DEMYSTIFIED EDUCATIONAL RESEARCH, INSTRUMENTATION AND TEST

First Edition

Omachi, Daniel, Ph.D Nigeria,
Educational Measurement, Evaluation and Research
Department of Educational Foundations
Peaceland College of Education, 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.

PREFACE

In education, there are many issues revolving around research, instrumentation and test as many scholars will normally come up with their different views. In an attempt to deal with this expression of individual opinions or understanding as the case may be, some authors puts students into more confusion as some of the text are unable to give adequate explanations to some phenomenon clearly to the learner. It is against this background that the book was written to demystify the mysteries of research, how to construct instruments and testing respectively. The book has been written therefore with a view of providing skills necessary to achieve a high degree of objectivity.

The book started with a historical background of research. Chapter 2 treated research design. Chapter 3 treated literature review. Chapter 4 dealt with population, sample and sampling technique. Chapter 5 discussed data collection and analysis. Chapter 5,6,7 and 8 dealt with measurement of central tendencies, measurement of relationship and measurement of disparities. Chapter 9 treated inferential statistics. Chapter 10 non parametric statistics while chapter 11 looked at taxonomies. Chapter 13 treated the development of research instrument. Chapter 14, 15 and 16 discussed validity, reliability and item analysis. Chapter 18 dealt with scoring of research instrument. Chapter 18 discussed the nature of measurement and evaluation. Chapter 19 treated continuous assessment and types of test. Chapter 20 treated the goals and learning targets of instruction. Finally, chapter 21 discussed the stages in classroom test construction. Students are usually scared whenever research, test and instrumentation is mentioned as a result of its technicalities. The author believes that such fear can be allayed by being down to earth and using simple

illustrative examples in the teaching of the various techniques. This is what had been done in this book.

The book is intended for use by students in the universities and colleges of education as well as undergraduates who are undergoing courses in education, psychology, guidance and counseling and other social sciences. It has therefore been written to cover topics in research, instrumentation and test stipulated by the National Commission for Colleges of Education and the National Universities Commission (NUC) minimum standards. The book has been written in such a simple language that a beginner in testing, measurement and evaluation and other social sciences will find easily comprehendible.

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 test and instrumentation in educational research and testing. 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 developing and using test instruments as well as that of reporting test results (2) the paucity of commercially produced and validated, useful, relevant and appropriate educational research, test and instrumentation.

It is against these backgrounds that I am pleased to write the foreword to this invaluable book on research, instrumentation and test. This book addresses the three areas exceptionally well and at the classroom level. This is because the presentation of educational and psychological 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 21 chapters adequatelycovers each of the subject matters components. The content of each of the chapter flows into the other chapter, smoothly and coherently. In totality, the 21 chapters adequately cover the essential subject matter of educational research, instrumentation and test.

The substance and style of presentation of the content are quiet commendable. Illustrative examples teachers are familiar with, are 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 and psychological instrument, test and research. I therefore recommend this text strongly to colleges of education and university students and their teachers to demystify the mysteries of research, instrumentation and test in education and other social fields.

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

TABLE OF CONTENT

Chapter 6
Data collection and analysis
Elements of frequency distribution
Guidelines on how to construct frequency distribution table
Graphical representation of data
    Pie chart
    Bar Charts

Chapter 7
Measurement of central tendency
Arithmetic mean
Median
Mode

Chapter 8
Measurement of relationship/Association/Correlation
Pearson's correlation
Spearman's correlation

Chapter 9
Measurement of variability/Spread/Dispersion
Range
Variance
Standard deviation

Chapter 10
Inferential Statistics
T-test and Z-test

Chapter 11
Non parametric statistics
Run's test
Chi-square

Chapter 12
Taxonomies of educational objectives and test development
Cognitive domain
Affective domain
Psychomotor domains

Chapter 13
Validation of research instrument
Practical illustration of content validation of a test

Chapter 14
Reliability of research instrument
Estimation of stability
Estimation of equivalence
Estimation of internal consistency
Scorers reliability

Chapter 15
Item analysis
- Item difficulty
- Item discriminatory
- Item distracters

Chapter 16
The nature of measurement and evaluation
Testing

Assessment
Functions of assessment

Chapter 17
Stages in Classroom Test Construction
Preparing a test blue print or table of specification
Advantages of the Test Blue Print
Writing the individual items
Revieiwng the items
Preparing the scoring key or the marking scheme
Writing the Test
valuating the Test
Administering the Test

Chapter 18
Major types of test
Essay test
Objective test
True or false
Multiple choice tests

CHAPTER 1

INTRODUCTION

Originally, no knowledge is completely new any longer; old things kept on re-occurring to new people in different form. For the fact that understanding and interpretation differs from one person to another, or lack of satisfaction with the discovery of some people, or as a result of new ideas coming up due to changes and other forces, people decided to go into research.

Research comprises creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of humans, culture and society, and the use of this stock of knowledge to devise new applications.

It is used to establish or confirm facts, reaffirm the results of previous work, solve new or existing problems, support theorems, or develop new theories. A research project may also be an expansion on past work in the field. To test the validity of instruments, procedures, or experiments, research may replicate elements of prior projects, or the project as a whole. The primary purposes of basic research (as opposed to applied research) are documentation, discovery, interpretation, or the research and development (R&D) of methods and systems for the advancement of human knowledge.

Definitions of research

Research has been defined in a number of different ways.

- Research can be defined as the process of making enquiries into the unknown with the knowledge of the known in an attempt to solve an identified problem.

- Martyn &Shuttleworth (2014) Defines researchas any gathering of data, information and facts for the advancement of knowledge.
- Creswell (2004)sees research as a process of steps used to collect and analyze information to increase our understanding of a topic or issue.
- The Merriam-Webster Online Dictionary defines research in more detail as "a studious inquiry or examination; especially : investigation or experimentation aimed at the discovery and interpretation of facts, revision of accepted theories or laws in the light of new facts, or practical application of such new or revised theories or laws".
- Ojoma (2020) defines research as the systematic application of scientific principles in solving an identified problem.

It is an agreeable fact that research adopts scientific approaches in solving an identified problem, however, it is not all fields that adopts scientific approaches. Educational research is scientific in nature.

10 Most Salient Features of Scientific Research

1- Systematics

The systematization of scientific research is linked to the need for it to be rigorous in procedures.

This is not a random observation, but is the result of a well-structured plan, with specific objectives.

The processes must be standardized, always be sought to execute the actions in the same way, so that the result can be reliable as a result of having always followed the same guidelines.

The systematic plan that must guide a scientific investigation must consider all the aspects and moments of this research: from the objects of study and the variables to be taken into account, to the rhythm of work that must be followed in order to arrive at conclusions in time expected.

2- Controlled

Scientific research must avoid chance, and the process must be supported by control mechanisms that allow it to obtain truthful results.

Chance has no place in scientific research: all actions and observations are controlled, according to the researcher's criteria and according to the object investigated, through well-defined methods and rules.

3- Empirical

The results of a scientific investigation must deal with the aspects of reality related to the subject under investigation. The aspects that characterize a particular research must be observable in the real world.

Scientific research refers to issues that can be measured and identified as facts.

Is about Experiment with evidence . In this way it is possible to test the research hypothesis, and thus be able to affirm, deny or supplement it, as the case may be.

4- Rational

Science in general is characterized by being rational and logical. In a scientific investigation must emphasize the rationality on the subjectivity.

Its empirical characteristic makes it necessary to be based on real and verifiable facts, and demands from the researcher a critical attitude and a dispossession of his personal conceptions or judgments of value.
Some scientists and philosophers maintain that it is precisely the rational and critical character of an investigation that generates progress in the intellectual field and an important development of knowledge.

5- Reproducible
The findings obtained through scientific research should be able to be reproduced under the same conditions established in the study.
Given the systematized nature of scientific research, it must be verifiable. The fact of having controlled the variables that were part of the process, allows to be able to reproduce the results achieved.

6- Consider everyday problems
In a scientific investigation, the hypotheses constitute the nucleus of the study, and must be generated of problems and situations of the daily life, that affect the people of habitual form.
It is hoped that scientific research will solve a problem that ideally affects several groups of people.
By critically observing this problem and making it an object of study, it is possible to find an answer that, hopefully, can improve the quality of life of many people in different areas.

7- Objective
Just as rationality and critical character must be emphasized in scientific research, it must also be objective.
The goal of the investigator is not to justify own postures, but to expose the facts in the purest way possible.

The explanation arising from scientific research must be legitimate for people with different inclinations of thought. The results of scientific research must be universal.

8- Provisional

Science is constantly expanding. Scientific research is considered provisional because it must be open to further studies that reaffirm, refute or complement the findings obtained in that research.

The debate is a fundamental part of the scientific field. Therefore, a scientific investigation must be able to be questioned and, if there is some subsequent research proving contrary hypotheses, it must be able to rectify.

9- Original

There is no sense in focusing scientific research on proven facts. A scientific investigation must treat new or little studied aspects, so that the result of the study implies a true contribution to the science and the humanity.

If it is based on an existing research, the researcher should focus on a different area of the problem, look for alternative results to those presented in the first place, or refute the research hypothesis as mistaken.

In any case, it is essential that scientific research brings something new and useful for people.

10- Ordered

Scientific research needs rigorous planning so that it can yield true results. This planning must have a specific order, which responds to the interests of the study.

In a scientific investigation it is necessary that the processes are designed and ordered in such a way that they reach secondary objectives that, in the last instance, can help to verify the main objectives raised by the researcher.

Educational research

Educational research can be defined as the application of scientific and systematized approaches in solving identified educational problem.

Educational researchers have come to the consensus that, educational research must be conducted in a rigorous and systematic way, although what this implies is often debated. There are a variety of disciplines which are each present to some degree in educational research. These include psychology, sociology, anthropology, and philosophy. The overlap in disciplines creates a broad range from which methodology can be drawn. The findings of educational research also need to be interpreted within the context in which they were discovered as they may not be applicable in every time or place.

Characteristics of Educational research

Anderson (2019) x-rayed ten aspects of educational research:

- Educational research attempts to solve a problem.
- Research involves gathering new data from primary or first-hand sources or using existing data for a new purpose.
- Research is based upon observable experience or empirical evidence.
- Research demands accurate observation and description.
- Research generally employs carefully designed procedures and rigorous analysis.

- Research emphasizes the development of generalizations, principles or theories that will help in understanding, prediction and/or control.
- Research requires expertise—familiarity with the field; competence in methodology; technical skill in collecting and analyzing the data.

Educational Research Processes

The following are the fundamental educational research approaches and sequence of operations:

- Problem identification and definition
- Review of related literature
- Formulating research hypothesis and research questions
- Designing a study to gather necessary data
- Gathering of data for the purpose of answering the research questions and testing the hypothesis
- Data analysis
- Decision making/inference

Approaches

There are two main approaches in educational research. The first is a basic approach.This approach is also referred to as an academic research approach. The second approach is applied research or a contract research approach. Both of these approaches have different purposes which influence the nature of the respective research.

Basic approach

Basic, or academic research focuses on the search for truth or the development of educational theory. Researchers with this background "design studies that can test, refine, modify, or develop theories". Generally, these researchers are affiliated with

an academic institution and are performing this research as part of their graduate or doctoral work.

Applied approach

The pursuit of information that can be directly applied to practice is aptly known as applied or contractual research.] Researchers in this field are trying to find solutions to existing educational problems. The approach is much more utilitarian as it strives to find information that will directly influence practice. Applied researchers are commissioned by a sponsor and are responsible for addressing the needs presented by this employer. The goal of this research is "to determine the applicability of educational theory and principles by testing hypotheses within specific settings".

Methodology

The basis for educational research is the scientific method.] The scientific method uses directed questions and manipulation of variables to systematically find information about the teaching and learning process. In this scenario questions are answered by the analysis of data that is collected specifically for the purpose of answering these questions. Hypotheses are written and subsequently proved or disproved by data which leads to the creation of new hypotheses. The two main types of data that are used under this method are qualitative and quantitative.

Qualitative research

Qualitative research uses data which is descriptive in nature. Tools that educational researchers use in collecting qualitative data include: observations, conducting interviews, conducting document analysis, and analyzing participant products such as journals, diaries, images or blogs,.

Types of qualitative research

- Case study
- Ethnography
- Phenomenological Research
- Narrative Research
- Historical Research

Quantitative research

Quantitative research uses data that is numerical and is based on the assumption that the numbers will describe a single reality. Statistics are often applied to find relationships between variables.

Types of quantitative research

- Descriptive Survey Research
- Experimental Research
- Single — Subject Research
- Causal — Comparative Research
- Correlational Research

Combination methods

There also exists a new school of thought that these derivatives of the scientific method are far too reductionistic in nature,. Since educational research includes other disciplines such as psychology, sociology, anthropology, science, and philosophy and refers to work done in a wide variety of contexts [3] it is proposed that researchers should use "multiple research approaches and theoretical constructs". This could mean using a combination of qualitative and quantitative methods as well as common methodology from the fields mentioned above. In social research this phenomenon is referred to as triangulation (social science). This idea is well summarized by the work of Barrow in his text An introduction to philosophy of education:

"Since educational issues are of many different kinds and logical types, it is to be expected that quite different types of research should be brought into play on different occasions.

Types of combined methods

- Action Research
- Program Evaluation

Variables in research methodology

Research papers will mention a variety of different variables, and, at first, these technical terms might seem difficult and confusing. But with a little practice, identifying these variables becomes second nature. Because they are sometimes not explicitly labeled in the research writeup, it is useful to have a real research paper on hand as you learn these terms, so you can get some hands-on practice at identifying them.

Variables in research

The word variable is derived from the root word "vary", meaning, changing in amount, volume, number, form, nature or type. These variables should be measurable, i.e., they can be counted or subjected to a scale.

Variables are those simplified portions of the complex phenomena that you intend to study.

Variable can also be define as any characteristics, trait or attribute of interest to the research that can change or produce different precision at a given period of time and in a given geographical area.

The following are examples of variables in research:

- Independent variables
- Dependent Variables
- Moderator variables
- Covariate is a variable
- Intervening variables
- Extraneous variables

Independent Variable

- Another name for this variable is called the criterion variable. It is the variable that the researcher can actually manipulate. The variable explains the cause of any actions and that explains why it could in other words be regarded as ‘‘the cause’’. Example is the effects of *teaching methods* on the academic achievement of secondary school students in Biology in Enugu State. In this, situation, teaching methods becomes the independent variable as it accounts for what happens to the academic achievement of the learner in biology.

Dependent Variable

Another name for the dependent variable is called the effects. It is the variable that reflects the result of the manipulation of the independent variables. For example effects of teaching methods on the *academic achievement* of secondary school students in Biology in Enugu State. In this situation, academic achievement is regarded as the dependent variable since its function at this point is to reflect the effects of teaching methodology on the academic achievement of the students.

Moderator Variable

A moderator variable is a variable, which is thought to temper or modulate the magnitude of the effect of an independent variable on a dependent one**.** The major work of this variable is categorization of data for identification purpose. Moderators may be characteristics of people or characteristics of situations. In either case, they affect the magnitude of the relationship between an independent variable and a dependent one.

Covariate variables

A covariate is a variable that the researchers include in an analysis to determine whether the IV is able to influence the DV over and above any effect the covariate might have. The classic example is

when researchers take a baseline measurement, perform some manipulation, then take the measurement again. When they analyze this data, they will enter the baseline scores as a covariate, which will help cancel out any initial differences between the participants.

Intervening Variable

An intervening variable is a hypothetical variable used to explain causal links between other variables. Intervening variables cannot be observed in an experiment (that's why they are hypothetical). For example, there is an association between being poor and having a shorter life span. Just because someone is poor doesn't mean that will lead to an early death, so other hypothetical variables are used to explain the phenomenon.

Extraneous Variables

An extraneous variable is a little different from the rest because it is not directly measured, or often even wanted, by the researchers. It is a variable that has an impact on the results of the experiment that the researchers didn't anticipate. For example, the heat of the room might be different between two groups of IQ tests, and the hot room might annoy people and affect their scores.

CHAPTER 2

RESEARCH DESIGN

A **research design** is the "blue print" of the study. In other words, research design can be described as the skeletal structure or architectural framework upon which the research is conducted. The design of research can also be referred to as the process that help the collection of data, analysis of data for decision making. The design of a study defines the study type (descriptive, correlational, semi-experimental, experimental, review, meta-analytic) and sub-type (e.g., descriptive-longitudinal case study), research question, hypotheses, independent and dependent variables, experimental design, and, if applicable, data collection methods and a statistical analysis plan. Research design is the framework that has been created to seek answers to research questions.

Characteristics of research design

There are four key characteristics of research design:

Validity: There are multiple measuring tools available. However, the only correct measuring tools are those which help a researcher in gauging results according to the objective of the research. The questionnaire developed from this design will then be valid.

Generalization: The outcome of your design should apply to a population and not just a restricted sample. A generalized design implies that your survey can be conducted on any part of a population with similar accuracy.

Neutrality: When you set up your study, you may have to make assumptions about the data you expect to collect. The results projected in the research design should be free from bias and neutral. Understand opinions about the final evaluated scores and

conclusion from multiple individuals and consider those who agree with the derived results.

Reliability: With regularly conducted research, the researcher involved expects similar results every time. Your design should indicate how to form research questions to ensure the standard of results. You'll only be able to reach the expected results if your design is reliable.

Types Research Design

There are two main types of research design, namely the experimental research design and the non experimental research design.

Experimental research is a scientific approach to research, where one or more independent variables are manipulated and applied to one or more dependent variables to measure their effect on the latter. The effect of the independent variables on the dependent variables is usually observed and recorded over some time, to aid researchers in drawing a reasonable conclusion regarding the relationship between these 2 variable types.

The experimental research method is widely used in physical and social sciences, psychology, and education. It is based on the comparison between two or more groups with a straightforward logic, which may, however, be difficult to execute.

Mostly related to a laboratory test procedure, experimental research designs involve collecting quantitative data and performing statistical analysis on them during research. Therefore, making it an example of quantitative research method.

Quasi-experimental Research Design

The word "quasi" means partial, half, or pseudo. Therefore, the quasi-experimental research bearing a resemblance to the true experimental research, but not the same. In quasi-experiments, the participants are not randomly assigned, and as such, they are used in settings where randomization is difficult or impossible.

This is very common in educational research, where administrators are unwilling to allow the random selection of students for experimental samples.

Some examples of quasi-experimental research design include; the time series, no equivalent control group design, and the counterbalanced design.

True Experimental Research Design

The true experimental research design relies on statistical analysis to approve or disprove a hypothesis. It is the most accurate type of experimental design and may be carried out with or without a pretest on at least 2 randomly assigned dependent subjects.

The true experimental research design must contain a control group, a variable that can be manipulated by the researcher, and the distribution must be random. The classification of true experimental design include:

- **The posttest-only Control Group Design:** In this design, subjects are randomly selected and assigned to the 2 groups (control and experimental), and only the experimental group is treated. After close observation, both groups are post-tested, and a conclusion is drawn from the difference between these groups.
- **The pretest-posttest Control Group Design:** For this control group design, subjects are randomly assigned to the 2 groups, both are presented, but only the experimental group is treated.

After close observation, both groups are post-tested to measure the degree of change in each group.

- **Solomon four-group Design:** This is the combination of the pretest-only and the pretest-posttest control groups. In this case, the randomly selected subjects are placed into 4 groups.

The first two of these groups are tested using the posttest-only method, while the other two are tested using the pretest-posttest method.

Characteristics of Experimental Research

- **Variables**

Experimental research contains dependent, independent and extraneous variables. The dependent variables are the variables being treated or manipulated and are sometimes called the subject of the research.The independent variables are the experimental treatment being exerted on the dependent variables. Extraneous variables, on the other hand, are other factors affecting the experiment that may also contribute to the change.

- **Setting**

The setting is where the experiment is carried out. Many experiments are carried out in the laboratory, where control can be exerted on the extraneous variables, thereby eliminating them.
Other experiments are carried out in a less controllable setting. The choice of setting used in research depends on the nature of the experiment being carried out.

- **Multivariable**

Experimental research may include multiple independent variables, e.g. time, skills, test scores, etc.

Reasons for the use of Experimental Research Design

Experimental research design can be majorly used in physical sciences, social sciences, education, and psychology. It is used to make predictions and draw conclusions on a subject matter.

Some uses of experimental research design are highlighted below.

- **Medicine:** Experimental research is used to provide the proper treatment for diseases. In most cases, rather than directly using patients as the research subject, researchers take a sample of the bacteria from the patient's body and are treated with the developed antibacterial

  The changes observed during this period are recorded and evaluated to determine its effectiveness. This process can be carried out using different experimental research methods.
- **Education:** Asides from science subjects like Chemistry and Physics which involves teaching students how to perform experimental research, it can also be used in improving the standard of an academic institution. This includes testing students' knowledge on different topics, coming up with better teaching methods, and the implementation of other programs that will aid student learning.
- **Human Behavior:** Social scientists are the ones who mostly use experimental research to test human behaviour. For example, consider 2 people randomly chosen to be the subject of the social interaction research where one person is placed in a room without human interaction for 1 year.

  The other person is placed in a room with a few other people, enjoying human interaction. There will be a difference in their behaviour at the end of the experiment.

- **UI/UX:** During the product development phase, one of the major aims of the product team is to create a great user experience with the product. Therefore, before launching the

final product design, potential are brought in to interact with the product.

For example, when finding it difficult to choose how to position a button or feature on the app interface, a random sample of product testers are allowed to test the 2 samples and how the button positioning influences the user interaction is recorded.

Disadvantages of Experimental Research

- It is highly prone to human error due to its dependency on variable control which may not be properly implemented. These errors could eliminate the validity of the experiment and the research being conducted.
- Exerting control of extraneous variables may create unrealistic situations. Eliminating real-life variables will result in inaccurate conclusions. This may also result in researchers controlling the variables to suit his or her personal preferences.
- It is a time-consuming process. So much time is spent on testing dependent variables and waiting for the effect of the manipulation of dependent variables to manifest.
- It is expensive.
- It is very risky and may have ethical complications that cannot be ignored. This is common in medical research, where failed trials may lead to a patient's death or a deteriorating health condition.

The non-experimental research is one in which the variables of the study are not controlled or manipulated. To develop the research, the authors observe the phenomena to be studied in their natural environment, obtaining the data directly to analyze them later.

In other words, non-experimental research is research that lacks the manipulation of an independent variable. Rather than manipulating an independent variable, researchers conducting non-experimental research simply measure variables as they naturally occur (in the lab or real world).

Most researchers in psychology consider the distinction between experimental and non-experimental research to be an extremely important one. This is because although experimental research can provide strong evidence that changes in an independent variable cause differences in a dependent variable, non-experimental research generally cannot. As we will see, however, this inability to make causal conclusions does not mean that non-experimental research is less important than experimental research.

When to Use Non-Experimental Research

As we saw in the last chapter, experimental research is appropriate when the researcher has a specific research question or hypothesis about a causal relationship between two variables—and it is possible, feasible, and ethical to manipulate the independent variable. It stands to reason, therefore, that non-experimental research is appropriate—even necessary—when these conditions are not met. There are many times in which non-experimental research is preferred, including when:

- the research question or hypothesis relates to a single variable rather than a statistical relationship between two variables (e.g., How accurate are people's first impressions?).
- the research question pertains to a non-causal statistical relationship between variables (e.g., is there a correlation between verbal intelligence and mathematical intelligence?).
- the research question is about a causal relationship, but the independent variable cannot be manipulated or participants cannot be randomly assigned to conditions or orders of

conditions for practical or ethical reasons (e.g., does damage to a person's hippocampus impair the formation of long-term memory traces?).

- the research question is broad and exploratory, or is about what it is like to have a particular experience (e.g., what is it like to be a working mother diagnosed with depression?).

Characteristics of Non-Experimental

As previously mentioned, the first characteristic of this type of research is that there is no manipulation of the variables studied. Normally, these are phenomena that have already occurred and are analyzed a posteriori. Apart from this characteristic, other peculiarities present in these designs can be pointed out:

- Non-experimental research is widely used when, for ethical reasons (such as giving drink to young people), there is no option to conduct controlled experiments.
- No groups are formed to study them, but these are already pre-existing in their natural environments.
- The data is collected directly, and then analyzed and interpreted. There is no direct intervention on the phenomenon.
- It is very common that non-experimental designs are used in applied research, since they study the facts as they occur naturally.
- Given the characteristics presented, this type of research is not valid to establish unequivocal causal relationships.

Types of Non Experimental Design

- Survey research design
- Causal comparative (Ex-post-facto) research design
- Case study research design
- Historical research design
- Correlational research design

Survey Research Design

Technically, a survey is a method of gathering and compiling information from a group of people, more often known as the sample, to gain knowledge by organizations, businesses, or institutions. This information or opinion collected from the sample is more often generalization of what a large population thinks.

In survey research, respondents answer through surveys or questionnaires or polls. They are a popular market research tool to collect feedback from respondents. A study to gather useful data should have the right survey questions. It should be a balanced mix of open-ended questions and close ended-questions. The survey method can be conducted online or offline, making it the go-to option for descriptive research where the sample size is enormous.
Consider hypothetically, an organization conducts a study related to breast cancer in America, and they choose a sample to obtain cross-sectional data. This data indicated that breast cancer was most prevalent in women of African-American origin. The information is from one point in time. Now, if the researcher wants to dwell more in-depth into the research, he/she can deploy a longitudinal survey.

Types of survey research design

a. Descriptive survey research: Descriptive research is defined as a research method that describes the characteristics of the population or phenomenon studied. This methodology focuses more on the "what" of the research subject than the "why" of the research subject.
The descriptive research method primarily focuses on describing the nature of a demographic segment, without focusing on "why" a

particular phenomenon occurs. In other words, it "describes" the subject of the research, without covering "why" it happens.

Characteristics of descriptive survey research

The term descriptive research then refers to research questions, design of the study, and data analysis conducted on that topic. We call it an observational research method because none of the research study variables are influenced in any capacity.

Some distinctive characteristics of descriptive research are:

i. **Quantitative research:** Descriptive research is a quantitative research method that attempts to collect quantifiable information for statistical analysis of the population sample. It is a popular market research tool that allows us to collect and describe the demographic segment's nature.
ii. **Uncontrolled variables:** In descriptive research, none of the variables are influenced in any way. This uses observational methods to conduct the research. Hence, the nature of the variables or their behavior is not in the hands of the researcher.
iii. **Cross-sectional studies:** Descriptive research is generally a cross-sectional study where different sections belonging to the same group are studied.
iv. **The basis for further research:** Researchers further research the data collected and analyzed from descriptive research using different research techniques. The data can also help point towards the types of research methods used for the subsequent research.

a. Longitudinal surveys:

Longitudinal surveys are those surveys that help researchers to make an observation and collect data over an extended period. There are three main types of longitudinal studies: trend surveys, panel surveys, cohort surveys.

Trend surveys are deployed by researchers to understand the shift or transformation in the thought process of respondents over some time. They use these surveys to understand how people's inclination change with time.

Another longitudinal survey type is a panel survey. Researchers administer these surveys to the same set or group of people over the years. Panel surveys are expensive in nature and researchers to try to stick to their panel to gather unbiased opinions.

The third type of longitudinal survey is the cohort survey. In this type, categories of people that meet specific similar criteria and characteristics form the target audience. The same people don't need to create a group. However, people forming a group should have certain similarities.

b. Retrospective survey:

A retrospective survey is a type of study in which respondents answer questions to report on events from the past. By deploying this kind of survey, researchers can gather data based on past experiences and beliefs of people. This way, they can save the cost and time required, unlike a longitudinal survey.

- **Causal comparative research design/ex-post-facto**

Most often, in experimental research, when a researcher wants to compare groups in a more natural way, the approach used is causal design. On the other hand, in a non-experimental setting, if a researcher wants to identify consequences or causes of differences between groups of individuals, then typically causal-comparative design is deployed.

Causal-comparative, also known as *ex post facto* (after the fact) research design, is an approach that attempts to figure out a causative relationship between an independent variable & a dependent variable. It must be noted that the relationship between the independent variable and dependent variable is a

suggested relationship and not proven as the researcher do not have complete control over the independent variable.

This method seeks to build causal relationships between events and circumstances.

Types of causal comparative

This research design is further segregated into:

- **Retrospective causal-comparative research** – In this method, a research question after the effects have occurred is investigated. The researcher aims to determine how one variable may have impacted another variable.
- **Prospective causal-comparative research** – This method begins with studying the causes and is progressed by investigating the possible effects of a condition.

- **Case study method**

  Case studies involve in-depth research and study of individuals or groups. Case studies lead to a hypothesis and widen a further scope of studying a phenomenon. However, case studies should not be used to determine cause and effect as they can't make accurate predictions because there could be a bias on the researcher's part. The other reason why case studies are not a reliable way of conducting descriptive research is that there could be an atypical respondent in the survey. Describing them leads to weak generalizations and moving away from external validity.

Historical research design

This is an example of non experimental research design that is employed if the researcher is interested in historical studies.

The purpose of a historical research design is to collect, verify, and synthesize evidence from the past to establish facts that defend or refute a hypothesis. It uses secondary sources and a variety of

primary documentary evidence, such as, diaries, official records, reports, archives, and non-textual information [maps, pictures, audio and visual recordings.

Correlational research design: Correlational research is a non-experimental research design technique that helps researchers establish a relationship between two closely connected variables. This type of research requires two different groups. There is no assumption while evaluating a relationship between two different variables, and statistical analysis techniques calculate the relationship between them.

A correlation coefficient determines the correlation between two variables, whose value ranges between -1 and +1. If the correlation coefficient is towards +1, it indicates a positive relationship between the variables and -1 means a negative relationship between the two variables.

Differences between Experimental and Non-Experimental Research

- In experimental research, the researcher can control and manipulate the environment of the research, including the predictor variable which can be changed. On the other hand, non-experimental research cannot be controlled or manipulated by the researcher at will.

  This is because it takes place in a real-life setting, where extraneous variables cannot be eliminated. Therefore, it is more difficult to conclude non-experimental studies, even though they are much more flexible and allow for a greater range of study fields.

- The relationship between cause and effect cannot be established in non-experimental research, while it can be established in experimental research. This may be because many extraneous variables also influence the changes in the research subject, making it difficult to point at a particular variable as the cause of a particular change
- Independent variables are not introduced, withdrawn or manipulated in non-experimental designs, but the same may not be said about experimental research.

CHAPTER 3

LITERATURE REVIEW

Literature review is a search and evaluation of the available literature in your given subject or chosen topic area. It documents the state of the art with respect to the subject or topic you are writing about.

A literature review has four main objectives:

- It surveys the literature in your chosen area of study
- It synthesises the information in that literature into a summary
- It critically analyses the information gathered by identifying gaps in current knowledge; by showing limitations of theories and points of view; and by formulating areas for further research and reviewing areas of controversy
- It presents the literature in an organised way

A literature review shows your readers that you have an in-depth grasp of your subject; and that you understand where your own research fits into and adds to an existing body of agreed knowledge.

Reasons/importance of reviewing literature

- Literature review helps in identifying gaps in knowledge to be filled;
- Literature review helps in avoiding unnecessary duplication in research;
- demonstrates a familiarity with a body of knowledge and establishes the credibility of your work;
- summarises prior research and says how your project is linked to it;
- integrates and summarises what is known about a subject;

- demonstrates that you have learnt from others and that your research is a starting point for new ideas.
- to ensure you have a thorough understanding of the topic;
- to identify potential areas for research;
- to identify similar work done within the area;
- to compare previous findings;
- to critique existing findings and suggest further studies.

Sources of literature review

There are basically two sources of literature review in research. There are the primary source and the secondary source.

Primary source

A primary source is an original material created during the time under study. Primary sources can be original documents (such as letters, speeches, diaries), creative works (such as art, novels, music and film), published materials of the times (newspapers, magazines, memoirs, etc.), institutional and government documents (treaties, laws, court cases, marriage records) or relics and artifacts (such as buildings, clothing, or furniture).

Primary sources of information include:

- literary works
- original documents such as diaries, letters, original manuscripts
- archival material, such as official documents, minutes, etc. recorded by government agencies and organizations
- original research studies, also called empirical studies

Examples of primary sources include:

- Artifacts
- Books
- Diaries
- Ephemera

- Journals
- Ledgers
- Maps
- Letters
- Manuscripts
- Newsletters
- Pamphlets
- Photographs
- Videos

Secondary sources

Secondary sources put primary sources in context. They comment, summarize, interpret or analyze information found in primary sources. Secondary sources are usually written by individuals who did not experience firsthand the events about which they are writing.

Examples of secondary sources include:

- academic books
- biographies
- journal articles
- magazine articles
- dissertations
- theses
- essays
- encyclopedia articles

Dimensions of literature review

In research, there are five basic sub-heading upon which literature review is conducted. These sub headings are:

- the conceptual framework;
- theoretical framework;
- theoretical studies;
- empirical studies and

- summary of the review of related literature.

Conceptual framework

A conceptual framework is used to clarify concepts, organize ideas, and identify relationships with which to frame a study. Concepts are logically developed and organized to support an overall framework and often exhibited graphically within dissertation research.

A conceptual framework represents the researcher's synthesis of the literature on how to explain a phenomenon. It maps out the actions required in the course of the study, given his previous knowledge of other researchers' point of view and his observations on the subject of research.

In other words, the conceptual framework is the researcher's understanding of how the particular **variables** in his study connect. Thus, it identifies the variables required in the research investigation. It is the researcher's "map" in pursuing the investigation.

As McGaghie *et al*. (2001) put it: The conceptual framework "sets the stage" to present the particular research question that drives the investigation being reported based on the problem statement. The problem statement of a thesis gives the context and the issues that caused the researcher to conduct the study.

The conceptual framework lies within a much broader framework called a theoretical framework. The latter draws support from time-tested theories that embody many researchers' findings on why and how a particular phenomenon occurs.

Theoretical framework

A theoretical framework is a broad and established set of rules, truths, or principles on which the study is founded. Examples

include Newton's laws of motion in physical sciences and Maslow's hierarchy of needs in social sciences. Thus, for instance, a physicist *could* use Newton's laws of motion, or one of the laws, to study the appearance of comets, the speed of asteroids, or the gravitational pull of a black hole. Similarly, a sociologist *could* use Maslow's hierarchy of needs to study the life cycle of social media platforms. Note that you can use multiple theoretical frameworks as needed for your study.

Theoretical studies

A theoretical study is one that does not depend upon an experiment, manipulation of variables or empirical evidence. It is based on testing, exploring or developing theories, and it generally involves observation or the compilation of information

Empirical Studies

Empirical study articles are reports of original research. They can also include secondary analyses that test hypotheses by presenting novel analyses of data not considered or addressed in previous reports. Empirical studies are generally published in academic, peer-reviewed journals and generally consist of distinct sections that reflect the stages in the research process.

References

A reference list is a list of the resources that you used when writing your assignment or doing your research. These resources may include:

- books, including electronic books, journals (online an paper based)
- online sources including websites, blogs and forums
- speeches
- conference papers, proceedings and theses
- other sources of information such as film, television, video, etc.

Reference lists come at the end of an assignment, and are arranged in alphabetical order, usually by author or editor. If there isn't an author or an editor, the title is used.

Citations

Citations or in-text citations are similar to references, but occur in the body of the text with direct quotes and paraphrases to identify the author/publication for the material you have used. Citations are used:

- to show which reference supports a particular statement
- for direct quotes – when you repeat a passage from a text (or speech, video, etc.) in your assignment without changing any words
- when you paraphrase – this is when you use your own words to restate the meaning of a text in your assignment.

Importance of referencing and citation

- Helps show that you have been thorough and careful (or rigorous) in your academic work
- Indicates what material is the work of another person or is from another source
- Indicates what material is your original work since you have provided a citation for work that is not your own
- Allows the reader to refer back to any external material (i.e., not your own) that you have stated or discussed
- Provides the reader with an indication of the quality and authority of the material you are referencing (e.g., published article in a respected journal, unpublished opinion piece on a popular online website) Of course the relevance and importance of material is dependent on your topic
- Lets the reader see if you have included up-to-date work, seminal (early and influential) work, and material central to your research topic

CHAPTER 4

POPULATION, SAMPLE AND SAMPLING TECHNIQUES

A **population** is the entire group that you want to draw conclusions about.In research, a population doesn't always refer to people. It can mean a group containing elements of anything you want to study, such as objects, events, organizations, countries, species, organisms, variables, attribute, characteristics etc.

Sample
Sample can be described as a small subset carved out of the population to bear a representativeness of the population.
In other words, **sample** is the specific group that you will collect data from. The size of the sample is always less than the total size of the population.

Reasons for sampling

- **Necessity**: Sometimes it's simply not possible to study the whole population due to its size or inaccessibility.
- **Practicality**: It's easier and more efficient to collect data from a sample.
- **Cost-effectiveness**: There are fewer participant, laboratory, equipment, and researcher costs involved.
- **Manageability**: Storing and running statistical analyses on smaller datasets is easier and reliable.
- To help in minimizing error from the despondence due to large number in the population
- Sampling helps the researcher to meet up with the challenge of time.

Features of good sampling

The Features of good Sampling are stated below

- Sample design must result in a truly representative sample.
- Sample design must be such which results in a small sampling error.
- Sample design must be viable in the context of funds available for the research study.
- Sample design must be such so that systematic bias can be controlled in a better way.
- Sample should be such that the results of the sample study can be applied, in general, for the universe with a reasonable level of confidence.

Sampling Techniques

There are basically two sampling techniques namely, probability and the non probability sampling technique.

Probability sampling methods

Probability sampling means that every member of the population has a chance or equal opportunity of being selected. It is mainly used in quantitative research. If you want to produce results that are representative of the whole population, you need to use a probability sampling technique.

There are four main types of probability sample.

1. Simple random sampling

In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.

To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.

Example

You want to select a simple random sample of 100 employees of Company X. You assign a number to every employee in the

company database from 1 to 1000, and use a random number generator to select 100 numbers.

2. Systematic sampling

Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

Example

All employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point: number 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.

If you use this technique, it is important to make sure that there is no hidden pattern in the list that might skew the sample. For example, if the HR database groups employees by team, and team members are listed in order of seniority, there is a risk that your interval might skip over people in junior roles, resulting in a sample that is skewed towards senior employees.

3. Stratified sampling

Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.

To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (e.g. gender, age range, income bracket, job role).

Based on the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or systematic sampling to select a sample from each subgroup.

Example

The company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of

the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people.

4. Cluster sampling

Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.

If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above.

This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It's difficult to guarantee that the sampled clusters are really representative of the whole population.

Example

The company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don't have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.

Non-probability sampling methods

In a non-probability sample, individuals are selected based on non-random criteria, and not every individual has a chance of being included.

This type of sample is easier and cheaper to access, but it has a higher risk of sampling bias, and you can't use it to make valid statistical inferences about the whole population.

Non-probability sampling techniques are often appropriate for exploratory and qualitative research. In these types of research, the

aim is not to test a hypothesis about a broad population, but to develop an initial understanding of a small or under-researched population.

1. Convenience sampling

A convenience sample simply includes the individuals who happen to be most accessible to the researcher.

This is an easy and inexpensive way to gather initial data, but there is no way to tell if the sample is representative of the population, so it can't produce generalizable results.

Example

You are researching opinions about student support services in your university, so after each of your classes, you ask your fellow students to complete a survey on the topic. This is a convenient way to gather data, but as you only surveyed students taking the same classes as you at the same level, the sample is not representative of all the students at your university.

2. Voluntary response sampling

Similar to a convenience sample, a voluntary response sample is mainly based on ease of access. Instead of the researcher choosing participants and directly contacting them, people volunteer themselves (e.g. by responding to a public online survey).

Voluntary response samples are always at least somewhat biased, as some people will inherently be more likely to volunteer than others.

Example

You send out the survey to all students at your university and a lot of students decide to complete it. This can certainly give you some insight into the topic, but the people who responded are more likely to be those who have strong opinions about the student support services, so you can't be sure that their opinions are representative of all students.

3. Purposive sampling

This type of sampling, also known as judgement sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research.

It is often used in qualitative research, where the researcher wants to gain detailed knowledge about a specific phenomenon rather than make statistical inferences, or where the population is very small and specific. An effective purposive sample must have clear criteria and rationale for inclusion.

Example

You want to know more about the opinions and experiences of disabled students at your university, so you purposefully select a number of students with different support needs in order to gather a varied range of data on their experiences with student services.

4. Snowball sampling

If the population is hard to access, snowball sampling can be used to recruit participants via other participants. The number of people you have access to "snowballs" as you get in contact with more people.

Example

You are researching experiences of homelessness in your city. Since there is no list of all homeless people in the city, probability sampling isn't possible. You meet one person who agrees to participate in the research, and she puts you in contact with other homeless people that she knows in the area.

CHAPTER 5

BASIC STATISTICS FOR DATA ANALYSI IN RESEARCH

Meaning of Statistics

Statistics may be defined as a science that studies numerical values assigned to objects, attributes or variables in order to characterize and make the variables more meaningful. It is, therefore, the science of classifying, organizing, and analyzing data (King & Minium, 2003). Usually, in statistics, information is presented not in words but in numerical form- numbers.Statistics deals with populations and samples. The average or mean or other descriptions of the population are referred to as parameters but the descriptions of the sample (such as the mean or standard deviation) are referred to as statistics.

The subject, statistics may be divided into two broad areas:

a). Descriptive Statistics:

This refers to the statistical procedures that are used to describe and summarize data for easy comprehension. Examples of descriptive statistics are the mean, standard deviation, range etc which describe and give you information about a sample or population.

b). Inferential Statistics

This area of statistics is often based on the information derived from descriptive statistics. It is often referred to as analytical statistics and enables us draw conclusions or make informed statistical inferences from a mass of data. The purpose of inferential statistics is to draw a conclusion about conditions that exist in a population from study of a sample. Therefore, a researcher uses it when he bases his conclusions of a research on limited information (information from 1 sample of a population). Examples of inferential statistics are: t-test, ANOVA, Chi square etc. Inferential statistics starts with formulation of hypothesis - null

hypothesis or alternative hypothesis, which in fact are guesses. At the end of the analysis, we confirm or reject the hypothesis.

Significance of Statistics

Statistics to Mendenhall (1993, p.3) is about describing and predicting. It is a response to the limitations of our ability to know. Any complex aspect of human knowledge is either difficult for us to understand, or lends itself to misunderstanding - unless we possess the right tool with which to examine and analyze its complexity.

Statistics is that tool.

Statistics is basic to the world of manufacturing, social sciences, business, politics, pure sciences, behavioral sciences, and research in all spheres of life because:

a. It provides the necessary exact descriptions.
b. It forces researchers by its methods to follow defined procedures in their work and thinking.
c. Its method provides a means of summarizing results of research in meaningful and convenient form.
d. It enables researchers to make predictions based on the data available at hand.
e. It provides easy means of analyzing cause(s) of complex events.

Variables in Statistics

As stated earlier, in statistics, we use numbers to describe, measure or give other information e.g. about performance of students, intelligence level of students, height of class pupils etc. The characteristics being measured is referred to as **"variables"**.If we are studying the scores of students in mathematics, the scores become the variable. Sometimes we study the relationship between two variables. One of the two or more variables is called independent variable, while the other variable is referred to as dependent variable.

Continuous – Variables

These are variables that can take on infinite number of values within a range and may not necessarily be determined by mere counting. For example, given the range of scores 60 to 70, there is every possibility that some students may have scores between two scores like 60.5, 61.2, 62.4, 62.8, 64.7 etc. The fact that students' score are rounded up to whole numbers does not remove the possibility of occurrence of fractional scores.

Discontinuous or Discrete Variables

These are variables which take on finite number of values within a range. These refer to data that can be fully counted and are not fractional. Discontinuous variables refer to variables that are tangible and can be measured directly.

Variate in Educational Statistics

A variate is the single value of a variable. This means that if a variable of interest is test scores of students in statistics, a variate refers to one single score of a student in the distribution. While analyzing any data, we deal with a number of variates which we may refer to as raw scores or data.

Scales of Statistical Measurement

Once more, it is emphasized that variables are expressed in numerical form. The process used in determining or assigning these numbers informs the interpretation that can be made from them and the statistical procedures that can be used meaningfully. This means that during the process of data collection, different instruments are used and each instrument enables the researcher to collect a particular type of data. Each set of data collected lends itself to a particular statistical analysis. There are basically the following scales of measurement, starting with the least rated:

Nominal Scale

This is the lowest scale of measurement and involves placing objects or individuals into categories which are qualitative rather than quantitative. For example, you may categorize individuals into male or female; people you meet in the university can be divided into students or lecturers; or success in an examination as either pass or fail. There is no order of magnitude involved in this categorization. In filling forms sometimes you tick "1" for male and "2" for female. These numbers do not indicate any magnitude but are used for convenience; in fact they are arbitrarily assigned and can be changed at will. The numbers cannot be added, subtracted, divided or subjected to any mathematical treatment.

Ordinal Scale

The ordinal scale shows you that things are different but does not signify the direction or degree of the difference. It indicates the order of magnitude or rank e.g. 1^{st}, 2^{nd}, 3^{rd}, 4^{th}. In other words, ordinal scale categorizes variables in order of magnitude. It does not Je the exact interval or difference between the categories. When we rank students (without their marks) or when we use grades like grade 1, grade 2, etc, in WAEC examinations we are using ordinal scale. We shall use ordinal scale in correlation studies.

Interval Scale

In this scale, the exact difference or interval between categories is specified. Suppose we have four students with scores A, 90%; B, 80%, C, 70%; D. 60%, by ordinal scale, A is first, B is second, C is third, D is fourth. When we state their scores in percentages, we are using interval scale. Interval scales can be subjected to multiplication, addition or subtraction. For instance, we can subtract 10 marks from each student. Interval scale provides order as well as difference between the intervals. To get the difference, we subtract one category from another.

Ratio Scale

Ratio scale is the highest scale of measurement. This scale has absolute zero. Ratio scale is used more in the physical sciences than in behavioural sciences. For instance, when we measure temperature in degrees Kelvin, we are certain of the real meaning of zero degree on the Kelvin scale i.e. the temperature at which water freezes. Other examples of ratio scale measurements are length, weight, and measures of elapsed time.

CHAPTER 6

DATA COLLECTION AND ANALYSIS

Frequency-Distribution Table

When summarizing masses of raw data, it is often useful to distribute the data into classes, or categories and to determine the number of individuals belonging to each class called the class frequency. A tabular arrangement of data b y classes together with the corresponding class frequencies is called a frequency distribution or frequency table.

Elements of Frequency Distribution Table

Class Interval and Class Limit

A symbol defining a class such as 60-62, 63-65…. This is called interval. The end no 60 and 62, are called limit, the smaller one is called lower class limit and the larger one is called the upper limit. The term class limit and class interval is used interchangeably, although the class interval is always a symbol for the class.

Class Boundary

This shows the actual boundary between two classes and is always obtainable when you subtract 0.5 from the lower limit and add 0.5 to the upper limit. This is done to enable some classes to be represented. For example in the class interval above, (60-62) 50.5 and 62.5 falls within that class interval and unless there is a class boundary such group like 50.5 and 62.5 cannot be represented.

The Size, or Width of Class Interval

The size of width of a class interval is the difference between the lower and the upper boundaries. If all the interval of a frequency distribution has equal widths, this common width is denoted b y c. in such case c is equal to difference between two successive lower

class limits or two successive upper limits. For example, the class size of the above class boundary is 2.

Organisation of Data

This involves the use of tables more especially the frequency distribution tables. It is the one that represents information related to the number of items of each score in a given distribution.

There are two variations

a. Ungrouped frequency distribution
b. Grouped frequency distribution.

Ungrouped Frequency Distribution

Consider the following scores obtained from the boutique record of Gladys a store of 30 products

11 18 10 15 0 8 7 3 18 22
15 6 15 3 23 5 18 3 18 10
8 4 29 11 14 13 5 18 14 12

The above set of scores does not make sense as they appeared. For that reason, there is need to organize the data to make it have more meaning at a glance.

Guidelines on how to construct an ungrouped frequency distribution table

i. Build a column for the scores
ii. Arrange the scores from the highest to the lowest or from the lowest to the highest.
iii. The next column will be for tally to help you determine the number of time any score appeared.
iv. Add all the numbers in each line of the tally and write them down in the frequency.
v. From the tally you then determine your cumulative frequency.

From the above data, prepare a frequency distribution table for the above scores.

Table 1. Ungrouped Frequency Distribution Table

Score (X)	Tally (T)	Frequency (F)	Cum. Frequency (CF)
0	*I*	1	1
3	*III*	3	4
4	*I*	1	5
5	*II*	2	7
6	*I*	1	8
7	*I*	1	9
8	*II*	2	11
10	*II*	2	13
11	*II*	2	15
12	*I*	1	16
13	*I*	1	17
14	*II*	2	19
15	*III*	3	22
18	*IIII*	5	27
22	*I*	1	28
23	*I*	1	29
29	*I*	1	30
		30	

Organization of data (Example 1) (Ungrouped data)

~~1~~ ~~2~~ ~~1~~ ~~6~~ ~~8~~

~~3~~ ~~2~~ ~~7~~ ~~5~~ ~~4~~

~~2~~ ~~4~~ ~~5~~ ~~2~~ ~~3~~

Scores (x)	Tallies	Frequency (F)
1	II	2
2	III	4
3	II	2
4	II	2
5	1	1
6	I	1
7	I	1
8	I	1

$\Sigma F = 14$

Example 2

~~10~~ ~~11~~ ~~12~~ ~~10~~ ~~20~~

~~10~~ ~~20~~ ~~13~~ ~~14~~ ~~11~~

~~20~~ ~~18~~ ~~15~~ ~~17~~ ~~16~~

~~17~~ ~~19~~ ~~19~~ ~~10~~ ~~16~~

Score (x)	Tallies	Frequency (F)	Cumulative frequency (C. F)
10	III	3	3
11	II	2	5
12	I	1	6
13	I	1	7
14	I	1	8
15	I	1	9
16	II	2	11
17	II	2	13
18	II	2	15
19	II	2	17
20	III	3	20
		$\Sigma f = 20$	

Group Frequency Distribution

Consider the same data given above, we can compress the data and group the data and then prepare a frequency distribution table for it.

Guidelines to follow when preparing a frequency distribution table for grouped data are:

i. Decide on the number of classes you wish to have. This should neither be too small nor too large. This could be between 10 and twenty.
ii. Determine the range of scores (highest score-Lowest score) +1. In the case of the above score (29-0) +1=30.
iii. Divide the range by the number of class interval and approximate to the nearest whole number. This will give you the class size. Assuming we want to group the above data into 10 the class will be 30/3=3.

Therefore in each class interval we shall have three scores. Beginning with the lowest score, the first interval will be 0-2, followed by 3-5and so on, as shown below.

Table 2: Group frequency distribution table

Class interval	**Class Limit**	**Class mark (X)**	**Tally (T)**	**Frequency (F)**
0-2	-0.5-2.5	1	*I*	1
3-5	2.5-3.5	4	*IIII I*	6
6-8	3.5-8.5	7	*IIII*	4
9-11	8.5-11.5	10	*IIII*	4
12-14	11.5-14.5	19	*IIII*	4
15-17	14.5-17.5	16	*III*	3
18-20	17.5-20.5	19	*IIIII*	5
21-23	21.5-23.5	22	*II*	2
24-26	23.5-26.5	25	*0*	0
27-29	26.5-29.5	28	*I*	1
				30

Example 2

Consider the data below, being scores generated from a test administered by biology in Queens College Enugu

~~10~~ ~~20~~ ~~55~~ ~~12~~ ~~73~~ ~~19~~ ~~33~~ ~~14~~ ~~21~~ ~~54~~

~~20~~ ~~28~~ ~~80~~ ~~38~~ ~~64~~ ~~22~~ ~~16~~ ~~73~~ ~~32~~ ~~65~~

~~30~~ ~~40~~ ~~71~~ ~~50~~ ~~72~~ ~~30~~ ~~39~~ ~~62~~ ~~43~~ ~~76~~

~~41~~ ~~44~~ ~~20~~ ~~62~~ ~~12~~ ~~53~~ ~~18~~ ~~44~~ ~~54~~ ~~28~~

~~55~~ ~~60~~ ~~15~~ ~~33~~ ~~21~~ ~~60~~ ~~10~~ ~~30~~ ~~65~~ ~~39~~

Required categorize the data into five classes

Since the data is above 30 (N>30), they is need o group the data into the number of required classes.

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

Where R= Range, (H – L) => Highest observation – lower observation

$$C = \frac{H-L}{K}$$

$$C = \frac{80-10}{5}$$

$$C = \frac{70}{5}$$

$$C = 14$$

The implication is that each of the classes will contain 5 elements

C. I	Tallies	Frequency
10 – 24	~~𝍸~~ ~~𝍸~~ ~~𝍸~~ II	17
25 – 39	~~𝍸~~ IIII	9
40 – 54	~~𝍸~~ IIII	9
55 – 69	~~𝍸~~ IIII	9
70 - 84	~~𝍸~~ I	6
		Σf = 50

Example 3

Consider the scores obtained from biology achievement test administered by a teacher in Elyon's College Enugu

10	20	55	12	75	19	35	14	21	54
20	28	80	38	64	22	16	73	32	65
30	40	71	50	72	30	39	62	43	76
41	44	20	62	12	53	18	44	54	28
55	60	15	33	21	60	10	30	65	39

Categorize the data into seven classes and expand more on the table

Solution

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

$$C = \frac{80-10}{7}$$

$$C = \frac{70}{7}$$

$$C = 10$$

In this case, there will be 10 elements in each class frequency distribution table,

C. I	X	Tallies	Frequency	Cum. Freq. (C.F)	FX
10 – 20	15	𝍸 𝍸 II	12	12	180
21 – 31	26	𝍸 III	8	20	208
32 – 42	37	𝍸 III	8	28	296
43 - 53	48	𝍸	5	33	240
54 – 64	59	𝍸 IIII	9	42	531
65 – 75	70	𝍸 I	6	48	420
76 - 86	81	II	2	50	162
			Σf = 50		Σfx = 2037

Cumulative Frequency

To ascertain cumulative frequency, the first frequency is always added to the next frequency while the first cumulative frequency and the first frequency are the same. For example, in the above table, the first frequency is 12, the first cumulative frequency is therefore 12.

The next cumulative frequency will be 12 + 8 = 20

Next 20 + 8 = 28

Next 28 + 5 = 33

Next 33 + 9 = 42

Next 42 + 6 = 48

Next 48 + 2 = 50

Note that the last cumulative frequency must be equal to the summation of all the frequency.

X = Midpoint

This is obtained by adding the lower class interval to the upper class interval, divided by 2. In other words, X (midpoint) is

obtained by finding the average of the lower and upper class interval.

For example, in table above, to get the (X) midpoint of the first role is simply add $\frac{10+20}{2} = \frac{30}{2} = 15$

Second role $\frac{21+31}{2} = \frac{52}{2} = 26$

Third role $\frac{32+42}{2} = \frac{74}{2} = 37$

FX => Frequency (F) multiply by observation (X)

For example, in the first role

First FX => $12 \times 15 = 180$

Second role => $8 \times 26 = 208$

Third role => $8 \times 37 = 296$

Fourth role => $5 \times 48 = 240$ etc

Graphical Representation Of Data

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. Pie chart
    ii. Frequency polygon
    iii. Histogram
    iv. Cumulative frequency graph or ogive

Pie Chart

Pie chart is one of the ways by which business information can be represented in a summarized and interpreted form. This is a circular graph and is used when one want to show the relationship

between part and the whole, to determine the number of items sold from a boutique of Gladys Nigeria Limited among all the stock of cloth she has in the market.

Example 1

You have been invited by Mrs. Gladys and the following information is provided to you

Table 3. Boutique Items

S/N	Items sold	Frequency of sales
1	Hand bag	100
2	Amingo attachment	250
3	High hills	300
4	Skirts	200
5	Trousers	150
	Total	1000

Required: prepare a pie chart for Gladys Nigeria Ltd as a business expert in a pie chart.

Solution

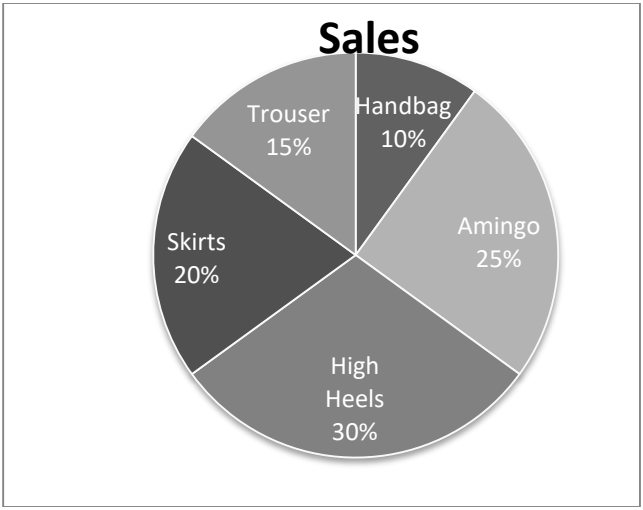
Hand bag $\frac{100}{1000} \times \frac{360}{1} = 36$

Amingo $\frac{250}{1000} \times \frac{360}{1} = 90$

High Hills $\frac{300}{1000} \times \frac{360}{1} = 108$

Skirts $\frac{200}{1000} \times \frac{360}{1} = 72$

Trouser $\frac{150}{1000} \times \frac{360}{1} = 54$



Example 2

Ralu who is a teacher with Urban Girls' Primary School Trans-Amadi conducted a test during the midterm. The following averages were obtained.

English language	20
Arithmetic	12
Elementary science	10
Verbal Reasoning	15
Health Education	18
Agricultural Science	20

You are required to present the information on a pie chart

Solution

S/N	Items	Scores	Scores restricted to 360°
1	English language	20	$\frac{20}{95} \times \frac{360}{1} = 75.79$
2	Arithmetic	12	$\frac{12}{95} \times \frac{360}{1} = 45.47$
3	Elementary science	10	$\frac{10}{95} \times \frac{360}{1} = 37.89$
4	Verbal Reasoning	15	$\frac{15}{95} \times \frac{360}{1} = 56.84$
5	Health Education	18	$\frac{18}{95} \times \frac{360}{1} = 68.21$
6	Agricultural Science	20	$\frac{20}{95} \times \frac{360}{1} = 75.79$
		95	$= 359.99 \cong 360°$

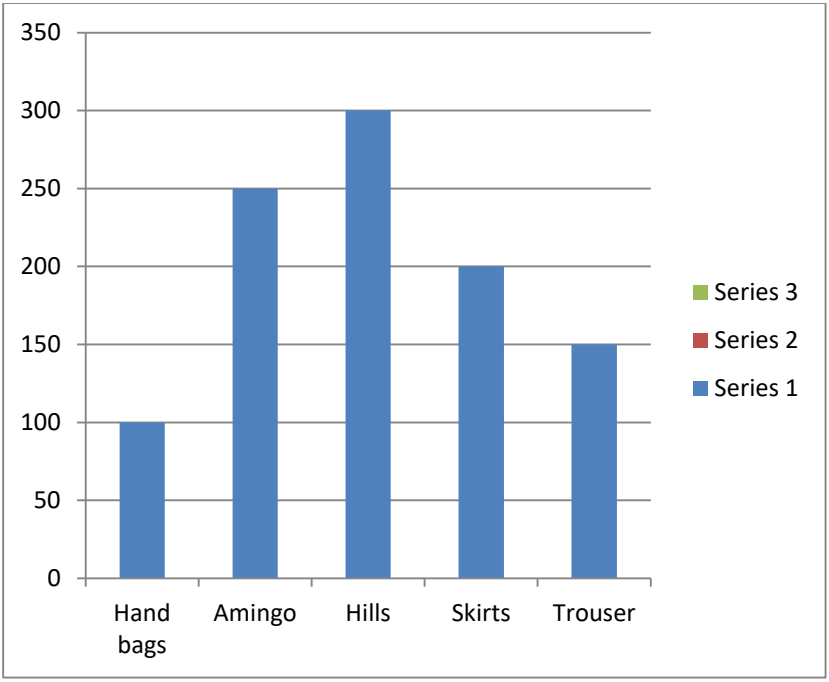
Bar Chart

Bar chart is another graphical and basic tool of statistical analysis that helps in presenting information in a summarized form using bars.In bar chart, bar's heights correspond to the frequency of each score or interval are used to represent information.

Illustration

Consider the data given from Gladys' Nigeria limited and present the information on a bar chart.

Solution



From the above information on the bar chart, all the items sold at Gladys' Nig. Ltd are shown at a glance and the frequencies at which they were sold at ease.

There are some situations, where the bar chart is constructed otherwise, for example when the bars are laying horizontally.

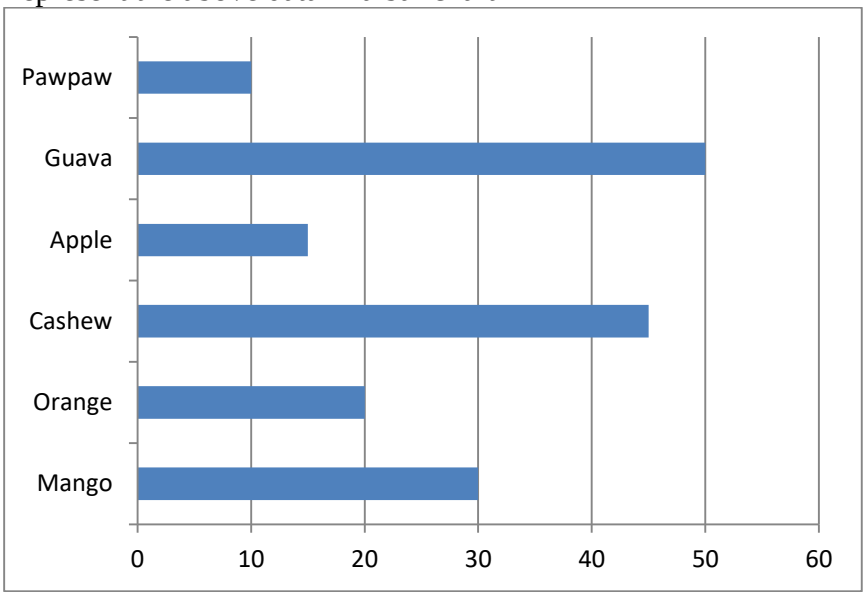
Illustration

Consider the data below obtained from the Federal Government on the Palliative fruits supplied to the citizen of Abuja during the Covid 19 lockdown to ameliorate hunger.

S/N	Fruits	Quantity
1	Mango	30
2	Orange	20

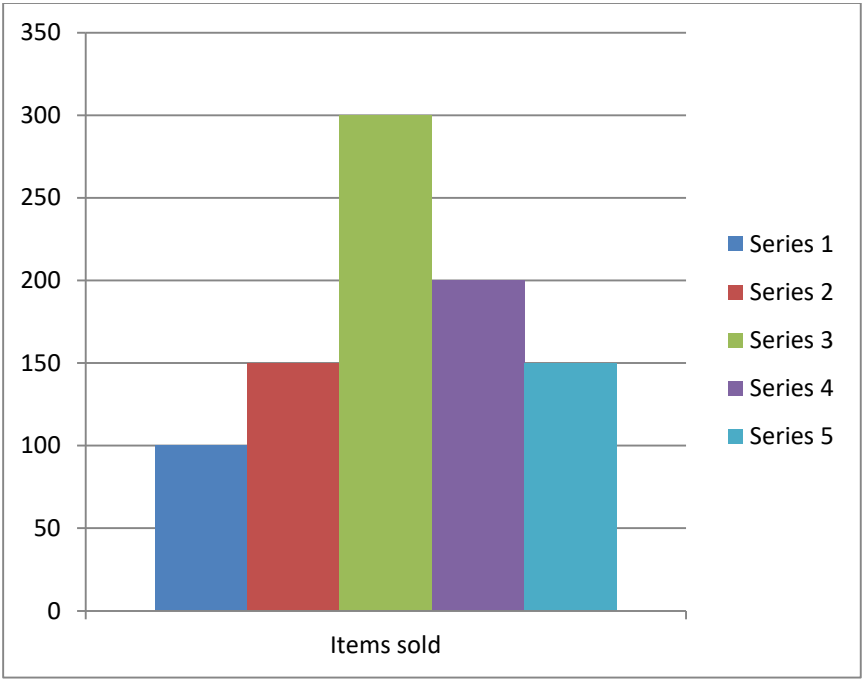
3	Cashew	45
4	Apple	15
5	Guava	50
6	Pawpaw	10
		Total = 170

Represent the above data in a bar chart



Histogram

Histogram is a series of rectangles with bases equal to the interval ranges and areas proportional to the frequencies. The polygon in ahistogram is drawn by connecting with straight lines the interval midpoints of a cumulative frequency histogram.



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:

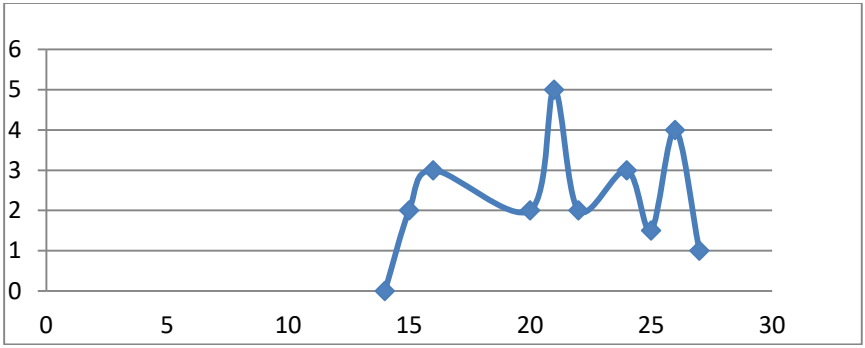


Fig. 1: Frequency Polygon of ungrouped data

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 6: Marks (Midpoints) of a Grouped Data

Classes	F	Cf	Class Marks (x)
55-59	1	50	51
50-54	1	49	52
45-49	3	48	47
40-44	4	45	42
35-39	6	41	37
30-34	7	35	32
25 - 29	12	28	27
20-24	6	16	22
15-19	7	10	17
10-14	3	3	12

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.

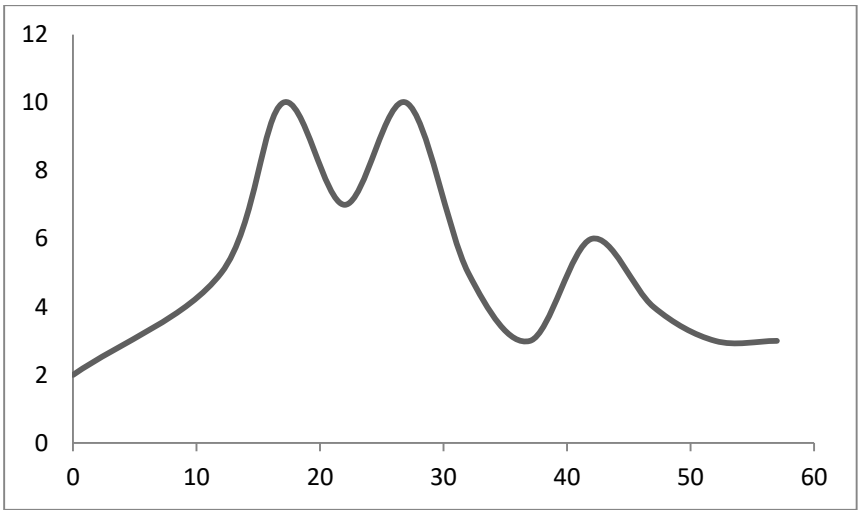


Fig. 2: Frequency Polygon of a Grouped Data

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, themedian, 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 orright, 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.

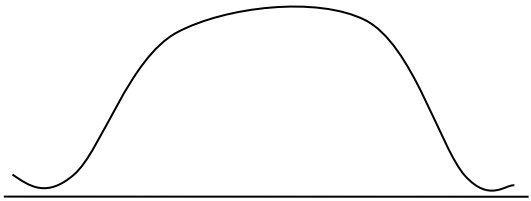
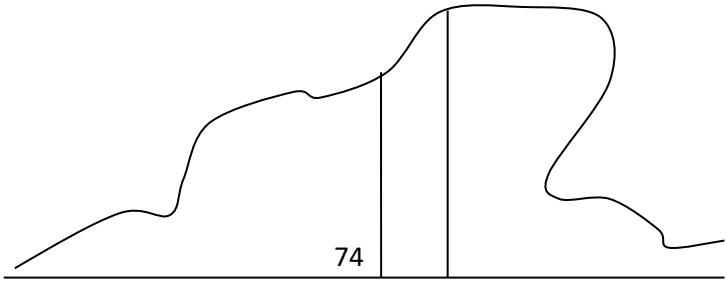
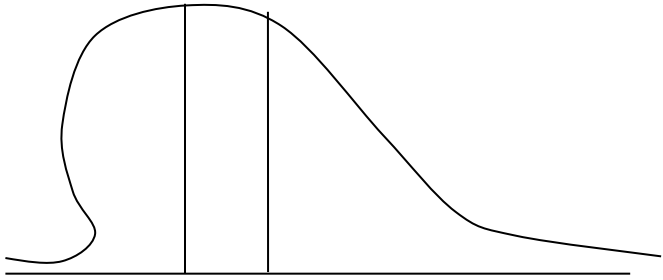


Fig. 3: Showinga dumb bell shape (symmetrical) distribution



Mean Median

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:

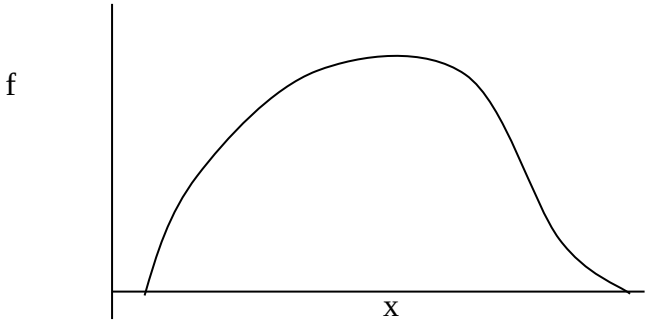


Fig. 6: A Mesokurtic Distribution (Normal Curve)

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.

CHAPTER 7

MEASUREMENT OF CENTRAL TENDENCY

After data have been collected and tabulated, analysis begins with the calculation of a single number, which will summarize or represent all the data. Because data often exhibit a cluster or central point, this number is called a measure of central tendency.

Let $x_1, x_2, \ldots, x_n$ be the n tabulated (but ungrouped) numbers of some statistic; the most frequently used measure is the simple arithmetic average, or mean, written □, which is the sum of the numbers divided by *n:*

If the x's are grouped into k intervals, with midpoints $m_1, m_2, \ldots, m_k$ and frequencies $f_1, f_2, \ldots, f_k$, respectively, the simple arithmetic average is given by
with$i = 1, 2, \ldots, k$.

The median and the mode are two other measures of central tendency. Let the x's be arranged in numerical order; if n is odd, the median is the middle x; if n is even, the median is the average of the two middle x's. The mode is the x that occurs most frequently. If two or more distinct x's occur with equal frequencies, but none with greater frequency, the set of x's may be said not to have a mode or to be bimodal, with modes at the two most frequent x's, or trimodal, with modes at the three most frequent x's.

For the purpose of this course we shall consider the following measures of central tendencies:Arithmetic Mean, Median and Mode.

Arithemetic Mean

Iwuji (2014) and Nwana (2005) defined arithmetic mean as the average or value obtained by adding up all the values and dividing by the total of scores. This is often denoted by X or M

Guidelines for the calculation of arithmetic mean:
If the data given is less than 30 we can make use of raw score.

i. Organize the scores into your frequency distribution table
ii. Multiply each score by its frequency
iii. Fine the sum of the product ∑fx then apply the formula

Example 1

Calculate the mean of the following data
1, 2, 3, 4, 5.

Mean = $\frac{\Sigma fx}{\Sigma f}$

Solution

$\frac{1+2+3+4+5}{5} = \frac{15}{5}$

Mean (X) = 3

Computation of mean from group data

When you have large data to deal with, the essential thing to do is to group the scores into class interval in order to reduce the number to a manageable size. There is no much difference in the calculation of group mean; the only difference is that the class mark must be calculated in other to obtain the score which could in other words be called the 'X'.

Example 1

Consider the data below obtained from the sale day book of Mr. Okafor who operates a 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.

From the above data, calculate the mean score of the products of Mr. Okafor.

Solution

Table 6. Frequency Distribution Table

Class interval	Class boundary	Class mark (X)	Tally (T)	Frequency (F)	FX
0-2	-0.5-2.5	1	~~*IIII*~~ *I*	6	6
3-5	2.5-5.5	4	~~*IIII*~~ *III*	8	32
6-8	5.5-8.5	7	~~*IIII*~~ *III*	8	56
9-11	8.5-11.5	10	~~*IIII*~~ *II*	7	70
12-14	11.5-14.5	13	~~*IIII*~~ *III*	8	104
15-17	14.5-17.5	16	~~*IIII*~~ *I*	6	96
18-20	17.5-20.5	19	~~*IIII*~~ *II*	7	133
21-23	20.5-23.5	22	*IIII*	4	88
24-26	23.5-26.5	25	*III*	3	75
27-29	26.5-27.5	28	*III*	3	84
Total				∑F =60	∑FX =744

Mean = $\frac{\sum fx}{\sum f}$

Where ∑FX =744

∑F = 60

Mean (X)= $\frac{744}{60}$

Mean (X) =12.4

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

$= 46.24$

Calculation of mean using the assume mean

This method involves assuming or guessing a mean for the distribution.

The computation is in the following form.

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.

Solution:

Class Interval	**Class Boundary**	**Class Mark (X**	**Tally (T)**	**Freq (F)**	**X-X (D)**	**FD**
0-2	-0.5-2.5	1	~~*III*~~ *I*	6	-7	-42
3-5	2.5-5.5	4	~~*III*~~ *III*	8	-4	-32
6-8	5.5-8.5	7	~~*III*~~ *III*	8	-1	-8
9-11	8.5-11.5	10	~~*III*~~ *II*	7	2	14
12-14	11.5-14.5	13	~~*III*~~ *III*	8	5	40
15-17	14.5-17.5	16	~~*III*~~ *I*	6	8	48
18-20	17.5-20.5	19	~~*III*~~ *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				∑F =60		∑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

Median (mathematics) is the value of the middle member of a set of numbers when they are arranged in order. Like the mean (or average) and mode of a set of numbers, the median can be used to get an idea of the distribution or spread of values within a set when examining every value individually would be overwhelming or tedious. The median of the set {1, 3, 7, 8, 9}, for example, is 7, because 7 is the member of the set that has an equal number of members on each side of it when the members are arranged from lowest to highest. If a set contains an even number of values, there is no single middle member. In such cases the median is the mean of the two values closest to the middle. The median of the set {1, 3, 9, 10}, for example, is (3 + 9)/2 = 6.

Consider the table below, calculate the mean using the assume mean formula and compare your answer to the previous calculation

$$\overline{X}_a = A + \left(\frac{\Sigma fd}{N}\right)$$

$$\overline{X}_a = 40$$

C.I	8-20	21-33	34-46	47-59	60-72	73-85	86-98
Freq	10	11	7	4	8	5	5

C. I	Midpoint (X)	Frequency	d ($x-\overline{x}_a$)	fd
8 – 20	14	10	-26	-260
21 – 33	27	11	-13	-143
34 – 46	40	7	0	0
47 – 59	53	4	13	52
60 - 72	66	8	26	208
73 - 85	79	5	39	195
86 - 98	92	5	52	260
		50		$\Sigma fd = 312$

$\overline{X}_a = 40 + \left(\frac{312}{50}\right)$

$40 + (6.24)$

$40 + 6.24$

$= 46.24$

Characteristics of the Mean

1. It is a high level statistics (it applies to interval and ratio scales)
2. The mean suffers from the problem of absolute extremism in value.
3. The mean is most stable of all the averages and most representative of the population.
4. The mean serve as basics for other statistical decisions.

5. The sum of the deviation from the mean is zero. This means that if the mean is subtracted from each score and the deviation are added, the result will be zero.
6. The mean makes use of every value in the distribution. A change in value of any score affects the mean.

Calculation of Median of Ungroup Data

For odd number of cases, the median is just the number at the middle of the distribution arranged in the order of magnitude.

In situation where even numbers are involved, the median if found as the average of the two scores occupying the midpoint.

Example

Consider for instance, a business man who procures goods and the frequency of the procurement is given as follows: 5, 6, 8, 9, 7, 4, 8, 6, 10, 9, 5, 6, 8.

Solution

4, 5, 5, 6, 6, 6, (7), 8, 8, 8, 9, 9, 10.

From the arrangement, the middle number is 7, therefore the median is 7.

Example 2

2,1,3,3,7,6,7,3.

Solution

1, 2, 3, 3, 3, 6, 7, 7.

In the above distribution, the median becomes

$\frac{3+3}{2} = \frac{6}{2}$

Median = 3.

Computation of Median from Group data

The formula for the calculation of media is

$$\text{Median} = L_1 + \frac{(\frac{N}{2} - \frac{CFb}{1})I}{f}$$

Where Lm = the lower class mark of the class containing the median class.

i = the width of the interval

F = cumulative freq. of the class above the median class.

f = Freq. of the median class

N = total number of scores in the distribution

Example

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 Eight

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

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

$37.5 + (5/12) \times 3$

$37.5 + (15/12)$

$37.5 + 1.25$

Median = 38.75.

Example 2

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

Mode

This is another measure of central tendency; it is the number that has the highest occurrence in any given data.

For example

Given the following data 1,2,3,4,2,5,2,5,2,4,2,3,2.
From the above data, determine the mode of the distribution.

Solution
1,2,2,2,2,2,2,3,3,4,4,5,5
2 have the highest occurrence. Therefore, 2 is the mode of the distribution.
The formula below can be use for calculating the mode of a group data.

$$M_o = L_1 + \left[\frac{D1}{D1+D2}\right]I$$

Where: Li= Lower class limit of the modal class
Di= Excess of the modal frequency over the freq of next lower class.
D_2 = Excess of the modal freq over freq of next higher class.
C= Size of class interval
From the above

Calculating the mode of a group 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	∑ 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 = $33.5 + \left(\frac{6}{6+13}\right)13$

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

$33.5 + (0.32)13$

$33.5 + 4.16 \Rightarrow 37.66$

CHAPTER 8

MEASUREMENT OF RELATIONSHIP/ASSOCIATION/CORRELATION

These are measures used to ascertain the extent or degree of relationship and direction between two or more variables. There are different types of relationships which may exist between variables.

i. Linear Relationship

This refers to a straight line relationship between variables where variable x may increase as variable y decreases or variable x and y may go on the same direction.

ii. Curve Linear Relationship

This refers to a relationship between two variables that are curved.

We determine relationship among things that relate. Different measures have been developed to analyze data collected on relationship existing between variables. The index of this measure is called coefficient.

Correlation coefficient ranges from positive (+1) through zero to negative (-1)

+1 perfect relationship (positive)

Scatter graph or scatter gram is the graph of correlation.

Uses of the Measures of Relationship

i. It is used to predict purpose
ii. It is used to determine progression
iii. It is used to estimate the reliability of an instrument
iv. It is used in the determination of relationship between variables

The major correlation coefficient was designed by Karl Pearson; it is used for interval and ratio coefficient represented by

Pearson correlation

Pearson r = $\frac{N\Sigma xy - \Sigma x\Sigma y}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$

Example 1:

Consider the table below being the data obtained from the scores of female and male students in mathematics' test:

X	6	5	8	6	7	5
Y	5	7	9	4	3	1

Solution:

S/N	X	Y	XY	X^2	Y^2
1	6	5	30	36	25
2	5	7	35	25	49
3	8	9	72	64	81
4	6	4	24	36	16
5	7	3	21	49	9
6	5	1	5	25	1
	35	29	187	235	181

Pearson r = $\frac{N\Sigma xy - \Sigma x\Sigma y}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$

r = $\frac{6(187) - (35 \times 39)}{\sqrt{[6(235) - (35)^2][6(181) - (29)^2]}}$

r = $\frac{1122 - 1073}{\sqrt{(1410 - 1369)(1086 - 84)}}$

r = $\frac{49}{\sqrt{(41)(245)}}$

r = $\frac{49}{\sqrt{10045}}$

r = $\frac{49}{100.224747}$

r = 0.49 The relationship is a positively weak one.

Example 2:

Consider the table below, determine and interprete the correlation index.

X	5	3	2	4	3
y	9	4	7	8	6

Solution

	x	Y	x2	y2	xy
	5	9	25	81	45
	3	4	9	16	12
	2	7	4	49	14
	4	8	16	64	32
	3	6	9	36	18
Total	**17**	**34**	**63**	**246**	**121**

$\Sigma x^2 = 63$

$(\Sigma x)^2 = 289$

$\Sigma y^2 = 246$

$(\Sigma y)^2 = 1156$

N = 5

$$r = \frac{5(121) - 17 \times 34}{\sqrt{[5(63) - 17^2)][5(246) - (34)^2]}}$$

$$r = \frac{605 - 578}{\sqrt{(315 - 289)(1230 - 1156)}}$$

$$r = \frac{27}{\sqrt{(26)(74)}}$$

$$r = \frac{27}{\sqrt{1924}} = \frac{27}{43.86}$$

$$r = 0.616$$

Example 3

The table below was extracted from the scores of student taught in the urban and rural area after the same test has been administered.

Urban (X)	20	60	85	72	45	50	32
Rural (Y)	33	52	62	40	81	30	71

Using Pearson's correlation, is there any relationship between the scores.

Solution

X	**Y**	**X^2**	**Y^2**	**XY**
20	33	400	1089	660
60	52	3600	2704	3120
85	62	7225	3844	5270
72	40	5184	1600	2880
45	81	2025	8561	3645
50	30	2500	900	1500
32	71	1024	5041	2272
364	**369**	**21958**	**23739**	**19347**

$\Sigma X = 364$, $\Sigma Y = 369$, $\Sigma X^2 = 21958$, $\Sigma Y^2 = 23730$, $\Sigma XY = 19347$, $N = 7$

$$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[N(\Sigma X^2) - (\Sigma X)^2][N(\Sigma Y^2) - (\Sigma Y)^2]}}$$

$$r = \frac{7(19347) - (364 \times 369)}{\sqrt{[7(21958) - (364)^2][7(23739) - (369)^2]}}$$

$$r = \frac{135429 - 134316}{\sqrt{(153706 - 132496)(166173 - 136161)}}$$

$$r = \frac{1113}{\sqrt{(21210)(30012)}}$$

$$r = \frac{1113}{\sqrt{636554520}}$$

$$r = \frac{1113}{25230.03}$$

$r = 0.0441$

$r = \approx 0.04$

The interpretation is that the relationship is positive but very low.

Spearman's Ranking Order Correlation

$$rho = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

X	40	50	60	40	58	55	59	20
Y	65	80	80	85	70	82	65	50

The data above are scores obtained from Jeba International College Wukari, showing the scores of male and female students in an examination. Using Spearman's ranking order correlation, ascertain the correlation index.

Solution:

$$rho = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

S/N	x	y	Rx	Rx	D	d^2
1	40	65	6.5	6.5	0	0
2	50	80	5	3.5	1.5	2.25
3	60	80	1	3.5	-2.5	6.25
4	40	85	6.5	1	5.5	30.25
5	58	70	3	5	-2	4
6	55	82	4	2	2	4
7	59	65	2	6.5	4.5	20.25
8	20	50	8	8	0	0
						67

$$rho = 1 - \frac{6(67)}{8(8^2-1)}$$

$= 1 - \frac{402}{8\times63}$

$= 1 - \frac{402}{504}$

$= 1 - 0.798$

$= 0.20$

Example 2:

Consider the scores below generated from the scores of test administered to students taught in the rural and urban area.

Urban (X)	20	60	85	72	45	50	32
Rural (Y)	33	52	62	40	81	30	71

X	**Y**	**R_X**	**R_y**	**d**	**d^2**
20	33	7	7	0	0
60	52	3	4	-1	1
85	62	1	3	-2	4
72	40	2	5	-3	9
45	81	5	1	4	16
50	30	4	6	-2	4
32	71	6	2	4	16
					Total = 50

$rh_o = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$

$rh_{o = 1 -} \frac{6\times50}{7(7^2-1)}$

$rh_{o = 1 -} \frac{300}{7(49-1)}$

$rh_o = 1 - \frac{300}{7(48)}$

$rh_o = 1 - \frac{300}{336}$

$rh_o = 1 - 0.8929$

$rh_o = 0.12$

The relationship is positive but very low

CHPATER 9

MEASUREMENT OF VARIABILITY

This is employed when the true picture of a data is given, to show how the scores are placed or demarcated even if the measurement has been taken on the central tendency.

Range

This is the difference between the highest number and the lowest number.

Ungrouped data Hs – highest score

Ls – Lowest score

Grouped data

UcLHs – LcLLs

UcLHs = upper class limit highest score

LcLLs = lower class limit lowest score

Or

Hs – Ls + 1

Range is a nominal statistics. Range has the challenge of extremism.

Quartile statistics is an ordinal statistics.

Variance is defines as mean of square deviation.

S^2 or $S^2 = (\frac{x-\bar{x}}{N})^2$ (without frequency)

$\frac{\sum f(x-\overline{x})2}{N}$ (with frequency)

Standard deviation is the square root of variance $\sqrt{\frac{\sum(x-\bar{x})2}{N}}$

Calculation of standard deviation using the raw score method

$f^2 = \frac{\sum FX^2}{N} - \left(\frac{\sum FX}{N}\right)^2$

Consider the table below

Table 15.

Classes	**F**	**X**	**FX**	**FX^2**
82-90	2	86	172	14792
73-81	6	77	462	35574
64-72	10	68	680	46240
55-63	12	59	608	35872
46-54	18	50	900	45000
37-45	6	41	246	10086
28-36	8	32	252	8192
19-27	4	23	92	2116
10-18	2	14	28	392
	68		3544	204164

$$\delta^2 = \frac{\sum FX^2}{N} - \left(\frac{\sum FX}{N}\right)^2$$

$$\delta^2 = \frac{204164}{68} - \left(\frac{3544}{68}\right)^2$$

$3002.41176 – (52.1176471)^2$

3002.41176 – 2716.24914

= 286.16262

s.d = $\sqrt{286.166}$

s.d = 16.92

Example 2

C. I	5 - 15	16 - 26	27 - 37	38 - 48	49 - 59	60 - 70
Frequency	2	3	4	7	5	8

Solution

C. I	Frequency	X	FX	FX^2
5 – 15	2	10	20	200
16 – 26	3	21	63	1323
27 – 37	4	32	128	4096
38 – 48	7	43	301	12943
49 – 59	5	54	270	14580
60 - 70	8	65	520	33800
	Total = 29		Total = 1302	= 66942

$S^2 = \frac{\Sigma fx^2}{N} - \left(\frac{\Sigma fx}{N}\right)^2$

$S^2 = \frac{66942}{29} - \left(\frac{1302}{29}\right)^2$

$S^2 = 2308.34 - (44.90)^2$

$S^2 = 2308.34 - 2016.01$

$S^2 = 292.33$

S. D = $\sqrt{292.33}$

S. D = 17.10

Standard deviation using arbitrary origin

C.I	F	X	d	Fd	Fd^2
82-90	2	86	3	9	18
73-81	6	77	2	12	24
64-72	10	68	1	10	10
55-63	12	59	0	0	0
46-54	18	50	-1	-18	18
37-45	6	41	-2	-12	24
28-36	8	32	-3	-24	72
19-27	4	23	-4	-16	64
10-18	2	14	-5	-10	50
				-52	280

$$\delta^2 = c\sqrt{\frac{\sum Fd^2}{N} - \left(\frac{\sum Fd}{N}\right)^2}$$

$$\delta^2 = 9\sqrt{\frac{280}{68} - \left(\frac{-52}{68}\right)^2}$$

$$9\sqrt{4.12 - 0.58}$$

$$9\sqrt{3.54}$$

$$9 \text{ x } 1.88$$

$$= 16.93$$

Calculation of standard deviation using mean deviation method

Consider the data below

Table 16

X	F	FX	$X - \bar{X}$	$(X - \bar{X})^2$	$F(X - \bar{X})^2$
10	2	10	4.2	17.64	35.28
8	5	40	2.2	4.84	24.2
6	12	72	0.2	0.04	0.48
4	4	16	-1.8	3.24	12.96
2	4	8	-3.8	14.44	37.76
	27	156			130.68

Variance $\frac{f(x-\bar{x})^2}{N}$

$= \frac{130.68}{27}$

$= 4.84$

$\delta^2 = \sqrt{4.84}$

$\delta^2 = 2.2$

Example 2

Consider the table below and determine the standard deviation.

C. I	5 - 15	16 - 26	27 - 37	38 - 48	49 - 59	60 - 70
Frequency	2	3	4	7	5	8

Solution

C. I	Frequency	X	FX	X - X	$(X-X)^2$	F$(X-X)^2$
5-15	2	10	20	-349	1218.01	2436.02
16-26	3	21	63	-23.9	571.21	1713.63
27-37	4	32	128	-12.9	166.41	665.64
38-48	7	43	301	-1.9	3.61	26.27
49-59	5	54	270	9.1	82.81	414.05
60-70	8	65	520	20.1	404.01	3232.08
	Total = 29		Total = 1302			Total = 8487.69

Step 2. Determine the mean (x)

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

$= \frac{1302}{29}$

= 44.90

$S^2 = \frac{\Sigma f(x-x)^2}{\Sigma f}$

$S^2 = \frac{8487.69}{29}$

S^2= 292.68

S. D = $\sqrt{292.68}$

S. D = 17.12

CHAPTER 10

INFERENTIAL STATISTICS

HypothesisTesting

This is the process of determining the truthfulness or otherwise of the Hypothesis through collection and analysis of data.

Possible decisions a researcher might make during testing of hypothesis are:

i. Rejecting a true null hypothesis X (Type 1 error)
ii. Accept a false null hypothesis X (Type II error)
iii. Reject a false null hypothesis
iv. Accept a true null hypothesis

Type one error = α (Alpha)/confidence level

Type two error = β (beta)

Level of significance refers to the amount of risks a researcher gives to himself as problem of committing type one error.

Degree of freedom is the amount of freedom numbers have to vary in a distribution. Normally, (N - 1) in a table that has column and row (R -1)(C - 1)

Parametric statistics is the statistics obtained when data is collected from normal distribution, or if the data is arranged in interval or ratio scale.

Non parametric is obtained when data is not collected from normal distribution normally nominal or ordinal scale.

T-TEST AND Z-TEST

This is an inferential statistics that compare samples when decision is to be taken about difference or relationship between two means.

T-Test is applied when you have two samples such data must be interval or ratio.

The data must have been collected from a normal distribution. T-Test is applicable when the sample is small.

$T = \frac{\bar{x}1 - \bar{x}2}{\sqrt{\frac{s2}{n1} + \frac{s2}{n2}}}$

Df = n1 + n2 – 2

A researcher is interested in finding whether gender is a relevant factor in student achievement in English language and obtained the following data.

N1=61

N2=55

$\bar{x}1=66$

$\bar{x}2=59$

$S_1^2 = 2.46$

$S_2^2 = 1.3$

5% level of sig, the mean performance if male and female in the test is just by chance.

There is no significant relationship in the achievement of male and female students in the test.

$$\frac{66 - 59}{\sqrt{\frac{(2.46)^2}{61} + \frac{(1.3)^2}{55}}}$$

$$\frac{7}{\sqrt{\frac{6.0516}{61} + \frac{1.69}{55}}}$$

$$\frac{7}{\sqrt{0.099+0.0307}} = \frac{7}{\sqrt{0.129722}} = \frac{7}{0.360176} = 19.4349$$

Df = n1 + n2 – 2

= 61 + 55 – 2 =114

$\alpha = 0.05$

ab = 1.960

When your sample is too large and cannot be located in the table, you are advised to take your degree of freedom, take your figure from ∞.
From the above value obtained, the Tcal>Ttab. We therefore reject the Null hypothesis and accept the alternative that $\overline{x_1} \neq \overline{x_2}$
T-test for small sample

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sum(\overline{X_1} - \overline{X_1})^2 + \sum(\overline{X_2} - \overline{X_2})^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Example 1
Two groups of students were taught shorthand with two different methods. The data collected were as follows:

Group I	Group II
19	17
14	16
12	15
16	12
17	10
10	11
13	12
11	14
14	11
12	17
20	13

Test the hypothesis that the two methods are equally effective.
H_o: There is no significant difference between the two methods
i.e $H_o = \overline{X_1} = \overline{X_2}$
$H_1 = \overline{X_1} \neq \overline{X_2}$

Table 17.

Group 1 (x_1)	$(X_1 - \overline{\overline{X}})$	$(X - \bar{X})^2$	Group 2 (x_2)	$(X_2 - \bar{\bar{X}}_2)$	$(X_2 - \overline{X_2})^2$
19	4.64	21.5296	17	3.5500	12.6025
14	-0.36	0.1296	16	2.5500	6.5025
12	-2.36	5.5696	15	1.5500	2.4025
16	-1.64	2.6896	12	-1.4500	2.1025
17	-2.64	6.9696	10	-3.4500	11.9025
10	-4.36	19.0096	11	-2.4500	6.0025
13	-1.36	1.8496	12	-1.4500	2.1025
11	-3.36	11.2896	14	0.550	0.3025
14	-0.36	0.1296	11	-2.4500	6.0025
12	-2.36	5.5696	17	3.5500	12.6025
20	-5.64	31.8096	13	-0.4500	0.2025
158		106.5456	148		

$\bar{X}_1 = \frac{158}{11} = 14.36$

$\bar{X}_2 = \frac{148}{11} = 13.45$

Putting the data into formula

$$\frac{14.36 - 13.45}{\sqrt{\frac{106.545+62.7275}{11+-2}\left(\frac{1}{11}+\frac{1}{11}\right)}}$$

$$= \frac{0.91}{\sqrt{\frac{169.2731}{20}(0.1818)}}$$

$$= \frac{0.91}{\sqrt{8.463655}(0.1818)}$$

$t = \frac{0.91}{\sqrt{1.5387}}$

$t = \frac{0.91}{1.2404}$

t = 0.7336

t≅ 0.734

Degree of freedom

$n_1 + n_2 - 2$

= 11 + 11 – 2

= 20

(a) α = 0.05, df = 20

ttab = 2.086

Recall: the decision rule says reject the H_o when the Tcal > Ttab

The Tcal = 0.734

Ttab = 2.086

On this note, the tab is greater than the tabulated; we therefore accept the null hypothesis that there is no significant difference in the twp method used for teaching the students.

When testing r, we use the small formula t which is $t = r\sqrt{\frac{n-2}{1-r^2}}$

r = 0.68

n = 2

Test the hypothesis that

Example 2

t- test for linear relationship

$t = \sqrt{\frac{n-2}{1-r^2}}$

r = 0.68

n = 2

$0.68\sqrt{21-2}$

Example 2

The scores below were generated from a test administered in Degree year 2 Biology department, Peaceland College of Education, Enugu.

Test the hypothesis at 0.05 alpha that the two methods are effective.

Group 1	60	38	45	70	25
Group 2	66	41	22	36	75

H_O: the two methods used are effective

Group x_1	$x_1 - \bar{x}_1$	$(x_1 - \bar{x}_1)^2$	Group x_2	$x_2 - \bar{x}_2$	$(x_2 - \bar{x}_2)^2$
60	12.4	153.76	66	18	324
38	-9.6	92.16	41	-7	49
45	-2.6	6.76	22	-26	676
70	22.4	501.76	36	-12	144
25	-22.6	510.76	75	27	729
$\Sigma x_1 = 238$ $\bar{x}_1 = 47.6$		1265.2	$\Sigma x_2 = 240$ $\bar{x}_2 = 48$		1922

Mean $(\bar{x}_1)$

$\bar{x}_1 = \frac{\Sigma fx}{N}$

$\bar{x}_1 = \frac{238}{5}$

$\bar{x}_1 = 47.6$

Mean $(\bar{x}_2)$

$\bar{x}_2 = \frac{\Sigma fx}{N}$

$\bar{x}_2 = \frac{240}{5}$

$\bar{x}_2 = 48$

$S.D_1 = \sqrt{\frac{1265.2}{5}}$

$S.D_1 = \sqrt{253.04}$

$S.D_1 = 15.90$

$S.D_2 = \sqrt{\frac{1922}{5}}$

$S.D_2 = \sqrt{384.1}$

$S.D_2 = 19.59$

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$T = \frac{47.6 - 48.0}{\sqrt{\frac{15.9}{5} + \frac{19.59}{5}}}$$

$$T = \frac{-0.4}{\sqrt{3.18 + 3.918}}$$

$$T = \frac{-0.4}{\sqrt{7.098}}$$

$$T = \frac{-0.4}{2.66}$$

$T = -0.15$

$df = n_1 + n_2 - 2$

$df = 5 + 5 - 2$

$df = 8$

$\alpha = 0.05$

Tcrit = 1.960

Tcal = -0.15

In this situation, we do not reject the null hypothesis since the Tcrit is greater than Tcal. It is concluded therefore that the two methods used are effective

CHAPTER 11

NON PARAMETRIC STATISTICS

Chi-Squared Test

A **chi-squared test**, also written as χ^2 **test**, is any statisticalhypothesis test wherein the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. Without other qualification, 'chi-squared test' often is used as short for Pearson's chi-squared test.

Chi-squared tests are often constructed from a sum of squared errors, or through the sample variance. Test statistics that follow a chi-squared distribution arise from an assumption of independent normally distributed data, which is valid in many cases due to the central limit theorem. A chi-squared test can be used to attempt rejection of the null hypothesis that the data are independent.

Also considered a chi-squared test is a test in which this is asymptotically true, meaning that the sampling distribution (if the null hypothesis is true) can be made to approximate a chi-squared distribution as closely as desired by making the sample size large enough. The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories.

Conditions for the use of chi-square

For Chi-square statistics to be applied, the following conditions must be met:

- The data must be at the nominal level
- The observations must be independent
- When the assumptions of parametric statistics are greatly violated.

Procedure for computing Chi-square

i. Note the number of cases in each category normally called the observed frequencies
ii. Following a line of speculation, a theory or hypothesis, determine the number of cases that should be in each category (expected frequency) assuming that the hypothesis, theory or speculation is true
iii. Find how much, each observed frequencies deviates from the expected frequencies
iv. Sum the deviations
v. Ordinarily, the observed frequency will deviate from the expected frequency by chance. If however, the difference is much, that cannot be attributed to chance, the hypothesis , theory or speculation upon which the expected frequency is generated will be rejected.

Ways of using the Chi-square

The Chi-square applied in the number of ways. Two popular ways of using it are:

1. Test of goodness of fit
2. Test of independence.

Goodness of fit test

This may be called one dimensional chi-square or single nominal variable test. In this situation the hypothesis is stated thus:
H_o: There is no significant difference between the observed and the theoretical frequencies.

Example

Supposed there is a government policy about admission which says 60% for science and 40% should be for the non-science.

	Science	Non-science
Observed Frequency	457	543
Expected frequency	600	400

Admission quota
Assume the following Quota:
Merit----------40%
Catchment area----------30%
Education disadvantaged area-----20%
Undefined ----------10%
The Vice chancellor admitted 500 students and the following data was given:

	Merit	Catchment area	Education disadvantaged	Undefined
Observed Freq	180	193	70	57
Expected Freq	200	150	100	50

Example chi-squared test for categorical data
Suppose there is a country of 1 million residents with four neighborhoods: *A*, *B*, *C*, and *D*. A random sample of 650 residents of the city is taken and their occupation is recorded as "white collar", "blue collar", or "no collar". The null hypothesis is that each person's neighborhood of residence is independent of the person's occupational classification. The data are tabulated as:

	A	**B**	**C**	**D**	**TOTAL**
White collar	90	60	104	95	349
Blue collar	30	50	51	20	151
No collar	30	40	45	35	150
Total	150	150	200	150	650

Required: test the hypothesis and make decision either to reject or accept the null.

Solution

Step 1: State the hypothesis

There is no significant difference the ways the students are distributed and the way they were expected to be distributed.

Step 2: Determine the expected frequencies

Step 3: Chose an alpha level

Whole 1 million people live in neighborhood *A*. Similarly we take 349/650 to estimate what proportion of the 1 million people are white-collar workers. By the assumption of independence under the hypothesis we should "expect" the number of white-collar workers in neighborhood *A* to be. The sum of these quantities over all of the cells is the test statistic. Under the null hypothesis, it has approximately a chi-squared distribution whose number of degrees of freedom are

150× 349/650

If the test statistic is improbably large according to that chi-squared distribution, then one rejects the null hypothesis of **independence**.

Chi-Square for Test Of Independence

This is also called two dimensional Chi-square or Chi-square for contingency term. This is used when we have two nominal variables and we want to know if one variable depends on the other.

Procedure for computing Chi-square under this condition:

i. State the hypothesis
ii. Chose an alpha level
iii. Determine the degree of freedom
iv. Obtain the critical value from the table
v. Compute the chi-square

Example 1

A group of people were asked to assess the performance of the President. Their responses are recorded in the table below:

	Very Good	Good	Poor	Total
Male	46	88	56	190
Female	63	52	35	150
Total	109	140	91	340

Solution

Step 1: State the Hypothesis

H_o: There is no significant difference between the rating of the President by male and female in the group

Step 2:

Alpha level 0.05 (5%)

Step 3: Determine the degree of freedom

Step 4: Obtain the critical value from the table

To determine the degree of freedom, using (R-1) (c-1) =(2-1) (3-1)=2

Prepare the contingent table.

	Very Good	Good	Poor	Total
Male	46 (60.91)	88(78.24)	56(50.85)	190
Female	63(48.09)	52(61.76)	35(40.15)	150
Total	109	140	91	340

The figure in the brackets is known as the expected frequency. It was obtained using this formula

$$E = \frac{Rat}{T}$$

Where R= Total number of roles

T= Total Number of column

$$X^2 = \frac{\sum(e-o)^2}{e}$$

Step 5: Compute the Chi-square

O	E	o-e	$(o-e)^2$	$\frac{(o-e)^2}{e}$
46	60.91	-14.91	222.159	3.65
88	78.24	9.76	95.26	1.22
56	50.85	5.15	26.52	0.52
63	48.09	14.91	222.31	4.62
52	61.76	-9.76	95.26	1.54
35	40.15	-5.15	26.52	0.66
				$\sum\frac{(e-o)^2}{e}$= 12.21

Do= 2
Alpha = 0.05

X^2tab= 5.991
X^2cal= 12.21
We reject the Null hypothesis since the calculated value of chi-square is greater than the critical value.

Example 2
Over the Christmas period, the traditional ruler of Udi Community decided to give out 60 scholarships. And he has invited you for your services as a 21st century scholar to help in distributing the scholarships.

	HSE	ASE	LES	Total
Male	3	7	10	20
Female	17	13	10	40
Total	20	20	20	60

Step 1: State your null hypothesis
H_o: there is no significant difference between the distribution of the scholarships.

Step 2: Make your decision rule

Reject the null hypothesis

Reject H_o if the χ^2 calculated is greater than the χ^2 critical

Step 3: Calculate the expected frequency

ef = $\frac{Rt \times G}{Tot}$

1.
2. $\frac{20 \times 20}{60} = 6.66$
3. $\frac{20 \times 20}{60} = 6.66$
4. $\frac{20 \times 20}{60} = 6.66$
5. $\frac{40 \times 20}{60} = 13.33$
6. $\frac{40 \times 20}{60} = 13.33$
7. $\frac{40 \times 20}{60} = 13.33$

Step 4: Calculate the degree of freedom

df = (R-1) (C-1)

= (2-1)(3-1)

= 2

Step 5: choose an alpha level

$\alpha = 0.05$

Step 6: Calculate

$\chi^2 = \Sigma\left(\frac{(of-ef)^2}{ef}\right)$

Make another table

of	Ef	of-ef	$(of-ef)^2$	$\left(\frac{(of-ef)^2}{ef}\right)$
3	6.66	-3.66	13.40	2.01
7	6.66	0.34	6012	0.02
10	6.66	3.34	11.16	1.68
17	13.33	3.67	13.47	1.01

13	13.33	-0.33	0.11	0.01
10	13.33	-3.38	11.09	0.83
				5.56

$\chi^2 = \Sigma\left(\frac{(of-ef)^2}{ef}\right) = 5.56$

We do not reject the null hypothesis.

CHAPTER 12

RESEARCH INSTRUMENT

Research Instruments are measurement tools (for example, questionnaires or scales) designed to obtain data on a topic of interest from research subjects**.**It helps the researcher to obtain, measure, and analyze data from subjects around the research topic. You need to decide the instrument to use based on the type of study you are conducting: quantitative, qualitative, or mixed-method. For instance, for a quantitative study, you may decide to use a questionnaire, and for a qualitative study, you may choose to use a scale. While it helps to use an established instrument, as its efficacy is already established, you may if needed use a new instrument or even create your own instrument.

Characteristics of a good research 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 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

Types of Data Collection Instruments

Questionnaire

Questionnaire is the best If you need to collect data from a large number of people, then you would need to consider using

questionnaires. 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 Research 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
- The nature of the population
- The design of the study
- The expected tool for data analysis

The purposes of the study

The researcher should state in general and specific terms what the purpose of his/her search/study is. Based on that he/she can formulate the research questions and/or hypotheses, which therefore, guides the construction of the instrument. The instrument, as a matter of fact, must be based on the purpose of the study and the research questions.

The nature of the population

Another issue of consideration is the nature of population. A researcher must understand his population very clearly before ever developing any type of instrument/test for them. The nature includes their educational status, cultural background, personality status and ability. Many researchers have adapted Standardized instruments without due consideration to these facts.

The design of the study

A researcher must, consider the design of his/her study before ever going into instrumentation. In often cases, the design guides the

choice of instrument. We have many types of designs, which range from survey to experimental designs. Some studies may lend themselves to physical measurements; some may specifically require cognitive tests while there are others that may require simple observations. Some designs require repeated measurement and this raises an obvious issue about the instruments and their use for subsequent measurements vis-a-vis their implications on the internal validity of the findings. It is always advisable to conclude with the designs before proceeding with the instrument.

The expected tool for data analysis

When a researcher has chosen a tool for data analysis based on the research question and hypothesis proposed for his/her study, the choice of instrument must be restricted.

There is a common mistake which researchers, research instructors and authors have fallen victim. In many cases they do not take into consideration the type of data expected from an instrument and how such data lend themselves to a stipulated method of data analysis.

Sampling Procedure

In developing a research instrument, a number of researchers have ignored the sampling approach. This takes us back to the issue of population. In a heterogeneous population, efforts are usually made to ensure that the sample represents the entire population. The extent to which the sample represents the population is strictly dependent on the sampling procedure.

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 2.1:

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

Culled *from Santrock (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.

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.

Sample Result Booklet indicating guides for ratings of the effective behavior

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

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 **Knowledge of Specifics:** Facts, factual information, (sources), (names), (dates), (events), (persons), (places), (time periods), properties, examples, phenomena. **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 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	**Translation:** meaning(s), definitions, abstractions, representations, words, phrases. **Interpretation:** relevancies, relationships, essentials, aspects, new view(s), 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 compare, to distinguish, to deduce	**Analysis of elements:** elements, hypothesis/hypotheses, conclusions, assumptions, statement (of facts), statement (of intents), arguments, particulars. **Analysis of relationships:** 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.

Test blueprint for a Newly Developed Geography Achievement Test (GAT)

Content	**Know-ledge 8%**	**Compre-hension 4%**	**Appli-cation 30%**	**Analys is 20%**	**Synthes is 26%**	**Evaluation 12%**	**Total**
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/La titude 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 14

DEVELOPMENT OF RESEARCII INSTRUMENTS

In the first unit, the two major categories of research instruments were discussed. The development strategies for both ability and personality tests varied alittle because of the nature of behaviour or trait involved. This unit discussed under two major sections: development of ability tests and development of personality tests.

Ability Tests

A number of researchers complicate the meaning of the two major dimensions of ability: achievement and aptitude, in most cases we find researchers trying to create a big gulf between the two concepts. In its strict sense achievement refers to the effects of learning that occurred uncle partially known and controlled conditions while aptitude is the effect o learning under relatively uncontrolled and unknown conditions (Anastasi and Urbina, 1992). Anastasi and Urbina further distinguished between these two terms in relation to their usage. While aptitude test is designed to "predict an individual's subsequent performance, estimate the extent to which the individual will profit from a specified course of training or forecast his/her quality of achievement in a new situation, an achievement test determine; what the individual can do at that particular time".

Although there are minor differences in these two concepts, aptitude and achievement tests play similar roles in many cases and as such an error is generated when distinctions between the two concepts are rigidly applied. Whereas some achievement tests cover a "broad and unstandardized educational experience" a number of aptitude tests depend strictly on specific and uniform prior learning. We should remember that the Universities

Matriculation Examination (UME), even though it is an achievement test, is designed as a predictor of students' performance in the university. Because both achievement and aptitude tests point directly to an individual's ability, it is most convenient and less ambiguous to use the term ability to represent the two constructs.

Procedures for Developing Ability Tests

In unit 2, emphasis was placed on educational objectives as basic guides in test construction. Test developers should bear in mind that whichever procedure being discoursed here takes into cognizance the appropriate objectives in the corresponding domains. While we discuss basic stages in test/instrument construction it is assumed the reader has grasped the applications of test blueprints or constructs specifications as the case may be. The general procedure in constructing ability test is:

- Planning the test
- Preparing the test
- Trying out the test
- Evaluating the test

Planning the Test

For any test to be successful, careful planning must precede its construction. The planning involves a thorough analysis of the objectives of the test and the purpose the testscores are meant to serve. The planning stage should also take into consideration the testing conditions. Testing conditions in behavioral research are very important in instrumentation. Take, for example, what we experience in the West African Senior School Certificate Practical Examinations in schools where there are very large numbers of students. In chemistry, there are alternatives I and II. In this case, because of the constraints arising from inadequate facilities, the test developer, understanding the condition at the planning stage,

decides to develop two forms of a test. These two tests are equivalent forms that make it possible for the same test to be administered to students in batches at different periods without giving any batch an undue advantage over the other.

At the planning stage also, the researcher takes decisions about the form the test will take. Some ability tests can be of the multiple-choice type, essay type, word completion type etc. Planning a test also involves decisions on whether the test is to be an individual test, test for a special population or a group test. This will also determine the form the test will take. A test designed for few individuals may require rigorous tasks and time consuming intensive tasks which practically impossible in group tests that may involve thousands of students. At this stage the planner takes into considerations the availability of facilities for effective administration of the test (i.e. feasibility).

The most fundamental aspect of the planning stage is the development of the test blueprint. This is very important for all tests that are based on specified content areas. The test should reflect the approximate proportion of emphasis in the behaviour being measured and ensure that all aspects of behaviour are evaluated. Our understanding of unit 2 is very important at this stage.

Preparing the Test

The preparation of an ability test involves making a preliminary draft of the test. During the preliminary the first step is item generation. A good item writer according to Mehrens and Lehman (1991) must master the subject matter thoroughly, know and understand the pupils being tested, be skilled m verbal expression, and thoroughly familiar with various item formats. In addition, he/she must be persevering and creative. The - nature and the

objective of the test guide item generation. As the test developer generates the items he examines their suitability and relevance and the extent to which responses to the items meet the objectives of the entire test.

Another important aspect of item writing is the phrasing of the items. The items should be phrased in such a way that the content rather than form of the statement determines the response. There are cases where the phrasing provides unwarranted clues to the answer or even obscure the proper understanding of the items in such a way that it is answered incorrectly by respondents who ordinarily could have got the answer correctly. The test developer must also ensure that many items are generated to make provision for item mortality as the items are tried out. In cases where the test is a multiple choice test the test developer should avoid a regular sequence in the patterns of correct responses. This is very important because clever unintelligent students may discover the patterns and answer the rest of the items correctly without even reading through the questions. For every test the instruction should be as clear and concise as possible to ensure that even the least able candidate in the class knows what he/she is expected to do.

Trying out the Test

Having prepared the test according to plan, the test is then given a trial. One important factor the researcher should consider in trying out a test is the testing condition. It is very necessary because the response to any test item is not just a function of the test itself but also a function of the surrounding condition. A test developer should be experienced enough to determine a normal condition for a given test. What determines a normal testing condition are the objectives of the test. A condition that may be said to be normal for one test may be abnormal for another test. In research, the term pilot testing is most frequently used in place of trying out. The

essence of this practice is to have a picture of the actual qualities of the instrument in practice and further obtain data that will be used at the evaluation stage to determine the psychometric indices of the tests.

After the administration, the researcher then develops a scoring guide that enables him to score the instrument. Let it be stated categorically here that scoring guides that will accompany the instrument are developed after pilot testing. The reason is to enrich the guide. Based on responses made at the trying out stage, the test developer will be better equipped with alternative responses to various items and this will minimize stringency of the assessor.
As we will discuss in detail in subsequent units, scoring an instrument is not quite an easy exercise. It is a function of the nature and objectives of the instrument. In cases where the instrument, especially the essay type, will be scored by a cohort of examiners, it is advised that the test developer established the scorer reliability of the test before proceeding to evaluate the obtained from it.

Evaluating the Test

The test developers need to evaluate both the quality of the respondent's response and the quality of the test itself. A good test developer starts with a preliminary evaluation of the test. In that case, he goes into a conference with the respondents immediately after administering the test to them. During the conference, the respondents express their views about major flaws in the test and make suggestions on how to improve the quality of the test generally.

After this preliminary assessment the test developer, using the scores obtained from the test at the trial stage then determines the validity, reliability and other psychometric properties of the test.

Test validity pertains to the extent to which a test measures accurately what it is designed to measure. The procedure for determining the validity of an instrument depends on the nature of the instrument. There are many procedures for determining the validity of a test. They are discussed in detail in unit 4. It is also necessary to establish the reliability of the test at the evaluation stage. This will determine the extent to which the test is consistent in measuring what it purports to measure. The specific procedure the test developer will employ is also dependent on the nature and objectives of the test. This is also treated in detail in unit 5.

Some ability tests are of the dichotomously scored multiple-choice tests. For such tests, it is necessary to subject the items to item analysis. This involves a detailed assessment of the difficulty and discrimination indices of the items. An assessment of the distraction indices of the options of the items of the test is also an aspect of item analysis. These procedures will determine the adequacy of the items of a test. For details of item analysis, please refer to unit 6.

Personality Tests

Personality tests, as we already know, are designed to measure individual's emotional, motivational, interpersonal, and attitudinal characteristics as distinguished from abilities". Personality tests are the most complex forms of instruments in behavioural research. This is due to the nature of the traits involved. Personality traits are usually composed of minor constructs, which most test developers find very difficult to identify. An instrument, as we will discuss in the subsequent units, is valid when it measures what it purports to measure. This implies that for personality tests all the relevant construct of a trait must be included in an instrument that measures that given trait. Take, for example, a test that is designed to assess students' interest in

science. A researcher who wishes to develop a test of this nature must appreciate the fact the students' interest in science may include minor constructs (latent factors) such as academic interest in science, vocational interest in science, and leisure interest in science in addition to other identifiable latent groups that may be isolated within a given population of respondents.

Within the personality tests we have the test and the non-test techniques. The test techniques include self-report personality tests, general attitude and interest tests, and projective tests. On the other hand the non-test techniques encompass naturalistic observations, interviewing, ratings and analysis of life history data.

Development of Personality Tests
The most outstanding forms of personality tests in terms of structure are the projective and non-projective forms.

Non-projective forms
As in the ability tests the procedures in constructing a non projectve also involve planning the test, preparing the test, trying out the test and evaluating the test.

Planning the Test
Planning involves a clear analysis of the personality trait that will be measured and also the purpose the test scores are mean to serve. The test developer must bear in mind the various constructs that make up the trait for which he wants to develop a measuring instrument so as to ensure that all dimensions of the behaviour involved are taken into consideration in the test. To say that an individual is neurotic, instrument you used for the assessment must have measured successfully all component behaviour aspects that clearly depict the individual as neurotic. To what extent has your instrument such irrationality, phobia, obsession, and anxiety which constitute neuroticism? What I am trying to say here is that at the

planning *stage you must think* critically about the objectives of your study, the behavior you want to measure, the nature of the behaviour and how such behavior is manifested.

Preparing the Test

As in ability tests, the preparation of personality tests also involve making a preliminary draft of it. As we discussed in the previous section, during the preliminary draft, the first step is item generation. The items must represent the various constructs that make up the trait or behaviour that will be measured with the instrument. The nature and objectives of the study also guide the process. As the test developer generates the items he/she examines their suitability and relevance and the extent to which they represent the various constructs that constitute the trait that will be measured with the proposed instrument. In preparing a personality test another important writing is the phrasing of the items. The nature of the population is a vital issue to be considered in phrasing a personality test. Unlike in achievement tests where the population seems to be more homogenous, for personality tests the nature of the population is obviously heterogenous, varying in cognitive structure and affective disposition. In a searcher may observe that some members of the population are literate while others are illiterate and they are expected to respond to of an instrument, maybe on a social issue like "Politics in Nigerian Context". We may also observe that there are some obvious variations in other attributes of the population, which make it imperative that the test developer strikes a balance in the item structure to ensure that the items of the instrument can be generalized. Because the will be subjected to construct validation procedures, the test developer should also make provision for item mortality. The number of items must be relatively large so that as the items are dropped, there will be them remaining to cover the various dimensions of the behavior in question.

Trying out the Test

The procedure for trying out a non-projective personality test is the same as the procedure discussed in ability tests. The test developer should, however, understand that for a personality test the nature and behavior of the population is somewhat more heterogeneous. As such, in trying out the test, he/she must ensure representativeness so that the validity of the test may not be compromised.

Evaluating the Test

The processes of evaluating personality tests do not vary much from the procedures involved in ability tests. As in ability tests, the test developer needs to evaluate both the quality of the responses of the respondents and the quality of the test itself. A good test developer, be it an ability or a personality test, starts with a preliminary evaluation of the test. In that case, he goes into conference with the respondents immediately after administering the test to them. This provides the test developers an opportunity to interact with the testees so that they can express their views about major flaws in the test and make suggestions on how to improve the quality of the test.

After the preliminary assessment the test developer, using the scores obtained from the trying out exercise, then determines the validity and reliability of the test. As I noted earlier, the procedure for determining the validity and reliability of any instrument is a function of the instrument itself: its nature and purpose.

Projective Tests

The projective test is another form of personality test. The projective test is the type of test designed in such a way that individuals offer responses to ambiguous scenes, words or images. The use of projective test is based on the assumption that "the ways in which the individual perceives and interprets the test

material or "structures" the situation will reflect fundamental aspects of her or his psychological functioning. According to Anastasi and Urbina (1997) it is expected that the test materials will serve as a sort of screen on which respondents "project characteristic thought, processes, needs, anxieties and conflicts". The projective technique is specifically designed to permit the respondent an unlimited and diversified response about his/her psychological disposition as it pertains to an event, a condition or situations that would otherwise be impossible to explore without an X-ray of the inner disposition of the respondent. Projective techniques could appear as an inkblot (e.g. Rorschach Inkblot, Holt/man Inkblot), pictorials (e.g. Thematic Apperception Test TAT, Rosenzweig Picture-Frustration. Test), verbal techniques, autobiographical memories, and performance techniques (drawing techniques, play techniques, & toy tests).

There is no specific for developing a projective test. The reason is because there is no definite response expected from the respondents and also, because of the fact that the projective test is a generalized test that can be given to different groups of individuals with varying antecedents. The basic attribute of a projective test is that it must contain a number of emotional indicators and thought-sensitizing features that enables a respondent a wide range of responses on the basis of which the researcher can draw inferences. This technique is specifically designed for clinical purposes and because of the sensitivity of the results of projective tests, it is not advised that amateur researcher develop their own projective tests and proceed to collect data with them. There are a number of standardized commercial projective tests which researchers could readily adapt. In situation where amateur researchers have developed a projective test**,** they must ensure that repeated evaluation of the test is done. The test developer must ensure concordance of data obtained with the test.

CHAPTER 15

VALIDATION OF RESEARCH 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.

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**
		Knowl. 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

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.

Abuse of Construct Validity

Construct validity has been abused by researchers and instrument developers. Many researchers jump into construct validation of their instrument without taking a look at the nature of the instrument. Variables that are not construct-based should not be subjected to construct validation. For example a researcher once investigated the *"Availability and Utilization of Instructional* Resources in Nigerian Secondary Schools". The researcher developed a rating scale for assessing the availability and extent of utilization of various resources that were itemized in the rating scale. The researcher rated the extent of availability on a 4-point basis e.g. Very Available (VA), Available (A), Scarce (S), Not Available (NA). On the extent of utilization, for the various items listed, their extent of utilization was also rated on a 4-point basis thus: Very Often (VO), Occasionally (O), Rarely (R), Never (N). This researcher, after developing this instrument, went ahead to validate it using factor analysis.

The questions here are: Is that instrument construct based? Which constructs could be isolated from this type of instrument? Although the researcher generated loadings from what the .PC extracted based on random responses it is pertinent to note that the

entire exercise is an error and a manifestation of incompetence not only by the researcher but also on the part of the supervisors of such a research. This type of abuse is so rampart that one is beginning to think that the concept of construct validity is still strange to higher degree researchers. Take for example also of another researcher that explored "Factors that hinder the teaching of Biology in Secondary Schools in Ebonyi State". The purpose of the study includes exploration of financial factors, personnel related factors, infrastructure factors, and cultural factors. After generating items for the various sections that were clearly identified in the instrument, the researcher moved on straight to factor analysis as a way of validating the instrument Items of financial factors can never correlate with each other how much more with those of cultural factors or those of infrastructural factors. An important fact that must be understood here is that for any instrument to be subjected to factor analysis, the test developer must ensure that the constituent items of the instrument have in common one fundamental function or group of functions, whereas the remaining or specific clusters of the instrument seem in every case to be wholly different from that in all the others. This formed the basis for factor analysis as originated by Spearman in 1904. What is being said here -is that in any construct based instrument, the items must regress with each other while still retaining the potential to regroup as specifically identifiable component of the whole.

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.

Factors that Influence the Validity of a Test

From the preceding discussions we have deduced that validity coefficient is simply a correlation between test scores and criterion measures. Because we use test scores to draw inference and also because the accuracy of prediction depends to a great extent on criterion-related validity evidence it is necessary in all validity test that a number of factors that may influence both test scores and criterion related validity evidence be taken into consideration. Such factors are the nature of the population/group, mode of selection of the sample, nature of the test items and the test guide.

Nature of the population/group

For all tests of validity there is the need to clearly specify the "nature of the group on which validity coefficient is found" (Anastasi and Urbina, 1997). Within a group there could be variation in age, sex, IQ, occupation, job experience, or other characteristics that may influence response to an instrument/test. A test may be valid in assessing behaviour among a given age while it may not be valid in assessing another age group. In the same vein a -test designed to select employees for a given job may not

predict effectively the competency of the individual on the job when there are applicants with differing levels of job experience on the interview list. In this case individuals who had previous experience on similar or related jobs may score higher, not necessarily because they have a higher aptitude for the job than other applicants, but simply because they stand a higher chance of utilizing previous experiences in solving problems of the test. The point being made here is that a test is valid only for the group it is specifically designed and the features of the group must be specified.

Mode of selection of the sample

The mode of selection of the sample also should be determined during the validity test. If an instrument is designed for a heterogeneous population, the test of validity must take into consideration all the component units of the population. The point being made here is that during the pilot testing of any instrument, the test developer must ensure that a representative sample of the population for whom the instrument is being designed is used. The sample procedure to be employed will be determined by the nature of the population. The test developer must ensure that the sample he/she has taken compares with other samples that might be taken from the population. In addition, he must ensure a high probability that those other samples would behave in a similar way - exhibiting the characteristics similar to those in the sample used for the validity test.

Nature of the test items

A number of errors have been noted in tests designed by specialists. Most of these errors concern the nature of the items of the test. This has a lot of implications-on test scores and test validity in general. For some tests reading vocabulary and sentence structure is so complex that the test ends up measuring reading comprehension instead of what it is designed to measure. Care should be taken to make sure that statements in the test items are as

clear and unambiguous as possible to avoid misinterpretation and unnecessary confusion. For ability tests, identifiable patterns of answers usually lower the validity of a test. In the same vein, -test items should not be such that they .provide clues to the answers.

Test guide

A test is said to be valid not just because of the items alone but also due to other accompanying components, especially the test guide. The most fundamental aspect of the test guide is the instruction to respondents. In situations where the directions with regard to instruction pertaining to the test are not clearly spelt out, a lot of errors will be encountered in the respondents' response to the test. This automatically influences test validity. Within the test guide we also have scorer's guide. The scoring pattern of any test depends on the nature and objective of the test. When this is not clearly spelt out or provided, a number of errors may be encountered in scoring. This has serious implications on test validity.

CHAPTER 16

Reliability of Research 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.

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. $\sum D^2$

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

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

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

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

$$= 1\text{-}0.6576 = 0.3424 = 0.34$$

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

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.

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 5.4

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.

Variance of responses on items of a Likert-type *scats*

	Item	4 1 SA	3 2 A	2 3 D	1 4 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

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

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

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 17

ITEM ANALYSIS

Teachers are generally confronted with the task of designing measuring tools for their students. This is not just because of the paucity of standardized test but because in some cases a teacher-made test serves the specific objective of teachers most. We already know that when a teacher wants to develop a test for a given group of learners in a given field of knowledge or behaviour, he has to pass through the stages of item writing, production of the test, administration of the test, scoring and item analysis.

In test construction item analysis is the last step the researcher takes into consideration. Item analysis has to do with the assessment of the adequacy of each of the items that make up the test/instrument. During item analysis each of the items is assessed in terms of its difficulty, discrimination, and distractor index.

Item Difficulty

The difficulty of an item is viewed in terms of the proportion of testees who got the item correct. When an item is very easy the proportion of people that answers it correctly will be high but when the item is too difficult the proportion of people that answers it correctly will be very low. In some cases we encounter an item that none of the testees passes while in a few other cases every testee passes a given item. Such items are unnecessary in the test insofar as they do not present any information about individual differences of the testees. According to Anastasi and Urbina (1997) "the closer the difficulty of an item approaches 1.00 or 0 the less differential information about test takers it contributes. Conversely, the closer the difficulty level approaches 0.5, the more differential the item can -make". What Anastasi and her colleague were trying to explain is that the difficulty index of a test provides us with quantitative evidence of "paired comparisons or bits of differential information" about the testees. In their example, they noted that in a test supposing out of 100 persons,50 pass an item and 50 fail it (p

=0.50), then we have 50 x 50 or 2500 paired comparisons or bits of differential information. In the same vein an item passed by 70% of the persons provides 70 x 30 or 2100 bits of information, while an item passed by 90% provides 90 x 10 or 900 and one passed by 100% provides 100 x 0 or 0 bites of information (Anastasi and Urbina 1997: 173).

Although a test of moderate difficulty (p= 0.50) is recommended, it must be appreciated that items within a test tends to inter-correlate and when a test is very homogenous, the intercorrelation will be appreciably high. That is to say that if the difficulty index of all the items of a test is 0.50, then only 50%of the test takers will pass thereby making it possible for only fifty percent of the testees to get 100% and the other fifty percent 0. Jn order to avoid this problem, a test of moderate spread in terms of item difficulty is recommended. The spread for example could be say for a ten-item test 0.3, 0.5, 0.8, 0.4, 0.6, 0.5, 0.5, 0.7, 0.3, and 0.6 for items I - 10 respectively. In this case for the ten items the entire test has a mean difficulty index of 0.52. A good test, therefore, should have a few items that are of high difficulty, a few that are of low difficulty and more items of average difficulty. In other words the items should be normally distributed.

The statistical quantification of the difficulty level of an item is referred to as its difficulty index (p). The formulae for estimating the difficulty index of a test is:

$$P = \frac{R}{T}$$

Where
R = number of testees who got the answer to the item correctly
T = total number of testees

The percentage difficulty index therefore becomes

$$P = \frac{R}{T} X \frac{100}{1}$$

In cases where the number of testees is large it is recommended that 54% of the testees be used for item analysis. This 54% comprise the upper criterion group and the lower criterion group. The upper criterion group is made up of the best 27% of the testees in the total test while the lower criterion group is made up of 2.7% who scored least. This is achieved by listing the items and the number of students in each criterion group who scored the item correct. In this case the 'T in our formula is the number of students in the two criterion groups and not the total number of students who took the test.

Let us assume that we have a case where there are 25 students in the upper criterion group and another 25 students in the lower criterion group who responded to a 5 — item multiple choice test as in Table 6.1.

Responses of fifty students to a 5-item multiple choice test

Item	**Response Options**				
1.		A	B	C	D
	Upper Group	0	0	25	0
	Lower Group	3	5	14	3
2.		A	B	C	D
	Upper Group	22	0	2	1
	Lower Group	7	5	8	6
3.		A	B	C	D
	Upper Group	25	0	0	0
	Lower Group	10	3	5	7
4.		A	B	C	D
	Upper Group	15	5	1	4
	Lower Group	5	10	6	4
5.		A	B	C	D
	Upper Group	1	4	4	16
	Lower Group	5	10	6	4

For item I the difficulty index is: $^{39}/_{50} = {}^{0}/_{78}$

% item difficulty is: $^{39}/_{50}$ x $^{100}/_{1}$ = 78%
For item 2 the difficulty index is: $^{29}/_{50}$ = 0/58
% item difficulty is: $^{29}/_{50}$ x $^{100}/_{1}$ = 58%
For item 5 the difficulty index is: $^{35}/_{50}$ = 0.70
% item difficulty is: $^{35}/_{50}$ x $^{100}/_{1}$ = 70%
A very high index implies that the question seems to be very easy while a low difficulty index indicates high item difficulty. The estimation of item difficulty is very necessary in test revision. It guides the test developer on which item that should be modified or expunged entirely from the test. I wish to emphasize here that item difficulty index is limited *to* the ability range of the sample from where the test scores were derived. As such item difficulty is relative to a specified group of testees.

Item Discrimination
Item discrimination has to do with the extent to which a test discriminates between upper and lower criterion group. It refers to the "degree to which an item differentiates correctly among test takers in the behaviour that the test is designed to measure (Anastasi and Urbina 1997). Statistically speaking the discrimination index of an item of a, test is the difference between the proportion of the students in the upper criterion group who scored the item correct and those in the lower criterion group that scored the same item correct. Assuming "U" represents the number of students who scored the item correct in the upper criterion group and "L" represents the number of students who scored the same item correct in the lower criterion group, the discrimination index of an item is represented thus:

$$\frac{U-L}{1/2N}$$

where:
"U" = number of people in the upper criterion group that answered the item correct
"L" = number of people in the lower criterion group that answered the item correct
"N" = total number of students in both upper and lower criterion groups

In our previous illustration on item analysis we saw that for item I number of people in the upper criterion group that answered the item correct is 25 while the number of people in the lower criterion group that answered the item correct is 14.
For item I therefore the discrimination index is:

$$\frac{25-14}{25} = 0.44$$

For item 5 the discrimination index is

$$\frac{16-4}{25} = 0.48$$

item discrimination index lies between + 1.00 and -1.00
When the index is positive, it implies that the item discriminates in the right direction but when the index is negative it means that the item discriminates in the wrong direction and therefore need to be revised. When an item discriminates in the right direction or positively it implies that more students in the upper criterion group got the item correct. On the other hand when it discriminates negatively it implies that more of the lower ability students (i.e. students in the lower criterion group) scored the item correct. It is very unusual for more of the lower ability students to pass an item while only a few of the high ability students passed it.

Such an item obviously has a problem. It could be the item has more than one answer or that the item is poorly framed. The test developers need to conduct a supplementary analysis of the test to determine the problem of such an item.

The index of discrimination can also be computed using simple percentages. We can express U and L as percentages in which case the index of discrimination is the difference between % of U and % of L (percentage of U minus percentage of L). For example in item I of our example table, the number of people passing in upper group is 25. In that case the percentage of U is $^{25}/_{25} \times {}^{100}/_{1} = 100\%$. For the lower criterion group the number of students passing is 14. In that case also the percentage of L is $^{14}/_{25} \times {}^{100}/_{1} = 56\%$.

Since item discrimination expressed as percentage is the difference between the % for upper and lower criterion group, the % item discrimination for item I is 100-56 =44%.
You could recollect that in our previous example we got 0.44 when it was not expressed as a percentage. The procedure can be applied for other items and summarized as in Table 6.2

percentage item *discrimination of the 5-item* test

Items	% of pass in upper group	% of pass in lower group	% Discrimination
1 .	100	56	44
2	88	28	60
3	100	40	60
4	60	20	40
5	64	16	48

Item Distractor

When a test or instrument is of the multiple-choice type, the distractor index is employed to determine the extent to which each of the options distracts the respondents/testees from the correct answer i.e. the extent to which each of the options looks like the correct answer. It is used to determine the "fitness of the distractors." In a multiple-choice test, only one option represents the correct answer. The other options that are the non-correct answers are called the distractors. These distractors as a matter of fact must-be very reasonable to attract the respondents so that they will be distracted from choosing the correct answer.
Consider these examples:

1. What is the capital of Nigeria?
   (a). London (b). Abuja (c). New York
2. The capital of Enugu State is
   (a). Asaba (b). Kaduna (c). Enugu
3. The capital of Anambra State is
   (a). Onitsha (b). Awka (c). Nnewi

In these examples items number I & 2 have poor distractors while .item number 3 has good distractors. For example in item number I it is clear that London and New York are not in Nigeria so even if the respondent does not actually know the capital of Nigeria, he will know that the answer is option "b" (Abuja) because that is the only city in the options that can be found in Nigeria. The same thing is applicable in number 2 because neither Asaba nor Kaduna is located in Enugu State so they cannot be the capital of Enugu State. As for number 3, all the options are towns in Anambra State. In that case, the respondents will be well distracted by options A & C if he is not sure of the correct answer.

The formulae for estimating the distractor index of an option in a multiple choice test is:
PU – PL
Where
PU – proportion of respondents that chose the distractor in the upper group
PL – proportion of respondents that chose the distractor in the lower group.

When the distractor is positive, the option is bad. This implies that more intelligent students are distracted than the less intelligent ones. On the other hand, if the distractor index is negative, then the option is good indicating that only those who do not know the correct answer are distracted by it. In the same vein when the index is zero, the option is also bad because it does not differentiate between those who know the correct answer and those who do not know the correct answer in terms of its level of distraction.

Using our example table (Table 6.3) the distraction indices can be computed.

Distractor indices of a 5-item multiple choice test

Item	Response Options				
1.		A	B	C	D
	Upper Group	0	0	25	0
	Lower Group	3	5	14	3
2.		A	B	C	D
	Upper Group	22	0	2	1
	Lower Group	7	5	8	6
3.		A	B	C	D
	Upper Group	25	0	0	0
	Lower Group	10	3	5	7
4.		A	B	C	D
	Upper Group	15	5	1	4
	Lower Group	5	10	6	4
5.		A	B	C	D
	Upper Group	1	4	4	16
	Lower Group	5	10	6	4

For item 1 option A

Proportion in upper group 0/25 = 0
Proportion in lower group is 3/25 = 0.12
Distraction index = 0 - 0.12 = -0.12

For option B

Proportion in upper group 0/25 = 0
Proportion in lower group is 5/25 = 0.20
Distraction index = 0 - 0.20 = -0.20

For option D

Proportion in upper group 0/25 — 0
Proportion in lower group is 3/25 = 0.12
Distraction index = 0 - 0.12 = -0.12

This procedure is applied for all the distractor options. In any situation where the index is positive or zero the option must be revised. Take for example item 4, option D where the number of people who chose the option as the answer is 4 for both the upper and lower criterion group; the proportion for each of the group is 0.16. In this case the distractor index for the option is 0.16 - 0.16 = 0. This implies that the option distracts the two groups equally. The option should, therefore, be revised.

CHAPTER 20

THE NATURE OF MEASUREMENT AND EVALUATION

Introduction

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

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

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

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

Steps in Measurement

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

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

Definition of Evaluation

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

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

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

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

Continued evaluation is therefore very essential in educational programmes. Evaluation is more comprehensive than measurement for it deals with both qualitative and quantitative characteristics of

events of attributes. It is also judgemental in nature and deals with the worth, the goodness or badness of a performance or decision. Evaluation is therefore relatively subjective.

Process of Evaluation

Evaluation follows three steps:

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

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

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

Purpose of Evaluation

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

- One of the most important purposes of evaluation is to adapt instruction to the differing needs of individual pupils. Evaluation techniques help teachers to identify pupils needing specialized work and the kind of specialization required. Without evaluation techniques teachers may over-estimate or under-estimate the extent to which they should differentiate their treatment of pupils. Evaluation leads to better-directed and more effective methods of carrying out educational activities.

- Another use of evaluation is educational guidance. Evaluation provides information on how much aptitude pupils possess for scholastic work in which he is most likely to succeed.
- Furthermore, evaluation provides a basis for long-range counselling, placement and follow-up work as well as assistance in dealing with immediate problems of pupils.
- In personal guidance, evaluation is used to identify the most troublesome educational, vocational, social and emotional problems which pupils face.
- In addition to purposes pointed directly towards pupil needs, pupil evaluation helps in the overall appraisal of the total school programme by revealing specific strengths and weakness in an educational programme.
- Evaluation provides a basis upon which to compare one school's programme with another. It makes possible a study of a programme between different dates, school standards, school norms and the nature of needs in curriculum improvement.
- Pupil reports to parents and school patrons may also be used as a basis for the improvement of public relations and the mobilization of public opinion.

Differences between Measurement and Evaluation

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

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

The Importance of Testing In Education

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

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

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

Anxiety, unfairness, dishonesty, humiliation and presumptuousness are associated with tests. The process of examining and evaluating

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

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

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

Learner

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

Teacher

1. Tests help teachers to give more valid and reliable grades to students as the grades are intended to summarise concisely a comprehensive evaluation of the student's achievement.
2. Tests facilitate instructional directions – the process of constructing them if it is approached carefully, a test may cause teachers to think about instructional goals and help in linking pedagogy with educational objectives.
3. Tests provide feedback to teachers and help them to modify instructional methods.
4. Tests equip teachers with the knowledge of the strengths and weakness of pupils (diagnostic) and therefore encourage remediation and individualized attention.
5. Tests provide teachers with information on the entry behaviour of pupils and help them to determine the standing level of instructional programmes.

Guidance Decisions

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

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

Administrative Decisions (Social Functions)

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

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

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

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

Disadvantages of Testing

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

1. Tests create fear and anxiety in students and interfere with learning. The apprehension in students has the potential to destroy self-confidence and kill the desire to learn.

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

Assessment

Definition of assessment

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

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

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

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

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

- Clearly specifying, what is to be assessed: the effectiveness of assessment depends as much on a careful description of what to assess as it does on the technical qualities of the assessment procedures used. Thus, specification of the characteristics to be measured should precede the selection or development of assessment procedure. When assessing student learning the intended learning goals should be clearly specified before selecting the assessment procedures to use.
- An assessment procedure should be selected because of its relevance to the characteristics or performance to be measured. It is to be noted that assessment procedures are frequently selected on the basis of their objectivity, accuracy or convenience.
- Comprehensive assessment requires a variety of procedures. It is important to note that no single type of instrument or procedure can assess the vast array of learning and development outcomes emphasized in a school programme.

Multi-choice and short-answer tests of achievement are useful for measuring knowledge, understanding and application outcomes, but essay tests and other written projects that require students to formulate problems, accumulate information through library research or collect data (for example, through experimental observations and interviews) are needed to measure certain skills in formulating and solving problems.

- Proper use of assessment procedures requires an awareness of their limitations. Assessment procedures range from very highly developed measuring instruments, for example, achievement tests, to rather crude assessment devices, for example self-report technique. It is important to note even the best educational and psychological measuring instruments yield results that are subject to various types of measurement error.

Functions of Assessment

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

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

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

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

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

Differences between Assessment and Evaluation

BASIS FOR COMPARISON	ASSESSMENT	EVALUATION
Meaning	Assessment is a process of collecting, reviewing and using data, for the purpose of improvement in the current performance.	Evaluation is described as an act of passing judgement on the basis of set of standards.

Nature	Diagnostic	Judgemental
What it does?	Provides feedback on performance and areas of improvement.	Determines the extent to which objectives are achieved.
Purpose	Formative	Summative
Orientation	Process Oriented	Product Oriented
Feedback	Based on observation and positive & negative points.	Based on the level of quality as per set standard.
Relationship between parties	Reflective	Prescriptive
Criteria	Set by both the parties jointly.	Set by the evaluator.
Measurement Standards	Absolute	Comparative

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/Challenges of Continuous Assessment

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

i. Inadequate conceptualization

ii. Doubtful validity
iii. Inadequate structural and administrative support

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 a text blue print for a unit of instruction.

Example 1

Table of Specification for a fifty item test in Geography

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

Example 2

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

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

Advantages of the Test Blue Print

The test blue print is important for a number of reasons. Firstly, the test blue print helps one to plan adequately to set items to cover all the topics treated as well as the behaviours. That is to say, plunging into item writing without the specification table is likely to produce

a test which may be lopsided. Secondly, the procedure facilitates meaningful weighting of the items in each cell of the table in accordance with the importance attached to them. Thirdly, the blue print ensures content validity of the test. Content validity in this sense means the items adequately sample the universe content. This is achieved through the selection and writing of appropriate items in both behavioural and content areas.

Writing the individual items

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

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

Reviewing the items

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

Preparing the Scoring Key or the Marking Scheme

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

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

Writing Directions

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

orderly presentation of material and mode of identification of respondent.

Evaluating the Test

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

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

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

Administering the Test

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

1. Candidates must be made aware of rules and regulations governing the conduct of test. Penalties for malpractice such as cheating should be clearly spelt out.
2. The sitting arrangement must allow enough space so that candidates may not copy each others' work.
3. Adequate ventilation and lighting is expected in the lighting room
4. Candidates should start the test promptly and stop on time.
5. Announcement must be made about the time at regular intervals.
6. Invigilators are expected to stand at a point where they could view all students.
7. They should once in a while move among the students to check malpractices

8. Such movements should not disturb the students
9. Invigilators must be vigilant
10. Threatening behaviours should be avoided by the invigilators. Speeches like, if you don't write fast you will fail are threatening. Students should be made to feel at ease.
11. The testing environment should be free from distractions.
12. Noise should be kept at a very low level if it cannot be eliminated or removed
13. Interruptions within and outside the classroom should be reduced.
14. Expect and prepare for emergency.

CHAPTER 19

MAJOR TYPES OF TEST FORMAT

Standardized Test

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

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

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

Design and scoring

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

Standardized testing can be composed of multiple-choice questions, true-false questions, essay questions, authentic assessments, or nearly any other form of assessment. Multiple-choice and true-false items are often chosen because they can be given and scored inexpensively and quickly by scoring special

answer sheets by computer or via computer-adaptive testing. Some standardized tests have short-answer or essay writing components that are assigned a score by independent evaluators who use rubrics (rules or guidelines) and benchmark papers (examples of papers for each possible score) to determine the grade to be given to a response. Most assessments, however, are not scored by people; people are used to score items that are not able to be scored easily by computer (i.e., essays). For example, the Graduate Record Exam is a computer-adaptive assessment that requires no scoring by people (except for the writing portion).

Scoring issues

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

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

Scoring

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

- **Norm-referenced score interpretations** compare test-takers to a sample of peers. The goal is to rank students as being better or worse than other students. Norm-referenced test score interpretations are associated with traditional education.

Students who perform better than others pass the test, and students who perform worse than others fail the test.

- **Criterion-referenced score interpretations** compare test-takers to a criterion (a formal definition of content), regardless of the scores of other examinees. These may also be described as standards-based assessments, as they are aligned with the standards-based education reform movement. Criterion-referenced score interpretations are concerned solely with whether or not this particular student's answer is correct and complete. Under criterion-referenced systems, it is possible for all students to pass the test, or for all students to fail the test.

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

Aptitude Test

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

Intelligence

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

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

Combined aptitude and knowledge tests

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

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

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

Achievement Test

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

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

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

year, and therefore to increase achievement scores, and yield more "proficient" students than before.

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

Essay Type of Test

Classroom teachers construct and use a number of tests either to determine the achievement of their students motivates or encourage them to learn, identify their strengths and weaknesses, and prompt them to develop good study habits and so on. Most classroom test could be classified under two main types of test. These are the essay type and the objective type tests. Kubiszyn and Borich (1984) defined an essay test as one for which the student supplies rather than selects the correct answer and demands that the student composes a response often extensive to the question from which no single response or pattern of responses can be cited as correct to the exclusion of all other answers

Types of Essay Tests

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

Restricted Response Items (Controlled Response)

The restricted response essay item tends to limit the content, form or the number of words for testees. The limitations in terms of content and form of response are generally indicated in the

statement of the item or question. The item therefore, tends to be specific and task related. Example;

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

The Open or Extended Response Item

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

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

Suggestion for writing and using Essay test

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

I. Have clearly in mind what mental processes you want the testee (student) to use before starting to write the question. For instance, if you want the testee to analyse, judge, or think

critically include mental processes that involve analysis, judgement or critical thinking.

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

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

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

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

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

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

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

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

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

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

I. When to use Essay Test

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

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

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

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

Scoring of Essay Tests

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

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

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

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

Suggestions for improving Essay Scoring

1. Use the scoring scheme consistently. Do not favour one student over another or get stricter or more relaxed over time.
2. Remove or cover the names on the papers before scoring. This will help the ratter score the paper on its merit, rather than an overall impression of the student.
3. Score each student answer to the same question before going on the next answer. This avoids a student score having influence on another student score and it helps in maintaining the scoring criteria.

4. Keep scores for precious items hidden when scoring subsequent items to avoid straying from the scoring criteria.
5. Evaluate the papers before returning them. They help detect discrepant rating for correction.

Advantages of Essay items

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

Disadvantages of Essay Tests

1. Essay test encourage bluffing. By its length and score, essay test encourage verbosity and digression.
2. Essay test are difficult to score. By their nature there is a degree of high subjectivity in the hands of an inefficient examiner
3. It is tedious and time consuming in both writing and scoring. It is tedious to wade through pages and pages of students hand writing and students also spend a lot of time in writing.
4. Essay test suffer from limited sampling. The items are inadequate due to the needed time to respond to them.
5. Scorers of essay test are unreliable as it is difficult to maintain a common set of criteria for all the students

Objective Type Test

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

Types of objective tests

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

Selection type

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

Supply type

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

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

Suggestions for constructing supply test item

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

Advantages of supply test

i. It is used for testing knowledge on definitions and terminologies
ii. It facilitates computational skills in sciences- mathematics, etc.

iii. It is more discriminative than multi-choice and True-false test
iv. It reduces copying and guessing
v. It allows students to exercise their ability in thinking thoroughly for the answer
vi. It facilitates vocabulary and concept development in students.

Disadvantages of supply test

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

Selection type

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

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

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

Suggestions for writing True-False Items

I. The statement should be clear and lucid i.e. the statement should be definitely true or definitely false without additional qualifications.
II. Uses relatively short statement and eliminate extraneous material.

III. Keep true and false statement at approximately the same length
IV. Avoid using double – negative statement. Avoid verbal clues, specific determiners and complex sentences
V. Avoid broad general statements that are usually true and false without further qualifications

Advantages of True-False Test item

I. They provide simple, fundamental and direct test of students' knowledge
II. It is easy to score quickly and objectively
III. True and false items are quite efficient. The number of independently scrabble responses tend to be higher than multiple-choice test
IV. Most testers find the task of writing true-false items simple and less time consuming.
V. It can be adequate to most question areas

Disadvantages of True-False items

I. They are suspected with good reason of being particularly susceptible to chance error resulting from guessing
II. They are less reliable than multiple-choice test of equal length due to chance errors
III. They are usually judged to be trivial
IV. Most true-false items are based directly from textbooks and there is the danger that they might encourage and reward sheer verbal memory
V. True-false questions may lack background information or qualifications to enable even an expert to judge with assurance whether they are True/False
VI. They do not provide explicit alternative in relation to which the relative truth or falsity of the item can be judged.

Multiple-Choice Test (Mcqs)

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

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

Guidelines for constructing Multiple-Choice tests

I. The stem should contain the central issue of the item, and should be concise, clear to read and understand.
II. Options should be plausible. Distracters must be plausibly attractive to the uninformed.
III. All options for a given item should be homogenous in content, form and grammatical structure
IV. Avoid the repetition of words in the options
V. Avoid specific determiners, which are clues to the correct option.

Advantages of Multiple-Choice test items

i. Multiple-choice questions have considerable versatility in measuring objectives from knowledge to evaluation
ii. A substantial amount of course material can be sampled in a relatively short time.
iii. Scoring is highly objective acquiring only a count of the number of correct response
iv. The degree of discrimination among the correct options is high and this allows the student to select the best alternative and avoid the absolute judgement found in True/False test items.

v. The multiple options reduce the effect of guessing

Disadvantages

i. Multiple-choice questions can be time-consuming to construct
ii. Multiple-choice questions can at times have more than one defensible correct answer
iii. The error of guessing is only reduced by the discriminatory element but not eliminated
iv. To an extent promotes rote learning

Matching Test items

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

Suggestions for writing matching test items

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

Please, find below an example of a matching test:

Match the following names with their countries

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

Advantages

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

Disadvantages

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

Comparison of Essay and Objective test items

	Essay Test	**Objective test**
I.	Requires the student to plan and organise his answers	Requires the student to choose among several alternatives
II.	Consists of relatively few or more general questions which call for extended answers	Consists of many rather specific questions which require brief answers.
III.	Demands a lot of thinking and writing from students	Less time is spent on thinking and writing due to guessing and the options
IV.	It allows the student the	It affords freedom for a test

	freedom to express him/herself	constructor but limits the freedom of the student
V.	There is subjectivity in grading	It measures a minute aspect of the individual (Low level ability)
VI.	It is easy to construct but difficult to score	It is relatively tedious and difficult to prepare but easy to score.

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