.5 | 4.5 | 3.0 | |
| 25% | 4 | 5 | 6 | 3 | 4 | 3 | 25 |
| TOT | 15 | 20 | 25 | 10 | 18 | 12 | 100 |

The alphabet in each of the cells gives the cell label.

To ascertain the number of elements in each of the cells, the formula

$$\frac{Row\ total\ \times Column\ total}{Sum\ total}$$

Cell A $\frac{25\times15}{100}$= 3.75 Cell G$\frac{15\times15}{100}$=2.25 Cell M $\frac{35\times25}{100}$=2.25 Cell S $\frac{25\times15}{100}$=3.75

| | | | |
|---|---|---|---|
| Cell B $\frac{25\times20}{100}$= 5.0 | Cell H $\frac{15\times20}{100}$=3.0 | Cell N $\frac{35\times20}{100}$=7 | Cell T $\frac{25\times20}{100}$=5 |
| Cell C $\frac{25\times25}{100}$= 6.25 | Cell I $\frac{15\times25}{100}$=3.75 | Cell O $\frac{35\times25}{100}$=8.75 | Cell U $\frac{25\times25}{100}$=6.25 |
| Cell D $\frac{25\times10}{100}$=2.5 | Cell J $\frac{15\times10}{100}$=1.5 | Cell P $\frac{35\times10}{100}$=3.5 | Cell V $\frac{25\times10}{100}$=2.5 |
| Cell E $\frac{25\times18}{100}$=4.5 | Cell K $\frac{15\times18}{100}$=2.7 | Cell Q $\frac{35\times18}{100}$=6.3 | Cell W $\frac{25\times18}{100}$=4.5 |
| Cell F $\frac{25\times12}{100}$=3 | Cell L $\frac{15\times12}{100}$=1.8 | Cell R $\frac{35\times12}{100}$=4.2 | Cell X $\frac{25\times12}{100}$=3 |

**Construct Validity**

Construct *Validity*, according to Mehrens and Lehmann (1991) "is the degree to which one can infer certain constructs in a psychological theory from the test scores". Construct validity deals with the extent to which an instrument is said to measure a theoretical construct or traits such as verbal fluency, neuroticism, and anxiety (Anastasi & Urbina 1997). It refers to the degree to which scores on a measure permits inference about underlying traits. Nunnally (1972) observed that trait measures are constructs in the sense that they are things that scientists literally put together to account for phenomena in the world. They do not exist as visible event in daily life. For example intelligence, paranoia, compulsiveness, motivation, and anxiety do not represent simple observable events instead "they stand for devices which are employed to explain human behaviour".

Assuming a researcher is interested in studying a construct such as attitude to science, what he should do is to rely on theories about the various dimensions of attitude to science. This may include enjoyment of science, vocational interest in science, normality of science, and Leisure interest in science. When a researcher wants to develop an attitude to science scale, what he should do is to ensure that constructs covered by the theories are adequately reflected in the test scores.

Various techniques utilized in construct validation include age differentiation, correlates with other tests, factor analysis, internal consistency and effects of experimental variables on the test scores. Although a number of approaches have been employed in construct validation, the most outstanding, though explicitly complex is the factor analyses. This section provides a concise and understandable principles and basic applications of the factor analysis as a measure of construct validity.

Ferguson and Takane (1989) conceived factor analysis as "a multivariate statistical method which is used in the analysis of tables, of matrices, of correlation coefficients." The main focus of this procedure is to "simplify the description of data by reducing the number of necessary variables or dimensions. As Anastasi and Urbina (1997) rightfully pointed out, "if we find that five factors are sufficient to account for all the common variance in a battery of 20 tests, we can for most purposes substitute 5 scores for the original 20 without sacrificing essential information".

A researcher who has developed, say a Likert-type attitude scale for a particular study and wish to ascertain its construct validity should adopt the following procedures.

1. Administer the instrument to a representative sample of the population (sample size in this case will depend on the population).

2. Compute the correlations of each item with every other items in the test.
3. Apply orthogonal (varimax) rotation in rotating the axis.
4. Assess the factor loading from the resulting table of varimax rotated factor matrix.
5. Adopt a criteria for accepting an item in terms of its factor loading (this varies with authors e.g. 0.3 was recommended by Schuster and Miiland (1978), 0.35 by meredith (1969); 0.4 by Leak (1982) and 0.5 by Plake and Parker (1982). The most popular and accepted criterion in current literature is 0.35 by Meredith (1969).
6. Drop items that fail to attain the factor loading standard which you have adopted. Also drop items that are loaded on more than one factor. Such items are said to be factorially impure.

It must, however, be appreciated that the process of factor analysis is highly complex and requires the assistance of a computer. Instrument developer are therefore, advised to key in the responses of the pilot test of the instrument on individual item basis in the multiple purpose coding form - which should be fed into the computer by experts. A typical example of a computer printout of a factor matrix for analysis of principal component is shown in the table 4.7. Let us illustrate this using the work of Abonyi (2003).

Abonyi (2003) developed and validated a biology interest inventory. He employed construct validation .procedure.
A total of one hundred and fifty (150) Senior Secondary School students offering biology were used in this instrumentation research. The researcher initially generated a total of sixty items of the Likert-type. These items were intended to address all aspects of interest in biology. The following aspects of interests were taken into consideration during item generation;

- Vocational interest;
- Leisure interest;

- Academic interest; and
- General interest.

The researcher took the basic rules of Liker-type scales into consideration in structuring of the items, i.e. ensuring that there are equal numbers of positively and negatively directed items and also ensuring that the items are mixed up before subjecting them to construct validation.

After the item generation the 60-item inventory was given to three other specialists in measurement and Evaluation, two specialists in Science Education and one specialist in psychology for face validation. The specialists in measurement and evaluation screened the items in terms of general test format and appropriateness of the scale format while specialist in science education assessed the relevance and appropriateness of the items as it pertains to biology. On the other hand the psychologist took care of the item structures as it pertains to the specified group of respondents.

After the face validation 25 items were dropped. The remaining 35 items were re-structured in line with the recommendations of the Specialists.

**Criterion-related validity**

This type of validation has to do with the assessment of the extent to which test scores are related to some independent external measures which could be referred to as criteria. Simply put, Anastasi and his colleagues described this type of validity as "the procedure which indicates the effectiveness of test in predicting an individual's performance in specified activities." As Kerhnger (1992) rightfully pointed out, this aspect of validity is studied by comparing tests score

with other external variable which actually is believed to be an unbiased assessment of the attribute being studied.

There are too basic dimensions of criterion related validity. They are the concurrent validity and the predictive validity. The major difference between these types of validity simply lies in the time when the criterion data are collected. For criterion related validity the criterion data are collected at approximately the same time, while in predictive validity the criterion data are collected at a later time.

Mehren's and Lehmann (1991) further provided another interesting distinction between concurrent and predictive validity. They noted that the other distinction is a "logical rather than a procedural one, and is based not on time but on the purpose of testing or the inference we wish to make". Their explanation is that "in concurrent validity, we are asking whether the test score can be substituted for some less efficient way of gathering criterion data (such as using a score from a group scholastic aptitude test instead of a more expensive-to-gather individual aptitude test score)."
The concurrent validity may ask - Is Abonyi Psychopathic? while the predictive aspect will ask- Is Abonyi likely to be Psychopathic? What the predictive validity does not appreciate is the fact that time difference, increased learning, experience and accidental events may influence the correlation.

A student wishing to employ predictive validation procedure should ensure that his data are not contaminated by extraneous factors which were not envisaged between the time test data were collected and the time the criterion data were collected. A good example of the predictive validity is the use of S.S.C.E. examination result to predict performance in the university or the use of UME scores to predict performance in a specific course in the university. The common Entrance Examination is also a good predictor of performance of candidates in secondary

schools. The success of a predictor instrument depends entirely on the extent to which it correlates with some criteria.of successful performance.

Assuming a researcher wants to assess the predictive validity of a given test, he should adopt the following procedures:

- Administer the predictor instrument to your target population;
- Score the instrument and note individual scores in the instrument;
- At a later time (e.g. if the predictor instrument is a Common Entrance Examination later time should be at the end of the secondary education but if it is a test meant to assess all individual performance on a job which he has applied the later time should be when he is already on the job) you then look for the criterion measures;
- Then correlate the scores obtained from the predictive test from that of the criterion measure. This could be done using any measure of correlation. The correlation obtained from the two sets of scores is the index of the predictive validity.

**Face Validity**

Most often students confuse actual validation with face validity, which, in technical sense, is not validity. According to Anastasi and Urbina (1997) "face validity pertains to whether the test looks valid to the examinees who take it, the administrative personnel who decide on its use and the other technically untrained observers". They, however, advised that face validity should not be neglected because if a test should' look childish and inappropriate,,- it may generate poor cooperation among the students not minding that in actual sense, the test may be valid.

It is also very unfortunate that researchers, even doctoral researchers have resorted to gross abuse of face validation. Most of them simply write that the instrument has been given to experts in different fields

and measurement and evaluation without providing any evidence of such validation exercise in the appendices of their write-up. In many cases, it is apparent that the researcher did not even show the instrument to any expert. In some cases, researchers give their instruments to anybody they like because they do not want "unnecessary stress".

Researchers should be reminded that face validation does not prevent them from subjecting their instruments to other empirically valid procedures.

# CHAPTER 16

## Reliability of Instruments

### Concept of Reliability

Reliabilityis conceived in relation to the extent of consistency or dependence of a measuring instrument Mehrens and Lehmann (1991) defined *reliability* as "the consistency of scores obtained by the same persons when they are re-examined with the same test on different occasions, or with different sets of equivalent items, or under other variable examining conditions". Although this Anastasian concept of reliability looks quite elaborate, the insistence on "re-examination" tends to reduce the scope of reliability to simple measures of stability.

Reliability could be assessed (as we are going to experience practically in this section) without a repeated measure. What reliability measures are "the extent to which we can attribute individual differences in test scores to true differences" in the constructs or attributes being measured or whether the observed individual differences in the test scores are simply results of chance errors. We must bear in mind that "errors are random and uncorrelated with each other and with true scores" (Ferguson and Takane 1989). According to Ferguson and Takane (1989) the reliability coefficience is the "proportion of obtained variance that is true variance". This is based on the premise that the paired observations are measures of the same attribute and that $Oi = O_2$, that $\sum(Ti - \mu)^2 = NP\sigma^2T$ which implies that $\sigma_1 = \sigma_2 = \sigma_x$. Based on these derivations Ferguson and his colleague presented a quantitative view of reliability coefficience thus:

$$P_{xx} = \frac{Q^2T}{Q^2X}$$

Where $p_{xx}$ = reliability Coefficience

Reliability assessment according to Kerlinger (1992) ensures dependability, stability, consistency, predictability and accuracy.

Researchers should not be deceived into thinking that reliability ensures validity. An instrument may be very reliable and yet invalid. In any case, the reliability of any instrument must be ensured before its use can be approved for a given population. It is also worthy of note that no test is a perfectly reliable instrument unless the characteristics of the sample for which it should be used on is specified, together with the type of reliability that was measured.

The various ways through which the reliability of instruments can be estimated are:

- Measures of Stability
- Measures of equivalence
- Measures of internal consistency
- Scorer reliability

**Estimation of Stability**

An instrument is said to be stable when repeated measures obtained from the instrument for a given sample do not fluctuate. This approach is also called the test - retest *approach.* In this method the same test is given to the same group of testees on more than one occasion. Then the scores obtained by the group on the first administration are correlated with the scores obtained for the same group of testees on the second administration of the same test. The reliability coefficient in this case is simply the correlation between the two sets of scores by the same testees on the two administration of the same test. Such a reliability coefficient is known as coefficience of stability (Anastasi and Urbina 1997; Mehrens and Lehmann 1992). When the reliability index (from + 0 - + I) is above 0.50, then the instrument is reliable but when the index is below 0.50, then the instrument is unreliable.

The problem likely to be encountered with this method has to do with the time lag between the first and second administration of the test. If the time difference between the first and second administration of the instrument is short, there will be the likelihood that the subjects will

recall the items through sheer memorization. On the other hand if the time difference is so long to the extent that the testee's memory of the items will vanish completely, there is also the likelihood that the error variance arising from additional knowledge may set in.

This procedure is determined using correlation coefficience. When a change in one variable is associated with change in another variable, we then say that the two variables are correlated. The relationship may either be positive or negative. The stability of an instrument can be established using the following hypothetical example:

A test designed to assess Muslim Women's attitude to birth control was administered to ten Muslim women and repeated after two weeks on the same ten women. The scoring was done and recorded.

Table 11: Scores of two *administration of a test* on the some *sample of Christians*

| **Christian women** | **First testing (X)** | **Second testing (Y)** |
|---|---|---|
| A. | 78 | 65 |
| B. | 76 | 70 |
| C. | 68 | 72 |
| D. | 76 | 61 |
| E. | 75 | 77 |
| F. | 72 | 67 |
| G. | 53 | 53 |
| H. | 65 | 60 |
| I. | 61 | 63 |
| J. | 64 | 70 |

In assessing the stability of the instrument, which yielded these set of scores, we first list the procedures thus:

Determine the type of correlation you wish to apply. We are at liberty to employ any measure of relationship. In this case we may decide, just as a matter of choice to use the Spearman Rank Order correlation procedure. Having chosen the spearman Hank Order procedure, we then proceed with the necessary steps in this procedure by

(1) Ranking each of the scores i.e. XR and YR
(2) Finding the difference between the ranks XR -YR i.e. D
(3) Squaring each of the difference i.e. $D^2$
(4) Summing the squares of the difference i.e. $\sum D^2$

Table 12: Ranking of the scores *of* two *administration* of *a test on the same sample of Christian* Women

| **Christian Women** | **1st test (X)** | **2nd test (Y)** | **Rank for X(XR)** | **Rank for Y(YR)** | **XR-YR (D)** | **$D^2$** |
|---|---|---|---|---|---|---|
| A. | 78 | 65 | 1 | 6 | -5 | 25 |
| B. | 76 | 70 | 2.5 | 3.5 | -1 | 1 |
| C. | 68 | 72 | 6 | 2 | 4 | 16 |
| D. | 76 | 61 | 2.5 | 8 | -5.5 | 30.25 |
| E | 75 | 77 | 4 | 1 | 3 | 9 |
| F. | 72 | 67 | 5 | 5 | 0 | 0 |
| G. | 53 | 53 | 10 | 10 | 0 | 0 |
| H. | 65 | 60 | 7 | 9 | -2 | 4 |
| 1. | 61 | 63 | 9 | 7 | 2 | 4 |
| J. | 64 | 70 | 8 | 3.5 | 4.5 | 20.25 |
| | | | | | | $\sum D^2$= 109.5 |

Apply the formula having computed the necessary quantities. The formula is

$$rho = \frac{1 - 6\sum D^2}{N(N^2 - 1)}$$

$$rho = \frac{1 - 6 \times 109.5}{10(10^2 - 1)}$$

$$rho = \frac{1 - 657}{999}$$

$$= 1 - 0.6576 = 0.3424 = 0.34$$

Since the correlation index is below the average (0.50), it implies that the response of the Muslim women to the instrument is not stable which implies that the instrument is unreliable. Whenever a low reliability coefficient is obtained the researcher should discard the instrument and look for another instrument that will yield a higher reliability coefficient. A researcher should not merely assume that his instrument is reliable based on its face value unless he subjects it to empirical assessment using appropriate procedures and with the population for whom the instrument was designed.

**Estimation of Equivalence**

In behavioural research, most instruments are psychological in orientation. Unlike achievement tests, such tests may involve constructs whose measurement may not be dependent on a "specific set of questions". In the equivalent form approach the subjects are tested with one form of the test in one occasion and the alternate form in the second occasion (Anastasi 1997). The essence is to control for memorization of items.

A major limitation in this approach is the difficulty in developing tests that are truly parallel. In this case a parallel test should be strictly parallel and not near parallel. This procedure does not, however, control for practice effect. What we mean here is that subjects may bring carry over effect from the first test to the second form of the test.

In developing a parallel form of a test, attention should be paid to the degree of similarity in such aspects as content, item difficulty, item validity, means and variance as well as mental processes required for answering the .items correctly (Loevinger 1966; Lord 1970). It must also be noted that the second form must not necessarily be a reverse of the first test. What should be borne in mine is the relatedness of the items in the two tests. When this test is well developed it may reduce memory effects which is prominent in test re-test and which tend to influence reliability index.

In order to appreciate this method more clearly let us use the instrument designed to assess the self concept of secondary school biology students. An instrument designed to assess the self-concept of secondary school biology students was designed in two equivalent forms and administered to five biology students randomly sampled from SSII class in Boys' High School Orba in succession. The responses were scored and shown in table 5.3

Table 13: Scores of *a sample of SSII students on alternate forms* of a *Self Concept*

| **Students** | **Scores on first form** | **Scores on second form** |
|---|---|---|
| A | 60 | 58 |
| B | 46 | 50 |
| C | 39 | 43 |
| D | 71 | 66 |
| E | 53 | 59 |

In order to determine the equivalent form reliability index the researcher computes the correlation coefficient for the two groups of scores. The researcher may decide to use any type of correlation procedure in determining the reliability of the instrument. In this case

the researcher employed the Pearson's Product Moment Procedure as shown below. Table 14

| Students | X(scores of the 1st form) | Y(scores of the 2nd form | X~x | $(X-x)^2$ | Y-y | $(Y-y)^2$ | (X-x) (Y-y) |
|---|---|---|---|---|---|---|---|
| A | 60 | 58 | 6.2 | 38.44 | 2.8 | 7.84 | 17.36 |
| B | 46 | 50 | -7.8 | 60.84 | -5.2 | 27.04 | 40.56 |
| C | 39 | 43 | -14.8 | 219.04 | -12.2 | 148.84 | 180.56 |
| D | 71 | 66 | 17.2 | 295.84 | 10.8 | 116.64 | 185.76 |
| E | 53 | 59 | 5.2 | 27.04 | 3.8 | 14.44 | 19.76 |
| **TOTAL** | | | | **641.2** | | **314.8** | **444** |

X = 53.8
Y = 55.2

$$r = \frac{\sum(X - x)(Y - y)}{\sqrt{\sum(X - x)2 \sum(Y - y)2}}$$

Having gotten all the necessary quantities we then substitute thus:

$$r = \frac{44}{\sqrt{(641.2)(314.8)}}$$

$$r = \frac{444}{449.276} = 0.98$$

A correlation index of 0.98 indicates a very high reliability. The test therefore is very reliable in assessing the self-concept of senior secondary biology students.

It should be noted that the instrument or test should not be too brief as the example used in this text. More items should be included so as to

cover all dimensions of the construct under study. Moreover, the number of subjects to be used should be relatively large (at least 30 depending on the population size).

**Estimation of Internal Consistency**

We have already treated two approaches in the assessment of reliability which requires two testing sessions. In all measures of internal consistency, the test is only administered once. There are four approaches in the estimation of internal consistency. They are:

a. the spilt-half estimates
b. the Kuder - Richardson's estimates;
c. the Cronbach Procedure; and
d. the Hoyt's procedure.

**The spilt-half estimates**

The difference between the equivalent form and the Spilt half is just that the spilt half is given in a single administration while the equivalent form is given in two separate administrations. In addition, the spilt-half procedure has the two equivalent forms in one test.

The procedure in the spilt-half approach involves splitting one test into two in such a way that two scores are obtained for an individual from one test The correlation of the two set of scores obtained from each half of the single test is computed using any procedure for assessing correlation. The most popular approach is the Pearson's Product Moment Procedure. The computed V does not, however, represent the reliability estimate of the whole test rather it is "an estimate of the reliability of a test only half as long as the original." As Mehrens and Lehmann 1991) rightfully emphasized, a correlation factor is needed to be applied in determining the reliability of the whole test. This is achieved using the Spearman Brown Prophecy Formula.

$$R_n = \frac{2rt}{1 + rt}$$

where $r\pi$ = estimated reliability of the whole test
rt = reliability of the half test.

Assuming the computed relationship between the two halves of a test, say a self concept scale or a mathematics achievement test is 0.74, the reliability of the whole test is computed thus:

$$m = \frac{2x0.74}{1 + 0.74} = \frac{1.48}{1.74} = 0.85$$

The test whose two halve has a correlation coefficient of 0.74 is now shown to have a reliability index of 0.85.

**Kuder-Richardson Estimates**

Both the Kuder Richardson approach and Coefficient alpha otherwise known as the Cronbach alpha are tests of internal consistency. This approach is based on the consistency of the testee's response to the various items that make up the test. This provides a measure of both equivalence and homogeneity.

The K-R 20 procedure can only be applied to tests that are scored dichotomously e.g. either pass or fail, right or wrong. The approach was developed by Kuder and Richardson and used for estimating internal consistency, just as in the Spilt half estimate, this estimate is found from a single administration of an instrument. This technique does not require two halves of the test; rather, it involves a thorough examination of testee's response or performance on each item of the test.

There are two approaches devised by Kuder and Richardson. They are the K-R 20 approach and the K-R 21 approach. The formulas for these approaches are represented thus:

$$K - R20 = \frac{n}{n-1}\left(\frac{SD^2 - \sum pq}{SDt^2}\right)$$

*where;*

*n* = number of items in the test
$SDt^2$ = variance of the total test
p = proportion of people who answered the items correctly
q = proportion of the people who answered the items incorrectly

the formula for the K - R 21 approach is

$$K - R21 = \left(\frac{n}{n-1}\right)1 - \frac{xt(n - xt)}{(nSD^2t)}$$

where
n = number items in the test
xt = mean of the total test
$SD^2t$ = variance of the total test
For the avoidance of doubt, it must be re-emphasized that the K-R-21 assumes all items to be of equal difficulty, which is practically unattainable in test development. Although it is simple in operation, its use is bound to carry along a number of errors, which leads to the reduction of the reliability estimate.

Let us now take a practical example draw from a biology achievement test which was administered to ten Senior Secondary school Biology Students. The test is a multiple choice test generally called objective test. The number of items in the test is ten. The following procedures were employed in estimating the internal consistency of the test using the K-R 20 Procedure:

1. First and foremost the test was administered on a sample of the population under normal testing conditions.
2. The responses were scored like all multiple choice tests.
3. Based on the responses, the researcher then determined the number of respondents/testees failing and those passing each of the items.

4. The next step is to compute the proportion of testees passing (p) and those failing each item of the test (q).
5. Then, for each item, multiply the proportion that passed by the proportion that failed i.e. (pq).
6. Sum up the pq to get Spq.

For the Biology test the summary of the procedures can be represented in Table 15.

*Hypothetical data representing students' achievement in a Biology achievement* test *and reliability procedures for the K.-R 20 approach*

| Items | No. Passing | No. Failing | Proportion Passing (p) | Proportion Failing (q) | Pq |
|---|---|---|---|---|---|
| 1 | 6 | 4 | 0:6 | 0.4 | 0.24 |
| 2 | 5 | 5 | 0.5 | 0.5 | 0.25 |
| 3 | 8 | 2 | 0.8 | 0.2 | 0.16 |
| 4 | 7 | 3 | 0.7 | 0.3 | 0.2! |
| 5 | 6 | 4 | 0.6 | 0.4 | 0.24 |
| 6 | 5 | 5 | 0.5 | 0.5 | 0.25 |
| 7 | 5 | 5 | 0.5 | 0.5 | '0.25 |
| 8 | 4 | 6 | 0.4 | 0.6 | 0.24 |
| 9 | 5 | 5 | 0.5 | 0.5 | 0.25 |
| 10 | 7 | 3 | 0.7 | 0.3 | 0.21 |
| **∑pq= 2.30** | | | | | |

Having completed these steps, then determined the total score of each of the testees in the whole 10 - item test. For the ten testees used, we

have a group of ten scores. For these ten scores we will determine the variance of the total test scores (i.e. $SDt^2$). Variance is the standard deviation squared.

Assuming the scores for the ten candidates on the 10-item test are:

8 7 10 2 2 5 10 9 3 and 1

The variance is expected to be 12.46. This is so because the standard deviation for the scores is 3.5292. So if you square it you get 12.455.

When this is done, then you apply the formula and proceed with the necessary substitutions

$$K - R20 = \left(\frac{n}{n-1}\right)\left(\frac{SDt2 - \sum pq}{SDt^2}\right)$$

$n = 10$

$SDt^2 = 12.46$

$\sum pq = 2.30$

$$= \frac{10}{10-1} \times \frac{12.46 - 2.30}{12.46}$$

$$= 1.11X0.8154 = 0.91$$

**The Cronbach Procedure [Coefficient Alpha A],**

Cronbach L J. in 1951 sough for a procedure that could be applied in estimating internal consistency of tests that are not dichotomously scored. We appreciate the fact that some personality scales and some essay type questions could take on a range of values and as such could not be assessed in terms of proportion failing or passing an item. What Cronbach did was to substitute the-Zpq in the K-R 20 with £vi to take care of variability of responses in each item of the test.

Cronbach [1951] is of the opinion that although two halves of a test may look alike, there is every likelihood that the variation in the two halves may be very uncompromisable. The coefficient alpha

developed by Cronbach in 1951 is a generalization of the Kuder Richardson 20 approach when items are non-dichotomously scored.

The Kuder-Richardson method is an overall measure of internal consistency, but a test which is not internally homogeneous may nonetheless have a high correlation with a carefully planned equivalent form. In fact items within each test may correlate zero and yet the two tests may correlate perfectly if there is item to item correspondence of content.

Alpha, according to Cronbach is the average of all the possible split-half coefficient for a given test juxtaposed with further analysis of variance of split half coefficient from split to split and with an examination of the relation of alpha to item homogeneity. This relation led to the recommendation for estimating coefficient of equivalence and homogeneity.

In their assumption, Brown and Spearman sought to predict correlation with a test whose haves are 'c' and 'd' possessing data from a test whose halves are 'a' and V and that

Va = Vb = Vc = Vd and Vab = Vac = Vad = Vbc = Vbd = Vcd.

Cronbach argued that this assumption is far from general. According to him for many split halves Va * Vb and an equivalent form confirming to the data is practically impossible.

Kuder and Richardson assumed that corresponding items in test and parallel tests have the same common content and same specific content i.e. that they are as alike as two trials of the same item would be. Otherwise they took the zero internal retest correlation as their standard. Gutman also began his derivation by defining equivalent tests as identical.

In the coefficient of stability, variance in total scores between trials (within persons) is regarded as a source of error, and variance in specific factors (between items within persons) within trials is regarded as a true variance. In the coefficience of equivalence, such as alpha, this is just reversed. Variance in specific factors is treated as error.

Variation between trials is non-existent and does not reduce true variance (Cronbach 1946).

To be very analytic, alpha or any other coefficient of equivalence treats the specific content of an item as error but the coefficient of precision treats it as part of the thing being measured.

The coefficient alpha can be represented thus;

$$\infty = \left(\frac{K}{K-1}\right)\left(\frac{1 - \sum vi}{vt}\right)$$

where

k = number of items

*vi* = variance of individual items of the test

vt = variance of the total test

It must, be noted that the Cronbach procedure is allergic to the following:

- Instruments that are dichotomously scored
- Instruments that are not balanced (i.e. number of positively directed and negatively directed items)
- Poor representation of constructs.

Assuming we want to assess the internal consistency of a Likert-type instrument say a 20-item conflict resolution strategy scale for Nigerian universities, the following procedures are adopted:

- Administer the instrument to a sample of the population.
- Score responses on individual items of the scale (i.e. specify for each item the number of respondents while indicating the direction of the item).
- Calculate for each item the variance of responses i.e. vi (the direction of the item must be considered)
- Then sum up the variance for all the items to get Evi

For the hypothetical 20-item Likert-type scale on conflict resolution strategies, the procedure for determining the Evi is shown thus.

*Table 16 : Variance* of responses on items of a Likert-type *scale*

| | Item | 4<br>1<br>SA | 3<br>2<br>A | 2<br>3<br>D | 1<br>4<br>SD | Variance |
|---|---|---|---|---|---|---|
| Negative | 1 | 18 | II | 16 | 5 | 1.07 |
| Positive | 2 | 32 | 6 | 9 | 3 | 0.96 |
| Positive | 3 | 26 | 8 | II | 5 | 1.15 |
| Positive | 4 | 4 | 7 | 21 | 18 | 0.83 |
| Negative | 5 | 2 | 13 | 27 | 8 | 0.55 |
| Positive | 6 | 6 | 9 | 23 | 12 | 0.88 |
| Positive | 7 | 17 | 21 | 8 | 4 | 0.83 |
| Negative | 8 | 34 | 12 | 4 | 0 | 0.40 |
| Negative | 9 | 0 | 3 | 23 | 24 | 0.37 |
| Negative | 10 | 8 | 9 | 22 | 11 | 0.98 |
| Positive | II | 4 | 8 | 10 | 26 | 1.0 |
| Negative | 12 | 1 | 3 | 9 | 37 | 0.48 |
| Positive | 13 | 13 | II | 16 | 10 | 0.19 |
| Positive | 14 | 9 | 14 | 9 | 8 | 1.20 |
| Positive | 15 | 36 | 8 | 6 | 0 | 0.48 |
| Negative | 16 | 21 | 15 | 10 | 4 | 0.95 |
| Positive | 17 | 28 | 16 | 3 | 3 | 0.73 |
| Negative | 18 | 41 | 6 | 2 | 1 | 0.40 |
| Negative | 19 | 4 | 9 | 13 | 24 | 0.98 |
| Negative | 20 | 0 | 4 | 7 | 39 | 0.38 |
| | | | | | | **∑vi 16.16** |

Having achieved this, the next stage is to compute the variance of the total test i.e. vt. For each of the SO respondents, the sum of the responses (in raw scores) in all the 20 items is determined. This will yield a group of fifty scores. For the same instrument, the scores of the 50-respondents are shown below:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 57, | 72, | 79, | 76, | 59 | 55, | 86, | 76, | 61, | 85 |
| 70, | 58, | 64, | 55, | 74 | 74, | 61, | 69, | 72, | 78 |
| 75, | 39, | 79, | 46, | 62 | 81, | 70, | 71, | 78, | 69 |
| 53, | 81, | 85, | 86, | 67 | 61, | 58, | 96, | 56, | 70 |
| 68, | 79, | 78, | 39, | 85 | 83, | 80, | 68, | 77, | 78 |

The variance of the total test was calculated to *be* 149.
Applying the formula

$$\infty = \left(\frac{K}{K-1}\right)\left(1 - \frac{\sum vi}{vt}\right)$$

We have:

$$\infty = \left(\frac{20}{19}\right)\left(\frac{1 - 16.16}{149}\right) = 0.93$$

An alpha of 0.93 shows that the test has a high internal consistency. The essence of this reliability test is to ascertain whether the respondents are very sincere and consistent in their responses. A respondent who earlier agreed that dialogue is a good strategy for conflict resolutions is not expected to agree with the statement that dialogue inhibits conflict resolution. Contradictory responses make an instrument unreliable and consequently reduce the internal consistency index of the instrument.

**Scorer Reliability**

In some cases, the sample size for a particular research may be too large that no one researcher can assess or score all the instruments that will be administered to the sample. In often cases the researcher employs research assistants. If the instrument is not the type that requires a definite answer, or scoring pattern, there will be the likelihood of individual differences in scoring. This invariably introduces error in the study. To control for such an error, the researcher has to assess the extent of agreement of the scorers to ensure that the scoring pattern of all the scorers are the same. This is the essence of the scorer reliability. The scorer reliability is determined through a technique developed by Kendall otherwise known as *Kendall's Coefficient of Concordance* (W).

$$W = \frac{125}{N^2 (K^3 - K)}$$

Let us illustrate this with an interview organized in Ebonyi State University to select candidates for Diploma course.

In a brief interview organized by the department of Science Education of Ebonyi State University to select candidates for a Diploma Course four lecturers in the department were asked to rate six applicants using a rating scale developed by the Department. The procedures are:

- Each of the students were asked the same question at the panel
- Each of the four lecturers rate the students independently
- The scores given to each of the students were tabulated as shown in Table 5.6
- The scores were then ranked as shown in Table 5.7
- After ranking then sum the ranks for each applicant as shown in the last column of Table 5.7
- Then apply the formula as shown to determine the index of concordance of the raters/scorers

Table 17: Row scores of the *applicants per rater/scorer*

| **Lecturers/Raters** | **Raw scores of the applicants per rater/scorer** | | | | | |
|---|---|---|---|---|---|---|
| | **A** | **B** | **C** | **d** | **e** | **f** |
| A | 48 | 37 | 60 | 55 | 38 | 40 |
| B | 45 | 40 | 65 | 58 | 36 | 50 |
| C | 45 | 33 | 50 | 60 | 38 | 40 |
| D | 65 | 60 | 53 | 50 | 48 | 70 |
| | | | | | | |

| **Lecturers/Raters** | **Ranks of the scores of applicants per rater/scorer** | | | | | |
|---|---|---|---|---|---|---|
| | **A** | **B** | **C** | **D** | **E** | **F** |
| **A** | **3** | **6** | **1** | **2** | **5** | **4** |
| **B** | **4** | **5** | **1** | **2** | **6** | **3** |
| **C** | **3** | **6** | **2** | **1** | **5** | **4** |
| **D** | **2** | **3** | **4** | **5** | **6** | **1** |
| **Rj** | **12** | **20** | **8** | **10** | **22** | **12** |

Assuming you are asked to determine the reliability, agreement or concordance of the lecturers in the ratings of the students the Kendall procedures (i.e. Kendall W) will be applied. The Kendall W is popularly called the Kendall's Coefficience of concordance or Kendall's estimate of inter-rater reliability.

$$W = \frac{125}{N^2 (K^3 - K)}$$

Ri = sum of the ranks for each applicants ( for applicant "a" the Ri = 12, for "b" Ri = 20 etc

$\sum Ri/k$ = mean of the total ranks for all the applicants =12 + 20 + 8+10 + 22 + 12 divided by 6 = 14
N= number of raters i.e. in this case it is number of lecturers (4)
K = number of candidates being rated (in this case they are 6)
S therefore is $(12- 14)^2 + (20- 14)^2 + (8- 14)^2 + (10- 14)^2 + (22- 14)^2 + (12- 14)^2 = 160$

$$W = \frac{12X160}{4^2(6^3-6)} = 0.57$$

It must be emphasized that no matter how high a reliability index may be for any of the estimates of reliability, it cannot be said to be significant unless the test developer subject such a reliability index to appropriate tests of significance at a given alpha level.

**General Considerations in reliability Assessment**

A number of factors are taken into consideration in the reliability assessment of instruments. Such factors are individual and environmental factors, length of tests, reliability assessment procedure, interval between two tests, group homogeneity, difficulty of the items, and objectivity in scoring.

**Individual and environmental factors**

There are a number of factors from the individual himself, which influence the reliability of a test. They include fatigue, health, hunger or the general emotional disposition of the individual. When a test is administered to an individual when he/she is not disposed, there is the likelihood that his/her response will not be reliable. This matter takes us to the issue of testees' readiness as an ethical consideration in testing. A testee must be disposed and ready before being subjected to a test. The testing environment is also essential. To ensure the reliability of scores obtained from a test the test must be conducted in a conducive environment Conduciveness of the testing environment is however relative. It depends on the type of test and the objective of the test?

**Length of the test**

A test is said to be more reliable when it is long. A long test provides a more adequate sample of behaviour being measure. For a long test, Mehrens and Lehman (1997) point out that "random positive and negative errors within the test have a better chance of cancelling each other out thus making the observed score (X) closer to the true score (T)." This is applicable to all research instruments/tests whether they are rating scales or observational schedules. It is also necessary to emphasize that a test need not be too long; otherwise it will appear very boring and therefore generate fatigue on the testees/respondents.

**Reliability assessment procedure**

The approach a test developer or researcher employs in determining the reliability of a test is essential. Most test developers and researchers do not know that the nature and purpose of a test determines the procedure to employ in the reliability assessment of instruments. Some ability tests are classified as either power tests or speed tests. In a power test all the testees are given enough time to attempt all the items. In such a test it is ordinarily difficult for any testee to obtain a perfect score because of the difficulty level of the test. For a speed test all testees can get all items correct once they get to the item but the time allowed for the test is so limited that no testee can get at all the items. In this case, the score difference depends on the speed of the testee (i.e. in the number of items attempted).

The method to employ in determining the reliability of power tests will obviously differ from the method for assessing reliability of speed tests. For speed test it is advisable to employ measures of stability, while for power tests, tests of internal consistency are most appropriate. There are also different patterns of scoring among tests/instruments and this is also taken into consideration in reliability assessment.

**Interval between two tests**

When reliability assessment involves two testing intervals (e.g. test retest) the interval between the first and second administration should be taken into consideration. We are aware that testees cannot remain exactly the same in the first and second administration of a test. Some would have acquired new knowledge, while some may have forgotten what they learnt or knew previously. Because of this it is advisable that the interval 'between the first and second -administration be not too long or too short. It will not be too long to avoid the effect of new knowledge or the tendency to forget what one knew before. On the other hand, the time lag should not be too short to avoid memorization effects, especially for ability tests. The use of alternate forms solves this problem, except that it is difficult to develop a true alternate form of a test.

**Group homogeneity**

Another important factor to consider in the reliability assessment of a test is group homogeneity. Mehrens and Lehmann (I99I;259) note "there is no reason to expect the precision of a person's observed score to vary as a result of group characteristics" and provided a detailed explanation of the influence of group homogeneity on test reliability using this equation:

$$Rxx = 1 - \frac{Se^2}{S_x^2}$$

In their explanations they stress that "because $Se^2$ is conceptually thought of as the variance of a person's observed score about his true score, $Se^2$ should remain constant with changes in group heterogeneity but $Sx^2$ increases with group homogeneity". They then explained that if $So^2$ remains constant and $Sx^z$ increases, r*x increases. Just as we discussed in factors that influence validity, if a population is heterogeneous, comprising about five strata, you have to ensure that samples are drawn from each stratum; otherwise both the reliability index and validity estimate will be compromised.

**Difficulty of the items**

In ability tests the difficulty of the test influences the reliability of the test scores. Reliability of test scores is affected when the test is either too easy or too difficult. The reason is because it does not make for score variability on which reliability estimates depend. If a test is too easy, almost everybody gets the answer and gets the same score. On the other hand if the test is too difficult that all the students fail almost all the questions, all the candidates also arrive at the same score, making room for no variability in scores. Such tests do not provide good reliability measures.

**Objectivity in scoring**

When the scorers are not objective in scoring the test, a true measure of reliability of the test scores cannot be established. As I noted in the previous unit, every test must be accompanied by a scoring guide or what we popularly call the marking scheme. If the test is of the essay type, scorers must be given orientation on the scoring format This will ensure the reliability of the scores obtained from such a test.

## CHAPTER
## BASIC STATISTIC FOR CONTINUOUS ASSESSMENT
### PRESENTATION AND ORGANIZATION OF DATA

Data such as scores which are collected and recorded in the way they occur without any order or arrangements or processing are called raw data. The table below represents unorganized raw scores of 50 students.

**Table 18:** ***Raw Scores of 50 Students in Geography***

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 25 | 25 | 29 | 26 | 27 | 22 | 24 | 38 | 39 | 32 | 34 | 44 | 33 | 51 | 41 |
| 25 | 21 | 28 | 14 | 33 | 33 | 15 | 27 | 36 | 20 | 55 | 16 | 33 | 47 | 16 |
| 15 | 27 | 42 | 37 | 10 | 11 | 29 | 21 | 18 | 28 | 46 | 19 | 21 | 36 | 17 |
| 46 | 40 | 34 | 27 | 29 | | | | | | | | | | |

Before we can meaningfully analyze any data like the raw scores above, the data has to be carefully organized. One common way of organizing data such as in the table above is to arrange them in frequency distribution form and present them in graphic form as the case may be.

**Frequency Distributions**

Frequency distribution is a systematic arrangement of data (scores) from the lowest to the highest taking into consideration the number of occurrence of each datum. The process involved is simply writing out the scores and as we come across that score in our distribution we put a stroke. When four strokes have been made, a fifth stroke is used to cross the four strokes (𝍸). A crossed bundle is thus five strokes. The frequencies are the sum of the strokes for the range of scores.

Table 19: Frequency Distribution of Scores

| **Record (x)** | **Tallies** | **Frequency** |
|---|---|---|
| 10 – 14 | //// | 4 |
| 15 – 19 | 𝍸 //// | 9 |
| 20 – 24 | 𝍸 / | 6 |
| 25 – 29 | 𝍸 𝍸 // | 12 |
| 30 – 34 | 𝍸 𝍸 /// | 13 |

| | | |
|---|---|---|
| 35 – 39 | ~~///~~ / | 6 |
| 40 – 44 | //// | 4 |
| 45 – 49 | /// | 3 |
| 50 – 54 | / | 1 |
| 55 – 59 | /// | 3 |

We started with 50 raw scores of 50 students. When at the end we sum up the frequencies they make up 50. In statistics, N usually represents the number of scores or categories while f represents frequency. Thus we may write $\sum f = N$.

**Grouped and Ungrouped Data**

When scores are less than thirty, we can just write down the number and carry out the tallying. For numbers more than thirty e.g. 100, tallying without grouping the scores would be difficult. This leads to what is called grouped and ungrouped data or frequency distribution.

**Ungrouped Frequency Distribution**

This refers to systematic arrangement of scores from highest to lowest in such a way that each individual score is listed and the frequency recorded.

Suppose we have the following distribution:

20, 22, 25, 20, 15, 14, 24, 21
26, 27, 20, 26, 16, 15, 15, 21

We can put it in a frequency distribution table as presented in Table 3. Scores or data are presented in ungrouped form when they are less than thirty and when the range of the data is small.

**Table 20:** ***Ungrouped Frequency Distribution of Scores***

| **Scores (x)** | **Tally** | **Frequency** |
|---|---|---|
| 27 | / | 1 |
| 26 | /// | 3 |
| 25 | /// | 3 |
| 24 | / | 1 |
| 22 | // | 2 |
| 21 | // | 2 |
| 20 | ~~///~~ | 5 |

| | | |
|---|---|---|
| 16 | // | 2 |
| 15 | /// | 3 |
| 14 | // | 2 |

**Grouped Frequency Distribution**

When the raw scores obtained are more than thirty, there is a need to group them for convenience. Given the 50 raw scores of 50 students in Table 2 above, there is need to group them as they are more than 30. To group such data, the following steps will be taken.

1. **Determine the Class Interval**

   This means the number of groups or classes one wants to have in a distribution. Statistically, class interval is mostly denoted with letter 'K'. The decision on this class interval depends on the researcher which he takes considering the number of scores he has. For our data in Table 1, we will take class interval to be 10. K

2. **Compute the class or interval width or size:**

   Class (interval) width (size) refers to the number of scores which will be in each class. This is determined by dividing the range of scores with the class interval. Class width is statistically denoted with small letter 'c' or 'i'. The formula is:

   $$c = \frac{R}{K}$$

   Where R = Range, K = Class interval

   For the data in Table 1 above, class width is: $\frac{55 - 10}{10} = 4.5 = 5$

**NOTE:** Always approximate class width to the next higher whole number. Also prefer an odd numbered class width because it will reduce the tediousness of working with fractions later in the statistical analysis of grouped data. Where you have an even numbered class

width, adjust the class interval until it yields an odd numbered class width.

**3. Group the Data Starting from Lowest Score**

Ensure that each class has a class width of 5 each. This should be constant in all classes. Then determine the frequencies by using the tally column as in table 4 below.

**Table 21: Grouped Frequency Distribution of Scores**

| Record (x) | Tallies | Frequency |
|---|---|---|
| 10 – 14 | /// | 3 |
| 15 – 19 | ~~////~~ // | 7 |
| 20 – 24 | ~~////~~ / | 6 |
| 25 – 29 | ~~////~~ ~~////~~ // | 12 |
| 30 – 34 | ~~////~~ // | 7 |
| 35 – 39 | ~~////~~ / | 6 |
| 40 – 44 | //// | 4 |
| 45 – 49 | /// | 3 |
| 50 – 54 | / | 1 |
| 55 – 59 | / | 1 |

**Cumulative Frequency (cf) Distribution**

After data has been put in frequency distribution, there is need to determine the cumulative frequency distribution of such data. Cumulative frequency is the progressive summation of the frequencies from that of the lowest score to the highest. The frequency of the lowest score is carried over to form a base. This is then added to the frequency of the next higher score or class continuously until that of the last score. This is represented in Table 5 below.

**Table 22: Cumulative Frequency Distribution of a Grouped Data**

| Classes | F | Cf |
|---|---|---|
| 55-59<br>50-54 | 1<br>1 | 50<br>49 |
| 45-49<br>40-44 | 3<br>4 | 48<br>45 |
| 35-39<br>30-34 | 6<br>7 | 41<br>35 |
| 25 - 29<br>20-24 | 12<br>6 | 28<br>16 |
| 15-19<br>10-14 | 7<br>3 | 10<br>3 |

**Graphic Representation of Frequency Distributions**

In discussing this section, the entry behaviour assumed is that the reader has known how to draw graphs. If you do not know this or have forgotten, please revise arithmetic sections that deal with graphs. We plot graphs on graph paper. We may represent the frequency distribution in a graphic form because:

a) the pictorial effect easily catches the eye, that is
b) the graph acts as a seductive slogan that holds attention. There are three methods of representing a frequency distribution graphically which we shall consider.
   i. Frequency polygon
   ii. Histogram
   iii. Cumulative frequency graph or ogive

**Frequency Polygon of an Ungrouped Data**

For an ungrouped data, the frequency polygon is plotted by listing the scores (x) on the x axis of the graph and the frequencies of the scores

(f) on the y axis. A dot or mark is made at the intersect between each score and its frequency, after which the marks are connected with lines. Using the frequency distribution in Table 3 above, the frequency polygon is thus represented in fig. 1:

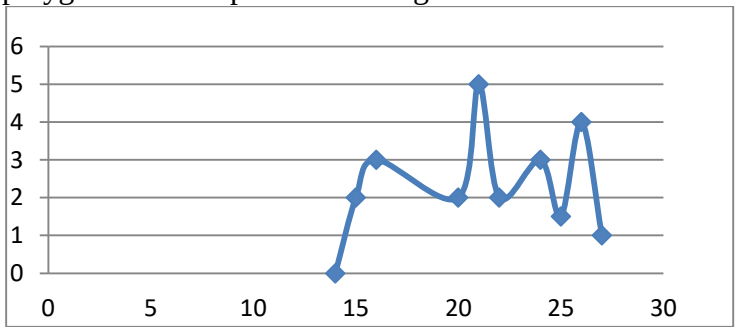



***Fig. 1: Frequency Polygon of ungrouped data***

**Frequency Polygon of a Grouped Data**

To construct the frequency polygon of a grouped data, we need to first of all determine the class marks or midpoints of each class interval. This is written on the y axis.

The class mark of any class is the midpoint of the class which is by listing determined by dividing the sum of the extreme scores by two. Using of the data in Table 5 above, the class marks are presented in the table 1 below as:

**Table 23: Marks (Midpoints) of a Grouped Data**

| **Classes** | **F** | **Cf** | **Class Marks (x)** |
|---|---|---|---|
| 10-14 | 4 | 4 | 12 |
| 15-19 | 6 | 10 | 17 |
| 20-24 | 2 | 12 | 22 |
| 25 – 29 | 8 | 20 | 27 |
| 30-34 | 4 | 24 | 32 |
| 35-39 | 6 | 30 | 37 |
| 40-44 | 4 | 34 | 42 |

| | | | |
|---|---|---|---|
| 45-49 | 4 | 38 | 47 |
| 50-54 | 8 | 46 | 52 |
| 55-59 | 9 | 55 | 57 |

Using the data in Table 6 above, we construct the frequency polygon using the x axis for the scores (mid-point of class intervals) and the y axis for the frequencies. We then join the lines neatly. We choose the scales for the graph making sure for a good graph, that the y unit is about 75% of the x units.

In the polygon graph, the total area of the polygon represents the total frequency N.

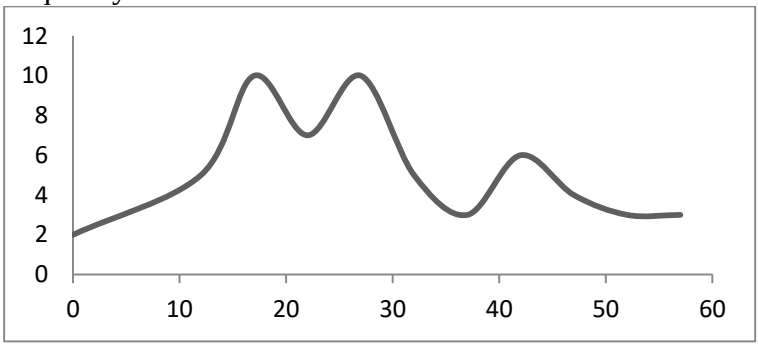



*Fig. 2: Frequency Polygon of a Grouped Data*

**Smoothing the Frequency Polygon**

The frequency polygon is usually very irregular and jagged in outline. To iron out the irregularities, the frequency polygon is often smoothened.

In smoothening, we use the adjusted values of "f". These values are found by adding together the frequency above the one we want to adjust, the particular frequency being adjusted plus the frequency below it. We divide the sum by 3. See table 7 below.

**Table 24: Smoothened Frequencies of a Grouped Data**

| Class Interval | F | Mid Point (x) | Cum f. | Smoothened F |
|---|---|---|---|---|
| 195 – 199 | 2 | 197 | 50 | 1.33 |
| 190 – 194 | 2 | 192 | 49 | 2.67 |
| 185 – 189 | 4 | 187 | 47 | 3.67 |
| 180 – 184 | 5 | 182 | 43 | 5.67 |
| 175 – 179 | 8 | 177 | 38 | 7.67 |
| 170 – 174 | 10 | 172 | 30 | 8.00 |
| 165 – 169 | 6 | 167 | 20 | 6.67 |
| 160 – 164 | 4 | 162 | 14 | 4.67 |
| 155 – 159 | 4 | 157 | 10 | 3.33 |
| 150 – 154 | 2 | 152 | 6 | 3.00 |
| 145 – 149 | 3 | 147 | 4 | 2.00 |
| 140 – 144 | 1 | 142 | 1 | 1.33 |

If we want to adjust ft. frequency 2 in the class interval 190 - 194, we take the given frequency which is 2, the frequency above this, which is also 2, the frequency below which is 4, we add up these and divide by 3. This gives us 2.67.

For the next frequency (for the class interval 195 - 199), we add the frequency of the next higher class, the frequency of the class and the frequency of the next lower class and divide by 3 i.e.

$$\frac{0 + 2 + 2}{3} = \frac{4}{3} = 1.33$$

Now a graph is plotted using the adjusted or smoothened "f" instead of the original ***f***.

**Exercise:** Now use the data in Table 7 to plot the frequency polygon. On the same graph paper plot the smoothened frequency polygon. What do you notice?

**The Shape of the Frequency Polygon**

In statistical measurements, all things being equal, when the frequency polygon of a set of data is plotted, we expect to get graph of the normal

curve which is symmetrical and dumb bell shaped. In practice, the graph we obtain may be skewed.

In the case where the frequency polygon is normally curved, the mean, the median, and the mode all coincide and there is perfect balance between the right and left halves of the polygon.

Where the mean and the median fall at different points in the distribution, and the balance is shifted to one side or the other-to left or right, the distribution is said to be skewed.

Distributions are said to be skewed negatively or to the left when scores are massed at the high end of the scale (the right end) and are spread out more gradually toward the low end (or left) as in fig 4 below. In such cases, the mode is greater than the median and the median is greater than the mean. Distributions are skewed positively or to the right when scores are massed at the low (or left) end of the scale and are spread out gradually toward the high or right end as shown in fig 5 below (Garrett, 1966). In such cases, the mean is greater than the median and the median is greater than the mode.

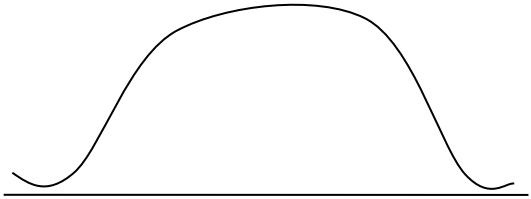


**Fig. 3: Showinga dumb bell shape (symmetrical) distribution**

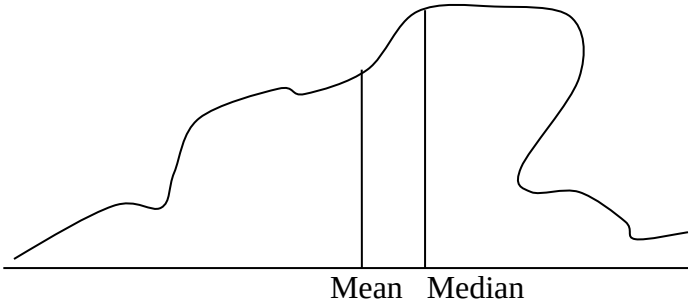
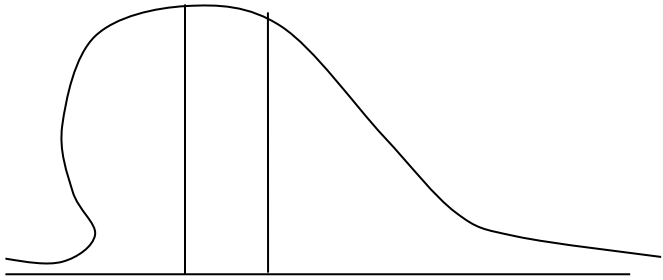


**Fig. 4: Showing negative skewness to the left**



Median Mean

**Fig. 5: Showing positive skewness to the right**

We will notice the figures above (figs 4 and 5) are skewed to the left and to the right respectively. To determine the direction of the skew, we also use the tail of the polygon. If the tail is on the left side, we say that it is negatively skewed. If the tail is on the right side, we say it positively skewed. To calculate the skewness of a distribution, we apply the formula:

$$SK = \frac{3(\text{mean} - \text{median})}{SD}$$

**Kurtosis of Data**

The term "Kurtosis," to Garrett (1966, p. 101), "refers to the 'peakedness' or flatness of a frequency distribution as compared with the normal."

Kurtosis is also described as the 'curvedness' of the graph of a distribution. Kurtosis is frequency used in a relative sense. There are different forms of curves or peaks which the frequency polygon of data distributions may take. These forms depend on the data collected, and they are:

**a. Mesokurtic**

This refers to a symmetrical shaped distribution or a normally curved distribution as represented in the figure below:

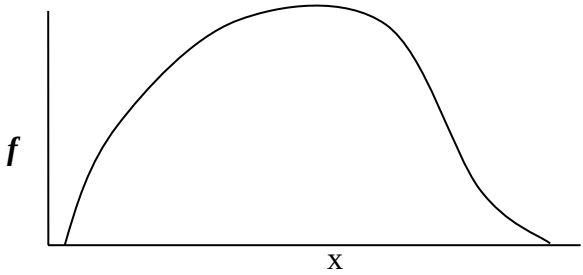
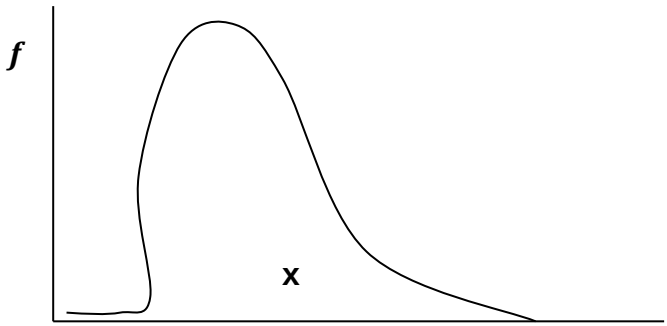



**Fig. 6: A Mesokurtic Distribution (Normal Curve)**

**b. Leptokurtic**

The Greek word lepto means thin, so leptokurtic implies a thin distribution. Another way of describing leptokurtic distribution is a distribution with a high peak as in figure 7 below.




**c. Platykurtic:**

Platy means flat, therefore, a platykurtic distribution is a distribution with a flatter curve than that of normal distribution. A platykurtic distribution is represented in fig. 8 below.

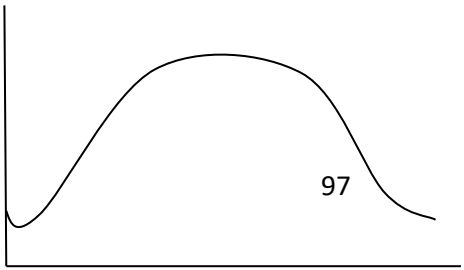


**Fig. 8: Platykurtic Distribution**

It is necessary to determine the kurtosis of a data when one wants to ascertain whether the data is normally distributed or not. Assumptions of a normally distributed population helps one to decide whether the test statistics will be parametric or non parametric. The formula for measuring kurtosis is:

$$Ku = \frac{Q}{(P_{90} - P_{10})}$$

where Q = Quartile deviation of the distribution.

For a normally distributed data, the Ku = .263. If Ku is greater than .263 the distribution is platykurtic, if less than .263 the distribution is leptokurtic.

**Histogram or Column Diagram**

This has the same shape as the frequency polygon. Instead of using the midpoint of the class interval to draw a line graph as in the frequency polygon, the frequencies in the histogram are represented by a column or a rectangle respectively.

For example, in Table 7, for the class interval 140 - 144, the frequency is 1 and should be plotted against the midpoint of the class interval which is 142.

In the histogram, the point, 142, can be represented by a line of 1cm.

The lower and upper limits are by the ends of the line while the middle is really 142. The frequency is indicated along the Y axis.

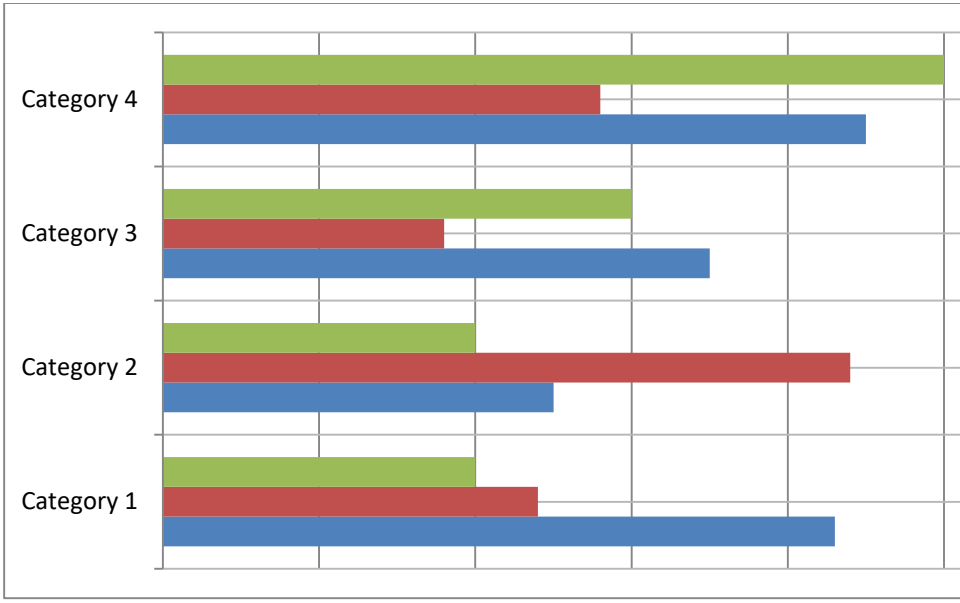



**Fig. 9: Histogram or column diagram of a grouped data**

The graph looks like series of rectangles of the same base (arbitrarily chosen) but different heights which depend on the frequency.

Note that the area of each rectangle in a histogram is directly proportional to the number of scores or measures within the interval. The histogram presents accurate relative proportions of the total number.

**Class Limits:**

Each class in a grouped data has two limits viz: lower-class limit and upper class limit. These class limits are the extreme scores in a distribution. Using the data in Table 8, for example, the class limits of class interval 1 are 8 and 10 respectively, 8 becomes the lower one while. 10 is the upper limit. These types of class limits are appropriate for discrete variables. However, for continuous variables to which education data belong, true or actual class limits are used in statistical analysis. These class limits are boundaries between one class and

another. .To determine the class boundaries, we simply subtract .5 from the lowest score in each class for lower class limits, and add .5 to the highest score for the upper class limit. The class limits are presented in the Table 9 below.

**Table 26: Upper and Lower Class Limits of a Grouped Data**

| **Classes** | **F** | **Lower Class Limit (LCL)** | **Upper Class Limit (UCL)** |
|---|---|---|---|
| 1-15 | 4 | 0.5 | 15.5 |
| 16-30 | 6 | 15.5 | 30.5 |
| 31- 45 | 8 | 30.5 | 45.5 |
| 46- 60 | 2 | 46.5 | 60.5 |
| 61- 75 | 6 | 60.5 | 75.5 |
| 76 – 90 | 8 | 75.5 | 90.5 |
| 91 – 104 | 9 | 90.5 | 104.5 |
| 105 - 119 | 4 | 104.5 | 119.5 |
| 120 - 134 | 6 | 119.5 | 134.5 |

**Percentage Frequencies and Cumulative Proportion Frequencies**

Percentage frequencies are determined by dividing each ***f*** by N, and then multiply the result by 100. The percentage frequency of class 2 in Table 9 is 4/86 x 100 = 4.7. Proportion frequency is obtained by simply dividing each *f* by N. For our data in Table 9, the proportion frequency for the class interval 11 - 13 with a frequency of 4 is 4/86 =0.047.

To get the proportion frequency for all other classes, we use the same procedure.

Cumulative percentage frequency and cumulative proportion frequency are computed by adding up progressively the percentage or proportion frequencies respectively already determined starting from the lowest class. This is because in adding progressively from the

bottom up, each cumulative frequency carries through to the exact upper limit of the interval. The cumulative frequency graph of the data in Table 27 is presented below.

**The Cumulative Frequency Graph** (Ogive)

**Table 9: Cumulative frequencies, percentages and proportions for memory test scores**

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| **Scores** | **Upper Limit** | **Lower Limit** | **F** | **CF** | **PF** | **CPF (Cumulative Percentage Frequency)** | **PF (Proportion Frequencies)** | **CPRF (Cumulative Proportion Frequencies)** |
| 41 – 43 | 42.5 | 40.5 | 1 | 86 | 1.2 | 100.1 | 0.012 | 1.00 |
| 38 – 40 | 40.5 | 37.5 | 4 | 85 | 4.7 | 98.9 | 0.047 | 0.99 |
| 35 – 37 | 37.5 | 34.5 | 5 | 81 | 5.8 | 94.2 | 0.058 | 0.943 |
| 32 – 34 | 34.5 | 31.5 | 8 | 76 | 9.3 | 88.4 | 0.093 | 0.885 |
| 29 – 31 | 31.5 | 28.5 | 14 | 68 | 16.3 | 79.1 | 0.163 | 0.792 |
| 26 – 28 | 28.5 | 25.5 | 17 | 54 | 19.8 | 62.8 | 0.198 | 0.629 |
| 23 – 25 | 25.5 | 22.5 | 9 | 37 | 10.4 | 43.0 | 0.105 | 0.431 |
| 20 – 22 | 22.5 | 19.5 | 13 | 28 | 15.1 | 32.6 | 0.151 | 0.326 |
| 17 – 19 | 19.5 | 16.5 | 8 | 15 | 9.3 | 17.5 | 0.093 | 0.175 |
| 14 – 16 | 16.5 | 13.5 | 3 | 7 | 3.5 | 8.2 | 0.035 | 0.082 |
| 11 – 13 | 13.5 | 10.5 | 4 | 4 | 4.7 | 4.7 | 0.047 | 0.047 |
| 8 – 10 | 10.5 | 7.5 | 0 | 0 | 0 | 0.0 | 0.000 | 0.000 |

**N = 86**

The construction of the ogive or cumulative frequency graph using the above data is simple. It is the graph which uses the upper limit of the class interval and the cumulative frequency, (cf) cumulative percentage frequencies (cpf) or cumulative proportion frequencies (cpfr).

However, the cumulative frequency graph can be constructed without the cpf and cpfr. Before we can plot a cumulative frequency polygon (graph), the scores of the distribution must be added serially or cumulated as in Table 9. We then write out the exact upper limits of the class intervals along the x axis and their cumulative frequencies along the y axis. This is because in adding progressively from the bottom, each cumulative frequency carries through to be exact upper limit of the interval. The cumulative frequency graph of the data in Table 9 is presented below:

Note that if we are plotting the frequency polygon, we use the midpoints of the class intervals with their respective frequencies instead.

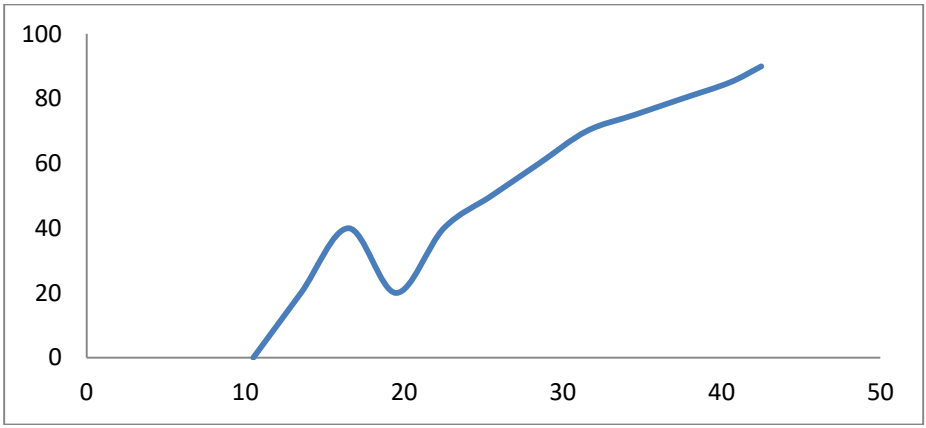



**Fig. 10: Cumulative Frequency Polygon**

**The Measures of Central Value or Tendency (Averages)**

The central value is of two folds. Firstly, it is an average of a group. 'It serves as a description of a mass of quantitative data from a sample. As an average, it stands for a group and saves us the problems of the details of each member of the group. The average mark in examination immediately gives an idea of the general group performance. Secondly, measures of central value help us to compare two or more groups in terms of typical performance. In statistics, there are three kinds of averages:

i) The mean (arithmetic average)
ii) The median
iii) The mode.

Note: The term 'average' is popularly used for the arithmetic mean. In statistics analysis, however, 'average' is the general term for any measure of central tendency.

**The Mean**

The mean is arithmetic average of a distribution. This equals dividing the summation of total scores in a distribution by number of the scores.

**Analysis of Mean for Ungrouped Data**

**Example 1:**

Suppose we have a set of marks:

45, 65, 94, 30, 75, 20, 10, 56, 45, 20

We may represent each of these numbers by x. The mean, X of all the numbers is given by:

$$\frac{\sum x}{N}$$

where X

$$= \frac{45 + 65 + 94 + 30 + 75 + 20 + 10 + 56 + 45 + 20}{10}$$

$$= \frac{460}{10} = 46$$

Where X = arithmetic mean.

**The Mean of a Frequency Distribution**

In the example above, each value of x was listed individually even when any occurs more than once. There may be a situation where each x may occur more than once but it is listed once with the frequency of occurrence recorded as below.

Table 28. FDT

| **X** | ***F*** | ***Fx*** |
|---|---|---|
| 18 | 1 | 18 |
| 17 | 2 | 34 |
| 16 | 2 | 32 |
| 15 | 3 | 45 |
| 14 | 2 | 28 |

| | | |
|---|---|---|
| 13 | 5 | 65 |
| 12 | 3 | 36 |
| 11 | 2 | 22 |

***∑f or N = 20*** ***∑fx = 280***

The score 18 has occurred once - frequency is one.
The score 17 has occurred twice - frequency is two
The score 12 has occurred three times, etc.
In the above illustration, x is the score, *f* is the frequency, fx is the product of the score and its frequency of occurrence.

$$X = \frac{\sum fx}{N}$$

In this case x will be replaced by fx
N remains as number of scores.
For such frequency distributions,

$$x = \frac{\sum fx}{N} \text{ i.e. } \frac{18 + 34 + 32 + 45 + 28 + 65 + 36 + 22}{280}$$

In general, where $X_1$, $x_2$, $x_3$, .... $x_4$ occur with frequencies f,, $f_2$, $f_3$ . . $f_4$
The arithmetic mean = $\frac{fx_1 + fx_2 + fx_3 + fx_4}{}$ .

In the example above = $\frac{280}{20}$ = 14

**Analysis of Mean for Grouped Data**
Using the distribution below, the mean computation is:
**Table 29: Frequency Distribution of Grouped Data for Mean Computation**

| **Class Interval** | **Mid Point x** | **Frequency (*f*)** | **Frequency Mid Point FX** |
|---|---|---|---|
| 45 – 49 | 47 | 1 | 47 |
| 40 – 44 | 42 | 2 | 84 |
| 35 – 39 | 37 | 3 | 111 |

| | | | |
|---|---|---|---|
| 30 – 34 | 32 | 6 | 192 |
| 25 – 29 | 27 | 8 | 216 |
| 20 – 24 | 22 | 17 | 374 |
| 15 – 19 | 17 | 26 | 442 |
| 10 – 14 | 12 | 11 | 132 |
| 5 – 9 | 7 | 2 | 14 |
| 0 – 4 | 2 | 0 | 0 |
| | | **76** | **1612** |

For a problem of this type involving the class interval, we use the mid point of the class interval as x

$$X = \frac{\sum fx}{\sum f} = \frac{1612}{76} = 21.2$$

***Example 2***

Consider the data below, been scores generated from a test administered by a Physics teacher in City girls secondary school

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ~~22~~ | ~~10~~ | ~~20~~ | ~~32~~ | ~~30~~ | ~~70~~ | ~~88~~ | ~~40~~ | ~~33~~ | ~~16~~ |
| ~~33~~ | ~~8~~ | ~~16~~ | ~~23~~ | ~~35~~ | ~~76~~ | ~~87~~ | ~~27~~ | ~~15~~ | ~~27~~ |
| ~~44~~ | ~~15~~ | ~~26~~ | ~~37~~ | ~~47~~ | ~~78~~ | ~~90~~ | ~~66~~ | ~~37~~ | ~~40~~ |
| ~~51~~ | ~~26~~ | ~~37~~ | ~~64~~ | ~~58~~ | ~~69~~ | ~~10~~ | ~~25~~ | ~~86~~ | ~~51~~ |
| ~~62~~ | ~~74~~ | ~~85~~ | ~~90~~ | ~~69~~ | ~~80~~ | ~~15~~ | ~~72~~ | ~~71~~ | ~~20~~ |

Categorize the distribution into seven classes, and ascertain the mean.

**Solution**

Step 1 $C = \frac{R}{K}$

$$C = \frac{90-8}{7}$$
$$C = 11.7$$
$$C \approx 12$$

Step 2. Construction of the Frequency Distribution Table

| C. I | Midpoint (X) | Tallies | Frequency | FX |
|---|---|---|---|---|
| 8 – 20 | 14 | 𝍸 𝍸 | 10 | 140 |
| 21 – 33 | 27 | 𝍸 𝍸 I | 11 | 297 |
| 34 - 46 | 40 | 𝍸 II | 7 | 280 |
| 47 – 59 | 53 | IIII | 4 | 212 |
| 60 – 72 | 66 | 𝍸 III | 8 | 528 |
| 73 – 85 | 79 | 𝍸 | 5 | 395 |
| 86 - 98 | 92 | 𝍸 | 5 | 460 |
| | | | ΣF = 50 | ΣFX = 2312 |

Mean $(\overline{X}) = \frac{\Sigma FX}{\Sigma F}$

Mean $(\overline{X}) = \frac{2312}{50}$

$= 46.24$

**Computation of Mean Using Assumed Mean Formula**

The formula used in the computation of the mean above is the ordinary mean formula. The mean can also be computed using the Assumed or Guessed mean formula. This formula is appropriate when the data is large and use of raw scores with ordinary mean formula becomes tedious.

In assumed mean formula one assumes that one of the scores or mid points is the mean. The differences between the assumed mean and each of the other scores or mid points are gotten. These differences are summed and divided by number. The difference can also be determined

arbitrarily starting from any origin. Examples of determining the mean with assumed mean formula are given below.

**Computation of Mean with Assumed Mean Formula Using Deviation Scores**

Using the distribution in Table 11 below, and taking 22 as the assumed mean, the mean becomes:

$$Xa = A + \left( \frac{\sum fd}{N} \right)$$

Where:

A = the assumed mean

D = difference between the assumed mean and other scores (x).

**Table 30: Computation of Mean with Assumed Mean Formula Using Deviation Scores**

| **Class Interval** | **Mid Point x** | **Frequency (*f*)** | **d** | **Fd** |
|---|---|---|---|---|
| 45 – 49 | 47 | 1 | 25 | 25 |
| 40 – 44 | 42 | 2 | 20 | 40 |
| 35 – 39 | 37 | 3 | 15 | 45 |
| 30 – 34 | 32 | 6 | 10 | 60 |
| 25 – 29 | 27 | 8 | 5 | 40 |
| 20 – 24 | 22 | 17 | 0 | 0 |
| 15 – 19 | 17 | 26 | -5 | -130 |
| 10 – 14 | 12 | 11 | -10 | -110 |
| | | **$\sum f$ = 74** | | **$\sum Fd$ = -30** |

**Illustration**

Consider the data below obtained from the sale day book of Mr. Okafor who operate in shop rite Enugu.

3, 6, 9, 12, 16, 19, 24, 22, 27, 1.
4, 8, 10, 14, 16, 19, 21, 25, 2, 1
29, 3, 7, 11, 13, 16, 18, 23, 26, 29
5, 7, 11, 13, 15, 18, 21, 1, 4, 1
8, 6, 9, 10, 9, 14, 17, 15, 20, 5
2, 4, 7, 14, 20, 12, 12, 14, 6, 3.

Given that the assume mean is 8, calculate the mean of the distribution and compare your answer with the mean calculated in the above example.

**Table 31. Solution:**

| **Class Interval** | **Class Boundary** | **Class Mark (X)** | **Tally (T)** | **Freq (F)** | **X-X (D)** | **FD** |
|---|---|---|---|---|---|---|
| 0-2 | -0.5-2.5 | 1 | ~~*IIII*~~ *I* | 6 | -7 | -42 |
| 3-5 | 2.5-5.5 | 4 | ~~*IIII*~~ *III* | 8 | -4 | -32 |
| 6-8 | 5.5-8.5 | 7 | ~~*IIII*~~ *III* | 8 | -1 | -8 |
| 9-11 | 8.5-11.5 | 10 | ~~*IIII*~~ *II* | 7 | 2 | 14 |
| 12-14 | 11.5-14.5 | 13 | ~~*IIII*~~ *III* | 8 | 5 | 40 |
| 15-17 | 14.5-17.5 | 16 | ~~*IIII*~~ *I* | 6 | 8 | 48 |
| 18-20 | 17.5-20.5 | 19 | ~~*IIII*~~ *II* | 7 | 11 | 77 |
| 21-23 | 20.5-23.5 | 22 | *IIII* | 4 | 14 | 56 |
| 24-26 | 23.5-26.5 | 25 | *III* | 3 | 17 | 51 |
| 27-29 | 26.5-27.5 | 28 | *III* | 3 | 20 | 60 |
| Total | | | | $\sum F$ =60 | | $\sum FD$= 264 |

Average deviation = $\frac{\sum FD}{\sum F}$

A.D = $\frac{264}{60}$

A.D = 4.4

Mean (X) = 8+4.4
Mean (X) =12.4

**Computation of Mean with Assumed Mean Formula Starting From Arbitrary Origin**

The assumed mean formula used above was done by obtaining deviation scores through subtraction of the assumed mean from each of the X. We can, however, compute the mean with assumed mean formula where we obtain the deviation scores by starting from arbitrary origin. We start from the centre of the distribution to obtain our deviations and in this case, we take the deviation score of that class interval to be 'o', all the subsequent higher class intervals, we affix 1, 2, 3, etc as the deviation scores and -1, -2, -3, etc-to the lower class intervals. However, the class mark of the class interval which is the starting
point will automatically be taken as the assumed mean. The formula of this procedure is:

$$Xa = A + \left(\frac{\sum fd}{N}\right)$$

Where;

A = the assumed mean
d = deviation scores from arbitrary origin
c = class size of the assumed mean.

We have to note that whenever we compute the mean using assumed mean formula where the deviation scores are obtained arbitrarily, the $\sum fd/N$ is always multiplied by c(class size) before we add the result to the assumed mean. Using the data in Table 11, the computation goes in this way.

Table 32. FDT

| **Class Interval** | **Mid Point x** | **Frequency (*f*)** | **d** | **Fd** |
|---|---|---|---|---|
| 45 – 49 | 47 | 1 | 4 | 4 |
| 40 – 44 | 42 | 2 | 3 | 6 |

| | | | | |
|---|---|---|---|---|
| 35 – 39<br>30 – 34 | 37<br>32 | 3<br>6 | 2<br>1 | 6<br>6 |
| 25 – 29<br>20 – 24 | 27<br>22 | 8<br>17 | 0<br>-1 | 0<br>-17 |
| 15 – 19<br>10 – 14 | 17<br>12 | 26<br>11 | -2<br>-3 | -52<br>-33 |
| | | $\sum f$ = 74 | | $\sum Fd$ = -80 |

**Mean of Combined Groups**

If we are given the mean of two or more groups and their numbers, N, we can calculate the mean of the combined groups. Let there be two groups A & B. Let the number in group A = 30; Mean = 16. Let the number in group B = 40, Mean =. 14., To find the mean of the groups combined.

Total number for A = 30 x 16 = 480

Total number for B = 40 x 14 = 560

Mean of the combined group

$$= \frac{480}{30} + \frac{560}{40} = \frac{1040}{70}$$

**= 14.86**

In general, let N for the groups A, B, C, etc be NI, $N_{2i}$ $N_3$, etc respectively and their means X1, X2, X3, etc. The mean of the combined groups =

$$\frac{N_1X_1 + N_2X_2 + N_3X_3}{N_1 + N_2 + N_3}$$

Sometimes, the reporting of the mean cannot give a needed picture of a distribution. Suppose in an examination involving 40 pupils, one scored 100%, two scored 80%, ten 0%, seven scored 20%, 20 scored 30%.

| **No. of Students (f)** | **Scores (x)** | **Fx** |
|---|---|---|

| | | |
|---|---|---|
| 1 | 100 | 100 |
| 2<br>10 | 80<br>0 | 160<br>0 |
| 7<br>20 | 20<br>30 | 140<br>600 |
| 40 | 230 | 1000 |

N =40; X = 25

This mean, does not tell much about the group and is not descriptive of the group scores. This is so because if we plotted the graph, we would find out that it is much skewed. In every skewed condition, the mean may not be a good measure of the group characteristic. The mean is supposed to be the center of symmetry of distribution in the group. If the mean is not the center of symmetry, it gives a distorted picture.

**The Median**

The median is a point in the scale of distribution wherein half of the scores falls below it. It is necessary that in discussion about the - median, the distribution must be arranged in an order, either ascending or descending order of magnitude. The median, therefore, is an ordinal statistics.

Suppose there is a record of marks.

2, 27, 20, 7, 19, 25, 16.

In order to determine the median, we arrange the marks in ascending or descending order.

2, 7, 16, 19, 20, 25, 27.

We look for the median, i.e. the score that is mid way. That is 19.

Three scores fall above 19 and three scores fall below it. **Note:** The median can be easily determined by taking the score at the middle after ranking the scores. If two scores fall in the middle in the case of even numbered distributions, add the two scores and divide by 2.

You may notice that ungrouped scores are arranged in order of size.

Two things can happen.

i. N can be odd.

ii. N can be even.

When N is odd, there is no problem finding the median. When N is even as in the series below:

7, 8, 9, 10, 11, 11

The median lies between 9 and 10 i.e. the $3.5^{th}$ score. Add scores 9 and 19 and divide by 2, which gives us 9.5. The median of the distribution is 9.5

**Calculation of the Median When Data are Grouped Table 33: Computation of the Median for Grouped Data**

| **Class Interval** | **f** | |
|---|---|---|
| 40 – 44 | 1 | 32 cases above the interval containing the median |
| 35 – 39 | 0 | |
| 30 – 34 | 3 | |
| 25 – 29 | 15 | |
| 20 – 24 | 13 | |
| 15 – 19 | 10 | Median lies here |
| 10 – 14<br>5 – 9 | 11<br>11 | |
| 0 – 4 | 4 | 26 cases below |
| | 68 | |

The median lies between $34^{th}$ case above or $34^{th}$ case below in the table. Counting from the top, the number of cases above the interval containing the median is 32. The median must, therefore, be contained in the frequency 10.

10 cases lie within 15-19 intervals more exactly between 14.5-19.5.

The score must lie within this.

The class interval width is 5

We divide this by 10 to get 0.5

We have 32 cases above the median class. To get 34 which is half the number of scores, we need 2 more marks.

To get 2 out of this we multiply 0.5 by 2. This gives 1.00.
When we subtract this from the upper limits of 19.5 we get the, median, which is 18.5.
We may approach the problem counting from below. We have 26 cases below ten. We need 8 more to obtain exactly half of our distribution and these lie within the frequency 10 in the class intervals 14.5 to 19.5.
If 10 contains 5 scores, 8 contains $\frac{5}{10} \times \frac{8}{1} = 4$

This time we add to the lower limit of 14.5.
i.e. 14.5 + 4 = 18.5. You must decide which way to consistently use.

**Calculation of the Median of Grouped Data Using Median Formula**
The median calculated above can be determined with a formula as:

$$Md = L_1 + \left( \frac{N/_2 - cfb}{f_1} \right) c$$

where $L_1$ = lower limit of the median class
N = number of scores
cfb = cumulative frequency of the next lower class to the median class.

**Procedure**
Steps in calculation of the median

1. Determine the cumulative frequencies of the scores in the distribution.
2. Divide the number of scores by 2 and look for the result in the cumulative frequency column. If not seen, take the one nearest to it but not below it and call the class with that cumulative frequency the median class.
3. Calculate the median, using the formula as in the example below.

**Table 34: Data for Computation of the Median**

| **Class Interval** | ***f*** | ***Cf*** |
|---|---|---|
| 40-44<br>35-39 | 1<br>0 | 68<br>67 |
| 30 – 34<br>25 – 29 | 3<br>15 | 67<br>64 |
| 20 – 24<br>15-19 | 13<br>10 | 49<br>36 median class |
| 10-14<br>5-9<br>0-4 | 11<br>11<br>4 | 26<br>15<br>4 |
| | **68** | |

$$Md = 14.5 + \left( \frac{68/_2 - 26}{10} 5 \right) = 18.5$$

***Example 2***

The below data, has been fetched from the scattered information kept by Nwatu Ltd. Who is into the sales of sandal at the modern market Enugu.

Required: Calculate the median of the distribution.

22, 23, 25, 31, 32, 33, 38, 41, 44, 47
51, 54, 55, 21, 23, 26, 30, 32, 36, 39
42, 44, 47, 52, 53, 58, 20, 25, 28, 30
32, 36, 39, 42, 44, 48, 51, 55, 24, 27
28, 33, 35, 38, 43, 46, 47, 50, 25, 29
34, 37, 39, 42, 46, 49, 47, 29, 33, 35

36, 40, 43, 46, 32, 36, 38, 43, 45, 45
33, 37, 37, 40, 41, 40, 41, 38, 42, 40

**Solution**

Table 35 FDT

| Class interval | Class Limit | Frequency | Cumulative frequency |
|---|---|---|---|
| 20-22 | 19.5-22.5 | 2 | 2 |
| 23-25 | 22.5-25.5 | 4 | 6 |
| 26-28 | 25.5-28.5 | 5 | 11 |
| 29-31 | 29.5-31.5 | 6 | 17 |
| 32-34 | 31.5-34.5 | 8 | 25 |
| 35-37 | 35.5-37.5 | 10 | 35 |
| 38-40 | 37.7-40.5 | 12 | 47 |
| 41-43 | 40.5-43.5 | 10 | 57 |
| 44-46 | 43.5-46.5 | 8 | 65 |
| 47-49 | 46.5-49.5 | 6 | 71 |
| 50-52 | 49.5-52.5 | 4 | 75 |
| 53-55 | 52.5-55.5 | 3 | 78 |
| 56-58 | 55.5-58.5 | 2 | 80 |
| | | $\sum$ F=80 | |

Median = $L_i + \frac{(\frac{N}{2} - \frac{CFb}{1})I}{f}$

Where Li =37.5

N = 80

F = 35

f=12

Median = $37.5 + (\frac{80}{2} - \frac{35}{12}) \times 3$

$37.5 + (40 - \frac{35}{12}) \times 3$

37.5+ (5/12) x 3
37.5 + (15/12)
37.5 + 1.25

Median = 38.75.

***Example 3***

Table 35. FDT

Consider the table below

| C.I | Frequency | C.L | Cumulative frequency |
|---|---|---|---|
| 1-5 | 3 | 0.5-5.5 | 3 |
| 6-10 | 2 | 5.5-10.5 | 5 |
| 11-15 | 8 | 10.5-15.5 | 13 |
| 16-20 | 4 | 15.5-20.5 | 17 |
| 21-25 | 7 | 20.5-25.5 | 24 |

To identify the median class

Divide Σf by 2 = $\frac{24}{2}$= 12

The median class lies on class 3

$L_1$ = 10.5

N = 24

Cbf = 5

C = 5

F = 8

$$= 10.5 + \frac{(\frac{24}{2} - \frac{5}{1})}{8}$$

$$= 10.5 + \frac{(12-5)5}{8}$$

$$=10.5 + \frac{35}{8}$$

Median = 14.9

**The Mode**

The mode is that single measure or score which occurs most frequently. In other words, the mode is said to be the most occurring score or variate. This measure of central tendency is the simplest and most unreliable. It is the statistical procedure applied when the data is on the nominal scale, and it determines the height of the peak of a distribution.

The mode for an ungrouped data can simply be determined by tallying the frequencies of the scores and taking the score that occurs most frequently. For example, in a distribution with the scores: 3, 4, 5, 5, 6, 7, 7, 7, 7, 8, 8, 9, 10, the most occurring score which is 7 becomes the mode.

There can be more than one mode in a distribution. Where the mode is only one, we call the distribution a uni-modal distribution. If the mode is two, it becomes a bi-modal distribution and where it is more than two, it becomes a multimodal distribution.

**Computation of the Mode for a Grouped Data**

We can compute the mode for a grouped data in two different ways. First, we can simply determine the mode of a grouped data by taking the class midpoint, 'x' of the modal class as the mode. Secondly, the mode of a grouped data can be determined using the formula.

Where:

$M_0$ = the mode

LI = lower limit of the modal class

D1 = the difference between the modal frequency and the frequency of the next lower class.

D2 = the difference between the modal frequency and the frequency of the next higher class.

c = class size (width) of the modal class.

**Note:** To compute the mode, you will first determine the modal class which is the class with the highest frequency.

**Table 36: Computation of the Mode for a Grouped Data.**
***Example 1***

| Class interval | F |
|---|---|
| 20-22 | 2 |
| 23-25 | 4 |
| 26-28 | 5 |
| 29-31 | 6 |
| 32-34 | 8 |
| 35-37 | 10 |
| 38-40 | 12 |
| 41-43 | 10 |
| 44-46 | 8 |
| 47-49 | 6 |
| 50-52 | 4 |
| 53-55 | 3 |
| 56-58 | 2 |
| Total | $\sum$ F=30 |

$M_o = L_1 + [\frac{D1}{D1+D2}]I$
Where: Li=37.5
Di=12-10
$D_2$=12-10
C=3

Mo= 37.5 + (2) x 3
2+2
Mo=37.5 + 6/4
Mo=37.5 +1.5
Mo=39

Example 2

Consider the table below, determine the mode

| C. I | 8 - 20 | 21 - 33 | 34 - 46 | 47 - 59 | 60 - 72 | 73 - 85 | 86 - 98 |
|---|---|---|---|---|---|---|---|
| Frequency | 10 | 11 | 17 | 4 | 8 | 5 | 5 |

Solution

| C. I | Class limit | Frequency |
|---|---|---|
| 8 - 20 | 7.5 – 20.5 | 10 |
| 21 - 33 | 20.5 – 33.5 | 11 |
| 34 - 46 | 33.5 – 46.5 | 17 |
| 47 - 59 | 46.5 – 59.5 | 4 |
| 60 - 72 | 59.5 – 72.5 | 8 |
| 73 - 85 | 72.5 – 85.5 | 5 |
| 86 - 98 | 86.5 – 98.5 | 5 |

Mode = $L_1 + \left(\frac{D_1}{D_1 + D_2}\right)C$

$L_1 = 33.5$

$D_1 = 17 - 11 = 6$

$D_2 = 17 - 4 = 13$

$C = 13$

Mode = $33.5 + \left(\frac{6}{6+13}\right)13$

$33.5 + \left(\frac{6}{19}\right)13$

$33.5 + (0.32)13$

$33.5 + 4.16 \Rightarrow 37.66$

**Example 3**

Table 37

**Consider the table below**

| **Class Interval** | **X** | ***f*** |
|---|---|---|
| 40-44<br>35-39 | 42<br>37 | 1 0 |
| 30-34<br>25-29 | 32<br>27 | 3<br>15 modal class |
| 20-24<br>15-19 | 22<br>17 | 13<br>10 |
| 10-14<br>5-9 | 12<br>7 | 11<br>11 |
| 0-4 | 2 | 4 |
| | | **68** |

$$\text{Mo} = 24.5 + \left( \frac{15 - 13}{(15 - 13) + (15 - 3)} \right) 5$$

$$= 25.2$$

**Uses of the Mean, Median and Mode**

1. Where the distribution appears balanced from either side, the mean should be used, and is most appropriate. Where there is obvious skewness the median is preferable. The mode is used to get an idea of where there is the greatest bunching of marks.

When there is even distribution, we get a normal curve in which the mean and the median coincide. The curve is abnormal to the extent of the differences between the mean and the median.

**Exercises**

1a. What are measures of central value (tendency) used for?

b. Given the distribution of scores below, compute the median.

c. What is mode of the distribution?

| X | 22 | 24 | 26 | 28 | 30 | 32 | 34 |
|---|---|---|---|---|---|---|---|
| F | 2 | 4 | 4 | 12 | 6 | 2 | 3 |

2a. Determine the mean of the following students' test scores in Physics using the assumed mean formula.

| **Classes** | **F** |
|---|---|
| 20-26 | 12 |
| 27-33 | 22 |
| 34-40 | 25 |
| 41 -47 | 8 |
| 48-54 | 10 |
| 55-61 | 6 |

2b. What are the median and mode of the distribution?

3a. A researcher collected different sets of data from three groups as follows:

Group A: X = 35; N = 20
Group B: X = 44; N = 15
Group C: X = 50; N = 32

Determine the grand mean for the three groups.

## References


Abonyi S.O. (2003) *Instrumentation in behavioural research*. Enugu: Fulladu Pub. Ltd

Agu, N. *Basic statistics for behavioural sciences*. Awka: Madonna Pub Ltd

Ayuba M.M (2010). *Factors affecting academic achievement of student in secondary school*. Maiduguri: University Printing Press

Ajanyi E.A (2010). *Educational measurement and evaluation*. Lagos: University Press Ltd

Ajoni P.A (2011). *Validity and reliability of research instrument*. Ife: Odua Ajojo Pub ltd

Bright N.k (2011). Psychological perspectives of testing: *American Psychologist*, 38(5), 1045-1050

Bakwo (2015). *Strategies of testing cognitive abilities*. Kaduna: Kwabo African Press Ltd

Berk H. (2000). Relationship *between aptitude test and intelligence*. (2nd Ed). New York: MacMillan Pub. Co.

Bester P.O (2009). *Causes of difference between aptitude and academic achievement*. (3rd Ed). New York: John Wiley & Sons Inc.

Coetzee K. (2001). Predictive power of aptitude test. *West African Journal of Education*, 18(2) 300-350.

Egwa O.A (2011, June, 8). Strategies for improving universities student academic achievement. *Sunnews*, P.3

Eno & Seldon (nd). *Academic achievement and attrition in universities.* New York: McGraw-Hill

Egbo O.A (2015). Rules of aptitude test. Enugu: Celex Pub. Ltd

Encarta Encyclopedia. (2015). *Aptitude test.* Encarta: Oxford University Press.

Eze N.B (2009). *Scholastic aptitude test*. Awka: Famous Pub. Ltd.

Ejeh et al (2009). Multiple admission in Nigerian universities. In P. Edo(ed) *Handbook of measurement and evaluation* vol 12, Oxford, UK: Trikles Ltd.

Founche and Verwey (2013). Consequencies of sampling and its implications. *American Psychologist,* 32(2), 200-215

Ghiselli S. (2000). Career choice among student in secondary schools in Nigeria. *Measurement in Education*, 2, 3-9

Joela S.K (2004). The influence of paretal socio-economic condition on the result of aptitude test. *Journal of Educational measurement.* 12, 121- 127

Kama T.O. (2010). *The relationship between crystallized and fluid intelligence. Introduction to psychology of learning*. Zaria: Sambo and Kofas

Kaka M.A (2010). *Evaluation of testing in Nigerian educational system*. Anyigba: Agaba Pub Ltd

Kelly H. (2014). *Predicting student's academic performance in mathematics*. Enugu: Holand Ltd

Masel N.P. (2010). *Problem of test construction*. Enugu: Celex Ltd

Maris, M. (2012). *Relationship between intelligence, aptitude and socio-economic factor as predictors of academic achievement. Issues of academic excellence.* Handbook I: *Cognitive domain.* New York: Greg Int.

Mussen, Conger & Huston (2004). The Evils associated with test. Ghana:University of Ghana Press Ltds.

Oxford, A. (2013). Oxford Advanced Learner's Dictionary. (2nd Ed). London: Oxford University Press

Okonkwo, N. (2013). *Another view of aptitude test. Abia*: Ukah Pub. Co

Okoye, R.O (2015). *Educational and psychological measurement and evaluation*. (2nd Ed). Awka: Erudition Pub.

Pearson, K. (nd). *Differential aptitude test for career selection selection*. New York: Winston and Brass.

Port P.P & Digresia G. (2010). *Academic achievement test*. Georgia: University of American Pub.Co.

Prediager, D., Waple, H & Nusbaum (2008). *Educational measurement and Evaluation.* New York: John Willy Press

Reber O.O (2010). *Problems of aptitude test*. London: Alpha Book Press

Stumph, T., Stanley, D. (2002). Principles of scholastic achievement test: An overview. *Theory and practice,* 30(3), 112-130

Steyn, R.B. (2008). Academic aptitude and its effects on the learner's academic performance. *A collection of papers*. New York: Winston Press.

Tko, O., & Tolu, O. (2012). *Measurement and evaluation in education.* Lagos. University Press.

Taylor, O.A. (2014). *Predicting academic achievement of students, using scholastic aptitude test.* A paper presented at the International Conference on measurement in education, Arizona II

Ugboduma, U. (2011). *Effectiveness of aptitude test in predicting academic achievement in Health Sciences*. Oxford: Peg-mound Pub.Co

Vosloo, R.O., Coetzee & Classen (2000). Test validity and ethics of assessment. *American Psychologist:* 30,1000-1124