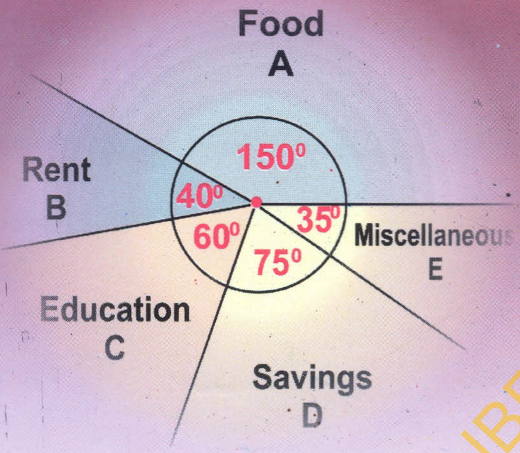


Designing Evaluation Plan
Designing Data Collection Tools
Collection Data
Analyzing Results
Reporting Findings
Planning Program Changes
Input From:
You, Your Staff, Program Participants,
Key Stakeholders, Wider Community

THE BASICS OF RESEARCH and Evaluation Tools

By
JOSHUA OLUWATOYIN ADELEKE

The Basics of Research

and

Evaluation Tools

By
Joshua Oluwatoyin Adeleke

The Basics of Research and Evaluation Tools

ISBN: 978-978-49876-4-6

Published by:
Somerest Ventures
Guinness Roundabout, by NEPA Gate,
Oba Akran Avenue End, Ogba-Ikeja.
P.O.Box 16440, Ikeja-Lagos.
Tel: 234-1-0802-306-6190, 01-8735469
e-mail: somerest_ventures@yahoo.com

DEDICATION

This book is dedicated to
My Almighty God, The Lifter of my head.

PREFACE

Understanding Basics of Research and Evaluation Tools has been my concern. Four years ago, I organized a workshop for young researchers like myself. Participants came from Lagos, Ogun, Oyo and Osun states. The plea at the end of the workshop was 'put these ideas in a text to assist researchers solve their problems'. Apart from the plea, Professor S.O. Ayodele, knowing my contributions to various research studies challenged me also to write a book that would be of help to young researchers. Considering my four years experience as a research consultant, and three years as a university don, a book like this is a necessity. I, therefore, give glory to God who had made it possible to have the book printed.

The book has twelve chapters. Chapter 1 dwells on continuous assessment practice, while issues related to measurement and Evaluation are discussed in chapter 2. Chapter 3 centres on curriculum evaluation. Chapter 4 addresses basics of research, while chapters 5 to 7 focus on evaluation tools (both cognitive and non-cognitive). Sample and sampling are discussed in chapter 8, while statistics is the focus of chapter 9. Issues related to hypothesis testing are addressed in chapter 10, statistical tools are discussed in chapter 11. Chapter 12 provides researchers with tips that guide research decisions. This book, therefore, becomes a need for college, undergraduate, postgraduate students and research driven minds.

I wish to register my appreciation to all those who have made the production of this book possible. They are many but very few are mentioned. I should specially thank Professor Mac Araromi, the Director Institute of Education, University of Ibadan for sparing time to write the FORWARD of this book. I should also thank Professor S.O Ayodele specially for the encouragement given to me while writing the book. I appreciate my Ph.D supervisor, Dr. G.N. Obaitan and Dr F.V. Falaye for the guidance they give from time to time. I must remember Prof. T.W. Yoloye for his assistance. Drs. Adenike Emeke, Ifeoma Isiugo-Abanihe, Gbenga Adewale, Biodun Adegbile, Modupe Osokoya, Adams Onuka, Monika Odinko, Felix Ibode, Benson Adegoke, Serifat Akorede and Lanre Jinaid are also highly appreciated for their support. I appreciate Drs. Fakeye and Opateye and Pastor A.S. Ogunsola for thorough editing of the book. Ayo Aibinuomo, Joshua Sijuade and Mike Adedigba are also appreciated for their assistance in typesetting the book.My wife has been a virtuous woman, giving me full support to achieve my goals. Shalom and Shammah are wonderful kids that usually make my day blissful, I appreciate you all. I cannot but appreciate all ICEE students.

Finally and most importantly, I thank God the Father, God the Son and Holy Spirit, the giver of wisdom and the savior of my soul. It is a true saying " without God, I can do nothing)".

Dr. Joshua O. Adeleke

FOREWARD

This is nearly a 220 page book dealing with the Basics of Research and Evaluation Tools which incidentally is the title of the book. The book consists of twelve chapters. The first chapter gives an overview of Continuous Assessment Practices wherein the following sub – topics were vividly discussed: Standards of Continuous Assessment Practice and Procedural steps in Continuous Assessment Practices. Chapters II deals with Measurement and Evaluation considering the importance of Testing and Evaluation, definition of Measurement, Assessment and Evaluation, Evaluation Strategies, Qualities of Evaluation Method, Validity, Reliability, Practicality and Cost, Nature and Relevance of Educational Measurement, Programme Evaluation, Planning Programme Evaluation, some major types of programme Evaluation, Goal Based Evaluation, Process Based Evaluation, Outcome Evaluation, Formative Evaluation, Summative Evaluation etc.

Chapter III is concerned with Curriculum Evaluation, where the following issues were well articulated: Evaluation of students' performance and outcome as influenced by the Curriculum, Evaluation of Curriculum Quality. Chapter IV deals with Basics of Research considering concept of Research, Research writing, Formation of the topic, Background to the study, Statement of problem, Objectives, Research Questions and Hypothesis, Conceptual Operational definitions of terms Literature Review, Research methodology, Result Presentation, Importance of Research, Types of Research, Classification of Research, Types of variables, Ethics of Research etc. Chapter V has to do with Research Tools: Achievement and Aptitude Tests, Types of Achievement Test, Guidelines for Constructing Multiple Choice items, Types of Analysis, Item selection, etc. Chapter VI deals with Non Cognitive Research Tools, considering functions, Observation schedule, Interview guide, Checklist, Construction of Schedules and Questionnaire, Overview of research Tools etc.

Chapter VII considers Measurement Scales and Indices, (the Nominal, Ordinal and Interval Scales), pilot studies and Pre – tests, Reliability and validity of Tests, Types of Reliability, Types of validity, Sources of variability, issues of Generalizability and Standardization.

Chapter VIII deals with Sample and Sampling procedures, Difference between Probability and non - probability Sampling , Simple Random Sampling, Systematic Sampling, Stratified Sampling, Cluster Sampling, Multi – Stage Sampling, Purposive Sampling, Modal Instance Sampling, Export Sampling, Quota Sampling, Heterogeneity Sampling and Snowball Sampling.

Chapter IX has to do with Statistics considering such issues as Data, Graphical Representative of Data, Bar Graphs, Pie Graph or Pie Chart, Frequency Polygon, Histogram, Measures of Central Tendency and Dispersion and Basic Statistical Terms etc. Chapter X deals specifically with Hypothesis Testing, Procedures in Hypothesis Testing, Errors in Hypothesis Tests, Hypothesis Testing for a population mean, Hypothesis Testing for a proportion and for a mean with unknown population standard Deviation, Hypothesis Testing for a Population Proportion, Hypothesis Testing of the difference between Two means and Hypothesis Testing for a Difference between proportions etc. Chapter XI deals with some Statistical Tools like Correlation, ANOVA and ANCOVA, Chi – Square significance of Tests etc. Chapter XII offered some tips for Researchers. Issues such as Chi – Square Test, Pearson Moment Correlation, T. Test independent, Mann – Whitney U. Test, Witcox in, Rank Biseral, ANOVA (One way), ANOVA (Two – way), Kruskal – Wallis, Fred man, Phi – Coefficient etc.

This book is a comprehensive work for anybody desirous of engaging in a research work. I went through and avidly read it page by page and indeed it refreshed my memory about the basic tools needed for carrying out educational research. It is indeed an important keepsake for both teachers and students of educational research. I recommend it highly.

Prof. Mac Araromi
Immediate past Director,
Institute of Education, University of Ibadan

TABLE OF CONTENT

CHAPTER ONE

CHAPTER TWO

CHAPTER THREE

CHAPTER FOUR

CHAPTER FIVE

CHAPTER SIX

CHAPTER SEVEN

CHAPTER EIGHT

CHAPTER NINE

CHAPTER TEN

CHAPTER ELEVEN

CHAPTER TWELVE

CHAPTER ONE

Continuous Assessment Practice

Assessment is a powerful diagnostic tool that enables learners understand the areas of strength and weaknesses. Many teachers erroneously assume assessment to be a strategy for generating scores and grading only. The roles of assessment go beyond such assumptions. Primarily, assessment is meant to diagnose areas where learners are having difficulty to allow for concentration of efforts in those areas. Assessment also allows teachers to monitor the achievement of set behavioural objectives. Teachers can evaluate their pedagogical strategies through effective assessment. Frequent interactions between learners and teachers mean that teachers know the strengths and weaknesses of their learners which is only possible through effective assessment. These exchanges foster a learner-teacher relationship based on individual interactions. Students learn that the teacher values their achievements and that their assessment outcomes have an impact on the instruction that they receive. Effective assessment practices can motivate students to continue attending school and to work hard to achieve higher levels of mastery. For assessment to be effective, it must be a continuous and comprehensive exercise as against one or two test(s) assessment. In today's policy environment, testing has become a critical component of education reform. Policy makers and education administrators often view test scores as measures of educational quality and, as such use test scores to hold schools accountable for teachers' performance. Continuous assessment, an alternative or supplement to high stakes testing of student achievement, offers a methodology for measuring learners performance and using those findings to improve the success of learners. The rationale behind continuous assessment is that the more you know about what and how students are learning, the better you can plan learning activities to structure your teaching. The technique is mostly simple, involving in-class and out-class activities that give both the teachers and the students useful feedback on the teaching-learning process.

What is Continuous Assessment?

Continuous assessment is a classroom strategy implemented by teachers to ascertain the knowledge, understanding, and skills attained by pupils. Teachers administer assessments in a variety of ways over time to allow them to observe multiple tasks and to collect information about what students know, understand, and can do. These assessments are curriculum-based tasks previously taught in class. Continuous assessment occurs frequently during the school year and is part of regular teacher-student interactions. Students receive feedback from teachers based on their performance that allows them to focus on topics they have not yet mastered. Teachers learn which students need review and remediation and which pupils are ready to move on to more complex tasks. Thus, the results of the assessments help to ensure that all pupils make learning progress throughout the school cycle thereby increasing their academic achievement.

In continuous assessment, teachers assess the curriculum as implemented in the classroom. It also allows teachers to evaluate the effectiveness of their teaching strategies relative to the curriculum, and to change those strategies as dictated by the needs of their learners. In addition, continuous assessments provide information on achievement of particular levels of skills, understanding, and knowledge rather than achievement of certain marks or scores. Thus, continuous assessment enables pupils to monitor their achievement of grade level goals and to visualize their progress towards those goals before it is too late to achieve them. Hence, the continuous assessment process supports a cycle of self-evaluation and student-specific activities by both students and teachers.

Continuous Assessment has been defined in various ways.

MSN Encarta Dictionary defines it as assessment of students' progress based on work they do or tests they take throughout the term or year, rather than on a single examination

Business Dictionary sees it as evaluation of a trainee or a training program carried out on a daily or fixed interval basis, instead of at the end of the training period.

Collins English Dictionary – 6th Edition 2003 conceptualizes it as the assessment of a pupil's progress throughout a course of study rather than exclusively by examination at the end of it.

Within the context of Nigeria Educational system, Federal Ministry of Education (1979) define it thus:

> Continuous assessment is a method of ascertaining what a pupil gains from schooling in terms of knowledge, skills, industry, and character development, taking account of all his/her performances in tests, assignments, projects, and other school activities during a given school period [term, year or entire period of an educational level].

Standards of Continuous Assessment Practice

Standards and statutes Continuous Assessment toss out the terms "validity" and "reliability" and require teacher educators to think about them in some meaningful way. But how does a teacher get there, making use of the *Standards for Educational and Psychological Testing (1999)*?

1. ***Identify the construct to be measured.*** The construct can be an aspect of the three domains, Cognitive, Affective and Psychomotor. A teacher may be interested in measuring students' attitude towards a school subject. In this case, attitude becomes the construct.

3. ***Define the purpose***. Why does the teacher need the measure of the construct? Any measure included in continuous assessment should assist in understanding the holistic development of a student in an educational system.

4. ***Relevance of the Construct.*** The knowledge and skills to be demonstrated in the assessment must be essential in nature. They must represent important work behaviors that are job-related and be authentic representations of what teachers do in the real world of work.

5. ***Determine the use.*** Continuous assessment should include variables that are useful to both students and the institution. Hence, there is need to decide variables that will serve as criteria for either promotion or certification. Outcome of students' assessment can also be used to determine the effectiveness of the teacher.

6. ***Identify the measurable conceptual framework.*** Continuous assessment is expected to cover expected *standards. Standards* here, therefore, refer to observation of knowledge, skills, and dispositions when discussing a conceptual framework, so the framework can be all the teacher standards that define competency in these three domains.

7. *Develop a blueprint or framework to guide the design process.* Continuous assessment demands building an assessment system, like any test, based on the domains to be measured – the conceptual framework. This is the reverse of what most teachers do. They construct items for measurement before considering the fitness of the items.

8. *Keep checking validity – both construct and content.* Ensure that the system that is being built measures the expected behavioural changes of the learners, through appropriate tasks (construct validity). Also show evidence that the set of assessments adequately represent the most important elements of the domains to be measured – with not too much and not too little and nothing irrelevant targeted for any given standard (content validity).

9. *Build assessments that can be studied for internal consistency.* A teacher is expected to assess all the students taught. Scoring students on a test especially theory type demands consistency. Developing a reliable marking scheme is recommended. Where two or more teachers teach different arms of a class, rater agreement is important.

10. *Develop systems to ensure fairness towards all those candidates assessed.* This includes the policies and procedures to implement and monitor the Continuous assessment as well as specified checks for bias in the way tasks are written and scoring is done.

11. *Check the consequences of the decisions.* Show evidence that (1) assessment grades are tenable (2) adequate feedbacks are given to the students at the right time and (3) arrangement is made for remediation.

12. *Build it once, and revise it.* The system may have branches or tracks to fit multiple purposes, but all standards and all purposes should be considered at one time. Then, revise based on experience, changes in institutional mission and standards, and problems identified related to validity, reliability, and fairness.

Procedural Steps in Continuous Assessment Practice

There are two major developmental or procedural steps that can help the designers of Continuous assessment systems progress towards achieving the above twelve recommendations.

Step 1: Define content, purpose, use, and other contextual factors.

In this step, designers begin by determining what they want to know (assessment content), why they want to know it (assessment purpose), and what they will do with the information once they get it (assessment use). Each purpose and use are conceptualized and evaluated separately as a matter of validity.

- ***Purposes*** will vary based on need. Institutions may have dual or triple purposes – such as student certification, program improvement, and student reflection or growth. The first would be more aligned with national program approval and would tend to be rather prescriptive and analytic in nature. The second and third would be more aligned with the institution's conceptual framework and could tolerate much more freedom of choice and more holistic decision-making. Thus, these are likely to require different data and data analysis strategies to fulfill the different purposes, although there may be substantial overlap.

- ***Content*** can be defined in many ways, but it is reasonable to be consistent with national standard . This would require defining content in terms of the knowledge, skills, and dispositions required of students by various groups – national, state, and institutional.

- ***Uses*** also dictate what will be done with the data. Some examples are: certify a student or allow him/her to graduate, identify certification-related program weaknesses and improve a program, identify weaknesses and make improvements in unit-defined areas of importance (e.g., conceptual framework).

Examples showing the relationship of purpose, content, and use are:

Table 1: Samples of Purpose, Use and Content

	Example 1	Example 2	Example 3
Purpose	The system ensures teacher competence in order to protect Nation's Educational goals	The system provides data on program quality for program improvement.	The system provides opportunities for teachers to demonstrate competency in students' assessment
Content	Specific tasks are required in the system for analysis of students' educational achievements. The content of each task varies but it is aligned with institutional, state, and national standards, agreed to by the stakeholders, showing depth and breadth of coverage.	Same as Example 1 plus any additional data targeted by the institution or district. Depth and breadth of coverage are important to the extent that they yield data on overall program quality that are useful for substantive improvement.	Selected evidence with reflections on demonstration of standards and future plans for growth.
Use	Aggregated data from scoring rubrics on tasks for each student determine whether the student successfully completes the program, and, therefore, can be certified.	Data are aggregated from scoring rubrics on tasks for each standard to identify standards with which students have more difficulty than anticipated in order to target program improvements.	Students receive feedback on individual work, and scoring rubrics for portfolios are aggregated for the unit to evaluate students' ability as a measure of curriculum strength.

Once the purpose(s), use(s), and content are determined, it is important to analyze all the local factors that would affect the system, e.g., conceptual framework, resources. It is also important to ensure that all aspects of the program to be measured, including the delineation of content and conceptual framework, are more motivated by school's ability to visualize what they look like when performed than any political or other motivations.

Step 2: Develop a valid sampling plan.

A critical next step is the identification of all relevant standards and the alignment of standards with each other into assessment domains. This is the beginning of the job analysis, and it is not as daunting a task as it sounds. There are common threads that run through all of the sets of standards because literature has identified several important characteristics of effective teaching consistently over time.

For example, virtually every set of standards has something about content knowledge, planning, and assessment. This can easily be demonstrated through the alignment process.

All of the standards could be correlated in a similar fashion, extending the matrix in both directions – vertically and horizontally. The point is that there is substantial overlap among standards sets. Aligning them allows for the creation of a solid set of assessment domains and reduces the overall workload in the end.

CHAPTER TWO

Measurement an Evaluation

Importance of Testing and Evaluation

Although the process of testing and evaluation is often treated as something separate from instruction, evaluation is an essential part of the instructional process. One way of looking at instruction is to note that it has three elements. The first element is introducing the learners to instructional goals and objectives, i.e. new ideas, processes, methods, etc. The second element is practice which provides the opportunity for learners to use the new ideas in appropriate contexts. The third element is feedback or evaluation when the learners are informed about whether or not they have used the new ideas in an appropriate fashion. The feedback phase of instruction is often neglected, so that learners fail to discover that they have not mastered ideas or skills. Even worse, learners may receive information from the feedback phase which is not helpful. For example, junior medical students taking a pediatrics course may be introduced to the enormously important set of concepts relating to fluid and electrolyte balance. If the students do not get a chance to practice these ideas and get appropriate information or feedback about their mastery of the concepts, the students may see children suffering from dehydration and have no idea about how to manage the children's problems.

Definition of Measurement, Assessment and Evaluation

Based upon the description of evaluation given above, measurement, assessment and evaluation can be conceptualized as providing learners with information about the results of their competency in acquiring knowledge and skills. The evidence that learning has taken place is based upon a sample of the behavior of learners, either directly observed behavior such as watching learners examine patients, or behavior on tests such as choosing, describing, etc. Behavior alone is not sufficient; the learners must be informed that the behaviors sampled have satisfied some criteria of effective behavior. Examinees must be informed about the results of the test; those observed must be told if they performed the task effectively. According to Ayodele, Adegbile and Adewale, (2008)

Measurement can be defined as the process of assigning numbers to human characteristics or attributes or that of objects or events based on certain rules or regulations. (through measurement the following scores could be generated: 45, 56, 77, 88, 59 etc.).

Assessment on the other hand could be referred to as the process of establishing the status of the performance of an individual or group in a given task usually with reference to the expected outcomes. Assessment can take both quantitative and non-quantitative forms (e.g Distinction = 3, Credit=2, Pass=1, Fail=0).

Evaluation is a sample of behavior which is used to make some value judgment. If the value judgment is not generalized beyond the particular context when it is gathered, we call it feedback. Medical students who examine patients and are told by preceptors that they did an effective or ineffective job of examining the patients are getting feedback. If the preceptors decide on the basis of a number of observations that certain students need to repeat a course or receive pass or honors, the students have been evaluated, i.e. they have received grades based upon generalizations derived from their accomplishment of a number of tasks.

Effects of Evaluation

The process of testing and evaluation has a large number of effects on education and educational systems beyond the task of providing feedback to learners in practice situations. Evaluation tells the learner what to study and how hard to study. It provides information about which learners need additional education and which ones should be dropped from an educational course of study. It helps education managers to develop curricula, and to select those who will be permitted to enter their programs. It helps learners to decide on their future careers. The effectiveness of decisions which are assisted by the various roles of testing and evaluation are considerably influenced by the strategies used by educational managers in the design of evaluation programs. Regardless of the particular approach taken in devising a testing instrument, it should simulate the work of a physician.

Evaluation Strategies

Instructors and course directors have a variety of strategies they can use to carry out their testing and evaluation functions. They can use *analytic* methods which gather samples of behavior in which each sample is only a small part of a more complex idea or concept, e.g. results of objective tests, to provide feedback and evaluation, or they can use *synthetic* methods in which the sample of behavior gathered is large and complex, e.g. observing a medical student gathering historical information, or they can use methods which lie between these types, e.g. simulation tests. They can give tests which are comprehensive, i.e. cover material learned over large blocks of time, or scatter tests throughout the course of study. They can give grades which focus on pass—fail, or provide a series of grades such as A, B, C.... They can base their selection of tests on a huge reference work, or on a defined set of instructional materials. They can use tests and testing personnel from outside the local system in designing the testing program or they can rely mainly on internal decision makers. They can rely on the learners choosing among a defined set of possible answers, multiple choice tests, or generating responses, essays and oral tests. All of these various strategies of approaching testing and evaluation have advantages and disadvantages based on their effects on learning and the costs of the evaluation system. Various instruments available to instructors and evaluators are sufficiently discussed in this book.

Qualities of Evaluation Method.

Evaluation specialists have developed four important criteria for choosing evaluation methods. Two of these criteria, *validity* and *reliability,* are technical attributes of all measurement instruments which should be taken into account in using measurements. The third criterion, *practicality and costs,* is essentially a management criterion which depends greatly upon the resources and values of those developing and using tests and evaluation methods. The fourth criterion is the *effect on learning*. This criterion is the most important for reasons which have already been discussed.

Validity

A test or evaluation has two attributes—the gathering of samples of behavior, and a decision based upon the sample. Validity is the characteristic of a test or a testing program which relates to the decisions made. For example students may know a large number of facts, and do well on fact based examinations. Based upon these data, it might be decided that these students can solve problems. However, this decision may or may not be valid because it goes beyond the data provided by the test. In many cases learners' knowledge of facts is a necessity but not sufficient condition for deciding that the learners have mastered the information required to meet the standards desired by the program. In clinical situations, the mastery of facts is often given overwhelming importance in deciding on whether students have satisfied program standards when a number of attributes of effective performance in clinical situations, e.g. problem-solving skills, interpersonal skills, technical skills such as those required in performing a physical examination or obtaining materials for laboratory examinations, work habits and attitudes are not assessed. In this case the test of facts may be valid, but the evaluation system is not valid.

Reliability

Validity is an attribute of the decisions made based on the results of evaluation methods, while reliability is a technical attribute of the measurement method itself as used on a particular population of individuals who are assessed by the instrument. Since evaluation is ultimately generalizing about a sample of behavior, reliability is an estimate of the amount of error which exists in a particular measurement. Error may be conceptualized as the likelihood that the results would be similar, if the measurement were repeated . Some types of error are typical of the evaluation method. For example, examinees with bad handwriting usually do less well on essay tests than they would on other types of tests. The most common type of error in tests and the most pervasive is sampling error. We all know, intuitively, that there is a great deal of error in small samples. A learner might get one question right and another wrong simply because of the choice of questions. The more questions that are asked, the more likely it is that a test is reliable. For this reason, certifying examinations such as the Medical Licensing Examination contain hundreds of questions. Evaluation exercises which can be used for feedback purposes, e.g. observing a medical student with one patient, cannot be used for decisions about promotion because of the error in small samples of behavior. Another source of error in examinations such as essays, orals or observations is rater error. Raters tend to value different attributes of performance, focus on different attributes of what is observed and if the evaluation is complex, weight elements of the sample of behavior, differently. Raters also have different standards, even if they agree on what is observed. Two raters may rank a group of examinees the same, but one might give higher grades than the other. Even though observer errors can be limited by examiner training, sampling error can still create great unreliability in any test which uses only a small number of exercises or observations.

Practicality and Cost

Some possible tests are impractical. It may be difficult to get enough oral examiners to conduct a certification examination, or a sufficient number of patients cannot be assembled to allow all the examinees to provide samples of behavior with patients for assessment purposes. Since it is essential that an educational program provide some assessment of clinical skills, in order to develop a valid evaluation system, issues of cost in terms of school time, the hiring of simulation patients, the utilization of support personnel, etc., go back to the values of those running the program. If faculty members receive little in the way of rewards or recognition for teaching, they will be reluctant to spend energy in evaluating learner performance. Sometimes an imaginative use of resources can modify the cost-benefit ratio of effective evaluation techniques. Examinees can be screened to see if they are at risk of marginal performance, and only the weakest performers given the more expensive techniques.

Effect on Learners

The possible effect on learners of evaluation methods has already been mentioned, so this section is quite brief. It is particularly important for instructors to realize that learners will focus on what is evaluated. If important attributes of instructional performance are not assessed, then the learners will neglect those aspects in favor of those which are assessed. Regardless of what is written in a course outline or syllabus, learners perceive that the objectives of the instruction are what is assessed.

Nature and Relevance of Educational Measurement

Decision making is a vital issue in Educational System. Such decision can be institutional in which a large number of comparable decision are made or individual where the choice confronting the decision made will rarely or never re-occur. Basically, institutional decisions are those made by school management concerning students (e.g. selection criteria for admission, requirements for selecting students for admissions, bases for advising students to withdraw).

Educational decisions can be viewed in a broader way as instructional, guidance, administrative or research. Though the classification may not be so definite, however they provide direction for decision making in Education system. For any decision to lead to favourable outcome, such a decision must be based on all relevant data. In fact there should be a variety of data from diverse sources in order to make the best decision possible. All these inform the relevance of Educational Measurement.

Measurement helps the Teacher

Pupils are put in teacher's care to achieve the set objectives on them through educational experiences or instruction. The teacher who had been exposing pupils to instruction for a given period of time (a week, 2 weeks, a term or a session) needs measurement to gauge the extent to which the instruction has been effective. The relationship among objectives, instruction and measurement is thus presented schematically as follows:-

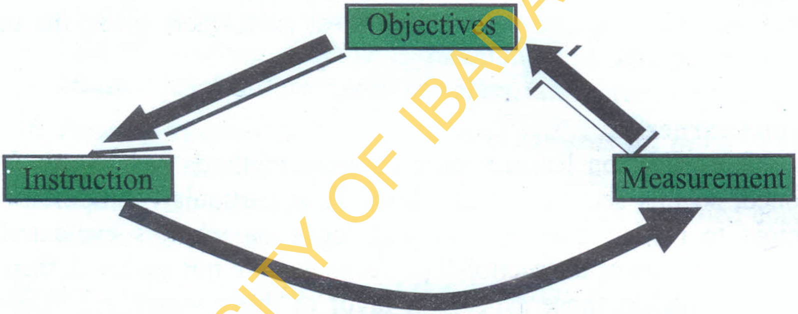


Figure 1. Relationship among Objectives, Instruction and Measurement

In specific ways, measurement assists the teacher in the following areas:

i. Assigning of grades
ii. Certification of skills and abilities
iii. Prediction of Success in subsequent courses
iv. Initiation point of instruction in a subsequent course.
v. Feedback to students
vi. Comparisons of outcomes of different groups

(i) Assigning of Grades

Contrary to views of many, measurement procedures do not make decisions; people make decisions they consider relevant at any time. At most, measurement procedures can provide useful information on some specific factors that are relevant to the decision an institution or individual is about making. University Matriculation Examination (UME) test can provide scores in Chemistry, Physics and Biology. Indications of how well Joshua is likely to perform in Medicine. The test scores of Joshua in the three Science subjects with information about how academically demanding Medicine as a course is at the University of Ibadan can be used to make a specific estimate of how well he is likely to perform in that course.

Teachers use measurement commonly from pre-school through primary, college and tertiary institutions to assign grades, represented by letters [A, B, C, D, E, F etc] or by numbers. They grade students by categorizing each learner in terms of amount or level of learning in relation to general performance. There is need to ensure adequacy (validity and reliability) of the instruments used for the measurement, otherwise, the information provided through measurement procedures will be highly misleading. Issues of validity and reliability shall be discussed in subsequent chapters in this book.

Item analysis shall be discussed later in this book, however, the applications of difficulty and discrimination indices in relation to assigning of grades shall be discussed in detail. In a normal situation, individual scores are expected to spread, a few in the top category, some more in a second group, a larger number who are 'average' and fewer in the lower categories. This intent can be ensured through construction process on the measuring instruments. Achievement test shall be used for all the illustrations under this section. The Proportion of the testees who passed an item is referred to as item difficulty index.

A goal spread in results can be obtained if the average difficulty of items is around 0.50 or 0.60 and if the items vary in difficulty from about 0.20 to 0.80. This should be the target anytime a teacher is preparing an achievement test. Test items passed or failed by nearly all the testees, do nothing to differentiate their level of achievement in the selected behaviours measured.

Similarly, the difference in the proportion of top (27%) students with that of poorer (27%) students who scored the same item correctly is referred to discrimination index. An item does not discriminate if at least 50 percent of each group got the item right.

According to Bloom, Hastings and Madaus (1971), "An alternative is possible. An Examination may be developed according to the best estimates available [experienced judgments of other teachers and past experience with items) and given to the students being evaluated. On the basis of the results, the indexes of difficulty and discrimination are computed for each item. Then the final score for each student is based only upon the items which meet the criteria of difficulty and discrimination. This solution of the problem is costly in terms of items dropped because they fail to meet the criteria. However, if what is wanted is spread of scores with most being in the centre category, the expense in lost items must be weighed against the purpose in the ordinary setting, you will not lose more than 10 percent of the items.

Certification of Skills and Abilities

Another reason why measurement is carried out is to certify that a given learner possesses, at least at that time certain skills, knowledge and abilities that make him relevant to himself and his world. This situation occur mostly in a formal learning centres such as secondary schools, technical colleges, Universities etc. Certificate possessed by learning can be used for further study or securing a job. At this point, it is necessary to emphasize that measurement that will be used to provide information for the certification of learners must be valid and reliable. The issue of validity and reliability of measuring instruments is discussed in another chapter of this book. There are problems associated with certification of skills and abilities using measurement of data. One of such problems is establishment of standards set by different learning centres vary considerably. Several attempts had been made by many nations to ensure standards through the establishment of examining bodies such as West African Examination Council (WAEC), Cameroon General Certificate of Education Board, Kenya National examination Council, Examination Council of Swaziland, Examination Council of Zambia, etc.

Prediction of Success in Subsequent Courses.

Information generated through adequate measurement assists in academic guidance. Such information can be used to predict success in subsequent courses. For example, in a university, it is recommended that MAT 111 (A Calculus Course) must be passed to a certain degree before a student takes MAT 211(ordinary differential equation). It is assumed in that university that MAT 111 may be quite helpful in the prediction of a student's success in MAT

211. Such prediction is expected to be based on the following assumptions:

- Empirical evidence of the relationship
- Subsequent course does not alter in method, content, or students' learning characteristics.

There is need to exercise caution while using measurement data to predict future performance. Basically, a student that performs well in Biology is likely to do well in Physics. Is Biology score predicting achievement in Physics? The performance in both subjects might be as a result of some factors resident in the learner. Such factors are intelligence, special ability in test taking, artistic skills (ability to draw diagrams correctly) etc. Nevertheless, these factors do not rule out the facts that certain measure can predict other measure(s).

Directive to Subsequent Course

Content of learning materials should be presented in hierarchy to the learners in such a way that links are established. Units in a particular subject should be interdependent instead of independent presentations. The dilemma of the teacher in next class on where to start can be overcome through revision of the last summative test the learners wrote. This exercise serves as the bedrock on which learning activities in the next class are built. Measurement can only maintain its relevance if there is co-operation among the teachers. There is need to reach agreement on table of specification for preparing measuring instruments for learners at different levels on each subject. This is important because ordinary scores will be of little use to the next teacher in establishing a behavior-content point at which to begin instruction.

Feedback to Students

Students have some rights within any educational system. One of such rights is to be fed back after any test is taken. If the intention of giving a test to students is for feedback, then scores generated from measuring activities should be interpretable in such a way that their attention is directed to useful things they may do to make up for deficiencies.

There is need for every teacher to incorporate the decision of the individual test items as part of instructional strategy test for a variety of reasons.

Mchrens and Lehmann (1984) provide these reasons:

- To correct errors in thinking on the part of the student, some of these errors being the result of poor study habits or sloppy test taking skills;
- To motivate pupils to do better on subsequent test;
- To demonstrate to pupils those instructional objectives that are being stressed and tested so that they can organize their study habits accordingly; and
- To permit students to discuss why a particular answer is keyed as the correct answer, that is, to clear up possible misconceptions and misunderstanding they may have.

Mehrens and Lehmann (1984) emphasize that feedback is extremely useful in the teaching-learning process. Although it might be considered by both teachers and pupils as time consuming and detracting from the time available for learning new material, however, it is more important to learn the material covered well before moving on to new areas. As of now, there appears to be some disagreement as to the time when this feedback should be given. Some feel that it is more effective if done as soon as possible after the test has been given so as to correct errors immediately, before they have a chance to solidify in the pupils' minds. Others say that feedback can be delayed for up to one week without serious dilatory effects. The truth of the matter is measurement provides information for feedback.

Comparison of Outcomes of Different Groups

Measurement of outcomes of behavioural objective of different groups for the purpose of comparing is important. The groups may differ based on instructional methods, materials or types of students. A Mathematics teacher may be interested in students' group performance in Science, commercial and Arts classes. It is often times the express purpose of examinations to compare groups of students. In such a case it is rather obvious immediately that certain kinds of antecedents are needed to make useful interpretations possible. For example one must know the entry behavior of students in terms of ability and motivation. (Bloom et al 1971)

Programme Evaluation

Evaluation helps you find out what works and what does not. It provides a road map for your program to improve processes, improve participant outcomes, and have a bigger impact on the organization and the community. With Up Front's experience, the stakeholders can be sure that all the information you need is collected.

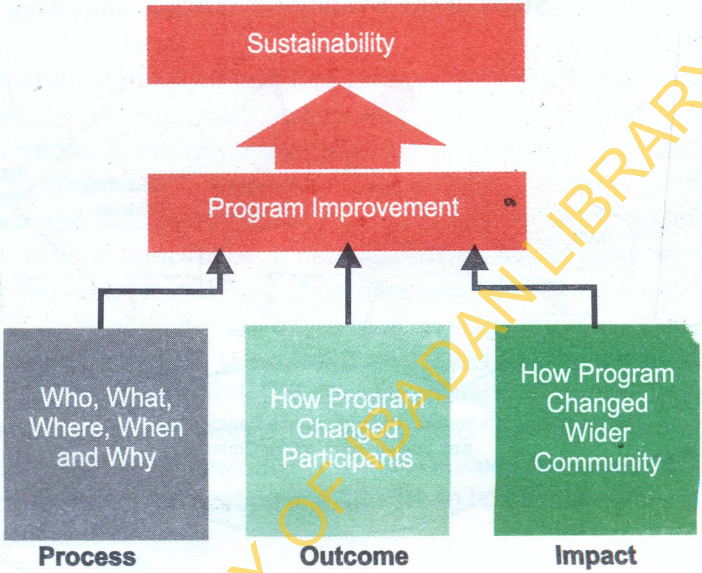

Sustainability
Program Improvement
Who, What, Where, When and Why
How Program Changed Participants
How Program Changed Wider Community
Process
Outcome
Impact

Figure 2 Up Front Evaluation Model 1

Process evaluation

Program evaluation starts with looking at the process. This is the who, what, where, when and why part of evaluation.

Outcomes evaluation

Outcomes evaluation measures changes in program participants. The question is, does your program in some way change the life of the people who participate? Outcomes are not always easy to measure. This information is critical to funders.

Impact evaluation

Impact evaluation looks at change beyond the program participants. Does your program help the bigger organization that includes your participants? Does it help the community? Though programs often lack the resources and expertise to measure impact, knowing the impact is very important in creating support for the program.

Steps in the Evaluation Process

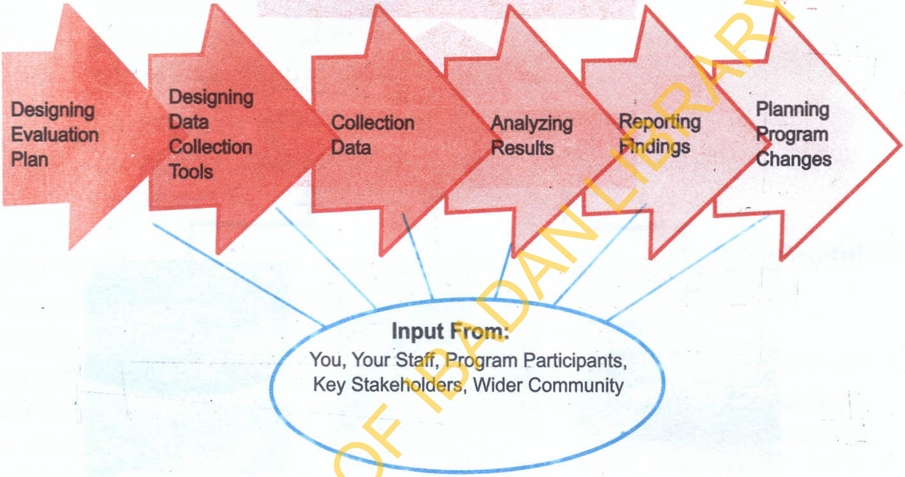

Designing Evaluation Plan
Designing Data Collection Tools
Collection Data
Analyzing Results
Reporting Findings
Planning Program Changes
Input From:
You, Your Staff, Program Participants, Key Stakeholders, Wider Community

Figure 3: Up Front Evaluation Model 2

Guiding Principles

The evaluators subscribe to the Guiding Principles for Evaluators, as defined by the American Evaluation Association:

- Systematic Inquiry—evaluators conduct systematic, data-based inquiries about what is being evaluated.
- Competence—evaluators provide competent performance to stakeholders.
- Integrity/Honesty—evaluators ensure the honesty and integrity of the entire evaluation process.
- Respect for people—evaluators respect the security, dignity, and self-

worth of the respondents, program participants, clients, and other stakeholders with whom they interact.

- Responsibilities for general and public welfare—evaluators articulate and take into account the diversity of interests and values that may be related to the general and public welfare.

Program Evaluation

Note that the concept of program evaluation can include a wide variety of methods to evaluate many aspects of programs in non-profit or for-profit organizations. There are numerous books and other materials that provide in-depth analysis of evaluations, their designs, methods, combination of methods and techniques of analysis. However, personnel do not have to be experts in these topics to carry out a useful program evaluation. The "20-80" rule applies here, that 20% of effort generates 80% of the needed results. It's better to do what might turn out to be an average effort at evaluation than to do no evaluation at all. (Besides, if you resort to bringing in an evaluation consultant, you should be a smart consumer. Far too many program evaluations generate information that is either impractical or irrelevant — if the information is understood at all.) This document orients personnel to the nature of program evaluation and how it can be carried out in a realistic and practical fashion.

What is Program Evaluation

First, "what is a program?" Typically, organizations work from their mission to identify several overall goals which must be reached to accomplish their mission. In non-profit, each of these goals often becomes a program. Non-profit programs are organized methods to provide certain related services to constituents, e.g., clients, customers, patients, etc. Programs must be evaluated to decide if the programs are indeed useful to constituents. In a for-profit, a program is often a one-time effort to produce a new product or line of products.

So, still, what is program evaluation? Program evaluation is carefully collecting information about a program or some aspect of a program in order to make necessary decisions about the program. Program evaluation can include any or a variety of at least 35 different types of evaluation, such as for needs assessments, accreditation, cost/benefit analysis, effectiveness, efficiency, formative, summative, goal-based, process, outcomes, etc. The type of evaluation you undertake to improve your programs depends on what you want to learn about the program. Don't worry about what type of evaluation you need or are doing — worry about what you need to know to make the program decisions you need to make, and worry about how you can accurately collect and understand that information.

Some Myths about Program Evaluation.

1. Many people believe evaluation is a useless activity that generates lots of boring data with useless conclusions. This was a problem with evaluations in the past when program evaluation methods were chosen largely on the basis of achieving complete scientific accuracy, reliability and validity. This approach often generated extensive data from which very carefully chosen conclusions were drawn. Generalizations and recommendations were avoided. As a result, evaluation reports tended to reiterate the obvious and left program administrators disappointed and skeptical about the value of evaluation in general. More recently (especially as a result of Michael Patton's development of utilization-focused evaluation), evaluation has focused on utility, relevance and practicality at least as much as scientific validity.

2. Many people believe that evaluation is about proving the success or failure of a program. This myth assumes that success is implementing the perfect program and never having to hear from employees, customers or clients again. The program will now run itself perfectly. This does not happen in real life. Success is remaining open to continuing feedback and adjusting the program accordingly. Evaluation gives you this continuing feedback.

3. Many believe that evaluation is a highly unique and complex process that occurs at a certain time in a certain way, and almost always includes the use of outside experts. Many people believe they must completely understand terms such as validity and reliability. They do not have to. They do have to consider what information they need in order to make current decisions about program issues or needs. And they have to be willing to commit to understanding what

is really going on. Note that many people regularly undertake some nature of program evaluation — they just don't do it in a formal fashion so they don't get the most out of their efforts or they make conclusions that are inaccurate (some evaluators would disagree that this is program evaluation if not done methodically). Consequently, they miss precious opportunities to make more of difference for their customers and clients, or to get a bigger bang for their buck.

Where Program Evaluation is Helpful

Program evaluation can:

1. Understand, verify or increase the impact of products or services on customers or clients - These "outcomes" evaluations are increasingly required by nonprofit funders as verification that the non-profits are indeed helping their constituents. Too often, service providers (for-profit or nonprofit) rely on their own instincts and passions to conclude what their customers or clients really need and whether the products or services are providing what is needed. Over time, these organizations find themselves in a lot of guessing about what would be a good product or service, and trial and error about how new products or services could be delivered.

2. Improve delivery mechanisms to be more efficient and less costly - Over time, product or service delivery ends up to be an inefficient collection of activities that are less efficient and more costly than need be. Evaluations can identify program strengths and weaknesses to improve the program.

3. Verify that you're doing what you think you're doing - Typically, plans about how to deliver services, end up changing substantially as those plans are put into place. Evaluations can verify if the program is really running as originally planned.

4. Facilitate management's really thinking about what their program is all about, including its goals, how it meets it goals and how it will know if it has met its goals or not.
5. Produce data or verify results that can be used for public relations and promoting services in the community.

6. Produce valid comparisons between programs to decide which should be retained, e.g., in the face of pending budget cuts.

7. Fully examine and describe effective programs for duplication elsewhere.

Basic Ingredients: Organization and Program(s)

This may seem too obvious to discuss, but before an organization embarks on evaluating a program, it should have well established means to conduct itself as an organization, e.g., (in the case of a non-profit) the board should be in good working order, the organization should be staffed and organized to conduct activities to work toward the mission of the organization, and there should be no current crisis that is clearly more important to address than evaluating programs.

To effectively conduct program evaluation, you should first have programs. That is, you need a strong impression of what your customers or clients actually need. (You may have used a needs assessment to determine these needs — itself a form of evaluation, but usually the first step in a good marketing plan). Next, you need some effective methods to meet each of those goals. These methods are usually in the form of programs.

It often helps to think of your programs in terms of inputs, process, outputs and outcomes. Inputs are the various resources needed to run the program, e.g., money, facilities, customers, clients, program staff, etc. The process is how the program is carried out, e.g., customers are served, clients are counseled, children are cared for, art is created, association members are supported, etc. The outputs are the units of service, e.g., number of customers serviced, number of clients counseled, children cared for, artistic pieces produced, or members in the association. Outcomes are the impacts on the customers or on clients receiving services, e.g., increased mental health, safe and secure development, richer artistic appreciation and perspectives in life, increased effectiveness among members, etc.

Planning Your Program Evaluation

Often, management wants to know everything about their products, services or programs. However, limited resources usually force managers to prioritize what they need to know to make current decisions.

Program evaluation plans depend on what information one needs to collect in order to make major decisions. Usually, management is faced with having to make major decisions due to decreased funding, ongoing complaints, unmet needs among customers and clients, the need to polish service delivery, etc. For example, do you want to know more about what is actually going on in your programs, whether your programs are meeting their goals, the impact of your programs on customers, etc? You may want other information or a combination of these. Ultimately, it's up to you.

But the more focused you are about what you want to examine by the evaluation, the more efficient you can be in your evaluation, the shorter the time it will take you and ultimately the less it will cost you (whether in your own time, the time of your employees and/or the time of a consultant).

There are trade offs, too, in the breadth and depth of information you get. The more breadth you want, usually the less depth you get (unless you have a great deal of resources to carry out the evaluation). On the other hand, if you want to examine a certain aspect of a program in great detail, you will likely not get as much information about other aspects of the program.

For those starting out in program evaluation or who have very limited resources, they can use various methods to get a good mix of breadth and depth of information. They can both understand more about certain areas of their programs and not go bankrupt doing so.

Key Considerations:

Consider the following key questions when designing a program evaluation.

1. For what purposes is the evaluation being done, i.e., what do you want to be able to decide as a result of the evaluation?
2. Who are the audiences for the information from the evaluation, e.g., customers, bankers, funders, board, management, staff, customers, clients, etc.
3. What kinds of information are needed to make the decision you need to make and/or enlighten your intended audiences, e.g., information to really understand the process of the product or program (its inputs, activities and outputs), the customers or clients who experience the product or program, strengths and weaknesses of the product or program, benefits to customers or clients (outcomes), how the product or program failed and why, etc.
4. From what sources should the information be collected, e.g., employees, customers, clients, groups of customers or clients and employees

together, program documentation, etc.

5. How can that information be collected in a reasonable fashion, e.g., questionnaires, interviews, examining documentation, observing customers or employees, conducting focus groups among customers or employees, etc.
6. When is the information needed (so, by when must it be collected)?
7. What resources are available to collect the information?

Major Types of Program Evaluation

When designing your evaluation approach, it may be helpful to review the following three types of evaluations, which are rather common in organizations. Note that you should not design your evaluation approach simply by choosing which of the following three types you will use — you should design your evaluation approach by carefully addressing the above key considerations.

Goal Based Evaluation

Often programs are established to meet one or more specific goals. These goals are often described in the original program plans.

Goal-based evaluations are evaluating the extent to which programs are meeting predetermined goals or objectives. Questions to ask yourself when designing an evaluation to see if you reached your goals, are:

1. How were the program goals (and objectives, is applicable) established? Was the process effective?
2. What is the status of the program's progress toward achieving the goals?
3. Will the goals be achieved according to the timelines specified in the program implementation or operations plan? If not, then why?
4. Do personnel have adequate resources (money, equipment, facilities, training, etc.) to achieve the goals?
5. How should priorities be changed to put more focus on achieving the goals? (Depending on the context, this question might be viewed as a program management decision, more than an evaluation question.)
6. How should timelines be changed (be careful about making these changes - know why efforts are behind schedule before timelines are changed)?
7. How should goals be changed (be careful about making these changes - know why efforts are not achieving the goals before changing the goals)?

Should any goals be added or removed? Why?

8. How should goals be established in the future?

Process-Based Evaluation

Process-based evaluations are geared to fully understanding how a program works — how does it produce that results that it does. These evaluations are useful if programs are long-standing and have changed over the years, employees or customers report a large number of complaints about the program, there appear to be large inefficiencies in delivering program services and they are also useful for accurately portraying to outside parties how a program truly operates (e.g., for replication elsewhere).

There are numerous questions that might be addressed in a process evaluation. These questions can be selected by carefully considering what is important to know about the program. Examples of questions to ask yourself when designing an evaluation to understand and/or closely examine the processes in your programs, are:

1. On what basis do employees and/or the customers decide that products or services are needed?
2. What is required of employees in order to deliver the product or services?
3. How are employees trained about how to deliver the product or services?
4. How do customers or clients come into the program?
5. What is required of customers or client?
6. How do employees select which products or services will be provided to the customer or client?
7. What is the general process that customers or clients go through with the product or program?
8. What do customers or clients consider to be strengths of the program?
9. What do staff consider to be strengths of the product or program?
10. What typical complaints are heard from employees and/or customers?
11. What do employees and/or customers recommend to improve the product or program?
12. On what basis do employees and/or the customer decide that the product or services are no longer needed?

Outcome Evaluation

Program evaluation with an outcomes focus is increasingly important for nonprofits and asked for by funders. An outcomes-based evaluation facilitates your asking if your organization is really doing the right program activities to bring about the outcomes you believe (or better yet, you've verified) to be needed by your clients (rather than just engaging in busy activities which seem reasonable to do at the time). Outcomes are benefits to clients from participation in the program. Outcomes are usually in terms of enhanced learning (knowledge, perceptions/attitudes or skills) or conditions, e.g., increased literacy, self-reliance, etc. Outcomes are often confused with program outputs or units of services, e.g., the number of clients who went through a program. To accomplish an outcomes-based evaluation, you should first pilot, or test, this evaluation approach on one or two programs at most (before doing all programs).

The general steps to accomplish an outcomes-based evaluation include to:

1. Identify the major outcomes that you want to examine or verify for the program under evaluation. You might reflect on your mission (the overall purpose of your organization) and ask yourself what impacts you will have on your clients as you work towards your mission. For example, if your overall mission is to provide shelter and resources to abused women, then ask yourself what benefits this will have on those women if you effectively provide them shelter and other services or resources. As a last resort, you might ask yourself, "What major activities are we doing now?" and then for each activity, ask "Why are we doing that?" The answer to this "Why?" question is usually an outcome. This "last resort" approach, though, may just end up justifying ineffective activities you are doing now, rather than examining what you should be doing in the first place.

2. Choose the outcomes that you want to examine, prioritize the outcomes and, if your time and resources are limited, pick the top two to four most important outcomes to examine for now.

3. For each outcome, specify what observable measures, or indicators, will suggest that you're achieving that key outcome with your clients. This is often the most important and enlightening step in outcomes-based

evaluation. However, it is often the most challenging and even confusing step, too, because you're suddenly going from a rather intangible concept, e.g., increased self-reliance, to specific activities, e.g., supporting clients to get themselves to and from work, staying off drugs and alcohol, etc. It helps to have a "devil's advocate" during this phase of identifying indicators, i.e., someone who can question why you can assume that an outcome was reached because certain associated indicators were present.

4. Specify a "target" goal of clients, i.e., what number or percent of clients you commit to achieving specific outcomes with, e.g., "increased self-reliance (an outcome) for 70% of adult, African American women living in the inner city of Minneapolis as evidenced by the following measures (indicators) ..."

5. Identify what information is needed to show these indicators, e.g., you'll need to know how many clients in the target group went through the program, how many of them reliably undertook their own transportation to work and stayed off drugs, etc. If your program is new, you may need to evaluate the process in the program to verify that the program is indeed carried out according to your original plans. (Michael Patton, prominent researcher, writer and consultant in evaluation, suggests that the most important type of evaluation to carry out may be this implementation evaluation to verify that your program ended up to be implemented as you originally planned.)

6. Decide how that information can be efficiently and realistically gathered. Consider program documentation, observation of program personnel and clients in the program, questionnaires and interviews about clients perceived benefits from the program, case studies of program failures and successes, etc. You may not need all of the above.

7. Analyze and report the findings.

Formative Evaluation

The gathering of data during the time the programme (e.g teaching learning process for a term, semester or session) is being developed for the purpose of guiding the developmental process is Formative Evaluation. According to Ayodele, Adegbile and Adewale (2008), formative evaluation also identifies strengths and weaknesses of a programme. It also gives feedback about an individuals and the extent to which each unit of instructional content is mastered. The intention of formative evaluation is to provide valid data that will provide reliable information useful enough to improve on an on-going programme. A tutor that gives short test at the end of a unit of instruction, scores the test and provide feedback to learners and relevant stakeholders such as parents has carried out formative evaluation. A person who is continually evaluating a programme will find many things that can be changed for the better during the operation of the programme. Outcome of formative evaluation is also useful for guiding and counseling of the learners.

Summative Evaluation

Making an overall assessment or decision with regards to the programme is a Summative Evaluation. This kind of evaluation always come at the end of a programme. Final examination on measurement and evaluation for students in Federal Training Centre, University College Hospital, Ibadan is an example of summative evaluation. Most often, summative evaluation takes the form of outcome evaluation where the success of the programme is being ascertained. Within the educational setting, summative evaluation is being carried out for the purpose of certification. In some situation, external bodies such as West African Examinations Council (WAEC), National Examinations Council (NECO) are engaged in carrying out summative evaluation of a programme. Summative examination can also take the form of terminal and semester examinations in secondary and tertiary institutions respectively. Summative evaluation is always in demand by sponsor or stakeholders at the end of any programme. According to *Garrison & Ehringhaus (2007)* the list is long, but here are some examples of summative assessments:

- State assessments
- District benchmark or interim assessments
- End-of-unit or chapter tests
- End-of-term or semester exams
- Scores that are used for accountability for schools and students (report card grades).

Ethics: Informed Consent from Program Participants

Note that if you plan to include in your evaluation, the focus and reporting on personal information about customers or clients participating in the evaluation, then you should first gain their consent to do so like in any other research. They should understand what you're doing with them in the evaluation and how any information associated with them will be reported. You should clearly convey terms of confidentiality regarding access to evaluation results. They should have the right to participate or not. Have participants review and sign an informed consent form.

Selecting Which Methods to Use

The overall goal in selecting evaluation method(s) is to get the most useful information to key decision makers in the most cost-effective and realistic fashion. Consider the following questions:

1. What information is needed to make current decisions about a product or program?
2. Of this information, how much can be collected and analyzed in a low-cost and practical manner, e.g., using questionnaires, surveys and checklists?
3. How accurate will the information be (reference the above table for disadvantages of methods)?
4. Will the methods get all of the needed information?
5. What additional methods should and could be used if additional information is needed?
6. Will the information appear as credible to decision makers, e.g., to funders or top management?
7. Will the nature of the audience conform to the methods, e.g., will they fill out questionnaires carefully, engage in interviews or focus groups, let you examine their documentations, etc.?
8. Who can administer the methods now or is a training required?
9. How can the information be analyzed?

Note that, ideally, the evaluator uses a combination of methods, for example, a questionnaire to quickly collect a great deal of information from a lot of people, and then interviews to get more in-depth information from certain respondents to the questionnaires. Perhaps case studies could then be used for more in-depth analysis of unique and notable cases, e.g., those who benefited or not from the program, those who quit the program, etc.

Four Levels of Evaluation

There are four levels of evaluation information that can be gathered from clients, including getting their:

1. reactions and feelings (feelings are often poor indicators that your service made lasting impact)
2. learning (enhanced attitudes, perceptions or knowledge)
3. changes in skills (applied the learning to enhance behaviors)
4. effectiveness (improved performance because of enhanced behaviors)

Usually, the farther your evaluation information gets down the list, the more useful is your evaluation. Unfortunately, it is quite difficult to reliably get information about effectiveness. Still, information about learning and skills is quite useful.

Analyzing and Interpreting Information

Analyzing quantitative and qualitative data is often the topic of advanced research and evaluation methods. There are certain basics which can help to make sense of reams of data.

Always start with your evaluation goals:

When analyzing data (whether from questionnaires, interviews, focus groups, or whatever), always start from review of your evaluation goals, i.e., the reason you undertook the evaluation in the first place. This will help you organize your data and focus your analysis. For example, if you wanted to improve your program by identifying its strengths and weaknesses, you can organize data into program strengths, weaknesses and suggestions to improve the program. If you wanted to fully understand how your program works, you could organize data in the chronological order in which clients go through your program. If you are conducting an outcomes-based evaluation, you can categorize data according to the indicators for each outcome.

Basic analysis of "quantitative" information (for information other than commentary, e.g., ratings, rankings, yes's, no's, etc.):

Reporting Evaluation Results

1. The level and scope of content depends on to whom the report is intended, e.g., to bankers, funders, employees, customers, clients, the public, etc.
2. Be sure employees have a chance to carefully review and discuss the report. Translate recommendations to action plans, including who is going to do what about the program and by when.
3. Bankers or funders will likely require a report that includes an executive summary (this is a summary of conclusions and recommendations, not a listing of what sections of information are in the report—that's a table of contents); description of the organization and the program under evaluation; explanation of the evaluation goals, methods, and analysis procedures; listing of conclusions and recommendations; and any relevant attachments, e.g., inclusion of evaluation questionnaires, interview guides, etc. The banker or funder may want the report to be delivered as a presentation, accompanied by an overview of the report. Or, the banker or funder may want to review the report alone.
4. Be sure to record the evaluation plans and activities in an evaluation plan which can be referenced when a similar program evaluation is needed in the future.

Contents of an Evaluation Plan

Develop an evaluation plan to ensure your program evaluations are carried out efficiently in the future. Note that bankers, donors or funders may want or benefit from a copy of this plan.

Ensure your evaluation plan is documented so you can regularly and efficiently carry out your evaluation activities. Record enough information in the plan so that someone outside of the organization can understand what you're evaluating and how. Consider the following format for your report:

1. Title Page (name of the organization that is being, or has a product/service/program that is being, evaluated; date)
2. Table of Contents
3. Executive Summary (one-page, concise overview of findings and recommendations)

4. Purpose of the Report (what type of evaluation(s) was conducted, what decisions are being aided by the findings of the evaluation, who is making the decision, etc.)

5. Background About Organization and Product/Service/Program that is being evaluated

   a) Organization Description/History

   b) Product/Service/Program Description (that is being evaluated)

   i) Problem Statement (in the case of non-profits, description of the community need that is being met by the product/service/program)

   ii) Overall Goal(s) of Product/Service/Program

   iii) Outcomes (or client/customer impacts) and Performance Measures (that can be measured as indicators toward the outcomes)

   iv) Activities/Technologies of the Product/Service/Program (general description of how the product/service/program is developed and delivered)

   v) Staffing (description of the number of personnel and roles in the organization that are relevant to developing and delivering the product/service/program)

6) Overall Evaluation Goals (eg, what questions are being answered by the evaluation)

7) Methodology

   a) Types of data/information that were collected

   b) How data/information were collected (what instruments were used, etc.)

   c) How data/information were analyzed

d) Limitations of the evaluation (eg, cautions about findings/conclusions and how to use the findings/conclusions, etc.)

8) Interpretations and Conclusions (from analysis of the data/information)

9) Recommendations (regarding the decisions that must be made about the product/service/program)

Appendices: content of the appendices depends on the goals of the evaluation report, eg.:

   a) Instruments used to collect data/information

   b) Data, eg, in tabular format, etc.

   c) Testimonials, comments made by users of the product/service/program

d) Case studies of users of the product/service/program

e) Any related literature

Pitfalls to Avoid

1. Don't balk at evaluation because it seems far too "scientific." It's not. Usually the first 20% of effort will generate the first 80% of the plan, and this is far better than nothing.
2. There is no "perfect" evaluation design. Don't worry about the plan being perfect. It's far more important to do something, than to wait until every last detail has been tested.
3. Work hard to include some interviews in your evaluation methods. Questionnaires don't capture "the story," and the story is usually the most powerful depiction of the benefits of your services.
4. Don't interview just the successes. You'll learn a great deal about the program by understanding its failures, drop-outs, etc.
5. Don't throw away evaluation results once a report has been generated. Results don't take up much room, and they can provide precious information later when trying to understand changes in the program.

CHAPTER THREE

Curriculum Evaluation

The primary objective of curriculum evaluation is the overall improvement of the student's education. Traditionally, curriculum evaluation has been limited to the appraisal of student's performance. There is a need to suggest that other areas of the curriculum should be evaluated to provide sufficient information with feedback in areas such as communication skills, problem solving, students' career choice and the influence of the curriculum in guiding the student in the evaluation and resolution of social and ethical issues.

Quantitative methods such as attitude scales and national exams, and qualitative methods such as interviews with school and direct observation of performance can be utilized in gathering the necessary data to determine the curriculum adequacy. Ideally, a combination of both methods should be used: quantitative methods to minimize error, ensure good sampling and control variables inherent to the particular training program; qualitative methods to ensure that the fluid,. dynamic, and real world nature of the instructional process is taken into account.

The following areas for curriculum evaluation should be explored, based on the premise that the purpose of the educational curriculum is to permit students to acquire both cognitive knowledge and skills adequate for the students' career development and to function in the society. It is also assumed that the curriculum does not simply provide knowledge, but encourages the practical application of that knowledge. Accordingly, curriculum should be measured not only from the student's point of view but also from the point of view of curriculum quality and implementation.

Evaluation of Student's Performance and Outcome as Influenced by the Curriculum

Knowledge Acquisition - Some form of testing for knowledge acquisition is done by most schools, but the form of test utilized varies depending upon the type of instruction philosophy utilized by that school. Brief description of the most commonly used methods will benefit evaluators that are in the stages of developing or modifying their current methods for curriculum evaluation.

- **Objective Testing -** This is probably the easiest of all the data to collect, and almost every program utilizes some form of objective evaluation. The multiple-choice examination permits internal comparison between different blocks and from year to year.

- **Subjective Testing -** These tests require testers with considerable experience in scoring these type of tests or with expertise in the subject tested. The greatest advantage of interpretative testing is that they test the student's holistic performance by determining the student's ability to apply factual knowledge to a learning situation. The disadvantages are that they are difficult to develop and standardize, are difficult to score, require scorers with a high degree of expertise and can be subject to scorer's bias. The latter can be partially obviated by having more than one examiner score the test.

What conclusions can be drawn regarding the usefulness of these two different approaches to test for knowledge acquisition? First of all, both methods are important but it should be recognized that they test for **different** aspects of knowledge acquisition. The choice of one over the other is mostly dictated by the type of teaching philosophy and by the availability and commitment of the school. Objective evaluations which are easy to develop and score are limited to test only of factual and recall knowledge. On the other hand, subjective evaluations, which test students by presenting them with situations which are close to real life conditions can be difficult to score and can be labor intensive. The choice of one over the other is not easy and to suggest only one of them will be an exercise in futility. This decision will be easier when all teachers of a subject adopt the same or similar teaching method. Until then, teachers should choose the method most appropriate for testing students in their program but strong consideration should be given to utilize a combined objective-subjective exam such as a Multiple Choice Test with Reasons Given, which includes open-ended thinking with objective testing. Regardless of the method used, the information obtained from the student's examination should be analyzed in a way that will permit modification of the curriculum in order to cover or expand areas in which students demonstrate deficiencies.

Career Differentiation - A well-designed curriculum should provide student contact with a full range of courses in the labour market and will permit exposure of students to role models that show career satisfaction and professional self-esteem. The adequacy of the curriculum in influencing students to go for certain courses can be.

Quality of Curriculum Evaluation

When evaluating courses of instruction, most evaluators focus on three specific areas: Program, Process, and Participants.

Program evaluation involves a critical look at the content, goals, objectives, and evaluation methods of a course. The usual tool utilized is a questionnaire completed by the students at the end of their rotation in which different aspects of instruction and student experiences are evaluated. The results of this questionnaire are often the only evaluation utilized to make changes in curriculum content and in curriculum implementation. The subjective nature of this questionnaire cannot be overemphasized; we are the ones who develop the questionnaire, ask the questions that we believe are important and those questions are answered by students with not much experience or information as to what a student's needs are. This is however, only the tip of the iceberg. Serious evaluators will also look at the content of their curriculum. Questions that should be asked include:

1. What are the overall goals of the curriculum?
2. What are the objectives and how do students reach them?
3. Are stated objectives relevant to the learner in the real world?
4. What is the quality of teaching?
5. Has the test bank of questions been reviewed for content and compatibility with what is being taught?
6. How effective is the admissions policy in attracting good learners?

Process evaluation refers to the analysis of the way the program is implemented. Questions to ask here include:

1. What characteristics of the learner are stressed: knowledge, problem-solving, self-learning, cost consciousness, or cultural sensitivity?
2. Do students receive feedback on their performance, and when?
3. What is the quality of the teaching and how is it measured?
4. What is the quality of the textbook(s) used?
5. How much does the school become involved in decision making?
6. Is the focus of the curriculum knowledge, skills, or attitudes? An often overlooked aspect of process evaluation is whether or not teaching time

is considered valuable and rewarded. Standardized questionnaires are available that will, in general, look at aspects of curriculum implementation and the learning environment. Modifications of the above source with inclusion of particular details from the individual evaluator can provide similar but much less biased data than that obtained by the previous example.

Participant evaluation includes an analysis of the attitudes and performance of students and school. Questions to ask are: How satisfied are participants with the curriculum? What is the performance (knowledge, skill acquisition and attitudes) of students who finish the course? What are the teachers' views on teaching and how do they see themselves as teachers (facilitator, lecturer, mentor)? What is the amount of school time devoted to teaching? Measurement of the outcomes of school graduates is also a part of participant evaluation. A review of all graduates' career choices, level of preparedness for next academic engagement or world of work, recent drop-out rates, certification and re-certification results, practice types and locations, and practice surveys can all be used to measure outcomes of the curricular plan.

Summary of Curricular Evaluation

Program evaluation tools

- Review of goals and objectives for relevance
- Review of teaching quality (direct observation and questionnaire)
- Student admission data
- Accreditation reports
- Alumni surveys
- Curriculum mapping

Process evaluation tools

- Questionnaires to assess attitudes
- Direct observations of the learning environment
- Interviews with school and students
- Debriefing sessions with students at end of course

- Review of test questions for validity and reliability

Participant evaluation tools

- Objective and subjective testing of students
- Grade distributions
- Feedback sessions for students and school.
- Peer evaluation
- Career differentiation
- Outcome studies
- Attitudes toward social responsibility

CHAPTER FOUR

Basics of Research

Concept of Research

Research can be defined as the search for knowledge or any systematic investigation to establish facts. The primary purpose for applied research (as opposed to basic research) is discovering, interpreting, developing methods and systems for the advancement of human knowledge on a wide variety of scientific matters of our world and the universe. Research can use the scientific method, but need not do so. The term *research* is also used to describe an entire collection of information about a particular subject.
Research is about finding out. It is about searching systematically for solutions to problems. It is about rules to guide search. It is also about helping to evaluate the research of others.

The term "research" has several meanings:

- Research is a systematic, formal rigorous and precise process employed to gain solutions to problems and/or to discover and interpret new facts and relationships. (Waltz and Bausell, 1981, p.1).
- Research is the process of looking for a specific answer to a specific question in an organized objective reliable way (Payton, 1979, p.4)
- Research is systematic, controlled, empirical and critical investigation of hypothetical propositions about the presumed relations among natural phenomena (Kerlinger, 1973, p.1).

Scientific research relies on the application of the scientific method, a harnessing of curiosity. This research provides scientific information and theories for the explanation of the nature and the properties of the world around us. It makes practical applications possible. Scientific research is funded by public authorities, by charitable organizations and by private groups, including many companies. Scientific research can be subdivided into different classifications according to their academic and application disciplines.

Artistic research, also seen as 'practice-based research', can take form when creative works are considered as both the research and the object of research itself. It is the debatable body of thought which offers an alternative to purely scientific methods in research in its search for knowledge and truth.

Research Writing

Scientific research

Generally, research is understood to follow a certain structural process. Though step order may vary depending on the subject matter and researcher, the following steps are usually part of most formal research, both basic and applied:

- Formation of the topic
- Background to the study
- Hypothesis
- Conceptual definitions
- Operational definition
- Review of related literature
- Gathering of data
- Analysis of data
- Test, revising of hypothesis
- Conclusion, iteration if necessary

Formation of the topic

A research proposal is expected to be prepared by any researcher before embarking on the study. Such proposal needs to be assessed by a project supervisor or the sponsor of a project. Much effort should be directed to getting a researchable topic. Searching for a good and appealing research topic is worthwhile. This is very important as a good title can be an eye catcher and will help in attracting readers and making them go through the study. A well written thesis but with an unappealing, unattractive title will find no takers and so will be rendered useless. A title relevant to the study and not too long will have to be thought of. The title you choose to work on should be interesting to the reader and researcher both. The topic should be such that there is enough research material available to work on. A vague and obscure topic would not be a good idea to work on, while a topic with too much of available material to work on, would also confuse the writer. A topic that is too broad, like the effects of teaching strategy on students' cognitive achievement seems to be indefinite. Major variables to be investigated should be reflected on the title. The title page usually consists of other information such, the names of the researcher and Supervisor(s); student number; department etc. Examples of Researchable Topics are:

(i) Impact of year of training and Communication Skill on job Effectiveness of University Lecturers in Nigeria.

(ii) The Complimentary Roles of Religious Institutions on the Fight Against Corruption in Nigeria.

A prudent researcher will consistently improve on the proposal until the study is completed.

Background to the study

Background to the study is part of Introduction (the heading of Chapter One) of any Dissertation or Thesis. Background to the study should emphasise the importance of the study and describe the research topic or theme in detail to justify the need for the study. The researcher's knowledge of the intended study has to be displayed. It is expected that the researcher supports his ideas with past studies; however this section must not be turned to mere review of existing literature. In all cases, it should be stated whether a relationship exists between the proposed research and research undertaken before. If no such research has been undertaken previously, this should be pointed out. To the end of this section, there is need to raise a major question that will direct the study. The question in the introduction is also very important as it is around this question that the statement of problem of the study is framed. To find a good question one has to be very familiar with the chosen topic. Having a good in depth knowledge on the topic will help to know the problems related to the topic, which will assist in framing a major question the study provides a answer to and finally lead to the thesis statement.

To develop a strong background to the study, the following must be considered:

- knowledge of the field and its literature;
- important research questions in the field;

- areas that need further exploration;
- the gap the proposed study will fill;
- great deal of research already been conducted in the topic area,
- identification of room for improvement in a research area;
- recency and relevance of the title. Is it a hot title, or obsolete one?
- fund;
- time;
- personnel
- disposition of target population to the proposed study; and
- significant impact of the study on the field.

Statement of Problem

This should come immediately the researcher has sufficiently justified the reason(s) for the study. There is need to present, as clearly as possible, the source of interest in the topic or theme. There is need to pose a problem that is visible to readers and indicate how the proposed research can provide solution to the problem raised. An Example of Statement of problem is presented below:

> *"The general achievement of students in mathematics nowadays is not encouraging. Available records from WAEC Chief Examiner reports of 2003,2004 2008 indicated that students' performance in mathematics is falling. Researchers need to provide explanation on such performance using essential variables that would make such explanation meaningful. Hence the study "Understanding Variation in Human Capital Development Through Path Analytic study of Gender, Mathematics Conception, Manipulative Skills, Learning readiness and Students' Achievement in Mathematics".*

Objectives

The objectives of the study in line with statement of problem should be clear, precise, and concise. Each objective should carry action word that gives direction to the study.

Objectives stated for a published paper is presented thus:

> *Based on the above background, the following form the objectives of the study; the study seeks to:*
> *-investigate corrupt tendencies in Nigeria*
> *-find out the effect of closeness to religious leaders on Nigerians' fight against Corruption.*
> *-investigate the impact of religious association membership on Nigerians' fight against Corruption.*
> *-establish the effect of Nigerians' involvement in religious activities on their fight against Corruption.*

Research Questions and Hypotheses

Write down specific questions you want answered. Each question must relate back to the main objective. Testable hypotheses can also be formulated in line with the main objectives. It is expected that these questions have been answered and the hypotheses tested the main objectives should be met. A researcher may stick to either research questions, hypotheses or both, however hypothesis is not expected to be a duplication of a research question or *vice versa*. Hypothesis is not always necessary for qualitative research. Research questions and hypotheses used to direct a completed study are thus presented:

> *The following research questions/hypotheses shall be ıswered/tested based on the stated objectives above.*
>
> ***Questions***
> *(1) How aggressive are Nigerians on their fight against corruption?*
> *(2) Is the aggressive fight of Nigerians against corruption a function of the quality of moral instruction received from affiliated religious group?*
>
> ***Hypotheses***
> *(1) Aggressive fight of Nigerians against corruption will not be a function of their closeness to religious Leaders?*
> *(2) Commitment to religious activities will not bring about aggressive fight against corruption in Nigeria?*

Conceptual/Operational definitions

Define key concepts and terms to clear up ambiguities and obscurities. The concepts clarified for the research proposal will eventually form part of the list of terms clarified for the research report.

When dictionary or literature definition of a word or phrase is given such is termed conceptual and when the definition is given based on the way the word or phrase is used within the context of a study, such is termed operational.

Literature Review

Literature survey and review should be undertaken and that enable the researcher to demarcate the research problem clearly. There should be sufficient evidence through literature review that relevant publications (books, legislation, documents, files, etc.) have been consulted to determine whether the envisaged research is viable and not a duplication of previous research. Review of literature is also important to adequately prepare for adequate discussion on the findings of the study.

Research methodology

This can be referred to as the strategy for research. It clearly indicates the design. Design must be appropriate for the type of data collected and type of questions and hypotheses raised. Methodology also indicates sample, methods of data collection, either within a quantitative or qualitative methodology; as well as the techniques for data collection, e.g. questionnaires, and measurement (the validation of the techniques). Indicate whether field workers will be used to collect data and whether computer programmes will be employed to analyse the data. The researcher should also indicate in this section of the study which strategies will be followed during the research (i.e. the actions and their sequence) .For example, a questionnaire will be constructed first, then the data collection and then data analysis techniques to be employed.

Result Presentation

The researcher is expected to analyze the data collected A computer will probably be helpful at this stage. Present the results clearly (Graphs and charts will be helpful), give detail interpretations directly from the results and discuss the findings of the study.

Conclusion

Conclude the Research Writing by giving implications of the research findings, recommendations for further study and conclusion.

Importance of Research

The purpose of research is to:

(1) inform action: thus, research should seek to contextualize its findings within the larger body of research. In most positions, some sort of research is required to support normal decision-making.
(2) produce knowledge that is applicable outside of the research setting with implications that go beyond the group that has participated in the research. Hence research must always be high quality.
(3) Have findings that should have implications for policy and project implementation.
(4) Fulfill requirements for course completion
(5) Earn recognition: publishing or sharing information via research publications or public meetings provides visibility and recognition.
(6) Satisfy curiosity: Curiousity is a crucial part of the human condition. Many professionals, including information ones, want to know more about something that interests them. Do religious leaders support the fight against corruption? Which instructional strategy will best support students Mathematics learning? There is an excitement in the discovery of new information and knowing more about some topic than anyone else. There is joy in sharing newly gathered and previously unavailable information.
(7) Form the foundation of program development and policies all over the world,
(8) Enhance how to best address the world's problems

Types of Research

Various authors present research type in different ways. The types herewith presented however are likely to cover a very wide range of research.

(1) **Applied Research** seeks the pecific knowledge necessary to improve the treatment of a particular disease or specific strategies to improve learning a subject.

(2) **Basic Biomedical Research** is conducted to increase understanding of fundamental life processes, such as discovering the molecular structure of deoxyribonucleic acid (DNA) — one-half of the genetic code of life — or investigating the genetics of lipid disease.

(3) **Basic Research** is a synonym for fundamental research, which is the study of life processes that are universal in their application to scientific knowledge.

(4) **Clinical Research** addresses important questions of normal function and disease using human subjects.

(5) **Directed Research** is conducted by an investigator in response to an outside request to explore a specific scientific area or question.

(6) **Fundamental Research** studies life processes that are universal in their application to scientific knowledge.

(7) **Investigator-Initiated Research** investigates a question or hypothesis that the researcher has defined.

(8) **Outcomes Research** focuses upon the end results of teaching strategies, tangible and quantifiable outcomes of the treatment upon learners and the determinants of these outcomes.

(9) **Population Research** is the science and art of studying the distribution and determinants of a phenomenon as influenced by social, economic and physical environments, human biology, health policy and services at the population levels.

(10)**Strategically Focused Research** focuses on science areas that the association has determined are important to achieving its mission and strategic objectives.

(11)**Targeted Research** is a synonym for directed research.

(12)**Translational Research** takes a result from basic or fundamental science and studies its applicability in the human situation. For example a translational research addresses the adoption of teaching strategies that have been demonstrated to be effective through educational research in the instructional deliveries to the students to teach important topic like mathematics

Classification of Research

Research studies can be broadly classified to two namely qualitative and **quantitative** research. The major categories can however be further classified into six sub categories.

Qualitative approach: The qualitative approach involves the collection of extensive narrative data in order to gain insights into phenomena of interest; data analysis includes the coding of the data and production of a verbal synthesis (inductive process)

1. Historical research
2. Qualitative research

Quantitative approach: The quantitative approaches involve the collection of numerical data in order to explain, predict, and/or control phenomena of interest; data analysis is mainly statistical (deductive process)

3. Descriptive research
4. Correlational research
5. Causal-comparative research
6. Experimental Research

Qualitative research approaches

Historical research and qualitative research are the two types of research classified as qualitative research approaches.

Historical research is involved with the study of past events.

The following are some examples of historical research studies:

1. Factors leading to militancy activities in the Niger-Delta Region.
2. Effect of Amnesty on the development of Niger-Delta Region.
3. Factors leading to the development and growth of Mastery learning between 1980 and 2000.

Qualitative research, also referred to as ethnographic research, is involved in the study of current events rather than past events. It involves the collection of extensive narrative data (non-numerical data) on many variables over an extended period of time in a naturalistic setting. Participant observation, where the researcher lives with the subjects being observed is frequently used in qualitative research. Case studies are also used in qualitative research.

Some examples of qualitative studies are:

4. A case study of parental involvement child career development.
5. Class interaction of teachers and students in public secondary school mathematics classes.
6. Pattern of Assignment marking in science class

Quantitative research approaches

Descriptive research involves collecting data in order to answer questions regarding the subjects of the study. In contrast with the qualitative approach the data are numerical. The data are typically collected through a questionnaire, an interview, or through observation.

In descriptive research, the investigator reports the numerical results for one or more variables on the subjects of the study.

Some examples of descriptive research studies are:

7. How do public secondary teachers spend their after school hours?
8. How will rural secondary students use the internet facility?
9. How do parents feel about provision of mid-day meal provided by the government?

Correlational research attempts to determine whether and to what degree, a relationship exists between two or more quantifiable (numerical) variables. However, it is important to note that significant relationship between two variables does not follow that one variable causes the other. Relationship does not imply causality. When two variables are correlated you can use the relationship to predict the value on one variable for a subject if you know that subject's value on the other variable. Correlation implies prediction but not causation. The investigator frequently uses the correlation coefficient to report the results of correlational research.

Some examples of correlational research are:

10. The relationship between Cognitive achievements in Physics and Mathematics.
11. The relationship between attitude and Mathematics achievement.
12. The use of spatial aptitude test to predict success in an algebra course.

Causal-comparative research attempts to establish cause-effect relationships among the variables of the study. The attempt is to establish that values of the independent variable have a significant effect on the dependent variable. This type of research usually involves group comparisons. The groups in the study make up the values of the independent variable, for example gender (male versus female), pre school attendance versus no pre school attendance, or children with a working mother versus children without a working mother. These could be the independent variables for the sample studies. However, in causal-comparative research, the independent variable is not under the experimenter's control, that is, the experimenter can't randomly assign the subjects to a gender classification (male or female) but has to take the values of the independent variable as they come. The dependent variable in a study is the outcome variable.

Here are some examples of causal-comparative research studies:

13. The effect of mid-day meal on pupil attendance in pre-primary schools.
14. The effect of teaching strategies on students achievement in Chemistry.
15. The effect of sex (gender) on Geometry achievement.

Experimental research like causal-comparative research attempts to establish cause-effect relationship among the groups of subjects that make up the independent variable of the study, but in the case of experimental research, the cause (the independent variable) is under the control of the experimenter. That is, the experimenter can randomly assign subjects to the groups that make up the independent variable in the study. In the typical experimental research design the experimenter randomly assigns subjects to the groups or conditions that constitute the independent variable of the study and then measures the effect this group membership has on another variable, i.e. the dependent variable of the study.

The following are some examples of experimental research mentioned by Gay.

16. The comparative effectiveness of personalized instruction versus traditional instruction on computational skill.
17. The effect of self-paced instruction on self-concept.
18. The effect of positive reinforcement on attitude toward school.

The Place of Literature in Research:

What is a literature review?

According to Cooper (1988) '... a literature review uses as its database reports of primary or original scholarship, and does not report new primary scholarship itself. The primary reports used in the literature may be verbal, but in the vast majority of cases reports are written documents. The types of scholarship may be empirical, theoretical, critical/analytic, or methodological in nature. Second a literature review seeks to describe, summarise, evaluate, clarify and/or integrate the content of primary reports.'

The review of relevant literature is nearly always a standard chapter of a thesis or dissertation. The review forms an important chapter in a thesis where its purpose is to provide the background to and justification for the research undertaken (Bruce, 1994). Bruce, who has published widely on the topic of the literature review, has identified six elements of a literature review. These elements comprise a list; a search; a survey; a vehicle for learning; a research facilitator; and a report (Bruce, 1994).

Why do a literature review?

A crucial element of all research degrees is the review of relevant literature. So important is this chapter that its omission represents a void or absence of a major element in research (Afolabi 1992). According to Bourner (1996) there are good reasons for spending time and effort on a review of the literature before embarking on a research project. These reasons include:

- to identify gaps in the literature
- to avoid reinventing the wheel (at the very least this will save time and it can stop you from making the same mistakes as others)
- to carry on from where others have already reached (reviewing the field allows you to build on the platform of existing knowledge and ideas)
- to identify other people working in the same fields (a researcher network is a valuable resource)
- to increase your breadth of knowledge of your subject area
- to identify seminal works in your area
- to provide the intellectual context for your own work, enabling you to position your project relative to other work
- to identify opposing views
- to put your work into perspective
- to demonstrate that you can access previous work in an area
- to identify information and ideas that may be relevant to your project
- to identify methods that could be relevant to your project

Variables

According to Spiegel and Stephens (1999), a variable is a symbol, such as X, Y, H, or B that can assume any of a prescribed set of values, called the domain of the variable. If the variable can assume only one value, it is called a constant. A *variable* is something that can change, such as 'academic qualification' and are typically the focus of a study. *Attributes* are sub-values of a variable, such as 'NCE', 'HND' First Degree etc. An *exhaustive* list contains all possible answers. An attitudinal item may have an exhaustive list presented below:

1 = strongly disagree
2 = disagree
3 = agree
4 = strongly agree

The role a variable plays within the context of a study will determine how it should be measured. For example if age plays the role of Independent variable in a study, the best approach to measure it should be by ordinal scale: 12-15 years, 16-19 years etc. and not by Interval Scale: 12, 13, 14, 15, 16, 17, 18 etc. However if parametric statistical tool e.g ANOVA is to be used and age serves as Dependent Variable the approach for measuring age will be the other way round i.e measure by interval scale.

Mutually exclusive attributes are those that cannot occur at the same time. Thus in a survey a person may be requested to select one answer from a list of alternatives (as opposed to selecting as many that might apply to the respondent).

Variables aren't always 'quantitative' or numerical. The variable 'gender' consists of two text values: 'male' and 'female'. We can, if it is useful, assign quantitative values instead of (or in place of) the text values, but we don't have to assign numbers in order for something to be a variable. It's also important to realize that variables aren't only things that we measure in the traditional sense. For instance, in much social research and in program evaluation, we consider the treatment or program to be made up of one or more variables (i.e., the 'cause' can be considered a variable). An educational program can have varying amounts of 'time on task', 'classroom settings', 'student-teacher ratios', and so on. So even the program can be considered a variable (which can be made up of a number of sub-variables-treatment levels).

Another important distinction having to do with the term 'variable' is the distinction between an ***independent*** and ***dependent*** variable. This distinction is particularly relevant when you are investigating cause-effect relationships. **Independent variable** is the variable manipulated by the researcher while the Dependent variable is what is affected by the independent variable. For example, if you are studying the effects of a new Teaching Strategy on student achievement in a Psychology course, the Teaching Strategy is the independent variable and your measure of achievement is the dependent variable.

Finally, there are two traits of variables that should always be achieved. Each variable should be ***exhaustive,*** it should include all possible answerable responses. For instance, if the variable is "Marital Status" and the only options are "Single", "Married", and "Divorced", there are quite a few categories I can think of that haven't been included. The list does not exhaust all possibilities. On the other hand, if you exhaust all the possibilities with some variables marital status being one of them you would simply have too many responses especially in these modern days when commitment to marriage is getting low. The way to deal with this is to explicitly list the most common attributes and then use a general category like "Other" to account for all remaining ones. In addition to being exhaustive, the attributes of a variable should be *mutually exclusive*, no respondent should be able to have two attributes simultaneously. While this might seem obvious, it is often rather tricky in practice. For instance, you might be tempted to represent the variable "Academic Qualification" with the attributes like "O Level" "NCE" "OND" "HND" "First Degree" "Master" etc. But these attributes are not necessarily mutually exclusive. A person who has both NCE and First degree certificates will check both attributes on the list. Best approach to handle this is to ask the respondent to "check all that apply" and then list a series of categories? Yes, we do, but technically speaking, each of the categories in a question like that is its own variable and is treated dichotomously as either "checked" or "unchecked", attributes that *are* mutually exclusive.

Types of Variable

Descriptive variables: are those variables that will be reported on, without relating them to anything in particular.

Categorical variables: result from a selection from categories, such as 'agree' and 'disagree'. Nominal and ordinal variables are categorical.

Numeric variables: give a number, such as age, achievement score, income etc..

Discrete variables: are numeric variables that come from a limited set of numbers. They may result from, answering questions such as 'number of Children', 'how often do you eat daily?', etc.

Continuous variables: are numeric variables that can take any value, such as weight, Height etc.

Ethics of Research

The ethical issues in human subjects research have received increasing attention over the last 50 years. Institutional Review Boards for the Protection of Human Subjects (IRB's) have been established at most institutions that undertake research with humans. These committees are made up of scientists, clinical faculty, and administrators who review research according to the procedures set out in the Federal Regulations. There are several ethical issues that must be considered when designing research that will utilize participants who are human beings.

- The primary concern of the investigator should be the safety of the research participant. This is accomplished by carefully considering the risk/benefit ratio, using all available information to make an appropriate assessment and continually monitoring the research as it proceeds.
- The scientific investigator must obtain informed consent from each research participant. This should be obtained in writing (although oral consents are sometimes acceptable) after the participant has had the opportunity to carefully consider the risks and benefits and to ask any pertinent questions. Informed consent should be seen as an ongoing process, not a singular event or a mere formality.
- The investigator must enumerate how privacy and confidentiality concerns will be approached. Researchers must be sensitive to not only

how information is protected from unauthorized observation, but also if and how participants are to be notified of any unforeseen findings from the research that they may or may not want to know.

- The investigator must consider how adverse events will be handled; who will provide care for a participant injured in a study and who will pay for that care are important considerations.
- In addition, before enrolling participants in an experimental trial, the investigator should be in a state of "equipoise," that is, if a new intervention is being tested against the currently accepted treatment, the investigator should be genuinely uncertain which approach is superior. In other words, a true null hypothesis should exist at the onset regarding the outcome of the trial.

Ethical Principles that Govern Research with Human Subjects

There are three primary ethical principles that are traditionally cited when discussing ethical concerns in human subjects research. (A more complete enumeration of these principles is available in the *Belmont Report*, written by The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research in 1979.)

- Autonomy, which refers to the obligation on the part of the investigator to respect each participant as a person capable of making an informed decision regarding participation in the research study. The investigator must ensure that the participant has received a full disclosure of the nature of the study, the risks, benefits and alternatives, with an extended opportunity to ask questions. The principle of autonomy finds expression in the informed consent document.
- Beneficience, which refers to the obligation on the part of the investigator to attempt to maximize benefits for the individual participant and/or society, while minimizing risk of harm to the individual. An honest and thorough risk/benefit calculation must be performed.
- Justice, which demands equitable selection of participants, i.e., avoiding participant populations that may be unfairly coerced into participating, such as prisoners and institutionalized children. The principle of justice also requires equality in distribution of benefits and burdens among the population group(s) likely to benefit from the research.

Components of an Ethically Valid Informed Consent for Research

For an informed consent to be ethically valid, the following components must be present:

Disclosure: The potential participant must be informed as fully as possible of the nature and purpose of the research, the procedures to be used, the expected benefits to the participant and/or society, the potential of reasonably foreseeable risks, stresses, and discomforts, and alternatives to participating in the research. There should also be a statement that describes procedures in place to ensure the confidentiality or anonymity of the participant. The informed consent document must also disclose what compensation and medical treatment are available in the case of a research-related injury. The document should make it clear whom to contact with questions about the research study, research subjects' rights, and in case of injury.

- **Understanding:** The participant must understand what has been explained and must be given the opportunity to ask questions and have them answered by one of the investigators. The informed consent document must be written in lay language, avoiding any technical jargon.
- **Voluntariness:** The participant's consent to participate in the research must be voluntary, free of any coercion or promises of benefits unlikely to result from participation.
- **Competence:** The participant must be competent to give consent. If the participant is not competent due to mental status, disease, or emergency, a designated surrogate may provide consent if it is in the participant's best interest to participate. In certain emergency cases, consent may be waived due to the lack of a competent participant and a surrogate.
- **Consent:** The potential human subject must authorize his/her participation in the research study, preferably in writing, although at times an oral consent or assent may be more appropriate.

CHAPTER FIVE

Research Tools: Achievement and Aptitude Tests

Tools of Research refer to instruments researchers use to collect data for research studies. They are alternatively called "tests". This chapter focuses Achievement and Aptitude Tests.

Achievement Test

An achievement test should be measuring fully the status of the individual in all the hierarchical levels of understanding as proposed in the Bloom's taxonomy of educational objectives. The test should measure:

i. Recall of information (knowledge),

ii. Understand the meaning, translation, interpolation, and interpretation of instructions and problems. State a problem in one's own words (comprehension),

iii. Use a concept in a new situation or unprompted use of an abstraction. Applies what was learned in the classroom into novel situations in the workplace. (application),

iv. Separates material or concepts into component parts so that its organizational structure may be understood. Distinguishes between facts and inferences (analysis),

v. Builds a structure or pattern from diverse elements. Put parts together to form a whole, with emphasis on creating a new meaning or structure. (synthesis), and

vi. Make judgments about the value of ideas or materials (evaluation).

Test experts have classified achievement tests using different parameters. Whereas, some classify tests on the basis of the behaviour that is being measured, others classify considering the types of items contained in the test, the purpose of tests, etc. However, achievement tests may be classified on the basis of the **essay-type** and the **objective-type.** The two major types are presented in table below:

Types of Achievement Test

Essay	Objectives	Others
1. Extended Response 2. Restricted Response	1.Fill in a. Short Answer b. Completion 2. Selective Type a. True or False b. Matching Items c. Multiple Choice Question	1. Oral 2. Student Portfolios 3. Performance

Essay Tests

The essay test has been a very popular type of achievement test. It is a test that allows the testees apply their ideas on the items in a personal way. The two forms of essay test are Extended Response and Restricted or Short-Answer Tests.

The Extended Response

In this type of essay test, the testee answers a small number of items. The examiner is expected to develop a valid marking Scheme to award marks on each item. The testees may be instructed to answer all the questions or to choose out of the number of items given. Instructions that guide the testees on how to write the test is very important.

Two examples of the extended response type are:

1. Differentiate between Bar chart and Histogram.
2. Discuss methods of establishing Reliability Coefficient of Achievement Test.

Extended Response test is good for:

- Application, synthesis and evaluation levels

Advantages:

- Students less likely to guess
- Easy to construct
- Stimulates more study
- Allows students to demonstrate ability to organize knowledge, express opinions, show originality.

Disadvantages:

- Can limit amount of material tested, therefore has decreased validity.
- Subjective, potentially unreliable scoring.
- Time consuming to score.

Tips for Writing Good Extended Response Items:

- State the instructional objectives in specific terms,
- Outline the course content,
- Prepare the table of specification (test blue print),
- Provide reasonable time limits for thinking and writing.
- Avoid letting them to answer a choice of questions (Instruct all the testees to answer the same set of items.)
- Give definitive task to student-compare, analyze, evaluate, etc.
- Use checklist point system to score with a model answer: write outline, determine how many points to assign to each part
- Score one question at a time-all at the same time.

The Restricted or Short-Answer Essay

In this kind of test, the candidate is given a number of questions to respond briefly to. It limits both the content and the type of learner's response. The following are examples of such a question:

1. (i) Mention any four functions of Achievement tes
   (iii) Write about six sentences on any 3 of them.

2. List the two major type of Achievement test and Discuss only one of them.

Short Answer items are good for:

- Application, synthesis, analysis, and evaluation levels

Advantages:

- Easy to construct
- Minimizes guessing
- Encourages more intensive study-student must know the answer vs. recognizing the answer.

Disadvantages:

- May overemphasize memorization of facts
- Testees may give different type of answers on an item.
- Scoring is laborious

Tips for Writing Good Short Answer Items:

- State the instructional objectives in specific terms,
- Outline the course content,
- Prepare the table of specification (test blue print),
- Provide reasonable time limits for thinking.
- For numbers, indicate the degree of precision/units expected.

Objective Test

It is a test consisting of factual questions requiring extremely short answers that can be quickly and unambiguously scored by anyone with an answer key, thus minimizing subjective judgments by both the person taking the test and the person scoring it.

Types of Objective Test Items

Objective tests are of various types. The commonly used among the types of Objective test are:

a. the true-false type,
b. the fill-in type , i.e. short answer or completion,
c. the matching type, and
d. the multiple choice type,

The True-False Type

In this type of test, the testee is given some statements to which he should respond. The statements have to be marked as either "true" or false". Let us consider the following examples:

i. Triangle is an example of polygon. (True or False).
ii. State capital of Rivers is Port Harcourt. (True or False).

True/False items are good for:

- knowledge level content
- evaluating student understanding of popular misconceptions
- concepts with two logical responses

Advantages:

- they can test large amounts of content
- students can answer 3-4 questions per minute

Disadvantages:

- They are easy to construct
- It is difficult to discriminate between students that know the material and students who do not
- Students have a 50-50 chance of getting the right answer by guessing
- Need a large number of items for high reliability

Tips for Writing Good True/False items:

- Avoid double negatives.
- Avoid long/complex sentences.
- Use specific determinants with caution: neve only, all, none, always, could, might, can, may, sometimes, generally, some, few.
- Use only one central idea in each item.
- Don't emphasize the trivial.
- Use exact quantitative language
- Don't lift items straight from the book.
- Make more false than true (60/40). (Students are more likely to answer true.)

The Fill-in Type (Short Answer or Completion)

This is another simple type of objective test where a testee is expected to provide short answers or complete some statements.

Examples.

1. What is the name of a side of a triangle that is opposite a right angle?
2. The name of a side of a triangle that is opposite a right angle is ....

Fill-in type of Objective test items are good for:

- Knowledge level content
- Evaluating student understanding of popular misconceptions

Advantages:

- Easy to construct
- Good for recalling date, idea, fact etc.
- Minimizes guessing
- Encourages more intensive study-student must know the answer.
- Scoring is easy.

Disadvantages:

- May overemphasize memorization of facts
- Take care - questions may have more than one correct answer

Tips for Writing Good Fill-in type Items:

- When using with definitions: supply term, not the definition-for a better judge of student knowledge.
- For numbers, indicate the degree of precision/units expected.
- Use direct questions, not an incomplete statement.
- If you do use incomplete statements, don't use more than 2 blanks within an item.
- Arrange blanks to make scoring easy.
- Try to phrase question so there is only one answer possible.

The matching type

As the name denotes matching type presents two column containing domains of elements that can be independently assigned.

Example

Use straight lines to match the items below

Question Column	goes to	Answer Column
Circle		Diagonal
Rectangle		edges
Cube		Average
		Mode
		HCF
		LCM
		radius

Matching Items are good for:

- Knowledge level
- Some comprehension level, if appropriately constructed

Types:

- Terms with definitions
- Phrases with other phrases
- Causes with effects
- Parts with larger units
- Problems with solutions

Advantages:

- Maximum coverage at knowledge level in a minimum amount of space/prepare time
- Valuable in content areas that have a lot of facts

Disadvantages:

- Time consuming for students
- Not good for higher levels of learning

Tips for Writing Good Matching items:

- Need 15 items or less.
- Give good directions on basis for matching.
- Use items iņ response column more than once (reduces the effects of guessing).
- Use homogenous material in each exercise.
- Make all responses plausible.
- Put all items on a single page.
- Put response in some logical order (chronological, alphabetical, etc.).
- Responses should be short.

The multiple choice type

The multiple-choice (MC) item is one of the most popular item formats used in educational assessment. A typical MC item has three parts: a stem that presents a problem; the correct or best answer; and several distractors (i.e., the wrong or less appropriate options).

MC items can be constructed to assess a variety of learning outcomes, from simple recall of facts to Bloom's highest taxonomic level of cognitive skills – evaluation (Osterlind, 1998). It is common knowledge that the correct answers should be distributed evenly among the alternative positions of MC items, but there are many other important guidelines for writing good items. Eight steps though not exhaustive are provided as guidelines to construction of Multiple choice questions.

Multiple Choice items are good for:

- Application, synthesis, analysis, and evaluation levels

Types:

- Question/Right answer
- Incomplete statement
- Best answer

Advantages:

- Very effective
- Versatile at all levels

- Minimum of writing for student
- Guessing reduced
- Can cover broad range of content

Disadvantages:

- Difficult to construct good test items.
- Difficult to come up with plausible distractors/alternative responses.

Tips for Writing Good Multiple Choice items:

- Stem should present single, clearly formulated problem.
- Stem should be in simple, understood language; delete extraneous words.
- Avoid "all of the above"—can answer based on partial knowledge (if one is incorrect or two are correct, but unsure of the third...).
- Avoid "none of the above."
- Make all distractors plausible/homoegenous.
- Don't overlap response alternatives (decreases discrimination between students who know the material and those who don't).
- Don't use double negatives.
- Present alternatives in logical or numerical order.
- Place correct answer at random (A answer is most often).
- Make each item independent of others on test.
- Way to judge a good stem: students who know the content should be able to answer before reading the alternatives
- List alternatives on separate lines, indent, separate by blank line, use letters vs. numbers for alternative answers.
- Need more than three alternatives, four is best.

Specific Guidelines for Constructing MC Items

1. State the objectives of the domain of knowledge to be assessed

The test developers can list major topics covered or expected to be covered in a term, semester or session if the focus of the test is summative assessment.

The components of a unit of instructional content can form the objective of the domain if formative assessment is the focus

2. Prepare Table of Specification.

The next task is to prepare table of specification (Test Blueprint) to cover appropriate levels of cognition using Bloom's taxonomy of learning outcomes: Knowledge, Comprehension, Application, Analysis, Synthesis and Evaluation. The cognition level of the testees should be considered while preparing the test blueprint. It ends up to be a table that has rows as the list of topics and columns as cognition levels. An example of test blueprint adaptable for developing multiple choice items on mathematics is presented below. The proportion of time spent on each topic can be a guide while distributing items on topics. Appendix I provides details on Bloom taxonomy of learning outcomes.

Test Blueprint

Content or Topic	Know 25%	Comp 20%	Appl 20%	Analy 15 %	Synt 15%	Eval 5%	Total 100%
Number and Numeration (24%)							24
Mensuration (10%)							10
Statistics/ Prob (16%)							16
Algebraic Proc (30%)							30
Plane Geo (8%)							8
Trigonometry (12%)							12
Total	25	20	20	15	15	5	100

Note: The percentages assigned to the cognition levels are suggested. In addition to Table of Specification, filling system should also be developed to ensure effective scoring and analysis of the scores. Test blueprint should include item number, level of cognition, topic from which each item was drawn and the key. The table below presents a sample of filling system.

Filling System

S/N	Topic / Content	Key	Level of Cognition
1.	Number and Numeration	B	Knowledge
2.			
3.			
4. • • • • • • • • 100			

3. The stem should be meaningful by itself and should present a definite problem.

A common fault in MC item writing is to have a brief, meaningless stem with problem definition revealed in the options. In such cases, it can be difficult to see the intent of the item after reading the stem. To write a focused item, we should include the central idea in the stem instead of the options. In Item 1, the stem does not present a definite problem.

ITEM 1

A triangle

A. Can be constructed by a ruler and a pencil only.

B. possess four interior angles and three exterior angles .

*C. has three sides.

D. a three dimensional shape.

The correct answer is indicated with an asterisk. Students are faced with four true-false options; each is about a triangle, but only option C is correct. Furthermore, the four options cover a set of widely dissimilar ideas about a triangle so that evaluation by comparison is not possible. The stem can be judged to be clearly presenting a problem if it forces the options to be parallel in type of content. Item 2 demonstrates one way to make the stem become a definite problem. Students can think about the correct answer rather than figuring out what the problem is. Also, the clearly stated problem in the stem has forced the four options to be parallel in content.

ITEM 2

How many sides has a triangle?

A. 1; B. 2; *C. 3; D. 4

4. The use of internal or beginning blanks in completion-type MC items should be avoided.

The stem may be written as an incomplete statement that needs to be completed by insertion of the correct option. Measurement specialists have advised not to use the completion format because a student has to retain the stem in short-term memory while completing the stem with each option. Test anxiety is even higher if the student is not a native English speaker. If the completion format is unavoidable, the omission should occur toward the end of the stem rather than in the middle or at the beginning of the stem as shown in Item 3. Item 4 shows an improved version.

ITEM 3

—— has four lines of symmetry.

A. Kite; *B. Square; C. Rectangle; D. Circle

ITEM 4

Which of the following has four lines of Symmetry?

A. Kite; *B. Square; C. Rectangle; D. Circle

5. Use a negatively stated stem only when significant learning outcomes require it.

Most students have difficulty understanding the meaning of negatively phrased items. They often read through the negative terms such as *not*, *no*, and *least*, and forget to reverse the logic of the relation being tested. For example, Items 5 and 6 assess the same concept of chemistry, but some students may answer Item 5 incorrectly merely because of the word *least*. Since least and concentrated are opposites, the phrase *least concentrated* is more difficult to understand than the phrase *most concentrated*. Research by Cassels and Johnstone (1984) has confirmed that the change from *least concentrated* to *most concentrated* will increase the percent of correct responses.

ITEM 5

Which of the following solutions is the least concentrated?

A. 50 g of calcium carbonate in 100 cm^3 of water

B. 60 g of sodium chloride in 200 cm^3 of water

C. 65 g of potassium nitrate in 100 cm^3 of water

*D. 120 g of potassium sulphate in 200cm^3 of water

ITEM 6

Which of the following solutions is the most concentrated?

A. 50 g of calcium carbonate in 100 cm^3 of water

B. 60 g of sodium chloride in 200 cm^3 of water

*C. 65 g of potassium nitrate in 100 cm^3 of water

D. 120 g of potassium sulphate in 200 cm^3 of water

Although negatively phrased stems should generally be avoided, they are useful if we want to assess whether students can identify dangerous laboratory practices that may damage expensive equipment or result in bodily injury, and which should not be carried out. Item 7 is an example of such an item. However, when a negative term is used, it should be emphasized by being underlined or capitalized. Replacing the negative term with the word *except* can sometimes improve clarity, as illustrated in Item 8. Few students would overlook the negative element in the stem because the word *except* is deliberately placed at the end of the stem and is capitalized.

ITEM 7

Water-type extinguisher is not suitable for putting out fire caused by burning

*A. alcohol. B. cotton. C. paper. D. wood.

ITEM 8

Water-type extinguisher is suitable for putting out fire caused by burning all of the following

EXCEPT

*A. alcohol. B. cotton. C. paper. D. wood.

6. Irrelevant difficulty should be avoided.

The difficulty of an item should not be increased by incorporating more complicated information in the stem than is necessary. For example, if we want to assess whether students can solve dilution problems using the concept of molarity, Item 9, contains confounding detail. The values used in Item 10 will assess the same learning outcome and will avoid irrelevant sources of difficulty and error.

ITEM 9

A pentagon like any other polygon has how many sides

*A. 5 sides; B. 6 sides; C. 7 sides; D. 8 sides

The words "like any other polygon" are irrelevant. Item 9 can be reworded as

ITEM 10

A pentagon how many sides

*A. 5; B. 6; C. 7; D. 8

7. All distractors should be plausible.

Designing plausible distractors is the most difficult part of MC item writing. A good distractor should be selected by low achievers and not by high achievers. To construct plausible distractors, teachers are encouraged to use common misconceptions. For example, the correct answer in both Items 11 and 12 is 7. Many students are familiar with number of sides of a triangle and a square, they can easily eliminate the distractors to pick the right answer unlike in question 12.

ITEM 11

A Heptagon has how many sides?

A. 2; B. 3; C. 4; D. 7

ITEM 12

A Heptagon has how many sides?

A. 5; B. 6; *C. 7; D. 8

8. Avoid the use of complex MC format.

Sometimes, teachers like to design complex MC items to make them harder. A complex MC item consists of a list of potentially correct answers called primary responses and a list of combinations of the primary responses called secondary options. Students have to select one of the secondary options in answering the item, as shown in Item 13. This item is equivalent to a set of four true-false items, but knowing that a particular primary response is correct or incorrect would help the examinee identify the correct secondary option by eliminating distractors (Ebel & Frisbie, 1991; Haladyna, 1999). For example, if students know that the primary response "rectangle" is untrue, they tend to pick option D because sulphur dioxide does not appear in options A and D and more than one primary response are usually included in the correct answer to a complex MC item. Although the complex MC format may make the items more difficult, research reports (Albanese, 1993; Rodriguez, 1997) reveal that it is less discriminating and reliable than the single-answer format.

ITEM 13

Which of the following shapes is/are 3 dimensional?

(1) Cuboid; (2) Rectangle; (3) Frustum; (4) Prism

A. (3) only; B. (1) and (2) only; C. (2) and (4) only; *D. (1), (3) and (4)

9. The relative length of the options should not provide a clue to the answer.

Teachers are mostly unaware of this item-writing principle (Rodriguez, 1997). It is common to express the correct response more carefully and at greater length than the distractors. However, research (Chase, 1964) has indicated that longer options tend to result in higher response rates. In Item 14, testwise students will notice that option B is much longer than the other options. Even without a good understanding of the concepts of Energy and Matter, they will guess that the correct answer is B because it stands out from the others. Note that the longest options may not be correct.

.ITEM 14

Energy like matter can

A. be destroyed and changed to matter.

B. never be destroyed but converted from one form to another.

C. be changed from solid to liquid. D. sublime

10. Avoid using “none of the above” or “all of the above” as an option.

The use of *none of the above* and *all of the above* as options in MC items is tempting to many teachers because they appear to fit easily into many items. However, many measurement specialists do not recommend the use of the option *none of the above*. For example, the correct answer for Item 15 is option D. A student may explain this way: “The correct answer is none of the above because, as everyone knows, hydrogen relights a glowing splint.” Another student may be surprised to hear that explanation: “What! The correct answer is not hydrogen, but sulphur dioxide.” It does not matter; neither gas is listed. Thus, the correct answer could be selected Using wrong ideas. This item may be modified to form Item 16.

ITEM 16

Which of the following substances would relight a glowing splint?

A. carbon dioxide; B. chlorine; C. nitrogen; *D. none of the above

ITEM 16

Which of the following substances would relight a glowing splint?

A. carbon dioxide; B. chlorine; C. nitrogen; *D. oxygen

Item Selection and Standardization: Practical Example

This section presents the results of foil analysis and item analysis, analytic procedures for item selection, psychometric properties and the standardization procedures of Mathematics Achievement Test (MAT).

Foil Analysis:

The testees responses on the first draft of MAT were subjected to descriptive analysis using SPSS software, through which the frequency counts and the percentages of students that chose each option under each item were obtained and presented in table below. The table also presents the proportion of the testees that did not respond to each item and the correct answers to each of the items were also presented in the table. These provide information on the clarity of each of the item and attractiveness of each option. This enhances modification of affected items and options.

Foil Analysis

S/N	A	B	C	D	NR	ANS.
1	156	59	121	44	20	C
	39.0	14.8	30.3	11.0	5.0	
2	188	100	58	36	18	A
	47.0	25.0	14.5	9.0	4.5	
3	130	73	89	94	14	D
	32.5	18.3	22.3	23.5	3.5	
4	122	112	83	69	14	C
	30.5	28.0	20.8	17.3	3.5	
5	195	89	59	37	20	A
	48.8	22.3	14.8	9.3	5.0	
6	109	145	86	45	15	D
	27.3	36.3	21.5	11.3	3.8	
7	105	139	82	49	25	B
	26.3	34.8	20.5	12.3	6.3	
8	191	84	64	42	19	A
	47.8	21.0	16.0	10.5	4.8	
9	115	146	54	68	17	B
	28.8	36.5	13.5	17.0	4.3	
10	129	89	71	90	21	C
	32.3	22.3	17.8	22.5	5.3	

Note:

NR means No Response while ANS means Answer.

Option E was not included because test development experts advise that the use of 4 options is better.

Item Analysis

SPSS software was used to mark the responses of the testees obtained from the first administration of the drafted version of MAT. The overall scores obtained by the testees were obtained by using computer to add up scores item by item. The data generated from the responses of 400 testees was sorted by their overall scores in descending (ascending order is also possible) order to obtained the best 27%(108 testees) and lower 27%(108 testees). The best 108 testees constitute the upper scorers while the worst 108 testees constitute the lower scorers. The number of the testees that scored each item correctly among the upper and lower scorers, and denoted by 'upperscorer' and 'lowerscorer' respectively are presented in the table below. The total number that scored each item correctly among all the testees were obtained and denoted by 'itemscore' in the table.

The above estimations were used to calculate the following:

(i) Item discrimination index = Diff/n
Where diff = difference between the upper and lower scorers
n = total No of the testees in either upper scorer or lower scorer group

(ii) Item Difficulty Index (P) = Itemsco/N
Where Itemscore = total Number of the testees that answer each item correctly
N = Addition of total number of testees in upper and lower scorer groups.

The table below presents discrimination and difficulty indices of all the items in MAT and they are denoted by 'dcdx' and 'dfdx' respectively in the table.

Item Analysis

S/N	itemno	Upper scorer (R_u)	Lower scorer (R_l)	Item score	Diff. (R_u- R_l)	Disc indx	Diff Indx(P)	q	pq
1	1	49	14	121	35	0.32	0.56	0.44	0.25
2	2	73	37	188	36	0.33	0.87	0.13	0.11
3	3	41	16	94	25	0.23	0.44	0.56	0.25
4	4	27	22	83	5	0.05	0.38	0.62	0.24
5	5	64	52	195	12	0.11	0.9	0.10	0.09
6	6	19	9	45	10	0.09	0.21	0.79	0.17
7	7	62	19	139	43	0.4	0.64	0.36	0.23
8	8	77	44	191	33	0.31	0.88	0.12	0.11
9	9	56	25	146	31	0.29	0.68	0.32	0.22
10	10	13	18	71	-5	-0.05	0.33	0.67	0.22
11	11	60	16	128	44	0.41	0.59	0.41	0.24
12	12	50	16	112	34	0.31	0.52	0.48	0.25
13	13	15	21	79	-6	-0.06	0.37	0.63	0.23
14	14	23	21	81	2	0.02	0.38	0.62	0.24
15	15	36	17	78	19	0.18	0.36	0.64	0.23
16	16	43	18	103	25	0.23	0.48	0.52	0.25
17	17	71	33	169	38	0.35	0.7	0.22	0.17
18	18	44	15	108	29	0.27	0.5	0.50	0.25
19	19	33	18	93	15	0.14	0.43	0.57	0.25
20	20	46	31	135	15	0.14	0.63	0.37	0.23
Total									4.21

Item Selection

In order to select good items that will constitute the final version on MAT, the following criteria were used.

(i) Difficulty indices ranged between 0.4 and 0.6

(ii) Discrimination indices ranged between 0.3 and above

(iii) Results of foil analysis were also used for the modification of items and options noticed to be ambiguous.

Based on the above stated criteria, the following items were discarded.

2 3 4 5 6 8 9 10 13 14

15 16 17 19 20

The items that satisfied the criteria and are retained to constitute the final version of MAT are:

1 7 11 12 18

Observation: Only five of twenty items were retained. Item developers are advised to develop five times the number of items they finally needed, if the final items are going to be good. For example a test developer who intends to have 100 good items at the end of item analysis should develop 500.

Validity of MAT

To ensure that the instrument (MAT) measures what it is purported to measure, content validity was established using test blue print. The nature of the test (objective achievement test) informed the choice of test blue print.

The test blue print ensures that:

(i) All the topics taught are covered

(ii) Expected cognition levels are equally covered out of Bloom's hypothesized six levels of cognition (knowledge, comprehension, application, analysis, synthesis and evaluation).

Hence, the Sample of test blue print for the final version of MAT is presented below.

Topic	Level of Cognition						Total
	Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation	
Number and Numeration	Q1						1
Basic Operation		Q7					1
Measurement							
Algebraic Process		Q11					1
Geometry and Mensuration	Q12						1
Everyday Statistics			Q18				1
Total	2	2	1				5

The limited number of items does not allow for adequate representation of topics in a test. It is suggested that the number of items in a test should be large enough to ensure content validity.

Reliability Coefficient of MAT

The reliability coefficient of MAT was established using Kuder Richardson 20 (KR 20) formula.

:. Reliability coefficient

$$r_{11} = \left(\frac{n}{n-1}\right)\left(\frac{SD_t^2 - \Sigma p_i q_i}{SD_t^2}\right)$$

Where SD_t
= variance of the testees' score
P_i = proportion of the testees' that answered each item correctly
q_i = proportion of the testees that answered each item wrongly.
n = the sum of testees in Upper and Lower groups.

$$R = \frac{216}{215}\left[\frac{83 - 4.21}{83}\right]$$

$$\therefore R = 0.9537$$

$$R^2 = 0.9095$$

The estimated reliability coefficient on MAT was 0.9537. This shows that the instrument is highly reliable and posses high internal consistency for measuring cognitive achievement of primary Junior Secondary One Students' Mathematics Achievement. R square (R^2) estimated was 0.9095 which implies that 90.95% variation in JSS 1 cognitive achievement in Mathematics was measured by MAT while the remaining 9.05 percent is traceable to other factors that can cause variation in the cognitive achievement of JS 1 Students.

Note:

Test development experts advise that only data generated through testees in upper and lower scorer groups should be used for item analysis.

Establishment of Norms on MAT

The basic norms established that constitute the normative data for comparing scores on MAT are presented under this section. The Complete data generated through 100 item-test on Mathematics administered to 400 testees was used in this section.

Gender Norms: This allows for the comparison of male and female testees on how they performed in the test.

Gender Norms

Gender	N	Mean	Std. Deviation
Male	207	39.58	16.73
Female	193	38.15	16.16
Total	400	38.89	16.45

The Table shows that the mean scores of male and female testees are 39.58 and 38.15 respectively. Male seems to achieve slightly better in test than their female counterparts. The table notwithstanding provides the normative data that form the bases for comparing individual score in MAT along the line of gender.

Age Norm: This determines the average score earned by individuals of a given age.

What is the Norm established on MAT based on Age

Age	N	Mean	Std. Deviation
10	93	34.33	13.15
11	243	33.75	13.10
12	35	31.49	11.63
13	10	39.50	11.90
14	9	25.89	5.35
15	4	29.00	5.35
16	1	31.00	-
17	3	32.67	4.62
18	1	45.00	-
19	1	39.00	-
Total	**400**	**38.89**	**16.45**

The table shows that the scores of the testees on MAT is not so sensitive to age however the table presents normative data for interpreting the scores of the testees base on their age.

School type Norm: This determines the performance of public and private primary schools.

What is the norm established on MAT based on school type?

School Type Norm

School type	N	MEAN	Std. Deviation
Private	174	49.44	17.03
Public		2260.77	10.22
Total	**400**	**16.45**	**16.45**

The table shows that the testees from private schools performed better than their female counterparts with means of 49.44 and 30.77 respectively. This form the normative data for comparing and interpreting individual score on MAT based on the type of school the testee attends.

Stanine Norms: This allows attention to be focused on differences that are large enough to matter.

	1st	2nd	3rd	4th	5th	6th	7th	8th	9th
% in stanine	4	7	12	17	20	17	12	7	4
Percentile equivalent	3	11	23	40	60	77	89	96	100
Score	17	21	26	30	40	51	61	73	90

The table presents the stanine Norm for determining the relative standing of individual testees on MAT. It also provides bases for interpreting the individual scores too.

Percentile Norm: This can be used whenever the relative standing of individual relative standing when the distribution is divided into 100. This is also useful while interpreting individual scores. An appropriate normative group can be obtained to serve as a yardstick.

Standard Score Norm (Z Score Norm): This shows scores that are reported as deviations away from the group mean in standard deviation units.

Other types of Achievement Test are Oral Exams, Student Portfolios and Performance Test

Oral Examination: It is a verbal response test useful as an instructional tool- allows students to learn at the same time as testing. Oral exam is also used during interview.

Oral Examamination Good for:

- Knowledge, synthesis, evaluation levels

Advantages:

- Allows teachers to give clues to facilitate learning.
- Useful to test speech and foreign language competencies.

Disadvantages:

- Time consuming to give and take.
- Could affect poor student's performance adversely, because they haven't had much practice with it.
- Provides no written record without checklists.

Student Portfolios

Portfolios are collections of students' work over time. A portfolio often documents a student's best work and may include other types of process information, such as drafts of the student's work, the student's self-assessment of the work, and the parents' assessment. Portfolios may be used for evaluation of a student's abilities and improvement.

In recent years, portfolios of students' performance and products have gained impressive degrees of support from educators, who view them as a way to collect authentic evidence of children's learning. For many early childhood educators, portfolios are an attractive alternative to more traditional assessment approaches. Often, however, teachers raise important questions about what portfolios contain, what benefits they will bring to the classroom and the children, and how they can be managed

Student Portfolios Good for:

- *Knowledge, application, synthesis, evaluation levels*

Advantages:

- Can assess compatible skills: writing, documentation, critical thinking, problem solving
- Can allow student to present totality of learning.
- Students become active participants in the evaluation process.

Disadvantages:

- Can be difficult and time consuming to grade.

Performance Test: It measures the skills of the testees

Good for:

- Application of knowledge, skills, abilities

Advantages:

- Measures some skills and abilities not possible to measure in other ways

Disadvantages:

- Can not be used in some fields of study
- Difficult to construct
- Difficult to grade
- *Time-consuming* to give and take

Aptitude Tests

Aptitude and ability tests are designed to assess logical reasoning or thinking performance. They consist of multiple choice questions and are administered under exam conditions. They are strictly timed and a typical test might allow 30 minutes for 30 questions. Test results are compared to that of a control group so that judgments can be made about individual's abilities. Aptitude test can be taken under normal examination condition or by online testing.

Aptitude and ability tests can be classified as speed tests or power tests

.

Speed Test: In speed tests the questions are relatively straightforward and the test is concerned with how many questions you can answer correctly in the allotted time. Speed tests tend to be used in selection at the administrative and clerical level.

Power Tests: A power test on the other hand will present a smaller number of more complex questions. Power tests tend to be used more at the professional or managerial level.

Types of aptitude Test.

There are at least 5000 aptitude and ability tests on the market. Some of them contain only one type of question (for example, verbal ability, numeric reasoning ability etc) while others are made up of different types of question. Few common ones are listed below:

Verbal Ability Test - measures ability to understand analogies and follow detailed written instructions e.g spelling and grammar. These questions appear in most general aptitude tests because employers usually want to know how well you can communicate.

Numeric Ability Test – Measures basic arithmetic, number sequences and simple mathematics abilities. In management level tests you will often be presented with charts and graphs that need to be interpreted. These questions appear in most general aptitude tests because employers usually want some indication of your ability to use numbers even if this is not a major part of the job.

Abstract Reasoning Test - Measures your ability to identify the underlying logic of a pattern and then determine the solution. Because abstract reasoning ability is believed to be the best indicator of fluid intelligence and your ability to learn new things quickly these questions appear in most general aptitude tests.

Spatial Ability Test - Measures your ability to manipulate shapes in two dimensions or to visualize three-dimensional objects presented as two-dimensional pictures. These questions not usually found in general aptitude tests unless the job specifically requires good spatial skills.

Mechanical Reasoning Test - Designed to assess your knowledge of physical and mechanical principles. Mechanical reasoning questions are used to select for a wide range of jobs including the military (Armed Services Vocational Aptitude Battery), police forces, fire services, as well as many craft, technical and engineering occupations.

Fault Diagnosis - These tests are used to select technical personnel who need to be able to find and repair faults in electronic and mechanical systems. As modern equipment of all types becomes more dependent on electronic control systems (and arguably more complex) the ability to approach problems logically in order to find the cause of the fault is increasingly important.

Data Checking - Measure how quickly and accurately errors can be detected in data and are used to select candidates for clerical and data input jobs.

Work Sample– Involves a sample of the work that you will be expected do. T hese types of test can be very broad ranging. They may involve exercises using a word processor or spreadsheet if the job is administrative or they may include giving a presentation or in-tray exercises if the job is management or supervisory.

CHAPTER SIX

Research Tools: Non Cognitive Research Tools

The various methods of data gathering involve the use of appropriate recording forms. These are called tools or instruments of data collection. They consist of

- Observation schedule
- Interview guide
- Interview schedule
- Mailed questionnaire
- Rating scale
- Checklist
- Document schedule/data sheet
- Schedule for institutions

Each of the above tools is used for a specific method of data gathering: Observation schedule for observation method, interview schedule and interview guide for interviewing, questionnaire for mail survey, and so on.

Functions

The tools of data collection translate the research objectives into specific questions/ items, the responses to which will provide the data required to achieve the research objectives. In order to achieve this pur-pose, each question/item must convey to the respondent the idea or group of ideas required by the research objectives, and each item must obtain a response which can be analysed for fulfilling the research objectives.

Information gathered through the tools provides descriptions of characteristics of individuals, institutions or other phenomena under study. It is useful for measuring the various variables pertaining to the study. The variables and their interrelationships are analysed for testing the hypothesis or for exploring the content areas set by the research objec-tives.

A brief description of the various tools of data collection is given below.

Observation schedule

This is a form on which observations of an object or a phenomenon are recorded. The items to be observed are determined with reference to the nature and objectives of the study. They are grouped into appropriate categories and listed in the schedule in the order in which the observer would observe them.

The schedule must be so devised as to provide the required verifiable and quantifiable data and to avoid selective bias and misinterpretation of observed items. The units of observation must be simple, and meticulously worded so as to facilitate precise and uniform recording.

Sample

Preparation:

		1	2	3	4	5
1.	How lesson agrees with scheme of work					
2.	Availability of adequate prepared lesson note					
3.	Availability of relevant teaching/learning materials					
4.	Application or use of teaching aids					
5.	Punctuality to class/lesson					

Interview guide

This is used for non-directive and depth interviews. It does not contain a complete list of items on which information has to be elicited from a respondent: it just contains only the broad topics or areas to be covered in the interview.

Interview guide serves as a suggestive reference or prompter during interview. It aids in focussing attention on salient points relating to the study and in securing comparable data in different interviews by the same or different interviewers.

Interview schedule and mailed Questionnaire

Both tools are widely used in surveys.

They are complete list of questions on which information is elicited from the respondents. The basic difference between them lies in recording responses. While the interviewer fills out a schedule, the respondent completes a questionnaire.

Sample

Interview Schedule

Nameof
Institution:...
Rank of the Interviewee:...

Question	Response
What are the major source(s) of fund to your institution?	
Is the generated fund sufficient to offset the budget of the institution?	
If no, how do you handle budget deficit?	
Does the society value the product of your institution?	

Rating Scale

This is a recording form used for measuring individual's attitudes, aspirations and other psychological and behavioural aspects, and group behaviour. A rating scale is an instrument that requires the rater to assign the rated object numerals. It provides commonly, data measured at the ordinal level

. Numbers indicate the relative position of items, but not the magnitude of difference. One example is a Likert scale:

Statement: I could not live without my computer.

Response options:

- 1. Strongly Disagree
- 2. Disagree
- 3. Agree
- 4. Strongly Agree

The rater can rate self or other(s)

Sample:

Rating Scale.

Instruction: Please kindly rate the teacher on the following classroom attitudes and behaviours putting a tick in the space provided in front of each statement.

Keys: Numerical

5=	Outstanding
4=	Above Average
3=	Average
2=	Below Average
1=	Unsatisfactory

Presentation/Development of content Communication

		1	2	3	4	5
10.	Your Maths teacher Speaks fluent English					
11.	He/She Communicates subject content in precise and clear terms					
12.	He/She Presents content step by step or point by point					
13.	He/She Lays emphasis on major points in the lesson					

Checklist

This is the simplest of all the devices. It consists of a prepared list of items pertinent to an object or a particular task. The presence or absence of each item may be indicated by checking 'yes' or 'no' or multipoint scale. The use of a checklist ensures a more complete consideration of all aspects of the object, act or task. Checklists contain terms, which the respondent understands, and which more briefly and succinctly express his views than answers to open-ended question. It is a crude device, but careful pre-test can make it less so. It is at best when used to test specific hypothesis. It may be used as an independent tool or as a part of a schedule/questionnaire.

Sample

Which of the following are available in your School for Teaching Mathematics?

Item	Available	Not Available
Graph Board		
Black Board Ruler		
Mathematical Set		
Shapes		

Document Schedule/Data Sheet.

This is a list of items of information to be obtained from documents, records and other materials. In order to secure measurable data, the items included in the schedule are limited to those that can be uniformly secured from a large number of case histories or other records.

Schedule for Institutions

This is used for survey of organisations like business enterprises, educational institutions, social or cultural organisations and the like. It will include various categories of data relating to their profile, functions and performance. These data are gathered from their records, annual reports and financial statements.

Construction of Schedules and Questionnaires

Schedule v. Questionnaire

Schedules and questionnaires are the most common instruments of data collection. These two types of tools have much in common. Both of them contain a set of questions logically related to a problem under study; both aim at eliciting responses from the respondents; in both cases the content, response structure, the wordings of questions, question sequence, etc. are the same for all respondents. Then why should they be denoted by the different terms: 'schedule' and 'questionnaires'? This is because the methods for which they are used are different. While a schedule is used as a tool for interviewing, a questionnaire is used for mailing.

This difference in usage gives rise to a subtle difference between these two recording forms. That is, the interviewer in a face-to-face interviewing fills a schedule, whereas the respondent himself fills in a questionnaire. Hence the

need for using two different terms.
The tool is referred to as a schedule when it is used for interviewing; and it is called a questionnaire when it is sent to a respondent for completion and return.

The process of construction

The process of construction of a schedule and a questionnaire is almost same, except some minor differences in mechanics. This process is not a matter of simply listing questions that comes to researchers mind. It is a rational process involving much time, effort and thought. It consists of the following major steps:

1. Data need determination: As an interview schedule or a mailed questionnaire is an instrument for gathering data for a specific study, its construction should flow logically from the data required for the given study.
2. Preparation of "Dummy" tables: The best way to ensure the requirements of information is to develop "dummy" tables in which to display the data to be gathered.
3. Determination of the respondents' level: Who are our respondents? Are they persons with specialized knowledge relating to the problem under study? Or are they lay people? What is their level of knowledge and understanding? The choice of words and concepts depends upon the level of the respondents' knowledge.
4. Data gathering method decision: Which communication mode is most appropriate - face-to-face interview or mailing? The choice of question structure depends largely on the communication mode chosen.
5. Instrument drafting: After determining the data required for the study, first, a broad outline of the instrument may be drafted, listing the various broad categories of data. Second, the sequence of these groupings must be decided. Third, the questions to be asked under each group heading must be listed. All conceivable items relevant to the 'data need' should be compiled.
6. Evaluation of the draft instrument: In consultation with other qualified persons, the researcher must rigorously examine each question in the draft instrument.
7. Pre-testing: The revised draft must be pre-tested in order to identify the weaknesses of the instrument and to make the required further revisions to rectify them.

8. Specification of procedures/instructions: After the instruction is finalised after pre-tests, the procedures or instructions, relating to its use must be specified.
9. Designing the format: The format should be suited to the needs of the research. The instrument should be divided into different sections relating to the different aspects of the problem.

Question Construction

A survey instrument - interview schedules or questionnaire - is useful for collecting various types of information, viz., (a) factual information - facts about the respondents: sex, age, marital status, education, religion, caste or social class, income and occupation; and facts about events and circumstances, (b) psychological information such as attitudes, opinions, beliefs, and expectations, and (c) behavioural information, like social participation, and so on.

Once the information need is determined as explained in the previous topic, we can begin question construction. This involves four major decision areas. They are: (a) question relevance and content, (b) question wording, (c) response form, and (d) question order or sequence.

Question relevance and content

Question to be included in the. instrument should pass c tain tests. Is it relevant to the research objectives? Can it yield significant information for answering an investigative question? If not, it should not be included in the instrument.

Question wording

This is a difficult task. The function of a question in a schedule/questionnaire is to elicit particular information without distortion. "Questioning people", says Oppenheim, "is more like trying to catch a particular elusive fish, by hopefully casting different kinds of bait at different depths, without knowing what goes on beneath the surface." As the meaning of words differs from person to person, the question designer should choose words which have the following characteristics:

a. Shared vocabulary.
b. Uniformity of meaning.
c. Exactness.
d. Simplicity.
e. Neutrality. The words to be used must be neutral ones, i.e., free from the distorting influence of fear, prestige, bias or emotion.

Certain other problem areas of question wording are
a. unwarranted assumptions,
b. personalization,
c. presumptions,
d. hypothetical question,
e. questions on embarrassing matters.

Some of the approaches to deal with this problem are:

i. to express the question in the third person; instead of asking the respondent for his views, he is asked about the views of others:
ii. to use a drawing of two persons in a certain setting with 'balloons' containing speech coming from their mouths, as in a cartoon - leaving one person's balloon empty and asking the respondent to put himself in the position of that person and to fill in the missing words; and
iii. to use sentence completion tests.

Response form or types of Questions

The third major area in question construction is the types of questions to be included in the instrument. They may be classified into open questions and closed questions. Closed questions may be dichotomous, multiple choice or declarative ones.

Types of questions to be avoided

The question designer should avoid the following types of questions: (a) Leading questions, (b) 'Loaded' questions, (c) Ambiguous questions, (d) Double-barrelled, (e) Long questions, (t) Avoid double negative.

Question order or Sequence

The order in which questions are arranged in a schedule/questionnaire is as important as question wording. It has two major implications. First, an appropriate sequence can ease the respondent's task in answering. Second, the sequence can either create or avoid biases due to context effects, i.e., the effects of preceding questions on the response to later questions.

Sample Questionnaire

Dear Respondent,
This questionnaire is designed to elicit your response to various items connected with the training workshop that you are participating in. UBEC is to use the collected data to determine the propriety and effectiveness, or otherwise, of the training, and for research purposes only. Your responses will be treated in confidence. Thanks.

A. PRELIMINARY INFORMATION:

1. State:...
2. LGEA:...
3. School:...
4. Gender: Male [ ] Female [ ]
5. Highest Educational Qualification:........................ ....
6. Duration of training/workshop/seminar: ....
7. Who is the Master-Trainer? ....
8. Venue of training: ....

B. PERCEPTION OF TRAINING WORKSHOP

Please rate (P) the various aspects of the training workshop in terms of the extent to which you agree with the various statements:
SA=Strongly Agree; A=Agree; D=Disagree;
SD=Strongly Disagree;

S/N	Statement	SA	A	D	SD
9.	The venue was conducive.				
10.	Registration procedure was cumbersome.				
11.	The presentations by resource persons were easily understood.				
12.	Training Manual was given to trainees.				
13.	The units treated will help me in my classroom delivery.				
14.	Activities were organized in a logical sequence for learning.				
15.	The methods used in the presentation of the units [illegible]				
16.	Adequate provision was made for trainees to practice what was taught.				
17.	The language used was clear and appropriate.				
18.	The resource persons ensured trainees' participation during the training.				
19.	Resource materials were adequate.				
20.	Resource materials were of good quality.				

Mechanics of the Schedule and Questionnaire

In addition to question wording and question construction, the mechanics of the form should also be considered in the design of a schedule/questionnaire. The mechanics of the form has several aspects: items of the form, instruction, pre-coding, sectionalisation, spacing, paper, printing, margins, etc.

Items of the form: The following items are mandatory for schedules and questionnaires.

1. The name of the organization collecting the data should appear at the top of front -page. The name of the sponsor, of the study, if any should also be shown.
2. The title of the study should appear in large print next to the name of the

organization on the first page. Below this title, the title of the tool - e.g., 'Schedule for-consumers; - may be noted. .

3. Assurance of confidentiality should be made clear.
4. A place for writing the date of filling in the form should be provided.
5. A serial number to each copy of the tool may be assigned.
6. The pages of the instrument should be numbered.

Instructions: In the face sheet below the title of the questionnaire, a brief statement of the objective of the study, the confidentialness of the data, and instructions relating to answering the questions may be provided.

Pre-coding: Items in the tool should be pre-coded so as to facilitate transcription of data.

Sectionalisation: There should be a separate section for each topical area.

Spacing: For each open-ended question, an adequate space should be provided for answer. There should, indeed more space than seems necessary, for some interviewers/ respondents may write in a large script for legibility. Moreover, liberal spacing is a stimulus for the questionnaire respondent to write more fully. Even short-answer questions should be spaced, so that the interviewer/respondent will not easily confuse the line, from which he is reading.

Paper: The paper used for mimeographing/printing should be of good quality.

Printing: Mailed questionnaire should necessarily be printed in order to make it attractive and to minimise the postal expenditure.

Margins: One inch margin on the left side of the sheet and one-half inch margin on other sides may be provided. If the instrument is to be bound, left-side margin should conform to the type of binding used.

Indentation: This is required for 'yes' or 'no' questions. If the respondent's answer is 'yes', then a series of questions is offered. If the answer is 'no' a different series of questions is offered.

Note of thanks: A final note or comment of thanks for the cooperation of the respondent should be included at the end of the instrument.

Overview of Uses of Research Tools

Method	Overall Purpose	Advantages	Challenges
Questionnaires Surveys, Checklists	when need to quickly and/or easily get lots of information from people in a non threatening way	-can complete anonymously -inexpensive to administer -easy to compare and analyze -administer to many people -can get lots of data -many sample questionnaires already exist	-might not get careful feedback -wording can bias client's responses -are impersonal -in surveys, may need sampling expert - doesn't get full story
Interviews	when want to fully understand someone's impressions or experiences, or learn more about their answers to questionnaires	-get full range and depth of information -develops relationship with client -can be flexible with client	-can take much time -can be hard to analyze and compare -can be costly -interviewer can bias client's responses
Documentation Review	when want impression of how program operates without interrupting the program; is from review of applications, finances, memos, minutes, etc.	-get comprehensive and historical information -doesn't interrupt program or client's routine in program -information already exists -few biases about information	-often takes much time -info may be incomplete -need to be quite clear about what looking for -not flexible means to get data; data restricted to what already exists
observation	to gather accurate information about how a program actually operates, particularly about processes	-view operations of a program as they are actually occurring -can adapt to events as they occur	can be difficult to interpret seen behaviors can be complex to categorize observations can influence behaviors of program participants can be expensive
Focus groups	explore a topic in depth through group discussion, e.g., about reactions to an experience or suggestion, understanding common complaints, etc.; useful in evaluation and marketing	-quickly and reliably get common impressions -can be efficient way to get much range and depth of information in short time - can convey key information about programs	can be hard to analyze responses -need good facilitator for safety and closure -difficult to schedule 6-8 people together
Case studies	to fully understand or depict client's experiences in a program, and conduct comprehensive examination through cross comparison of cases	-fully depicts client's experience in program input, process and results -powerful means to portray program to outsiders	-usually quite time consuming to collect, organize and describe -represents depth of information, rather than breadth

Concluding remarks

Question designing remains primarily a matter of common sense and experience and of avoiding known pitfalls, as there are no hard and fast rules relating to it. Hence alternative versions of questions must be rigorously tested in pre-tests. Test-revision-retests play a crucial role in questionnaire construction.

CHAPTER SEVEN

Measurement Scales and Indices

Scales are devised for measuring variables in social science research. During the past few decades thousands of scales have been designed by researchers in sociology, psychology, education, psychiatry, ethics, behavioural science, economics, administration and other fields.

All the scales can be classified into one of the following four classifications:-

Nominal Scale: Some data are measured at the nominal level
. That is, any numbers used are mere labels : they express no mathematical properties. Examples are Sex (1-Male, 2-Female); State of Origin (1-Abia, 2-Adamawa, …….,36- Zamfara)

Ordinal Scale: Some data are measured at the ordinal level. Numbers indicate the relative position of items, but not the magnitude of difference. One example is a Likert scale:

Statement: Mathematics will not be useful to me in the future.
Response options:

- 1. Strongly Disagree
- 2. Disagree
- 3. Agree
- 4. Strongly Agree

Interval Scale: Some data are measured at the interval level
. Numbers indicate the magnitude of difference between items, but there is no absolute zero point. Examples are attitude scales and opinion scales.

Ratio Scale: Some data are measured at the ratio level. Numbers indicate magnitude of difference and there is a fixed zero point. Ratios can be calculated. Examples include: age, income, price, costs, sales revenue, sales volume, and market share.
Indices and scales are often used interchangeably to refer to all sorts of measures, absolute or relative, single or composite, simple or elaborate.

"Scaling" refers to the procedure by which numbers or scores assigned to the various degrees of opinions, attitude and other concepts.

Pilot Studies and Pre-Tests

Pilot Study

The need for Pilot Study

It is difficult to plan a major study or project without adequate knowledge of its subject matter, the population it is to cover, their level of knowledge and understanding and the like. What are the issues involved? What are the concepts associated with the subject matter? How can they be operationalised? What method of study is appropriate? How long the study will take? How much money it will cost? These and other related questions call for a good deal of knowledge of the subject matter of the study and its dimensions. In order to gain such pre-knowledge of the subject matter of an extensive study, a preliminary investigation is conducted. This is called a pilot study.

Pre-test

Meaning

While a pilot study is a full-fledged miniature study of a problem, pre-test is a trial test of a specific aspect of the study such as method of data collection or data collection instrument - interview schedule, mailed questionnaire or measurement scale.

Need for Pre-testing

An instrument of data collection is designed with reference to the data requirements of the study. But it cannot be perfected purely on the basis of a critical scrutiny by the designer and other researchers. It should he empirically tested. As emphatically pointed by Goode and Hatt, "no amount of thinking, no matter how logical the mind or brilliant the insight, is likely to take the place of careful empirical checking". Hence pre-testing of a draft instrument is indispensable. Pre-testing-means trial administration of the instrument to a sample of respondents before finalising it.

Purposes of Pre-testing

Choosing appropriate instrumentation (surveys, questionnaires, observation schedule, rating scale etc.) is a vital part of conducting good quality empirical research and evaluation. Too often researchers fall vulnerable to 'availability bias' and simply select whatever they can get their hands on, or they default to using instruments that have commonly been used in the past. Poor instrument selection adds noise and error to your research.

A thorough search and evaluation of all possible measures is recommended. Time spent on critical reviewing of possible instruments for any research, is time well spent. Key factors to consider while searching for instrument for any research studies are:

Length and Complexity

On outdoor and experience-based programs, instruments are often administered in field settings (e.g. in the bush, on board a boat, in various weather conditions), on multiple occasions (e.g. pre-program, first day of program, last day of program and post-program follow-up) and to a wide range of participants (e.g. people with learning disabilities, people without English as their first language, school children, corporate managers). Hence, the shorter and simpler an instrument (reliability and validity aside), the greater the instrument's potential applicability. The aim was to develop an instrument which would provide a maximum amount and type of unique information in as short a time as possible (i.e. a maximum of about ten minutes). The instrument's instructions and layout also needed to be straightforward to allow people without research experience, such as group leaders or teachers, to administer the instrument consistently across different groups.

Relevance to Research Objectives

Generally, a major objective of many outdoor experiential research is to facilitate individuals' personal development in a broad range of life skills (e.g. self-confidence, initiative, communication skills, etc.), although different research may have more specific objectives such as the development of teamwork and leadership skills. Ideally, the instrument would encompass a wide range of life proficiency domains relevant to general and specific research objectives, so as to allow for within and between program comparison of different treatment outcomes.

Sensitivity to Change

The scoring system is an important aspect of an instrument's sensitivity to change. A dichotomous (yes/no) scoring does not provide much sensitivity, whereas a large range can reduce the instrument's reliability. A balance needs to be reached between sensitivity to change and reliability. Despite being a critical issue, a search of the literature revealed little research, for example, on the relative efficacy of likert-type scales with different numbers of responses for measuring change.

Two further issues to be considered in developing an instrument for measuring change are ceiling/floor effects and test-retest correlations. It is desirable that the wording of the items and the response scale tends to produce responses from participants that leave room for detecting shifts in self-perceptions either up or down; hence the means need to be examined during the instrument development. In addition, test-retest correlations give an indication of the stability of items and scales. If these correlations are low, then participants' responses to the item or scale may change for reasons other than those that can be attributed to an intervention experience.

Educational Exercise

The methods used to facilitate personal change during outdoor experiential programs include providing opportunities for self-assessment, goal-setting, and feedback on personal progress. An instrument which can be used to facilitate the processes of self-examination, goal-setting and feedback would give it added value.

Reliability and Validity:

The reliability and validity of the instrument should be well established via peer-reviewed publication and rigorous statistical procedures. The strengths and limitations of the instrument should be clearly indicated.

Ethical/Educational Issues:

If possible, the instrument should not be used only for the interests of the researcher, but also in the education/development of participants. For example, a self-assessment tool could be used not only for research purposes but also to lead onto a goal-setting and feedback session with participants.

Essentials of a Good Psychological Test

Reliability – An Overview

Reliability is the extent to which a test is *repeatable* and yields *consistent* scores. It is worthy to note that, in order to be valid, a test must be reliable; but reliability does not guarantee validity. All measurement procedures have the potential for error, so the aim is to minimize it. An observed test score is made up of the true score plus measurement error.

The goal of estimating reliability (consistency) is to determine how much of the variability in test scores is due to measurement error and how much is due to variability in true scores.

Measurement errors are essentially random: a person's test score might not reflect the true score because he was sick, anxious, in a noisy room, etc. loom

Reliability can be improved by:

- getting repeated measurements using the same test and
- getting many different measures using slightly different techniques and methods.

- e.g. Consider university assessment for grades involve several sources. You would not consider one multiple-choice exam question to be a reliable basis for testing your knowledge of "individual differences". Many questions are asked in many different formats (e.g., exam, essay, presentation) to help provide a more reliable score.

Types of Reliability

There are several types of reliability:

1. 1. Test-retest reliability

The test-retest method of estimating a test's reliability involves administering the test to the same group of people at least twice. Then the first set of scores is correlated with the second set of scores. Correlations range between 0 (low reliability) and 1 (high reliability) (highly unlikely they will be negative!)

Remember that change might be due to measurement error e.g if you use a tape measure to measure a room on two different days, any differences in the result is likely due to measurement error rather than a change in the room size. However, if you measure children's reading ability in February and again in June the change is likely due to changes in children's reading ability. Also, the actual experience of taking the test can have an impact (called reactivity).

2. Alternate/Parallel Forms

Develop or obtain two parallel tests A and B. Administer Test A to a group and then administer Test B to same group. Correlation between the two scores is the estimate of the test reliability

3. Split Half reliability

Relationship between half the items and the other half administer to the same group. A researcher can use odd and even item numbers to split the test to two.

4. Inter-rater Reliability

Compare scores given by different raters. Before using any rating scale or observation schedule, ensure inter-rater reliability.

5. Internal consistency

Internal consistence is commonly measured as Cronbach's Alpha (based on inter-item correlations) - between 0 (low) and 1 (high). The greater the number of similar items, the greater the internal consistency. That's why a researcher sometimes gets very long scales asking a question a myriad of different ways. If you add more items you get a higher cronbach's Alpha coefficient. Generally, alpha of .80 is considered as a reasonable benchmark

$$\alpha = \left(\frac{n}{n-1}\right)\left(\frac{SD_t^2 - \sum SD_i^2}{SD_t^2}\right)$$

Where: α is the estimate of reliability,

n is the number items in a test

SD_t is the standard deviation of the test scores

$\sum$ means "take the sum of" and covers the n items, and

SD_i is the standard deviation of the scores from a group of individuals. There are

n values of SD_i which are summed to give the second term in the numerator.

When each item is scored as either 1or 0, that is, as either right or wrong'

$SD_i^2 = p_i q_i$

$$r_{11} = \left(\frac{n}{n-1}\right)\left(\frac{SD_t^2 - \Sigma p_i q_i}{SD_t^2}\right).$$

Where r_{11} is the estimated reliability of the full-length test, the equation is called Kuder-Richardson Formula 20 (KR-20).

Reliability Guidelines:

.90 = high reliability
.80 = moderate reliability
.70 = low reliability

Reliability estimates of .80 or higher are typically regarded as moderate to high (approx. 16% of the variability in test scores is attributable to error)

Reliability estimates below .60 are usually regarded as unacceptably low.

Validity

Validity is the extent to which a test measures what it is supposed to measure. Validity is a subjective judgment made on the basis of experience and empirical indicators. Validity asks "Is the test measuring what you think it's measuring?"

For example, we might define "aggression" as an act intended to cause harm to another person (a conceptual definition) but the operational definition might be seeing:

- how many times a child hits a doll
- how often a child pushes to the front of the queue
- how many physical scraps a child gets into in the playground.

Are these valid measures of aggression? i.e., how well does the operational definition match the conceptual definition?

Remember: In order to be valid, a test must be reliable; but reliability does not guarantee validity, i.e. it is possible to have a highly reliable test which is meaningless (invalid).

Note that where validity coefficients are calculated, they will range between 0 (low) to 1 (high)

Types of Validity

Face validity: Face validity is the least important aspect of validity, because validity still needs to be directly checked through other methods. All that face validity means is:"Does the measure, on the face it, seem to measure what is intended?"

Sometimes researchers try to obscure a measure's face validity - say, if it's measuring a socially undesirable characteristic (such as aggression). The more practical point is to be suspicious of any measures that purport to measure one thing, but seem to measure something different. Perception of job effectiveness indicators is not necessarily a valid indicator of individual job effectiveness.

Construct validity: Construct Validity is the most important kind of validity. If a measure has construct validity *it measures what it purports to measure*. Establishing construct validity is a long and complex process. The various qualities that contribute to construct validity include:

- criterion validity (includes predictive and concurrent)
- convergent validity
- discriminant validity

To create a measure with construct validity, first define the domain of interest (i.e., what is to be measured), then construct measurement items are designed which adequately measure that domain. Then a scientific process of rigorously testing and modifying the measure is undertaken.

Criterion validity

Criterion validity consists of concurrent and predictive validity.

i. **Concurrent validity:** "Does the measure relate to other manifestations of the construct the device is supposed to be measuring?" To establish concurrent validity, correlate measures obtained from a developed instrument with that obtained through an existing equivalent validated instrument.

ii. **Predictive validity:** "Does the test predict an individual's performance in specific abilities?"

Convergent validity: It is important to know whether this tests returns similar results to other tests which purport to measure the same or related constructs. Does the measure match with an external 'criterion', e.g. behaviour or another, well-established, test? Does it measure it concurrently and can it predict this behaviour?

- Observations of dominant behaviour (criterion) can be compared with self-report dominance scores (measure)
- Trained interviewer ratings (criterion) can be compared with self-report dominance scores (measure)

Discriminant or Divergent validity: It is mportant to show that a measure doesn't measure what it isn't meant to measure - i.e. it *discriminates*. For example, discriminant validity would be evidenced by a low correlation between a quantitative reasoning test and scores on a reading comprehension test, since reading ability is an irrelevant variable in a test designed to measure quantitative reasoning.

Sources of Variability

Unreliability of the tool

- **Response sets:** Psychological orientation c bias towards answering in a particular way:

o Acquiescence: tendency to agree, i.e. say "Yes?. Hence use of half -vely and half +vely worded items (but there can be semantic difficulties with -vely wording)

o Social desirability: Tendency to portray self in a positive light. Try to design questions so that social desirability isn't salient.

o Faking bad: Purposely saying 'no' or looking bad if there's a 'reward' (e.g. attention, compensation, social welfare, etc.).

- **Bias**

o Cultural bias: Does the psychological construct have the same meaning from one culture to another? How are the different items interpreted by people from different cultures? Actual content (face) validity may be different for different cultures.

- o Gender bias may also be possible.
- o Test Bias
  - **Bias in <u>measurement</u>** occurs when the test makes systematic errors in measuring a particular characteristic or attribute e.g. many say that most IQ tests may well be valid for middle-class whites but not for blacks or other minorities. In interviews, which are a type of test, research shows that there is a bias in favour of good-looking applicants.

    Bias in <u>prediction</u> occurs when the test makes systematic errors in predicting some outcome (or criterion). It is often suggested that tests used in academic admissions and in personnel selection under-predict the performance of minority applicants. Also, a test may be useful for predicting the performance of one group e.g. males but be less accurate in predicting the performance of females.

Generalizability

Just a brief word on generalizability. Reliability and validity are often discussed separately but sometimes you will see them both referred to as aspects of generalizability. Often we want to know whether the results of a measure or a test used with a particular group can be generalized to other tests or other groups. So a test may be reliable and it may be valid but its results may not be generalizable to other tests measuring the same construct nor to populations other than the one sampled.

For example, if I measured the levels of aggression of a very large random sample of children in primary schools in the Niger-Delta region Nigeria where children observe physical violence frequently, I may use a scale which is perfectly reliable and a perfectly valid measure of aggression. But would my results be exactly the same had I used the same instrument on their counterpart in South-West where less violence is witnessed. Indicators of aggression may vary from region to region. Shouting while talking is regarded as aggression indicator but it is not so in some other regions. Indicators of aggression among children may differ from that of adolescents.

Standardization: Standardized tests are:

- administered under uniform conditions. i.e. no matter where, when, by whom or to whom it is given, the test is administered in a similar way.

- scored objectively, i.e. the procedures for scoring the test are specified in detail so that number of trained scorers will arrive at the same score for the same set of responses. So for example, questions that need subjective evaluation (e.g. essay questions) are generally not included in standardized tests.
- designed to measure relative performance. i.e. they are not designed to measure ABSOLUTE ability on a task. In order to measure relative performance, standardized tests are interpreted with reference to a comparable group of people, the standardization, or normative sample. e.g. Highest possible grade in a test is 100. Child scores 60 on a standardized achievement test. You may feel that the child has not demonstrated mastery of the material covered in the test (absolute ability) BUT if the average of the standardization sample was 55 the child has done quite well (RELATIVE performance).

The normative sample should (for hopefully obvious reasons!) be representative of the target population - however this is not always the case, thus norms and the structure of the test would need to be interpreted with appropriate caution.

CHAPTER EIGHT

Sample and Sampling

A sample

A finite part or a subset of a statistical population whose properties are studied to gain information about the whole. Typically, the population is very large, making a census or a complete enumeration of all the values in the population impractical or impossible. The sample represents a subset of manageable size. Samples are collected and statistics are calculated from the samples so that one can make inferences or extrapolations from the sample to the population. This process of collecting information from a sample is referred to as sampling.

Sampling

There are two major types of sampling namely, Probability and Non probability Sampling.

Probabilistic samples	Nonprobability samples
• Random Selection • Know the odds or probability that we have represented the population well. • Able to estimate confidence intervals for the statistic. • Generally more accurate and rigorous, but not always feasible	•No random Selection • May or may not represent the population well • It will often be hard to estimate the confidence interval for the statistics. • Very easy to obtain

Probability Sampling

Probability sampling technique ensures that *bias* is not introduced into research studies through who is included in the survey or experiment. Five common Probability sampling techniques are:

- Simple random sampling,
- systematic sampling,
- stratified sampling,
- cluster sampling, and
- multi-stage sampling.

Simple Random Sampling

With simple random sampling, each item in a population has an equal chance of inclusion in the sample. For example, each name on a list of SSII students in a school could be numbered sequentially. If the sample size was to include 150 students, then 150 numbers could be randomly generated by computer or numbers could be picked out of a hat. These numbers could then be matched to names on the list, thereby providing a list of 150 students.

The advantage of simple random sampling is that it is simple and easy to apply when small populations are involved. However, because every person or item in a population has to be listed before the corresponding random numbers can be read, this method is very cumbersome to use for large populations.

Systematic Sampling

Systematic sampling, sometimes called interval sampling, means that there is a gap, or interval, between each selection. This method is often used in industry, where an item is selected for testing from a production line (say, every fifteen minutes) to ensure that machines and equipment are working to specification.

Alternatively, the manufacturer might decide to select every 25th item on a production line to test for defects and quality. This technique requires the first item to be selected at random as a starting point for testing and, thereafter, every 25th item is chosen. This technique could also be used when questioning people in a sample survey. A market researcher might select every 7th person who enters a particular store, after selecting a person at random as a starting point; or interview occupants of every 6th house in a street, after selecting a house at random as a starting point.

It may be that a researcher wants to select a fixed size sample. In this case, it is first necessary to know the whole population size from which the sample is being selected. The appropriate *sampling interval*, I, is then calculated by dividing population size, N, by required sample size, n, as follows:

$$I = N/n$$

For example if a systematic sample of 250 students were to be carried out in Public secondary School with an enrolled population of 1,000, the sampling interval would be:

$$I = N/n = 1000/250 = 4$$

Note: if I is not a whole number, then it is rounded to the nearest whole number.

All students would be assigned sequential numbers. The starting point would be chosen by selecting a random number between 1 and 4. If this number was 3, then the 3^{rd} student on the list of students would be selected along with every following 4th student. The sample of students would be those corresponding to student numbers 3, 7, 11, 15,19 ....991, 995, 999.

The advantage of systematic sampling is that it is simpler to select one random number and then every 'Ith' (e.g. 4th) member on the list, than to select as many random numbers as sample size. It also gives a good spread right across the population. A disadvantage is that you may need a list to start with, if you wish to know your sample size and calculate your sampling interval.

Stratified Sampling

A general problem with random sampling is that you could, by chance, miss out a particular group in the sample. However, if you form the population into groups, and sample from each group, you can make sure the sample is representative.

In stratified sampling, the population is divided into groups called strata. A sample is then drawn from within these strata. Some examples of strata commonly used are States, Age and Sex. Other strata may be religion, academic ability or marital status. For example, a researcher can ensure that the population of students is divided into male and female groups before simple or systematic random sampling is employed to select from each group.

Stratification is most useful when the stratifying variables are simple to work with, easy to observe and closely related to the topic of the survey. An important aspect of stratification is that it can be used to select more of one group than another. You may do this if you feel that responses are more likely to vary in one group than another. So, if you know everyone in one group has much the same value, you only need a small sample to get information for that group; whereas in another group, the values may differ widely and a bigger sample is needed.

Cluster Sampling

It is sometimes expensive to spread your sample across the population as a whole. For example, travel can become expensive if you are using interviewers to travel between people spread all over the country. To reduce costs you may choose a cluster sampling technique.

Cluster sampling divides the population into groups, or clusters. A number of clusters are selected randomly to represent the population, and then all units within selected clusters are included in the sample. No units from non-selected clusters are included in the sample. They are represented by those from selected clusters. This differs from stratified sampling, where some units are selected from each group.

Suppose a company wishes to find out which brand of can milk Undergraduates in Nigeria Universities prefer. It would be too costly and take too long to survey every student, or even some students from every school. Instead, 20 universities can be randomly selected from all over Nigeria, and every level should be involved to ensure representativeness. Universities are considered to be clusters In effect, students in the sample of 20 universities represent all undergraduate students in Nigeria.

Cluster sampling has several advantages: reduced costs, simplified field work and administration is more convenient. Instead of having a sample scattered over the entire coverage area, the sample is more localised in relatively few centres (clusters).

Cluster sampling's disadvantage is that less accurate results are often obtained due to higher sampling error than for simple random sampling with the same sample size. In the above example, you might expect to get more accurate estimates from randomly selecting students across all universities than from randomly selecting 20 universities and taking every student in those chosen.

Multi-Stage Sampling

Multi-stage sampling involves selecting a sample in at least two stages. In the first stage, large groups or clusters are selected. These clusters are designed to contain more population units than are required for the final sample. In the second stage, population units are chosen from selected clusters to derive a final sample. If more than two stages are used, the process of choosing population units within clusters continues until the final sample is achieved.

- An example of multi-stage sampling is where, Nigeria is divided into Geo-political zones as population units. The selected units are also divided to states then to local government areas, and then to schools before sample students are drawn.

The advantages of multi-stage sampling are convenience, economy and efficiency. Multi-stage sampling does not require a complete list of members in the target population, which greatly reduces sample preparation cost. The list of members is required only for those clusters used in the final stage. The main disadvantage of multi-stage sampling is the same as for cluster sampling: lower accuracy due to higher sampling error.

Non probability Sampling

Does that mean that nonprobability samples aren't representative of the population? Not necessarily. According to Trochim (2006), it does mean that nonprobability samples cannot depend upon the rationale of probability theory. At least with a probabilistic sample, we know the odds or probability that we have represented the population well. We are able to estimate confidence intervals for the statistic. With nonprobability samples, we may or may not represent the population well, and it will often be hard for us to know how well we've done so. In general, researchers prefer probabilistic or random sampling methods over nonprobabilistic ones, and consider them to be more accurate and rigorous. However, in applied social research there may be circumstances where it is not feasible, practical or theoretically sensible to do random sampling. Here, we consider a wide range of nonprobabilistic alternatives.

Nonprobability sampling methods can be divided into two broad types: *accidental* or *purposive*. Most sampling methods are purposive in nature because we usually approach the sampling problem with a specific plan in mind. The most important distinctions among these types of sampling methods are the ones between the different types of purposive sampling approaches.

Accidental, Haphazard or Convenience Sampling

In many research contexts, we sample simply by asking for volunteers. Clearly, the problem with all of these types of samples is that we have no evidence that they are representative of the populations we're interested in generalizing to and in many cases we would clearly suspect that they are not.

Purposive Sampling

In purposive sampling, researchers sample with a *purpose* in mind. We usually would have one or more specific predefined groups we are seeking. When the target sample of a research is only pregnant women most likely they are conducting a purposive sample (and most likely they are engaged in research on anti-natal care). Any point such researcher stands he might be looking for a pregnant women passing. They size up the people passing by and anyone who looks to be in that category they stop to ask if they will participate. One of the first things they're likely to do is verify that the respondent does in fact meet the criteria for being in the sample. Purposive sampling can be very useful for situations where you need to reach a targeted sample quickly and where sampling for proportionality is not the primary concern. With a purposive sample, you are likely to get the opinions of your target population, but you are also likely to overweight subgroups in your population that are more readily accessible.

All of the methods that follow can be considered subcategories of purposive sampling methods. We might sample for specific groups or types of people as in modal instance, expert, or quota sampling. We might sample for diversity as in heterogeneity sampling. Or, we might capitalize on informal social networks to identify specific respondents who are hard to locate otherwise, as in snowball sampling. In all of these methods, we know what we want is sampling with a purpose.

Modal Instance Sampling

In statistics, the *mode* is the most frequently occurring value in a distribution. In sampling, when we do a modal instance sample, we are sampling the most frequent case, or the "typical" case. In a lot of informal public opinion polls, for instance, they interview a "typical" voter. There are a number of problems with this sampling approach. First, how do we know what the "typical" or "modal" case is? We could say that the modal voter is a person who is of average age, educational level, and income in the population. But, it's not clear that using the averages of these is the fairest (consider the skewed distribution of income, for instance). And, how do you know that those three variables — age, education, income — are the only or even the most relevant for classifying the typical voter? What if religion or ethnicity is an important discriminator? Clearly, modal instance sampling is only sensible for informal sampling contexts.

Expert Sampling

Expert sampling involves the assembling of a sample of persons with known or demonstrable experience and expertise in some area. Often, we convene such a sample under the auspices of a "panel of experts." There are actually two reasons you might do expert sampling. First, because it would be the best way to elicit the views of persons who have specific expertise. In this case, expert sampling is essentially just a specific subcase of purposive sampling. But the other reason you might use expert sampling is to provide evidence for the validity of another sampling approach you've chosen. For instance, let's say you do modal instance sampling and are concerned that the criteria you used for defining the modal instance are subject to criticism. You might convene an expert panel consisting of persons with acknowledged experience and insight into that field or topic and ask them to examine your modal definitions and comment on their appropriateness and validity. The advantage of doing this is that you aren't out on your own trying to defend your decisions, you have some acknowledged experts to back you. The disadvantage is that even the experts can be, and often are, wrong.

Quota Sampling

In quota sampling, you select people nonrandomly according to some fixed quota. There are two types of quota sampling: *proportional* and *non proportional*. In proportional quota sampling you want to represent the major characteristics of the population by sampling a proportional amount of each. For instance, if you know the population has 40% women and 60% men, and that you want a total sample size of 100, you will continue sampling until you get those percentages and then you will stop. So, if you've already got the 40 women for your sample, but not the sixty men, you will continue to sample men but even if legitimate women respondents come along, you will not sample them because you have already "met your quota." The problem here (as in much purposive sampling) is that you have to decide the specific characteristics on which you will base the quota. Will it be by gender, age, education race, religion, etc.?

Non proportional quota sampling is a bit less restrictive. In this method, you specify the minimum number of sampled units you want in each category. Here, you're not concerned with having numbers that match the proportions in the population. Instead, you simply want to have enough to assure that you will be able to talk about even small groups in the population. This method is the nonprobabilistic analogue of stratified random sampling in that it is typically used to assure that smaller groups are adequately represented in your sample.

Heterogeneity Sampling

Researchers sample for heterogeneity when we want to include all opinions or views, and we aren't concerned about representing these views proportionately. Another term for this is sampling for *diversity*. In many brainstorming or nominal group processes (including concept mapping), we would use some form of heterogeneity sampling because our primary interest is in getting broad spectrum of ideas, not identifying the "average" or "modal instance" ones. In effect, what we would like to be sampling is not people, but ideas. We imagine that there is a universe of all possible ideas relevant to some topic and that we want to sample this population, not the population of people who have the ideas. Clearly, in order to get all of the ideas, and especially the "outlier" or unusual ones, we have to include a broad and diverse range of participants. Heterogeneity sampling is, in this sense, almost the opposite of modal instance sampling.

Snowball Sampling

In snowball sampling, you begin by identifying someone who meets the criteria for inclusion in your study. You then ask them to recommend others who they may know who also meet the criteria. Although this method would hardly lead to representative samples, there are times when it may be the best method available. Snowball sampling is especially useful when you are trying to reach populations that are inaccessible or hard to find. For instance, if you are studying the homeless, you are not likely to be able to find good lists of homeless people within a specific geographical area. However, if you go to that area and identify one or two, you may find that they know very well who the other homeless people in their vicinity are and how you can find them.

CHAPTER NINE

STATISTICS

Statistics is concerned with scientific methods for collecting, organizing, summarizing, presenting, and analyzing data as well as with drawing valid conclusion and making reasonable decisions on the basis of such analysis.

Raw Data: Collected information or data that have not been organized numerically. Raw scores of 200 master students on EVE 707 written on a paper.

Arrays: An Array is an arrangement of raw numerical data in ascending or descending order of magnitude. When the scores of 200 Masters students on EVE 707 are carefully arranged from smallest to the highest, then the array of the scores is presented.

Graphical Representation of Data

The graphical representation of data makes the reading more interesting, less time-consuming and easily understandable. The disadvantage of graphical presentation is that it lacks details and is less accurate. There are so many ways data can be represented, only four common graphs will be presented in this book. They are 1. Bar Graphs 2. Pie Charts 3. Frequency Polygon 4. Histogram.

Bar Graphs

This is the simplest type of graphical presentation of data. The following types of bar graphs are possible:

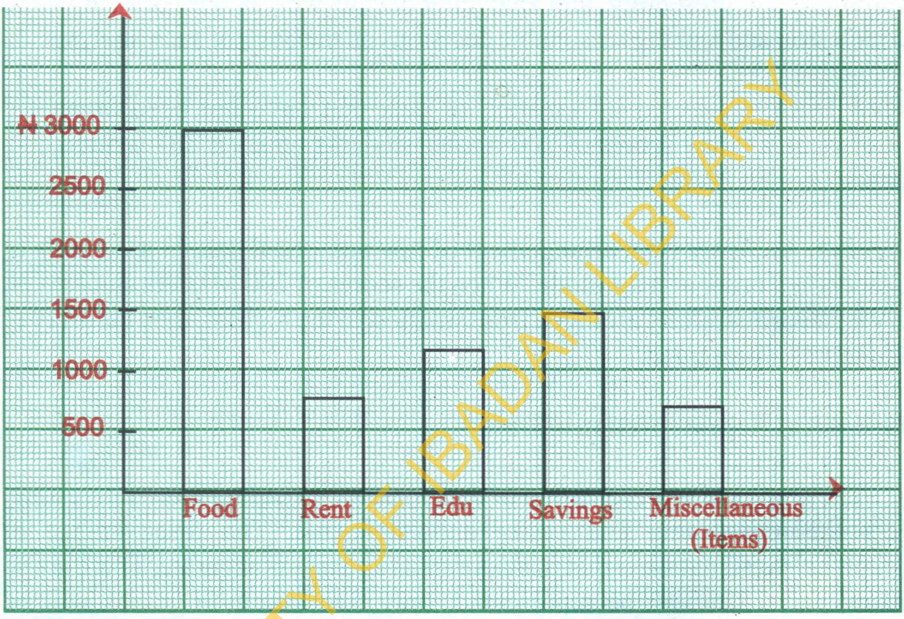
(a) Simple bar graph

(b) Double bar graph

(c) Divided bar graph.

Example1

Item	Food	Rent	Education	Savings	Misc.	Total
Amount (₦)	3000	800	1200	1500	700	7200



Bar Chart

Pie Graph or Pie Chart

Sometimes a circle is used to represent a given data. The various parts of it are proportionally represented by sectors of the circle. Then the graph is called a Pie Graph or Pie Chart

To find the angle of each sector

Total of data corresponds to 360°.

Let x° = the angle at the centre for item A, then

$$x^0 = \frac{\text{Value of item A}}{\text{Total value of all the items}} \times 360^0$$

Example 2

The data given in example 1 can be used to draw a pie graph.
Calculation of Angles

Food:

$$\text{Angle at centre} = \frac{360}{7200} \times 3000$$

$= 150^{\circ}$

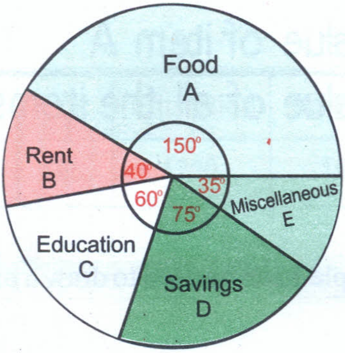
Rent:

$$\text{Angle at centre} = \frac{360}{7200} \times 800$$

$= 40^{\circ}$

Similarly we can calculat the remaining angles, and the total of angles column should always come to 360°.

Item	Amount (₦)	Angle
Food (A)	3000	150^0
Rent (B)	800	40^0
Education (C)	1200	60^0
Savings (D)	1500	75^0
Miscellaneous	700	35^0
Total	**7200**	**360^0**



Frequency Polygon

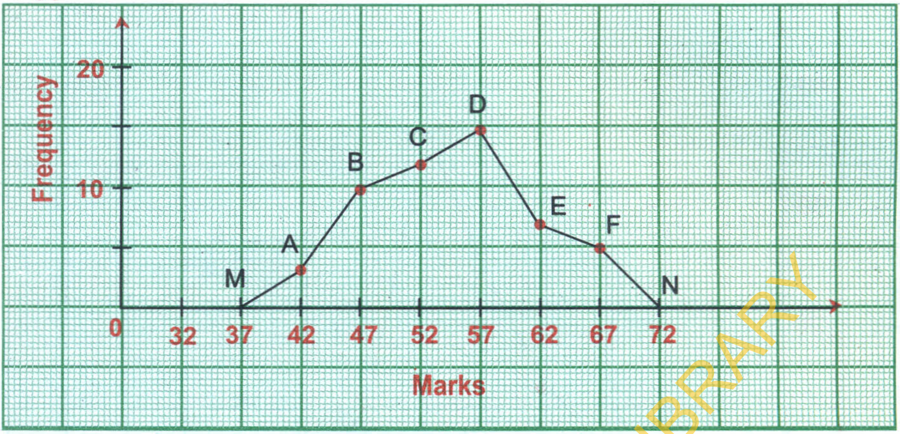
In a frequency distribution, the mid-value of each class is obtained. Then on the graph paper, the frequency is plotted against the corresponding mid-value. These points are joined by straight lines. These straight lines may be extended in both directions to meet the X - axis to form a polygon.

Example3

The weights of 50 students are recorded below. Draw a frequency polygon for this data.

Class	Mid-mark	Frequency
40 - 44	42	3
45 - 49	47	10
50 - 54	52	12
55 - 59	57	15
60 - 64	62	7
65 - 69	67	5

Suggested answer:

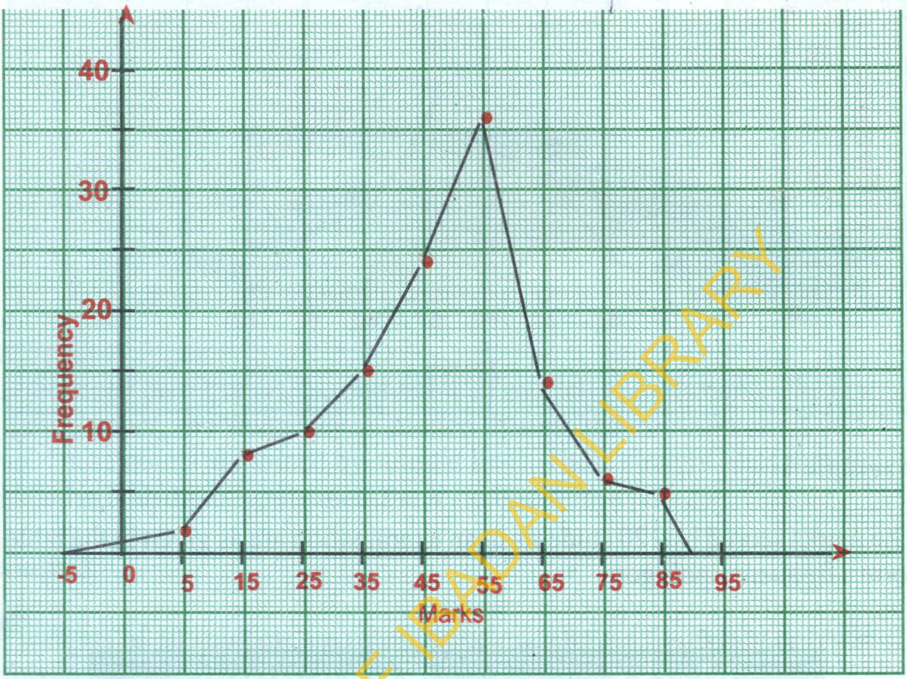


Example 4

The marks scored by 120 students in an examination are as given in the table form a frequency polygon.

Marks	Frequency
0 - 10	2
10 - 20	8
20 -30	10
30 -40	15
40 -50	24
50 -60	36
60 -70	14
70 -80	6
80 -90	5

Suggested answer:



Histogram

A histogramis a diagram which represents the class interval and frequency in the form of a rectangle.

To draw a histogram, follow the steps stated below

(1) Mark class intervals on X-axis and frequencies on Y-axis.

(2) The scales for both the axes need not be the same.

(3) Class intervals must be exclusive. If the intervals are in inclusive form, convert them to the exclusive form.

(4) Draw rectangles with class intervals as bases and the corresponding frequencies as heights.

The class limits are marked on the horizontal axis and the frequency is marked on the vertical axis. Thus a rectangle is constructed on each class interval.

If the intervals are equal, then the height of each rectangle is proportional to the corresponding class frequency.

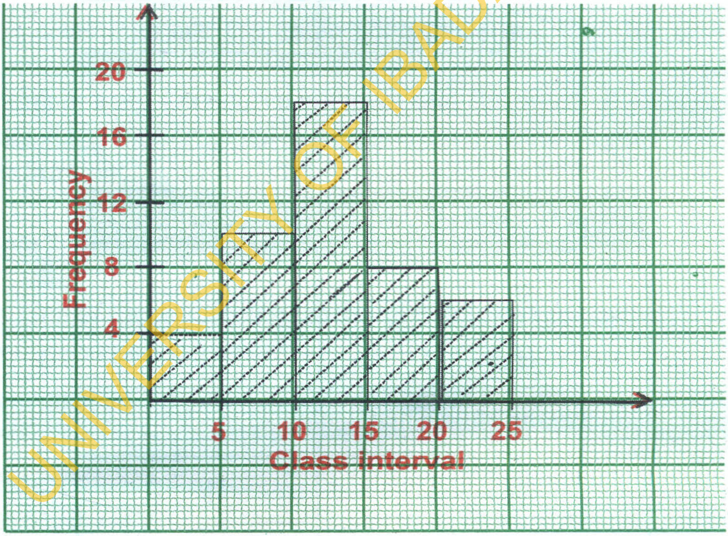
If the intervals are unequal, then the area of each rectangle is proportional to the corresponding class frequency.

Example 5:

Draw a histogram for the following data:

Class Interval	Frequency
0 - 5	4
5- 10	10
10 - 15	18
15 - 20	8
20 -25	6

Suggested answer:



Note

In the above example, the intervals are exclusive. Now, let us consider an example with inclusive intervals.

Example 6

The daily wages of 50 workers, in rupees, are given below:

In table (a), the class intervals are inclusive. So we convert them to the exclusive form as shown in table (b).

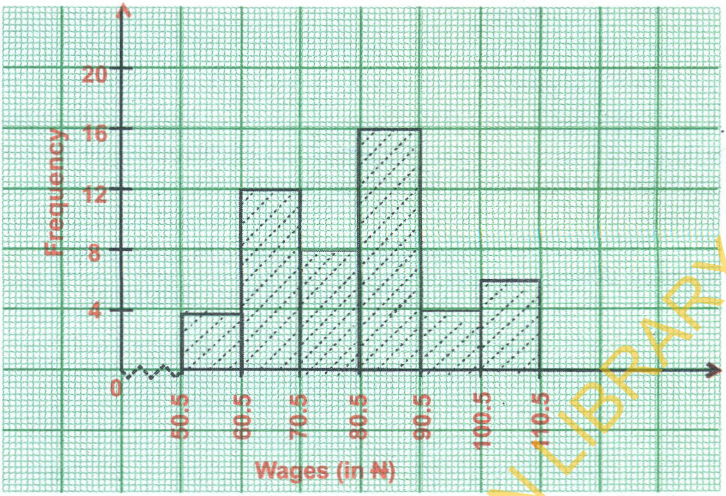
Table (a)

Wages (in ₦.)	Frequency
51 - 60	4
61 - 70	12
71 - 80	8
81 - 90	16
91 - 100	4
101 - 110	6

Table (b)

Wages (in ₦.)	Frequency
50.5 - 60.5	4
60.5 - 70.5	12
70.5 - 80.5	8
80.5 -90.5	16
90.5 - 100.5	4
100.5 110.5	6

Suggested answer:



Note:

(i) The class intervals are made continuous and then the histogram is constructed.

(ii) A kink or a zig - zag curve is shown near the origin. It indicates that the scale along the horizontal axis does not start at the origin.

(iii) The horizontal scale and vertical scale need not be the same.

Example 7

Distribution of shops according to the number of wage - earners employed at a shopping complex is given thus:

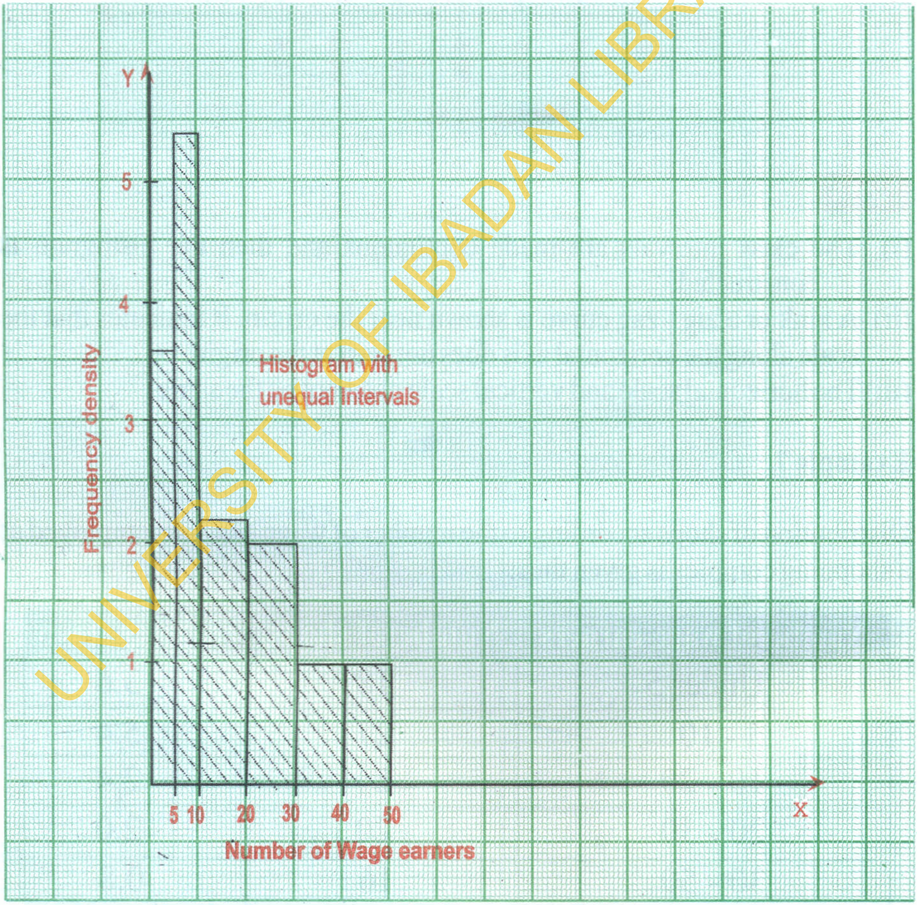
Numberof wage-earners	Numberof shops	Frequency density
Under 5	18	3.6
5 - 10	27	5.4
10 - 20	24	2.4
20 - 30	20	2.0
30 - 50	16	0.8

Illustrate the above table by a histogram, showing clearly how you deal with the unequal class intervals.

Suggested answer:

Note:
When the intervals are unequal, we construct each rectangle with the class intervals as base and frequency density as height.

$$\text{Frequency density} = \frac{\text{Frequency of the class interval}}{\text{Class size of the interval}}$$



Comparison of Histogram and Bar Graph

Histogram	Bar Graph
1. It con of resists of rectangles touching each other.	1. It consists rcctanglcs, normally scparatcd from cach othcr with cqual space.
2. The frequency is represented by the area of each rectangle.	2. The frequency is represented by height. The width has no significance.
3. It is two dimensional (width and height are considered.)	3. It is one dimensional (only height is considered.)
	4. It is used as a visual aid to represent data.

Measures of Central Tendency and Dispersion

The table below presents the distribution of the maximum loads in short tons (1 short ton = 2000 lb) supported by certain cables produced by a company. The information in the table was used to estimate different measures.

Frequency Table

Max. Load	No of Cable	Class mark	Class Boundaries	fx	$x-\bar{x}$	$(x-\bar{x})^2$	$f(x-\bar{x})^2$
9.3-9.7	2	9.5	9.25-9.75	19	-1.592	2.534464	5.068928
9.8-10.2	5	10	9.75-10.25	50	-1.092	1.192464	5.96232
10.3-10.7	12	10.5	10.25-10.75	126	-0.592	0.350464	4.205568
10.8-11.2	17	11	10.75-11.25	187	-0.092	0.008464	0.143888
11.3-11.7	14	11.5	11.25-11.75	161	0.408	0.166464	2.330496
11.8-12.2	6	12	11.75-12.25	72	0.908	0.824464	4.946784
12.3-12.7	3	12.5	12.25-12.75	37.5	1.408	1.982464	5.947392
12.8-13.2	1	13	12.75-13.25	13	1.908	3.640464	3.640464
Total	**60**			**665.5**			**32.24584**

$\bar{X}$ **= the mean of the distribution.**

Measures of Central Tendency

Mean

Mean usually denoted by $\bar{X}$

Mean

Mean usually denoted by $\bar{X}$

$$\text{Mean} = \bar{X} = \frac{\sum fx}{\sum f} \quad \text{or} \quad \frac{\sum fx}{N}$$

Where: X is the class mark
f is the frequency
N is total number of cases.

$$:.\ \bar{X} = \frac{665.5}{60} = 11.0917$$

$$\approx 11.09$$

Median

Median simply refers to the mark at the middle for grouped data.

$$\text{Median} = L + \left(\frac{\frac{N}{2}Cf_b}{f_w}\right)C$$

Where: L is the lower limit of the median group.
N is the total number of cases
Cf_b is the cumulative frequency before the median group.
f_w is the frequency within the median group
C is the class size.

Hence

$$\text{Median} = 10.75 + \left(\frac{30\text{-}19}{17}\right) 0.5$$

$$= 10.75 + \left(\frac{11}{17}\right)0.5$$

$$= 10.75 + 0.3235$$

$$= 11.07$$

Mode

Mode refers to the most occurred cases in a distribution.

$$\text{Mode} = L + \left(\frac{D_x}{D_x + D_y}\right)C$$

Where L = lower limit of the modal group.

D_x = difference in frequency between the modal group and the group before it

D_y = difference in frequency between the modal group and the group after it.

C = class size

Hence,

$$\text{Mode} = 10.75 + \left(\frac{5}{5+3}\right)0.5$$
$$= 10.75 + \left(\frac{5}{8}\right)0.5$$
$$= 11.0625$$
$$\approx 11.06$$

Quartiles

Quartiles refer to values which divide a distribution into four equal parts. Q_1, Q_2, and Q_3 are referred to as first, second and third quartiles respectively.

Hence,

$$\text{First Quartile} = Q_1 = L + \frac{N/4 - Cf_b}{f_w} C$$
$$= 10.25 + \left(\frac{15-7}{12}\right)0.5$$
$$= 10.25 + 0.3333 = 10.5833$$
$$\approx 10.58$$

$$\text{Third Quartile} = Q_3 = L + \frac{3N/4 - Cf_b}{f_w} C$$

$$= 11.25 + \left(\frac{45-36}{14}\right)0.5$$
$$= 11.25 + \left(\frac{9}{14}\right)0.5$$
$$= 11.25 + 0.3214$$
$$= 11.5714$$
$$\approx 11.57$$

$$\text{Inter Quartile Range} = Q_3 - Q_1$$
$$= 11.57 - 10.58$$
$$= 0.99$$

$$\text{Semi Inter Quartile Range} = \frac{Q_3 - Q_1}{2}$$
$$= \frac{0.99}{2}$$
$$= 0.495$$
$$\approx 0.5$$

Decile

Deciles refer to values which divide a distribution into ten equal parts .

Hence;

$$\text{Seventh Decile} = D_7 = L + \left[\frac{7N/10 - cf_b}{f_w}\right]C$$

$$= 11.25 + \left(\frac{42-36}{14}\right)0.5$$
$$= 11.25 + \left(\frac{6}{14}\right)0.5$$
$$= 11.25 + 0.2143$$
$$= 11.4643$$
$$\approx 11.46$$

Percentile

Percentiles refer to values which divide a distribution into four equal parts .

Hence;

$$\text{Sixty Seventh Percentile} = P_{67}$$

$$\therefore P_{67} = L + \left[\frac{67N/100 - cf_b}{f_w}\right]C$$

$$= 11.25 + \left(\frac{40.2-36}{14}\right)0.5$$

$$= 11.25 + \left(\frac{4.2}{14}\right) 0.5$$

$$= 11.25 + 0.15$$

$$= 11.4$$

Measure of Dispersion.

Range equals the difference between the highest and the lowest scores in a distribution.
Hence;
Range = 13.2 – 9.3 = 3.9

Variance and Standard Deviation.

Variance and standard deviation give the picture of how far apart are the scores. The square root of the estimated variance gives the standard deviation of the distribution.

Hence;

$$\text{Variance} = \frac{\sum f\,(\bar{X} - X)}{N}$$

$= 32.24584$
$= 0.5374$
≈ 0.54

$$\text{Standard Deviation} = SD = \sqrt{\textit{variance}}$$

$\therefore SD = \sqrt{0.5374}$
$= 0.7331$
≈ 0.73

Exercise

Frequency distribution of grades on a final examination in EVE 707 is presented below:

Grade	Number of Students
90-100	9
80-89	32
70-79	43
60-69	21
50-59	11
40-49	3
30-39	1

Find

1. The (i) mean (ii) Median and (iii) Mode
2. The first, second and third quartiles
3. 4th, 7th and 9th deciles
4. 27th, 55th, 75th and 83rd percentiles
5. Semi inter-quartile range
6. The (i) Variance and (ii) Standard deviation.

Basic Statistical Terms

Descriptive Statistics: - refers to procedure for organizing, summarizing and describing quantitative information or data. Examples of descriptive statistics are Mean, Median, Mode, Frequency Counts, Percentage, Standard Deviation e.t.c.

Inferential Statistics: - Concern the methods by which induction are made to a larger group (population) on the basis of observations made on a smaller subgroup (Sample).

Example of Inferential Statistics are:

Parametrics	Non - Parametrics
(I) t – test (dependent group)	(I) Wilcoxon Test for two corrected samples
(II) t – test (independent group)	(II) Mann w hitney U – test
(III) ANOVA (One way)	(III) Kruskal – Wallis test for K indep. Sample.
(IV) ANOVA (Two way)	(IV) Fres man's test
(V) ANCOVA	(V) Spearman Rank order correlation Coefficient
(VI) Pearson product moment correlation Coefficient.	(VI) Chi - Square

Statistics is the study of methods of handing quantitative information (data) on sample. These methods include techniques for organizing and summarizing as well as making generalizations and inferences from data. The two broad classes are descriptive and inferential Statistics. It concerns the characteristics of sample.

Parameter is the study of methods of handling quantitative information (data) on population. It concerns the characteristics of population.

Real Limits of a number are those points falling One – half a measure unit above and one – half a measurement unit below that number.

Number	Lower real limit	Upper real limit
(i) 4	3.5	4.5
(ii) 5.1	5.05	5.15
(iii) 6.84	6.835	6.845
(iv) 10	9.5	10.5
(v) 10004.6	10004.55	10004.65
(vi) 16.7	16.65	16.75

Population is an identifiable group of individuals or events. All the Students and the Staff of a school can form the population of the school.

Sample is a subset of a population. The group of SS 2 is a sample drawn from entire population of a Secondary School.

Standard Score are produced through the interpretation of scores within a distribution based upon their relative standings with respect to the mean and the variability (S.D) of the scores within the distribution. A standard score indicates how much below a given the remainder of the distribution was.

The Standard Score is denoted by Z.

$$Z = \frac{X - \overline{X}}{S}$$

Where X is the raw score to be transformed, $\overline{X}$ is the group mean and S is the Standard Deviation.

Area under a theoretical distribution signifies the proportion of frequencies, which occur between various of Z.

Idealized Experiment is an experiment in which a given phenomenon is repeatedly observed an indefinite number of under ideal conditions. For example, if somebody says the chance of tossing a fair die and obtaining a 4 is $^{1}/_{6}$, this implies that over an uncountable number of tosses of that fair die under ideal conditions, 4s will turn up ½ the time.

Probability Usually involves thinking about an idealized experiment. Hence probability of a given event E is defined to be the number of outcomes in E divided by the total number outcomes in the Sample Space S.

$$P(A) = \frac{n(A)}{n(S)}$$

Theoretical Relative Frequency is the relative frequency of score values in a theoretical distribution based on unlimited (infinite) number of cases. However probability is theoretical relative frequency.

Directional and Non – Directional Hypotheses.

Basis of deciding between a directional and a non – directional hypothesis is the availability of evidence or theory that might predict the result.

Alpha value: Selection of the value of alpha (á) depends on the following conditions.

i. How critical it is to be wrong in rejecting the null hypothesis

ii. Tolerance for error in decision making

iii. Desire for statistical power.

Statistical Assumptions are characteristics of a statistical situation or context. Guess to be true. For example, while testing the hypothesis that boys perform better in mathematics than girls, the assumptions were that:

(1) boys were randomly selected
(2) Sampling distribution of the mean is normal and
(3) N is relatively large.

Statistical Hypothesis represents a set of two or more contradiction and often exhaustive phenomena, only one of which can actually be the case. The hypothesis that is tentatively held to be true is called the hull hypothesis (H_0) while the other is alternative hypothesis (H_1). For example: H_0: boys do no perform better than girls in mathematics. H_1: boys do perform better than girls in mathematics.

Differences between Assumption and Hypotheses.

Assumption	Hypotheses
i. All assumption are expected to be true. ii. Need not to be tested.	i. Only one of the hypotheses is true. ii. Needed to be tested using statistical tool.

Decision A statistician never accept H_1 and H_0 respectively because H_1 and H_0 rejected at a particular á level say 0.05 may not be rejected at another á level, hence has no right to categorically accept H_0 and H_1. Also, H_0 and H_1 cannot be accepted at same time because if one is rejected the other will not be rejected.

Type I and Type II Error

A type I errors occurs when the null hypothesis is rejected. When it is in fact true then it is denoted by á. A type II error occurs when the null hypothesis is not rejected. When it is fact, false it is denoted by α. The four possible outcomes of a simple decision process and their associated probability.

	H_0 is true	**H_0 is false**
Decision: Reject H_0 Do not reject H_0.	Type I error ($P = \alpha$)	Correct decision $P = 1 - \beta$ (Power)
	Correct decision ($P = \alpha$)	Type II error ($P = \beta$)

Power of a Statistical Test.

The power of a statistical test is the probability that the test will correctly decided to reject H_0 when H_0 is indeed false. Thus, the power of a statistical test tends to increase as the significant level, á increase therefore a test performed at a significant level 0.05 has more power than at 0.01.

Assumption for Using Parametric Test: It must be assumed that the population distribution of the difference between means is normal to ensure that the two variance estimates are independent. The conditions for normality are (1) sample randomly selected to represent the population. (ii) Sufficient number of cases is sampled. Note: if these conditions are not met, non parametric test must be opted for.

Main and Interaction Effect:

when the researcher is investigating the effect of two factors A and B on a dependent variable, the possible differences between level of factor A or levels of factor B collapsed under the other factor are called main effects. The potential joint resultant effect of both factors A and B is known as an interaction effect.

Example of main effect: when a researcher is examining the effect of 3 teaching method on achievement in mathematics.

Example of interaction effect: a researcher examining effect teaching method across students gender(male and female groups) on achievement in mathematics.

CHAPTER TEN

Hypothesis Testing

Whenever we have a decision to make about a population characteristic, we make a hypothesis. Some examples are:

$\mu < 3$ (The mean is less than 3)

$\mu \neq 3$. (The mean is different from 3)

Suppose that we want to test the hypothesis that $\mu \neq 3$. Then we can think of our opponent suggesting that $\mu = 3$. We call the opponent's hypothesis the *null hypothesis* and write:

H_0: $\mu = 3$

and our hypothesis the alternative hypothesis and write

H_1: $\mu \neq 3$

For the null hypothesis we always use equality, since we are comparing m with a previously determined mean.
For the alternative hypothesis, we have the choices: <, >, or ≠

Procedures in Hypothesis Testing

When we test a hypothesis we proceed as follows:

1. Formulate the null and alternative hypothesis.
2. Choose a level of significance.
3. Determine the sample size. (Same as confidence intervals)
4. Collect data.
5. Calculate z (or t) score.
6. Utilize the table to determine if the z score falls within the acceptance region.
7. Decide to

a. reject the null hypothesis and therefore accept the alternative hypothesis or

b. fail to reject the null hypothesis and therefore state that there is not enough evidence to suggest the truth of the alternative hypothesis.

Errors in Hypothesis Tests

We define a *type I error* as the event of rejecting the null hypothesis when the null hypothesis was true. The probability of a type I error (a) is called the significance level.

We define a type II error (with probability b) as the event of failing to reject the null hypothesis when the null hypothesis was false.

Example

Suppose an Education expert has carried out an experiment by using a new teaching method to deliver instructional content to some mathematics. He then claimed that the students taught with the new method will perform better in Mathematics. Suppose the mean Mathematics Achievement is 74%.

You set the null hypothesis to be

$H_0: \mu = 74$

$H_1: \mu > 74$

Q. What is a type I error?

A. We agree with the claim of the education expert and encourage the teaching strategy.

Q. What is a type II error?

A. We disagree with the claim of the education expert

Hypothesis Testing For a Population Mean

The Idea of Hypothesis Testing

Suppose we want to show that only Fathers from the Northern Nigeria have the number of children higher than national average. Suppose the mean number of children of fathers in Nigeria is 5. Construct the relevant one tail hypothesis test:

H_0: $\mu = 5$ (sample mean not different from population mean)

H_1: $\mu > 5$ (sample mean greater than population mean)

We investigated only 100 Northern Fathers and find that

$\bar{X} = 8$

and suppose we know the population standard deviation

$\sigma = 2$.

Do we have evidence to suggest that Northern Nigeria fathers have an average number of children higher than the national average? We have

$$Z = \frac{\bar{X} - \mu}{\delta / \sqrt{n}}$$

Where: z is called the *test statistic*.

$\bar{X}$ = Sample mean

μ = Population mean

δ = Population Standard Deviation and

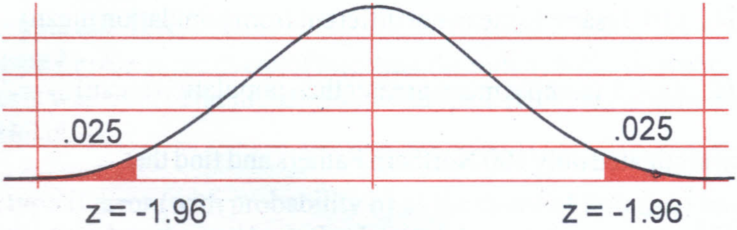
n = sample size

$$\text{Hence } Z = \frac{8-5}{2/\sqrt{100}} = 15$$

Since z is so high, the probability that Ho is true is so small that we decide to reject H_0 and accept H_1. Therefore, we can conclude that Fathers in the north have more children.

Rejection Regions

Suppose that a = .05. We can draw the appropriate picture and find the z score for -.025 and .025. We call the outside regions the rejection regions for two tail hypothesis.



We call the red areas the *rejection region* since if the value of z falls in these regions, we can say that the null hypothesis is very unlikely so we can reject the null hypothesis

Example

50 smokers were questioned about the number of hours they sleep each day. We want to test the hypothesis that the smokers need less sleep than the general public which needs an average of 7.7 hours of sleep. We follow the steps below.

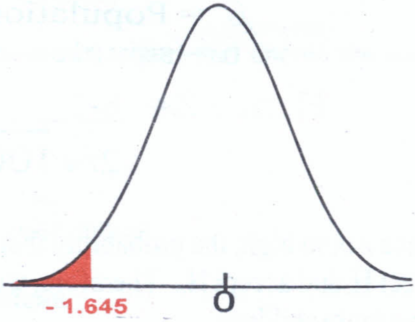
A. Compute a rejection region for a significance level of .05.

B. If the sample mean is 7.5 and the population standard deviation is 0.5, what can you conclude?

Solution

First, we write write down the null and alternative hypotheses

H_0: $\mu = 7.7$ H_1: $\mu < 7.7$



This is a left tailed test (one tailed test to the left). The z-score that corresponds to .05 is -1.645. The critical region is the area that lies to the left of -1.645. If the z-value is less than -1.645 there we will reject the null hypothesis and accept the alternative hypothesis. If it is greater than -1.645, we will fail to reject the null hypothesis and say that the test was not statistically significant.

We have

$$z = \frac{7.5 - 7.7}{.5/\sqrt{50}} = -2.83$$

Since -2.83 is to the left of -1.645, it is in the critical region. Hence, we reject the null hypothesis and accept the alternative hypothesis. We can conclude that smokers need less sleep.

p-values

There is another way to interpret the test statistic. In hypothesis testing, we make a yes or no decision without discussing borderline cases. For example with a = .06, a two tailed test will indicate rejection of H_o for a test statistic of z = 2 or for z = 6, but z = 6 is much stronger evidence than z = 2. To show this difference we write the *p-value* which is the lowest significance level such that we will still reject Ho. For a two tailed test, we use twice the table value to find p, and for a one tailed test, we use the table value.

Example:

Suppose that we want to test the hypothesis with a significance level of .05 that the climate has changed since industrialisation. If the mean temperature throughout history is 50 degrees, during the last 40 years, the mean temperature has been 51 degrees and suppose the population standard deviation is 2 degrees. What can we conclude?

We have

$$H_0: \mu = 50$$
$$H_1: \mu \neq 50$$

We compute the z score:

$$z = \frac{51 - 50}{2/\sqrt{40}} = 3.16$$

The table gives us .9992

so that $p = (1 - .9992)(2) = .002$

since

$$.002 < .05$$

we can conclude that there has been a change in temperature.

Note that small p-values will result in a rejection of H_0 and large p-values will result in failing to reject H_0.

Hypothesis Testing for a Proportion and for a Mean with Unknown Population Standard Deviation

Small Sample Hypothesis Tests For a Normal population

When we have a small sample from a normal population, we use the same method as a large sample except we use the t statistic instead of the z-statistic. Hence, we need to find the degrees of freedom (n - 1) and use the t-table at the back of the book.

Example

Is the temperature required to damage a computer on the average less than 110 degrees? Because of the price of testing, twenty computers were tested to see what minimum temperature will damage the computer. The damaging temperature averaged 109 degrees with a standard deviation of 3 degrees. Assume that the distribution of all computers' damaging temperatures is approximately normal. (use a = .05)

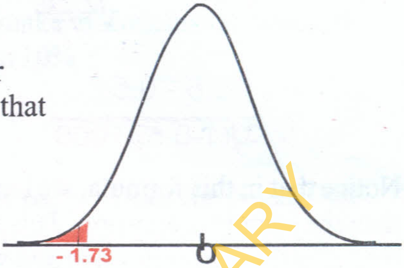
We test the hypothesis

$$H_0: \ \mu = 110$$
$$H_1: \ \mu < 110$$

We compute the t statistic:

$$t = \frac{109 - 110}{3/\sqrt{20}} = -1.49$$

This is a one tailed test, so we can go to our t-table with 19 degrees of freedom to find that



$t_c = 1.73$

Since $-1.49 > -1.73$

We see that the test statistic does not fall in the critical region. We fail to reject the null hypothesis and conclude that there is insufficient evidence to suggest that the temperature required to damage a computer on the average less than 110 degrees.

Hypothesis Testing for a Population Proportion

We have seen how to conduct hypothesis tests for a mean. We now turn to proportions. The process is completely analogous, although we will need to use the standard deviation formula for a proportion.

Example

Suppose that you interview 1000 exiting voters about who they voted for governor. Of the 1000 voters, 550 reported that they voted for the PDP candidate. Is there sufficient evidence to suggest that the PDP candidate will win the election at the .01 level? Suppose only PDP and Labour Party Candidates contested.

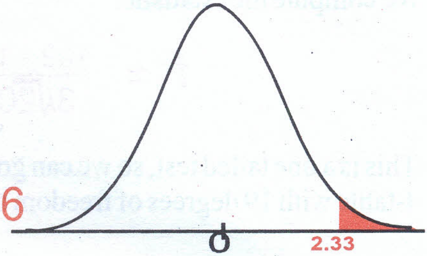
$H_0: p = .5$

$H_1: p > .5$

Since it is a large sample we can use the central limit theorem to say that the distribution of proportions is approximately normal. We compute the test statistic:

$$Z = \frac{\hat{P} - P}{\sqrt{\frac{pq}{n}}}$$

$$= \frac{0.6 - 0.5}{0.5\sqrt{(1-0.5)/1000}} = 3.16$$



Notice that in this formula, we have used the hypothesized proportion rather than the sample proportion. This is because if the null hypothesis is correct, then .5 is the true proportion and we are not making any approximations. We compute the rejection region using the z-table. We find that $z_c = 2.33$.

The picture shows us that 3.16 is in the rejection region. Therefore we reject H_0 so can conclude that the PDP candidate will win with a p-value of .0008.

Example

1500 randomly selected cocoa trees were tested for traces of the black pod infestation. It was found that 153 of the trees showed such traces. Test the hypothesis that more than 10% of the cocoa trees have been infested. (Use a 5% level of significance)

Solution

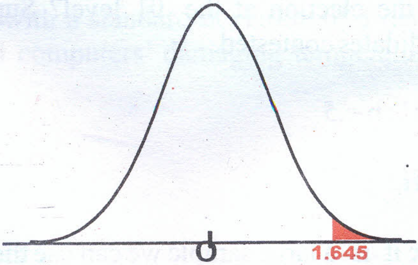
The hypothesis is

$H_0: p = .1$

$H_1: p > .1$

We have that

$$\hat{P} = \frac{153}{1500} = .102$$



Next we compute the z-score

$$Z = \frac{0.102 - 0.1}{\sqrt{0.1(1-0.1)/1500}} = 0.26$$

Since we are using a 95% level of significance with a one tailed test, we have $z_c = 1.645$. The rejection region is shown in the picture. We see that 0.26 does not lie in the rejection region, hence we fail to reject the null hypothesis. We say that there is insufficient evidence to make a conclusion about the percentage of infested cocoa being greater than 10%.

Exercises

A. If 40% of the nation is registered Labour party. Does the Niger Delta Region reflect the national proportion? Test the hypothesis that Niger-Delta residents differ from the rest of the nation in their affiliation, if of 200 locals surveyed, 75 are registered Labour party.

B. If 10% of City A married men are polygamist, test the hypothesis that married men who gamble are less likely to be polygamist. If the 120 people polled, 10 claimed to be a vegetarian.

Difference Between Means

Hypothesis Testing of the Difference Between Two Means

Do employees perform better at work with music playing. The music was turned on during the working hours of a business with 45 employees. There productivity level averaged 5.2 with a standard deviation of 2.4. On a different day the music was turned off and there were 40 workers. The workers' productivity level averaged 4.8 with a standard deviation of 1.2. What can we conclude at the .05 level?

Solution

We first develop the hypotheses

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 > 0$$

Next we need to find the standard deviation. Recall from before , we had that the mean of the difference is

$$\mu_x = \mu_1 - \mu_2$$

and the standard deviation is

$$\sigma_x = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

We can substitute the sample means and sample standard deviations for a point estimate of the population means and standard deviations. We have

$$\mu_{\overline{x}} \approx 5.2\text{-}4.8 = 0.4$$

Now we can calculate the t-score. We have

$$t = 0.4/0.405 = 0.988$$

To calculate the degrees of freedom, we can take the smaller of the two numbers $n_1 - 1$ and $n_2 - 1$. So, in this example, we use 39 degrees of freedom. The t-table gives a value of 1.690 for the $t_{.95}$ value. Notice that 0.988 is still smaller than 1.690 and the result is the same. Since the t-score is smaller than 1.690, we fail to reject the null hypothesis and state that there is insufficient evidence to make a conclusion about employees performing better at work with music playing.

Hypothesis Testing For a Difference Between Means for Small Samples Using Pooled Standard Deviations (Optional)

Recall that for small samples we need to make the following assumptions:

1. Random unbiased sample.
2. Both population distributions are normal.
3. The two standard deviations are equal.

If we know s, then the sampling standard deviation is:

$$s = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$$

$$= \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

If we do not know s then we use the pooled standard deviation.

$$S_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$$

Putting this together with hypothesis testing we can find the t-statistic.

$$t = \frac{x_1 - x_2 - \text{hypothesized difference}}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

and use $n_1 + n_2 - 2$ degrees of freedom.

Example

Nine dogs and ten cats were tested to determine if there is a difference in the average number of days that the animal can survive without food. The dogs averaged 11 days with a standard deviation of 2 days while the cats averaged 12 days with a standard deviation of 3 days. What can be concluded? (Use a = .05)

Solution

We write: H_0: $\mu_{dog} - \mu_{cat} = 0$

H_1: $\mu_{dog} - \mu_{cat} \neq 0$

We have: $n_1 = 9$, $n_2 = 10$

$x_1 = 11$, $x_2 = 12$

$s_1 = 2$, $s_2 = 3$

so that

$$S_p = \sqrt{\frac{(9-1)(4)+(10-1)(9)}{9+10-2}} = 2.58$$

and

$$t = \frac{12-11-0}{2.58\sqrt{\frac{1}{9}+\frac{1}{10}}} = 0.84$$

The t-critical value corresponding to á = .05 with 10 + 9 - 2 = 17 degrees of freedom is 2.11 which is greater than .84. Hence we fail to reject the null hypothesis and conclude that there is not sufficient evidence to suggest that there is a difference between the mean starvation time for cats and dogs.

Hypothesis Testing for a Difference Between Proportions

Inferences on the Difference Between Population Proportions

If two samples are counted independently of each other we use the test statistic:

$$z = \frac{p_1 - p_2}{\sqrt{\frac{pq}{n_1}+\frac{pq}{n_2}}} = 0.84$$

where $p=(r_1 + r_2)/(n_1 + n_2)$

$q = 1 - p$

Example

Is the severity of the drug problem in secondary school the same for boys and girls? 85 boys and 70 girls were questioned and 34 of the boys and 14 of the girls admitted to having tried some sort of drug. What can be concluded at the .05 level?

Solution: The hypotheses are

$$H_0: p_1 - p_2 = 0$$

$$H_1: p_1 - p_2 \neq 0$$

We have: $p_1 = 34/85 = 0.4 \qquad p_2 = 14/70 = 0.2$

$p = 48/155 = 0.31 \qquad q = 0.69$

Now compute the z-score

$$z = \frac{0.4 - 0.2}{\sqrt{\frac{(0.31)(0.69)}{85} + \frac{(0.31)(0.69)}{70}}} = 2.68$$

Since we are using a significance level of .05 and it is a two tailed test, the critical value is 1.96. Clearly 2.68 is in the critical region, hence we can reject the null hypothesis and accept the alternative hypothesis and conclude that gender does make a difference for drug use. Notice that the P-Value is

P=2(1 - .9963)=0.0074

is less than .05. Yet another way to see that we reject the null hypothesis.

CHAPTER ELEVEN

Statistical Tools

Correlation

Correlation is a bivariate measure of association (strength) of the relationship between two variables. It varies from 0 (random relationship) to 1 (perfect linear relationship) or -1 (perfect negative linear relationship). It is better reported in terms of its square (r^2), interpreted as percent of variance explained. For instance, if r^2 is .34, then the independent variable is said to explain 34% of the variance in the dependent variable.

According to Garson, (2008), there are several common pitfalls in using correlation. Correlation is symmetrical, not providing evidence of which way causation flows. If other variables also cause the dependent variable, then any covariance they share with the given *independent variable in a* correlation may be falsely attributed to that independent. Also, to the extent that there is a nonlinear relationship between the two variables being correlated, correlation will understate the relationship. Correlation will also be attenuated to the extent there is measurement error, including use of sub-interval data or artificial truncation of the range of the data. Correlation can also be a misleading average if the relationship varies depending on the value of the independent variable ("lack of homoscedasticity"). And, of course, a theoretical or post-hoc running of many correlations runs the risk that 5% of the coefficients may be found significant by chance alone.

Beside Pearsonian correlation (r), the most common type, there are other special types of correlation to handle the special characteristics of such types of variables as dichotomies, and there are other measures of association for nominal and ordinal variables. Regression procedures produce multiple correlation, R, which is the correlation of multiple independent variables with a single dependent. Also, there is partial correlation, which is the correlation of one variable with another, controlling both the given variable and the dependent for a third or additional variables. And there is part correlation, which is the correlation of one variable with another, controlling only the given variable for a third or additional variables. Click on these links to see the separate discussion.

Pearson's r: This is the usual measure of correlation, sometimes called *product-moment correlation*. The most common correlation for use with two continuous variables. Pearson's r is a measure of association which varies from -1 to +1, with 0 indicating no relationship (random pairing of values) and 1 indicating perfect relationship, taking the form, "The more the x, the more the y, and vice versa." A value of -1 is a perfect negative relationship, taking the form "The more the x, the less the y, and vice versa."

Sample Data

Physics Score	Mathematics Score
73	54
64	47
42	34
68	67
77	89
56	66

Note: In older works, predating the prevalence of computers, special computation formulas were used for computation of correlation by hand. For certain types of variables, notably dichotomies, there were computational formulas which differed one from another (ex., phi coefficient for two dichotomies, point-biserial correlation for an interval with a dichotomy). Today, however, SPSS will calculate the exact correlation regardless of whether the variables are continuous or dichotomous.

Coefficient of determination, r^2: The coefficient of determination is the square of the Pearsonian correlation coefficient. It represents the percent of the variance in the dependent variable explained by the independent. Of course, since correlation is bidirectional, r^2 is also the percent of the independent accounted for by the dependent. That is, the researcher must posit the direction of causation, if any, based on considerations external to correlation, which, in itself, cannot demonstrate causality. It is not sufficient for a researcher to make conclusion based on the value of **r** but take a step further to estimate the coefficient of determination ($\mathbf{r}^2$).

Attenuation of correlation. The correlation coefficient and r^2 are expected to be lower when there is variance restriction. A common cause of variance restriction is binning of continuous data, by reducing the original full range to a finite set of categories such as high/medium/low. Measurement error is also a cause of attenuation.

Ordinal correlation

- **Correlation for ordinal and dichotmous data**. Variations of correlation have been devised for binary and ordinal data. Some studies suggest use of these variant forms of correlation rarely affects substantive research conclusions. Rank correlation is nonparametric and does not assume normal distribution. It is less sensitive to outliers.

- **Spearman's rho**: The most common correlation for use with two ordinal variables or an ordinal and an interval variable. Rho for ranked data equals Pearson's r for ranked data.where d is the difference in ranks. In SPSS, choose Analyze, Correlate, Bivariate; check Spearman's rho.

Academic Qualification	Knowledge Adequacy
1	2
2	3
2	3
3	2
3	1
1	1
1	2
3	1

Academic Qualification	Monthly Income (₦)
1	60,000
2	45,000
2	52,000
3	70.000
3	43,000
1	34,000
1	40,000
3	66,000

Academic Qualification:
Below First Degree= 1
First Degree = 2
Above First Degree = 3

Knowledge Adequacy:
Not Adequate= 1
Adequate = 2
Very Adequate = 3

- **Kendall's tau**: Another common correlation for use with two ordinal variables or an ordinal and an interval variable. Prior to computers, rho was preferred to tau due to computational ease. Now that computers have rendered calculation trivial, tau is generally preferred. Partial Kendall's tau is also available as an ordinal analog to partial Pearsonian correlation. In SPSS, choose Analyze, Correlate, Bivariate; check Kendall's tau.

Correlation for dichotomies

- **Point-biserial correlation** is used when correlating a continuous variable with a true dichotomy. It is a special case of Pearsonian correlation and Pearson's r equals point-biserial correlation when one variable is continuous and the other is a dichotomy. Thus, in one sense it is true that a dichotomous or dummy variable can be used "like a continuous variable" in ordinary Pearsonian correlation. (Special formulas for point-biserial correlation in textbooks are for hand computation; point-biserial correlation is the same as Pearsonian correlation when applied to a dichotomy and a continuous variable).

Sex	Mathematics Score
1	54
1	47
2	34
1	67
2	89
2	66

Sex:
Male= 1
Female = 2

However, when the continuous variable is ordered perfectly from low to high, then even when the dichotomy is also ordered as perfectly as possible to match low to high, r will be less than 1.0 and therefore resulting r's must be interpreted accordingly. Specifically, point-biserial correlation will have a maximum of 1.0 only for the datasets with only two cases, and will have a maximum correlation around .85 even for large datasets, when the independent is normally distributed. The value of r may approach 1.0 when the continuous variable is bimodal and the dichotomy is a 50/50 split. Unequal splits in the dichotomy and variable will both depress the maximum possible point-biserial correlation even under perfect ordering. Moreover, if the dichotomy represents a true underlying continuum, correlation will be attenuated compared to what it would be if the dichotomy were coded as a continuous variable.

- **Rank biserial correlation**: Used when an ordinal variable is correlated with a dichotomous variable. Rank biserial correlation is not supported by SPSS .

Sample Data

Sex	Academic Qualification
1	1
1	3
2	3
1	2
2	1
2	2
1	2
1	1

Sex:
Male= 1
Female = 2

Academic Qualification:

Below First Degree= 1

First Degree = 2

Above First Degree = 3

- **Phi**: Used when both variables are dichotomies. Special formulas in textbooks are for hand computation; phi is the same as Pearsonian correlation for two dichotomies in SPSS correlation output, which uses exact algorithms. Alternatively, in SPSS, select Analyze, Descriptive Statistics, Crosstabs; click Statistics; check Phi

Sample Data

Sex	Meeting Attendance
1	2
1	1
2	1
1	2
2	1
2	2
1	2

Sex:
Male= 1
Female = 2

Meeting Attendance:
Attended= 1
Not Attended = 2

- **Correlation ratio, eta**. Eta, the coefficient of nonlinear correlation, known as the correlation ratio, is discussed in the section on **analysis of variance.** Eta is the ratio of the between sum of squares to total sum of squares in analysis of variance. The extent to which eta is greater than r is an estimate of the extent to which the data relationship is nonlinear. In SPSS, select Analyze, Compare Means, Means; click Options; check Anova table and eta. Eta is also computed in Analyze, General linear model, Multivariate; and elsewhere in SPSS.

- The **coefficient of intraclass correlation, r**: This ANOVA-based type of correlation measures the relative homogeneity within groups in ratio to the total variation and is used, for example, in assessing inter-rater reliability. Intraclass correlation, r = (Between-groups MS - Within-groups MS)/(Between-groups MS + (n-1)*Within-Groups MS), where n is the average number of cases in each category of the independent. Intraclass correlation is large and positive when there is no variation within the groups, but group means differ. It will be at its largest negative value when group means are the same but there is great variation within groups. Its maximum value is 1.0, but its maximum negative value is (-1/(n-1)). A negative intraclass correlation occurs when between-group variation is less than within-group variation, indicating some third (control) variable has introduced nonrandom effects on the different groups. Intraclass correlation is discussed further in the section on **reliability.**

ANOVA, and ANCOVA

Analysis of variance (ANOVA) is used to establish the main and interaction effects of categorical independent variables (called "factors" or “Grouping variable”) on an interval dependent variable. For example the main effect of Treatment group (Independent variable) on student achievement in Mathematics (Dependent Variable). A "main effect" is the direct effect of an independent variable on the dependent variable. An "interaction effect" is the joint effect of two or more independent variables on the dependent variable.

The key statistic in ANOVA is the F-test of difference of group means, testing if the means of the groups formed by values of the independent variable (or combinations of values for multiple independent variables) are different enough not to have occurred by chance. If the group means do not differ significantly then it is inferred that the independent variable(s) did not have an effect on the dependent variable. Based on the level of significance set (α = .01 or .05) if significant difference of group means is observed then independent variable(s) has/have effect on the dependent variable.

If the data involve **repeated measures** of the same variable, as in before-after or matched pairs tests, the F-test is computed differently from the usual between-groups design, but the inference logic is the same. There are also a large variety of other ANOVA designs for special purposes, all with the same general logic.

Note that analysis of variance tests the null hypotheses that group means do not differ. It is not a test of differences in variances, but rather assumes relative homogeneity of variances. Thus some key ANOVA assumptions are that the groups formed by the independent variable(s) are relatively equal in size and have similar variances on the dependent variable ("homogeneity of variances"). Like regression, ANOVA is a parametric procedure which assumes multivariate normality (the dependent has a normal distribution for each value category of the independent(s)).

Analysis of covariance (ANCOVA) is used to test the main and interaction effects of categorical variables on a continuous dependent variable, controlling for the effects of selected other continuous variables which covary with the dependent. The control variable is called the "covariate." There may be more than one covariate. One may also perform planned comparison or post hoc comparisons to see which values of a factor contribute most to the explanation of the dependent.

In SPSS, select Analyze, General Linear Model, Univariate; enter the dependent variable, the factor(s), and the covariate(s); click the Model button and accept the default, which is Full Factorial (if you select Custom, your model should not include interactions of factors with covariates: that is used beforehand in testing the equality of regressions assumption discussed below in the "Assumptions" section, but not in the ANCOVA model itself). The Full Factorial model contains the intercept, all factor and covariate main effects, and all factor-by-factor interactions. For instance, for three Independent variables X, Y, and Z, it includes the effects X, Y, Z, X*Y, X*Z, Y*Z, and X*Y*Z on dependent variable that is continuous.

ANCOVA is used for three purposes:

- In quasi-experimental (observational) designs, to remove the effects of variables which modify the relationship of the categorical independents to the interval dependent.

- In experimental designs, to control for factors which cannot be randomized but which can be measured on an interval scale. Since randomization in principle controls for all unmeasured variables, the addition of covariates to a model is rarely or never needed in experimental research. If a covariate is added and it is uncorrelated with the treatment (independent) variable, it is difficult to interpret as in principle it is controlling for something already controlled for by randomization. If the covariate is correlated with the treatment/independent, then its inclusion will lead the researcher to underestimate of the effect size of the treatment factors (independent variables)..

- In regression models, to fit regressions where there are both categorical and interval independents. (This third purpose has become displaced by logistic regression and other methods. On ANCOVA regression models, see Wildt and Ahtola, 1978: 52-54).

All three purposes have the goal of reducing the error term in the model. Like other control procedures, ANCOVA can be seen as a form of "what if" analysis, asking what would happen if all cases scored equally on the covariates, so that the effect of the factors over and beyond the covariates can be isolated. ANCOVA can be used in all ANOVA designs and the same assumptions apply.

Data requirements. For both ANOVA and ANCOVA, the dependent(s) is/are numeric. The independents may be categorical factors (including both numeric and string types) or quantitative covariates. Data are assumed to come from a random sample for purposes of significance testing. The variance(s) of the dependent variable(s) is/are assumed to be the same for each cell formed by categories of the factor(s) (this is the homogeneity of variances assumption).

- *One-way ANOVA* tests differences in a single interval dependent variable among two, three, or more groups formed by the categories of a single categorical independent variable. Also known as univariate ANOVA , simple ANOVA, single classification ANOVA, or one-factor ANOVA, this design deals with one independent variable and one dependent variable. It tests whether the groups formed by the categories of the independent variable seem similar (specifically that they have the same pattern of dispersion as measured by comparing estimates of group variances). If the groups seem different, then it is concluded that the independent variable has an effect on the dependent (ex., if different treatment groups have different health outcomes). One may note also that the significance level of a correlation coefficient for the correlation of an interval variable with a dichotomy will be the same as for a one-way ANOVA on the interval variable using the dichotomy as the only factor. This similarity does not extend to categorical variables with greater than two values.

In SPSS, select Analyze, Compare Means, One-Way ANOVA; enter the dependent variable in the Dependent list; enter the independent variable as the Factor.

- *Two-way ANOVA* analyzes one interval dependent in terms of the categories (groups) formed by two independents, one of which may be conceived as a control variable. Two-way ANOVA tests whether the groups formed by the categories of the independent variables have similar centroids. Two-way ANOVA is less sensitive than one-way ANOVA to moderate violations of the assumption of homogeneity of variances across the groups. In SPSS, select Analyze, General Linear Model, Univariate; enter the dependent variable and the independents (factors); if you want to test interactions, click Model and select Custom, Model (Interaction) and enter interaction terms (ex. gender*race); click Plots to set plot options; click Options to set what predicted group and interaction means are desired.

- *Multivariate or n-way ANOVA*. To generalize, n-way ANOVA deals with n independents. It should be noted that as the number of independents increases, the number of potential interactions proliferates. Two independents have a single first-order interaction (AB). Three independents have three first order interactions (AB,AC,BC) and one second-order interaction (ABC), or four in all. Four independents have six first-order (AB,AC,AD,BC,BC,CD), three second-order (ABC, ACD, BCD), and one third-order (ABCD) interaction, or 10 in all. As the number of interactions increase, it becomes increasingly difficult to interpret the model. The MAXORDERS command in SPSS syntax allows the researcher to limit what order of interaction is computed.

- **Variables**

  - *Factors* are categorical independent variables, such as treatments. A factor is a *fixed factor* if all of its values (categories) are measured, which is the usual case.

Research designs. ANOVA and ANCOVA have a number of different experimental designs. The alternative designs affect how the F-ratio is computed in generating the ANOVA table. However, regardless of design, the ANOVA table is interpreted similarly as the significance of the F-ratio indicates the significance of each main and interaction effect (and each covariate effect in ANCOVA).

- *Full Factorial ANOVA:* Factorial ANOVA is for more than one factor (more than one independent hence for two-way ANOVA or higher), used to assess the relative importance of various combinations of independents. In a full factorial design, the model includes all main effects and all interactions among the factors but does not include interactions between the factors and the covariates. As such factorial anova is not a true separate form of ANOVA design but rather a way of combining designs. A **design matrix table** shows the intersection of the categories of the independent variables. A corresponding ANOVA table is constructed in which the columns are the various covariate (in ANCOVA), main, and interaction effects. See the discussion of two-way ANOVA below.

  - *Factors* are categorical independent variables. The categories of a factor are its groups or levels. When using factorial ANOVA terminology, 2 x 3 ("two-by-three") factorial design means there are two factors with the first having two categories and the second having three, for a total of six groups (levels). A 2x2x2 factorial design has three factors, each with two categories. The order of the factors makes no difference. If you multiply through, you have the number of groups (often "treatment groups") formed by all the independents collectively. Thus a 2x3 design has 6 groups, and a 2x2x2 design has 8 groups. In experimental research equal numbers of subjects are assigned to each group on a random basis. Note that if factors have many levels, the number of required groups in a factorial design may become an unwieldy number and number of subjects in a group may be too small to meet the minimum number assumed.

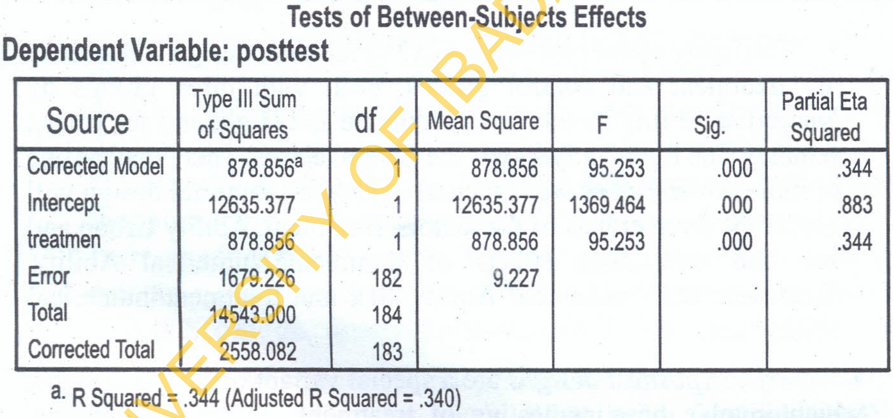
Group	Level of Numerical Ability	Sex	
		Male	Female
Treatment	Low		
	Moderate		
	High		
Control	Low		
	Moderate		
	High		

- The figure above represents a 2x3x2 factorial design where there are treatment and control groups, each with three groups by Numerical Ability Level (Low, Moderate and High) and sex (male, female). The figure only shows the design factors. There may be one or more covariates as well, such as age. A full factorial design will model the main effects of the factors Treatment, Ability Group and sex and interaction effects of treatment*Numerical Ability, Treatment*Sex, Numerical Ability*Sex and treatment*numerical ability*sex.

- *Pretest-posttest designs* are a special variant of mixed designs, which involve baseline testing of treatment and control groups, administration of a treatment, and post-test measurement. As Girden (1992: 57-58) notes, there are four ways of handling such designs: *One-way ANOVA* on the posttest scores. This involves ignoring the pretest data and is therefore not recommended.

1. *Split-plot repeated measures ANOVA* can be used when the same subjects are measured more than once. In this design, the between-subjects factor is the group (treatment or control) and the repeated measure is, for example, the test scores for two trials. The resulting ANOVA table will include a main treatment effect (reflecting being in the control or treatment group) and a group-by-trials interaction effect (reflecting treatment effect on posttest scores, taking pretest scores into account). This partitioning of the treatment effect may be more confusing than analysis of difference scores, which gives equivalent results and therefore is sometimes recommended.

2. *One-way ANOVA on difference scores*, where difference is the posttest score minus the pretest score. This is equivalent to a split-plot design if there is close to a perfect linear relation between the pretest and posttest scores in all treatment and control groups. This linearity will be reflected in a pooled within-groups regression coefficient of 1.0. Similarly effect size denoted by eta squared ranged 0 and 1. When this coefficient approaches 1.0, this method is more powerful than the ANCOVA method. An example of SPSS output on one way ANOVA is presented below. The result shows that there is significant main effect of treatment on posttest.



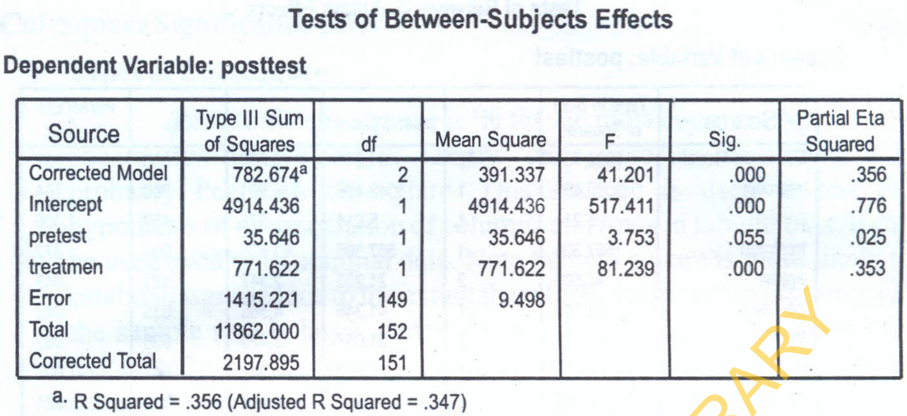
Tests of Between-Subjects Effects

Dependent Variable: posttest

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	878.856[a]	1	878.856	95.253	.000	.344
Intercept	12635.377	1	12635.377	1369.464	.000	.883
treatmen	878.856	1	878.856	95.253	.000	.344
Error	1679.226	182	9.227			
Total	14543.000	184				
Corrected Total	2558.082	183				

a. R Squared = .344 (Adjusted R Squared = .340)

3. *ANCOVA* on the posttest scores, using the pretest scores as a covariate control. When pooled within-groups regression coefficient is less than 1.0, the error term is smaller in this method than in ANOVA on difference scores, and the ANCOVA method is more powerful. An example of SPSS output on one way ANCOVA is presented below. The result shows that there is significant main effect of treatment on posttest.



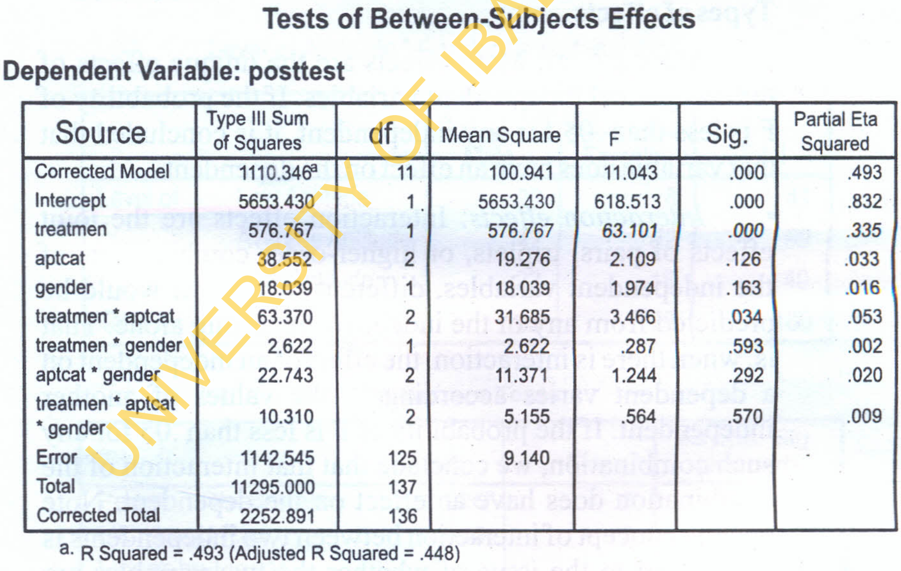
Tests of Between-Subjects Effects

Dependent Variable: posttest

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	782.674[a]	2	391.337	41.201	.000	.356
Intercept	4914.436	1	4914.436	517.411	.000	.776
pretest	35.646	1	35.646	3.753	.055	.025
treatmen	771.622	1	771.622	81.239	.000	.353
Error	1415.221	149	9.498			
Total	11862.000	152				
Corrected Total	2197.895	151				

a. R Squared = .356 (Adjusted R Squared = .347)

Examples of SPSS outputs on 3-way ANOVA and ANCOVA are presented below respectively. Significant interaction effect of Treatment and Aptitude was observed



Tests of Between-Subjects Effects

Dependent Variable: posttest

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	1110.346[a]	11	100.941	11.043	.000	.493
Intercept	5653.430	1	5653.430	618.513	.000	.832
treatmen	576.767	1	576.767	63.101	.000	.335
aptcat	38.552	2	19.276	2.109	.126	.033
gender	18.039	1	18.039	1.974	.163	.016
treatmen * aptcat	63.370	2	31.685	3.466	.034	.053
treatmen * gender	2.622	1	2.622	.287	.593	.002
aptcat * gender	22.743	2	11.371	1.244	.292	.020
treatmen * aptcat * gender	10.310	2	5.155	.564	.570	.009
Error	1142.545	125	9.140			
Total	11295.000	137				
Corrected Total	2252.891	136				

a. R Squared = .493 (Adjusted R Squared = .448)

Tests of Between-Subjects Effects

Dependent Variable: posttest

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	1092.134[a]	12	91.011	10.050	.000	.537
Intercept	3042.486	1	3042.486	335.961	.000	.764
pretest	5.214	1	5.214	.576	.450	.006
treatmen	557.305	1	557.305	61.539	.000	.372
aptcat	42.506	2	21.253	2.347	.101	.043
gender	41.346	1	41.346	4.566	.035	.042
treatmen * aptcat	62.115	2	31.057	3.429	.036	.062
treatmen * gender	1.912	1	1.912	.211	.647	.002
aptcat * gender	19.784	2	9.892	1.092	.339	.021
treatmen * aptcat * gender	4.534	2	2.267	.250	.779	.005
Error	941.831	104	9.056			
Total	9554.000	117				
Corrected Total	2033.966	116				

[a] R Squared = .537 (Adjusted R Squared = .484)

Types of effects

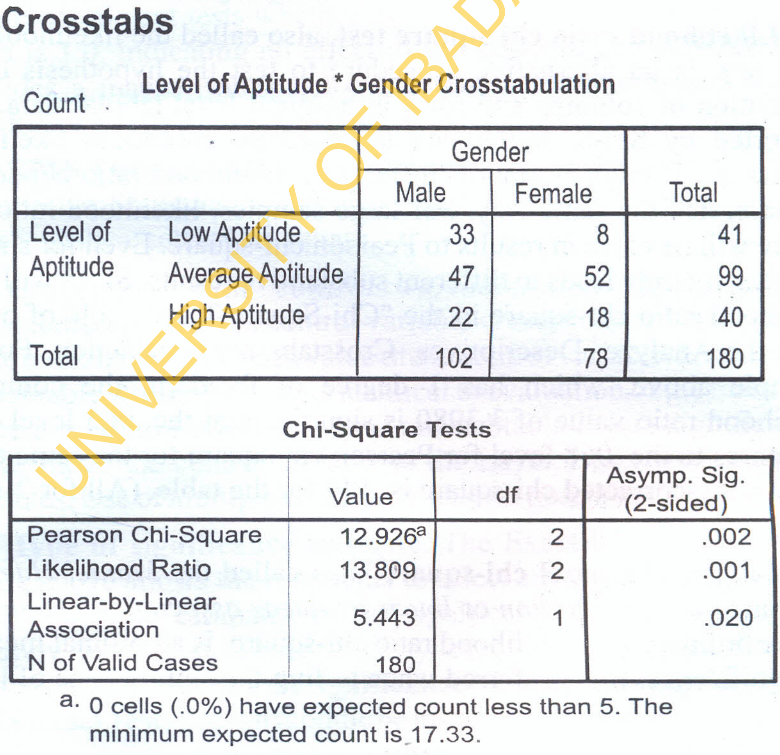
- *Main effects*: Main effects are the unique effects of the categorical independent variables. If the probability of F is less than .05 for any independent, it is concluded that that variable does have an effect on the dependent.

- *Interaction effects*: Interaction effects are the joint effects of pairs, triplets, or higher-order combinations of the independent variables, different from what would be predicted from any of the independents acting alone. That is, when there is interaction, the effect of an independent on a dependent varies according to the values of another independent. If the probability of F is less than .05 for any such combination, we conclude that that interaction of the combination does have an effect on the dependent. Note that the concept of interaction between two independents is not related to the issue of whether the two variables are correlated.

Chi-Square Significance Tests

- **Types of Chi-Square**

  - **Pearson's chi-square** is by far the most common type of chi-square significance test. If simply "chi-square" is mentioned, it is probably Pearson's chi-square. This statistic is used to test the hypothesis of no association of columns and rows in tabular data. It can be used even with nominal data. Note that chi square is more likely to establish significance to the extent that (1) the relationship is strong, (2) the sample size is large, and/or (3) the number of values of the two associated variables is large. A chi-square probability of .05 or less is commonly interpreted by social scientists as justification for rejecting the null hypothesis that the row variable is unrelated (that is, only randomly related) to the column variable. For this example, the SPSS output from the Analyze, Descriptive Statistics, Crosstabs menu choice looks like this:



Crosstabs

Level of Aptitude * Gender Crosstabulation

Count

		Gender		Total
		Male	Female	
Level of Aptitude	Low Aptitude	33	8	41
	Average Aptitude	47	52	99
	High Aptitude	22	18	40
Total		102	78	180

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	12.926[a]	2	.002
Likelihood Ratio	13.809	2	.001
Linear-by-Linear Association	5.443	1	.020
N of Valid Cases	180		

[a] 0 cells (.0%) have expected count less than 5. The minimum expected count is 17.33.

- **Chi-square goodness-of-fit test**. The goodness-of-fit test is simply a different use of Pearsonian chi-square. It is used to test if an observed distribution conforms to any other distribution, such as one based on theory (ex., if the observed distribution is not significantly different from a normal distribution) or one based on some other known distribution (ex., if the observed distribution is not significantly different from a known national distribution based on Census data). **The Kolmogorov-Smirnov goodness-of-fit test** is preferred for interval data, for which it is more powerful than chi-square goodness-of-fit.

An example of using the chi-square goodness-of-fit test to test if a sample distribution of response is different from the distribution expected on the basis of the population is located here, implemented as an Excel spreadsheet. Specifically, the example tests if the distribution of sample survey returns from field offices by region is not significantly different from what would be expected given the known (population) number of actual field offices by region.

- **Likelihood ratio chi-square test**, also called the likelihood test or G test, is an alternative procedure to test the hypothesis of no association of columns and rows in nominal-level tabular data. It is supported by SPSS output and is based on maximum likelihood estimation. Though computed differently, likelihood ratio chi-square is interpreted the same way. For large samples, likelihood ratio chi-square will be close in results to Pearson chi-square. Even for smaller samples, it rarely leads to different substantive results. SPSS will print likelihood ratio chi-square in the "Chi-Square Tests" table of output from the Analyze, Descriptives, Crosstabs menu selection. For the example above, which has 1 degree of freedom, the computed likelihood ratio value of 3.3980 is significant at the .065 level. This compares to the .068 level for Pearson chi-square for the same table. Continuity-corrected chi-square is .144 for the table. (All for 2-sided tests).

- **Mantel-Haenszel chi-square**, also called the *Mantel-Haenszel test for linear association* or *linear by linear association chi-square*, unlike ordinary and likelihood ratio chi-square, is an ordinal measure of significance. It is preferred when testing the significance of linear relationship between two ordinal variables because it is more powerful than Pearson chi-square (more likely to establish linear association). Mantel-Haenzel chi-square is not appropriate for nominal variables. If found significant, the interpretation is that increases in one variable are associated with increases (or decreases for negative relationships) in the other greater than would be expected by chance of random sampling. Like other chi-square statistics, M-H chi-square should not be used with tables with small cell counts.

- **Stratified analysis**, also called blocked analysis and matched analysis, is a form of control variable analysis conducted with the Mantel-Haenszel coefficient. For each of k categories of a control variable (called the *stratification* variable), a 2-by-2 table is created for the independent and dependent variables. The stratification variable need not be ordinal but it is assumed that the row and column marginals be the same for each of the k 2-by-2 tables, a circumstance which occurs mainly in experimental situations. The Mantel-Haenszel chi-square coefficient tests whether the common **odds ratio** across the k strata is 1.0, indicating no effect of the stratification variable. SPSS provides a macro (mh.sps) for Mantel-Haenszel stratified analysis which outputs M-H chi-square and its significance..

- **SPSS Output**. To obtain chi-square in SPSS: Select Analyze, Descriptive Statistics, Crosstabs; select row and column variables; slick Statistics; select Chi-square.

- *Chi-square with control variables*: In he context of crosstabulation, use of a control variable creates one subtable (similar to the overall table) for each value of the control variable. Evaluation of the subtable with chi-square is identical to evaluation of the main table. There is a control effect if at least one subtable is non-significant. In SPSS, move the control variable to the Layer 1 box when selecting variables. Use of multiple control variables is possible.

- **Type of significance estimate**. The Exact button in the SPSS dialog above allows the researcher to select among asymptotic, exact, or Monte-Carlo estimates of the significance of the Kolmogorov-Smirnov test value. These three types of estimates are discussed separately in the section on **significance testing**. This requires that the SPSS Exact Tests add-on module be installed.

Assumptions

- **Random sample data** are assumed. As with all significance tests, if you have population data, then any table differences are real and therefore significant. If you have non-random sample data, significance cannot be established, though significance tests are nonetheless sometimes utilized as crude "rules of thumb" anyway.

- **A sufficiently large sample size** is assumed, as in all significance tests. Applying chi-square to small samples exposes the researcher to an unacceptable rate of Type II errors. There is no accepted cutoff. Some set the minimum sample size at 50, while others would allow as few as 20. Note chi-square must be calculated on actual count data, not substituting percentages, which would have the effect of pretending the sample size is 100.

- **Adequate cell sizes** are also assumed. Some require 5 or more, some require more than 5, and others require 10 or more. A common rule is 5 or more in all cells of a 2-by-2 table, and 5 or more in 80% of cells in larger tables, but no cells with zero count. When this assumption is not met, Yates' correction is applied.

- **Independence**. Observations must be independent. The same observation can only appear in one cell. This means chi-square cannot be used to test correlated data (example before-after, matched pairs, panel data).

- **Similar distribution**. Observations must have the same underlying distribution.

- **Known distribution**. The hypothesized distribution is specified in advance, so that the number of observations that are expected to appear each cell in the table can be calculated without reference to the observed values. Normally this expected value is the crossproduct of the row and column marginals divided by the sample size.

- **Non-directional hypotheses** are assumed. Chi-square tests the hypothesis that two variables are related only by chance. If a significant relationship is found, this is not equivalent to establishing the researcher's hypothesis that A causes B, or that B causes A.

- **Finite values**. Observations must be grouped in categories.
- **Normal distribution of deviations** (observed minus expected values) is assumed. Note chi-square is a *nonparametric test* in the sense that is does not assume the parameter of normal distribution for the data -- only for the deviations.
- **Data level**. No assumption is made about level of data. Nominal, ordinal, or interval data may be used with chi-square tests.

Academic Qualification	Knowledge Adequacy
1	2
2	3
2	3
3	2

Academic Qualification:
Below First Degree= 1
First Degree = 2
Above First Degree = 3

Knowledge Adequacy:
Not Adequate= 1
Adequate = 2
Very Adequate = 3

CHAPTER TWELVE

Tips for Researcher

Conditions for using Statistical Tools.

Chi – Square Test

a. Two variables are involved.

b. The variables are categorical variables which make them non – parametric.

c. Researcher is interested in finding the impact / relationship / effect of two attributes either ordinarily or nominally.

d. Subjects are independently and randomly selected.

e. The nature of hypotheses or research question.

Pearson Moment.

a. Two variables are involved.

b. Two variables at metric variables which makes them parametric.

c. Researcher is interested in finding the relationship / impact / effect of two attributes at interval level.

d. Subjects are independently and randomly selected.

e. The nature of hypotheses or research questions.

T – Test Independent.

a. Subjects are randomly and independently selected.

b. The groups are independent of each other.

c. The researcher is only interested in comparing the means of two independent groups.

d. The hypotheses stated to be tested.

e. It is parametric in nature.

f. The nature and hypotheses or research questions.

Mann - Whitney U. Test.

g. Subjects are randomly and independently selected.

h. The groups are independent of each other.
i. The researcher is only interested in comparing the means of two independent groups.
j. The hypotheses stated to be tested.
k. The nature and hypotheses or research questions.
l. It is non parametric in nature.

T – Test (Dependent).

a. The samples are randomly and independently selected.
b. The sample is an integral part of population under consideration.
c. The researcher is interested in comparing the sample mean with the population means or the researcher is interested in comparing mean of two measures from same group.
d. The data in use is parametric in nature.

Wilcox in

a. The samples are randomly and independently selected.
b. The sample is an integral part of population under consideration.
c. . The researcher is interested in comparing the sample mean with the population means or the researcher is interested in comparing mean of two measures from same group.
d. The data is non – parametric in nature.

Rank Biseral.

a. Samples are independently and randomly selected.
b. The nature of hypotheses or researcher questions.
c. When the researcher is correlating two variables (measures), where one is nominal dichotomies and the other ordinal measure..

ANOVA (One Way)

a. The subject is randomly and independently selected.
b. The nature of hypotheses or research question.
c. The scale of measurement is interval.
d. The researcher is interested in comparing the mean performance of three or more independent groups. (One independent operating at three or more levels and one

dependent). It is interval in nature.

e. The data is parametric.

Kruskal – Wallis.

a. The subjects are randomly and independently selected.

b. The nature of hypotheses and research question.

c. The scale of measurement is ordinal or nominal in nature.

d. The researcher is interested in comparing the mean performance of three or more independent groups. (One independent operating at three or more levels and one dependent that either ordinal or nominal).

e. The data is non - parametric.

ANOVA. (Two – Way).

a. Samples are randomly and independently selected

b. The nature of hypotheses or research question.

c. The research is parametric in nature.

d. The researcher is interested on the effect of two independent variables on one dependent variable.

e. The data is parametric.

Fred man.

a. The researcher is interested on the effect of two independent variables on one dependent.

b. It is measured at nominal or interval scale.

c. It is non – parametric in nature.

d. Samples are randomly and independently selected.

e. The nature of hypotheses or research question.

Phi – Coefficient.

a. The samples are independently and randomly selected.

b. The nature of hypotheses and research question.

c. The researcher is interested in correlating two variables at one nominal dichotomous. E.g. one wants to find the correlation between gender and attendance to a meeting.

Attendance – Two present / One absent.

Spearman Rank Correlation.

a. Subjects are randomly sampled.
b. The nature of hypotheses or research question.
c. The two measures are at least ordinal in nature.
d. The researcher is interested in two measures that are ordinal in nature from the same sample.
e. It is parametric in nature.

Pearson Moment Correlation.

a. Subjects are randomly sampled.
b. The nature of hypotheses or research question.
c. The measures are at least interval in nature.
d. The researcher is interested in two measures that are interval in nature from the same sample.
e. It is parametric in nature.

Scheffe Test.

a. Subjects are randomly and independently selected.
b. The nature of hypotheses or research questions.
c. Anova produces F – ratio that is significant.
d. Researcher is interested in comparing the significant difference of mean amongst groups.

ANCOVA.

a. Samples are randomly in dependently selected.
b. The nature of hypothesis research questions.
c. The research is interested in effects of one more independent variables on dependent variable that is co – variated with pre – test.
d. The data is parametric in nature.

Multiple Regression.

a. Samples are randomly or independently selected.
b. The nature of research questions or hypothesis.
c. Interested in predicting prediction the predictive ability of the independent variable on the amounts of variability independent explained by independent (Composite accounted to
d. Relative contribution of each of the listed independent on

independent explained by independent (Composite accounted to

d. Relative contribution of each of the listed independent on dependent. (relative).

Analysis Procedure Using SPSS

S/N	Statistical Tool	Procedure
	Parametric	
1	Descriptive	Click on Analyse -Descriptive-Frequency-Select the variable(s) and finally click ok
2	One sample t-test	Click on Analyse - Compare means - One sample t-test - select the variable of interest and click ok
3	Independent Samples t-test	Click on Analyse -Compare means -Independent t-test – select the independent variable(s) and grouping variable of interest and define it then click ok
4	Paired t-test	Click on Analyse -Compare means -paired t -test – select the paired variables of interest and click ok
5	One Way Analysis	Click on Analyse-Compare means-One Way ANOVA – select the independent variable(s) of interest and a factor then click ok
6	Two Way ANOVA	Click on Analyse - General Lin ear Model - Univariate-Select the independent variable and fixed factors (more than one independent variables) then click ok
7	MANOVA	Click on Analyse - General Linear Model - Multivariate-Select the independent variables (more than one Dependent variable) and a fixed factor then click ok
8	ANCOVA	Click on Analyse - General Linear Model - Univariate-Select the independent variable, fixed factor(s) and the covariate (pretest) then click ok
9	Pearson product moment correlation	Click on Analyse –Correlate- Bivariate- Select the variable of interest and click ok
10	Multiple Regression	Click on Analyse –Regression- Linear- Select the dependent variable and the independent variables (predictors) of interest and click ok
	Non-Parametric	
11	Mann-Whitney-U	Click on Analyse-Non Parametric Tests-Independent samples-select test variable and grouping variable then click ok
12	Kruskal- Walis H	Click on Analyse -Non Parametric Tests -K independent samples-select test variable and grouping variable then click ok
13	Wilcoxon	Click on Analyse -Non Parametric Tests -2 related samples test-select pair(s) of variables of interest then click ok
14	Friedman	Click on Analyse -Non Parametric Tests - k related samples - select test variables then click ok

References

Albanese, M. A. (1993). Type K and other complex multiple-choice items: An analysis of research and item properties. *Educational Measurement: Issues and Practice*, 12, 28-33.

APA (1994). *Publication manual of the American Psychological Association*. Fourth ed., Washington, DC: American Psychological Association.

Ayodele, S.O., Adegbile, J.A., and Adewale, J.G. (2008) Evaluation Studies,The Powerhouse Press and Publishers, Ibadan

Bloom, B. S., Hastings, J.T. & Madaus G. F. (1971) Handbook on Formative and Summative Evaluation of Student Learning. McGraw-Hill Inc.

Box, G. E. P. (1954). "Some theorems on quadratic forms applied in the study of analysis of variance problems." *Annals of Statistics*, 25: 290-302. Cited with regard to robustness of the F test even in the face of small violations of the homogeneity of variances assumption.

Box, G.E. P. , Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters: An introduction to design and data analysis*. NY: John Wiley. General introduction.

Brown, Steven R. and Lawrence E. Melamed (1990). *Experimental design and analysis*. Thousand Oaks, CA: Sage Publications. Quantitative Applications in the Social Sciences series no. 74. Discusses alternative ANOVA designs and related coefficients.

Cassels, J. R. T., & Johnstone, A. H. (1984). The effect of language on student performance on multiple choice tests in chemistry. *Journal of Chemical Education, 61*(7), 613-615.

Chase, C. I. (1964). Relative length of option and response set in multiple choice items. *Educationaland Psychological Measurement*, 24(4), 861-866.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. NY: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* . Second ed., Hillsdale, NJ: Erlbaum.

Cortina, Jose M. and Hossein Nouri (2000). *Effect size for ANOVA designs*. Thousand Oaks, CA: Sage Publications. Quantitative Applications in the Social Sciences series no. 129.

Dunteman, George H. and Moon-Ho R. Ho (2005). *An introduction to generalized linear models*. Thousand Oaks, CA: Sage Publications. Quantitative Applications in the Social Sciences, Vol. 145.

Ebel, R. L. & Frisbie, D. A. (1991). *Essentials of educational measurement*. Englewood Cliffs, NJ:Prentice Hall.

Gardner, P. L. (1975). Scales and statistics. *Review of Educational Research*. 45: 43-57. Discusses assumptions of the t-test.

Garson, G.D. (2008) Online Text and Notes in Statistics for Economists. www.economicsnetwork.ac.uk/teaching

Girden, Ellen R. (1992). *ANOVA Repeated Measures*. Thousand Oaks, CA: Sage Publications. Quantitative Applications in the Social Sciences series no. 84.

Haladyna, T. M. & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1), 51-78.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items*. Mahwah: Lawrence Erlbaum.

Harold, G. & Levine, M.S. (2009) Basic Principles of Evaluation: An Overview.

Irby D. (1983), Peer review of teaching in medicine. J Med Educ. 73:459-461.

Iverson, G.R. and Helmut, N. (1987). *Analysis of variance*. Thousand Oaks, CA: Sage Publications. Quantitative Applications in the Social Sciences series no. 1.

Jaccard, J. (1998). *Interaction effects in factorial analysis of variance*. Quantitative Applications in the Social Sciences Series No. 118. Thousand Oaks, CA: Sage Publications.

Judy R. Wilkerson & William Steve Lang (2004). A standards-driven, task-based assessment approach for teacher credentialing with potential for college accreditation. *Practical Assessment, Research & Evaluation*, 9(12). Retrieved August, 2010 from http://PAREonline.net/getvn.asp?v=9&n=12 .

Kerlinger F.N. & Lee H.B. (2000) Foundations of Behavioral Research. 4th Edition, Wadsworth Publisher.

Levin, I. P. (1999). *Relating statistics and experimental design*. Thousand Oaks, CA: Sage Publications. Quantitative Applications in the Social Sciences series.

McCall, R.B. (   ) Fundamental Statistics for Psychology, Harcourt Brace Jovanovich. Inc.

Mehrens W. A., & Lehmann, I.J. (1984) Measurement and Evaluation in Education and Psychology 3rd Edition

Mitchell *R (1992). The Development of the Cognitive Behavior Survey to Assess Medical Student Learning*. Brown University.

Moore, D. S. (1995). *The basic practice of statistics*. NY: Freeman and Co.

Page, G. (1986). Essays on curriculum development and evaluation in medicine. Report of the second Cambridge conference June 21-28,. Vancouver, BC. University British Columbia. 1989.

Spiegel, M.R. & Stephens L.J. (1999) Theory and Problems of Statistics, 3rd Edition, McGraw-Hill, Inc.

Thondike, R.M (1997) Measurement and Evaluation in Psychology and Education, 6th Edition, Prentice-Hall, Inc.

UpFront Organization Development Consulting ( ) Programe Evaluation using Up Front Evaluation Model.

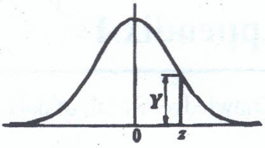
www.upfrontconsultingmn.com/utilization

INDEX

Appendix 1

Knowledge: Recall of data.	Key Words: defines, describes, identifies, knows, labels, lists, matches, names, outlines, recalls, recognizes, reproduces, selects, states.
Comprehension: Understand the meaning, translation, interpolation, and interpretation of instructions and problems. State a problem in one's own words.	Key words: comprehends, converts, defends, distinguishes, estimates, explains, extends, generalizes, gives examples, infers, interprets, paraphrases, predicts, rewrites, summarizes, translates.
Application: Use a concept in a new situation or unprompted use of an abstraction. Applies what was learned in the classroom into novel situations in the workplace.	Key Words: applies, changes, computes, constructs, demonstrates, discovers, manipulates, modifies, operates, predicts, prepares, produces, relates, shows, solves, uses.
Analysis: Separates material or concepts into component parts so that its organizational structure may be understood. Distinguishes between facts and inferences.	Keywords: analyzes, breaks down, compares, contrasts, diagrams, deconstructs, differentiates, discriminates, distinguishes, identifies, illustrates, infers, outlines, relates, selects, separates.
Synthesis: Builds a structure or pattern from diverse elements. Put parts together to form a whole, with emphasis on creating a new meaning or structure.	Keywords: categorizes, combines, compiles, composes, creates, devises, designs, explains, generates, modifies, organizes, plans, rearranges, reconstructs, rela tes, reorganizes, revises, rewrites, summarizes, tells, writes.
Evaluation: Make judgments about the value of ideas or materials.	Keywords: appraises, compares, concludes, contrasts, criticizes, critiques, defends, describes, discriminates, evaluates, exp lains, interprets, justifies, relates, summarizes, supports.

Ordinates(Y) of the Standard Normal

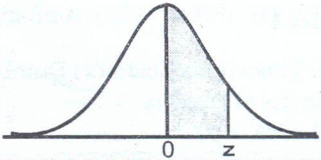


z	0	1	2	3	4	5	6	7	8	9
0.0	.3989	.3989	,3989	.3988	.3986	.3984	.3982	.3980	.3977	.3973
0.1	.3970	.3965	.3961	.3956	.3951	.3945	.3939	.3932	.3925	.3918
0.2	.3910	.3902	.3894	.3885	.3876	.3867	.3857	.3847	.3836	.3825
0.3	.3814	.3802	.3790	.3778	.3765	.3752	.3739	.3725	.3712	.3697
0.4	.3683	.3668	.3653	3637	.3621	.3605	.3589	.3572	.3555	.3538
0.5	.3521	.3503	.3485	.3467	.3448	3429	.3410	.3391	.3372	.3352
0.6	3332	.3312	.3292	.3271	.3251	.3230	.3209	.3187	.3166	.3144
0.7	.3123	.3101	.3079	.3056	.3034	.3011	.2989	.2966	.2943	.2920
0.8	.2897	.2874	.2850	.2827	.2803	.2780	.2756	.2732	.2709	.2685
0.9	.2661	.2637	.2613	.2589	.2565	.2541	.2516	.2492	.2468	.2444
1.0	.2420	.2396	.2371	.2347	.2323	.2299	.2275	.2251	.2227	.2203
1.1	.2179	.2155	.2131	.2107	.2083	.2059	.2036	.2012	.1989	.1965
1.2	.1942	.1919	.1895	.1872	.1849	.1826	.1804	.1781	.1758	.1736
1.3	.1714	.1691	.1669	,1647	.1626	.1604	.1582	.1561	.1539	.1518
1.4	.1497	.1476	.1456	.1435	.1415	.1394	.1374	.1354	.1334	.1315
1.5	.1295	.1276	.1257	.1238	.1219	.1200	.1182	.1163	.1145	.1127
1.6	.1109	.1092	.1074	.1057	.1040	.1023	.1006	.0989	.0973	.0957
1.7	.0940	.0925	.0909	.0893	.0878	.0863	.0848	.0833	.0818	.0804
1.8	.0790	.0775	.0761	.0748	.0734	,0721	.0707	.0694	.0681	.0669
1.9	.0656	.0644	.0632	.0620	.0608	.0596	.0584	.0573	.0562	.0551
2.0	.0540	.0529	.0519	.0508	.0498	.0488	.0478	.0468	.0459	.0449
2.1	.0440	.0431	.0422	.0413	.0404	.0396	.0387	.0379	.0371	.0363
2.2	.0355	.0347	.0339	.0332	.0325	.0317	.0310	.0303	.0297	.0290
2.3	.0283	.0277	.0270	.0264	.0258	.0252	.0246	.0241	.0235	.0229
2.4	.0224	.0219	.0213	.0208	.0203	.0198	.0194	.0189	.0184	.0180
2.5	.0175	.0171	.0167	.0163	.0158	.0154	.0151	.0147	.0143	.0139
2.6	.0136	.0132	.0129	.0126	.0122	.0119	.0116	.0113	.0110	.0107
2.7	.0104	.0101	.0099	.0096	.0093	.0091	.0085	.0086	.0084	.0081
2.8	.0079	.0077	.0075	.0073	.0071	.0069	.0067	.0065	.0063	.0061
2.9	.0060	.0058	.0056	.0055	.0053	.005!	.0050	.0048	.0047	.0046
3.0	.0044	.0043	.0042	.0040	.0039	.0038	.0037	.0036	.0035	.0034
3.1	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026	.0025	.0025
3.2	.0024	.0023	.0022	.0022	.0021	.0020	.0020	.0019	.0018	.0018
3.3	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014	.0013	.0013
3.4	.0012	.0012	.0012	.0011	.0011	.0010	.0010	.0010	.0009	.0009
3.5	.0009	.0008	.0008	.0008	.0008	.0007	.0007	.0007	.0007	.0006
3.6	.0006	.0006	.0006	.0005	.0005	.0005	.0005	.0005	.0005	.0004
3.7	.0004	.0004	.0004	.0004	.0004	.0004	.0003	.0003	.0003	.0003
3.8	.0003	.0003	.0003	.0003	.0003	.0002	.0002	.0002	.0002	.0002
3.9	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001

Appendix III

Areas Under the Standard Normal Curve

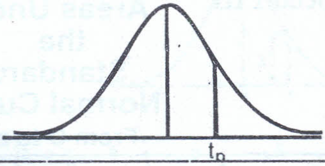
From O to z



Z	0	1	2	3	4	5	6	7	8	9
0.0	.3989	.3989	,3989	.3988	.3986	.3984	.3982	.3980	.3977	.3973
0.1	.3970	.3965	.3961	.3956	.3951	.3945	.3939	.3932	.3925	.3918
0.2	.3910	.3902	.3894	.3885	.3876	.3867	.3857	.3847	.3836	.3825
0.3	.3814	.3802	.3790	.3778	.3765	.3752	.3739	.3725	.3712	.3697
0.4	.3683	.3668	.3653	3637	.3621	.3605	.3589	.3572	.3555	.3538
0.5	.3521	.3503	.3485	.3467	.3448	,3429	.3410	.3391	.3372	.3352
0.6	3332	.3312	.3292	.3271	.3251	.3230	.3209	.3187	.3166	.3144
0.7	.3123	.3101	.3079	.3056	.3034	.3011	.2989	.2966	.2943	.2920
0.8	.2897	.2874	.2850	.2827	.2803	.2780	.2756	.2732	.2709	.2685
0.9	.2661	.2637	.2613	.2589	.2565	.2541	.2516	.2492	.2468	.2444
1.0	.2420	.2396	.2371	.2347	.2323	.2299	.2275	.2251	.2227	.2203
1.1	.279	.2155	.2131	.2107	.2083	.2059	.2036	.2012	.1989	.1965
1.2	.1942	.1919	.1895	.1872	.1849	.1826	.1804	.1781	.1758	.1736
1.3	.1714	.1691	.1669	,1647	.1626	.1604	.1582	.1561	.1539	.1518
1.4	.1497	.1476	.1456	.1435	.1415	.1394	.1374	.1354	.1334	.1315
1.5	.1295	.1276	.1257	.1238	.1219	.1200	.1182	.1163	.1145	.1127
1.6	.1109	.1092	.1074	.1057	.1040	.1023	.1006	.0989	.0973	.0957
1.7	.0940	.0925	.0909	.0893	.0878	.0863	.0848	.0833	.0818	.0804
1.8	.0790	.0775	.0761	.0748	.0734	,0721	.0707	.0694	.0681	.0669
1.9	.0656	.0644	.0632	.0620	.0608	.0596	.0584	.0573	.0562	.0551
2.0	.0540	.0529	.0519	.0508	.0498	.0488	.0478	.0468	.0459	.0449
2.1	.0440	.0431	.0422	.0413	.0404	.0396	.0387	.0379	.0371	.0363
2.2	.0355	.0347	.0339	.0332	.0325	.0317	.031	.0303	.0297	.0290
2.3	.0283	.0277	.0270	.0264	.0258	.0252	.02	.0241	.0235	.0229
2.4	.0224	.0219	.0213	.0208	.0203	.0198	.01	.0189	.0184	.0180
2.5	.0175	.0171	.0167	.0163	.0158	.0154	.015	.0147	.0143	.0139
2.6	.0136	.0132	.0129	.0126	.0122	.0119	.0116	.0113	.0110	.0107
2.7	.0104	.0101	.0099	.0096	.0093	.0091	.0085	.0086	.0084	.0081
2.8	.0079	.0077	.0075	.0073	.0071	.0069	.0067	.0065	.0063	.0061
2.9	.0060	.0058	.0056	.0055	.0053	.005!	.0050	.0048	.0047	.0046
3.0	.0044	.0043	.0042	.0040	.0039	.0038	.0037	.0036	.0035	.0034
3.1	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026	.0025	.0025
3.2	.0024	.0023	.0022	.0022	.0021	.0020	.0020	.0019	.0018	.0018
3.3	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014	.0013	.0013
3.4	.0012	.0012	.0012	.0011	.0011	.0010	.0010	.0010	.0009	.0009
3.5	.0009	.0008	.0008	.0008	.0008	:0007	.0007	.0007	.0007	.0006
3.6	.0006	.0006	.0006	.0005	.0005	.0005	.0005	.0005	.0005	.0004
3.7	.0004	.0004	.0004	.0004	.0004	.0004	.0003	.0003	.0003	.0003
3.8	.0003	.0003	.0003	.0003	.0003	.0002	.0002	.0002	.0002	.0002
3.9	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001

Appendix IV

Percentile Values (t_p)for Student's t Distribution
with v Degrees of Freedom
(shaded area = p)

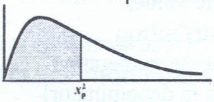


v,.....,.	$t_{.995}$	$t_{.99}$	$t_{.975}$	$t_{.95}$	$t_{.90}$	$t_{.80}$	$t_{.75}$	$t_{.70}$	$t_{.60}$	$t_{.55}$
1	63.66	31.82	12.71	6.31	3.08	1.376	1.000	.727	.325	.158
2	9.92	6.96	4.30	2.92	1.89	1.061	.816	.617	.289	.142
3	5.84	4.54	3.18	2.35	1.64	.978	.765	.584	.277	.137
4	4.60	3.75	2.78	2.13	1.53	.941	.741	.569	.271	.134
5	4.03	3.36	2.57	2.02	1.48	.920	.727	.559	.267	.132
6	3.71	3.14	2.45	1.94	1.44	.906	.718	.553	.265	.131
7	3.50	3.00	2.36	1.90	1.42	.896	.711	.549	.263	.130
8	3.36	2.90	2.31	'1.86	1.40	.889	.706	.546	.262	.130
9	3.25	2.82	2.26	1.83	1.38	.883	.703	.543	.261	.129
10	3.17	2.76	2.23	1.81	1.37	.879	.700	.542	.260	.129
11	3.11	2.72	2.20	1.80	1.36	.876	.697	.540	.260	.129
12	3.06	2.68	2.18	1.78	1.36	.873	.695	.539	.259	.128
13	3.01	2.65	2.16	1.77	1.35	.870	.694	.538	.259	.128
14	2.98	2.62	2.14	1.76	1.34	.868	.692	.537	.258	.128
15	2.95	2.60	2.13	1.75	1.34	.866	.691	.536	.258	.128
16	2.92	2.58	2.12	1.75	1.34	.865	.690	.535	.258	.128
17	2.90	2.57	2.11	1.74	1.33	.863	.689	.534	.257	.128
18	2.88	2.55	2.10	1.73	1.33	.862	.688	.534	.257	.127
19	2.86	2.54	2.09	1.73	1.33	.861	.688	.533	.257	.127
20	2.84	2.53	2.09	1.72	1.32	.860	.687	.533	.257	.127
21	2.83	2.52	2.08	1.72	1.32	.859	.686	.532	.257	.127
22	2.82	2.51	2.07	1.72	1.32	.858	.686	.532	.256	.127
23	2.81	2.50	2.07	1.71	1.32	.858	.685	.532	.256	.127
24	2.80	2.49	2.06	1.71	1.32	.857	.685	.53!	.256	.127
25	2.79	2.48	2.06	1.71	1.32	.856	.684	.531	.256	.127
26	2.78	2.48	2.06	1.71	1.32	.856	.684	.531	.256	.127
27	2.77	2.47	2.05	1.70	1.31	.855	.684	.531	.256	.127
28	2.76	2.47	2.05	1.70	1.31	.855	.683	.530	.256	.127
29	2.76	2.46	2.04	1.70	1.31	.854	.683	.530	.256	.127
30	2.75	2.46	2.04	1.70	1.31	.854	.683	.530	.256	.127
40	2.70	2.42	2.02	1.68	1.30	.851	.681	,529	.255	.126
60	2.66	2.39	2,00	1.67	1.30	.848	.679	.527	.254	.126
120	2.62	2.36	1.98	1.66	1.29	.845	.677	.526	.254	.126
∞	2.58	2.33	1.96	1.645	1.28	.842	.674	.524	.253	.126

Source: R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research* (5th edition). Table III, Oliver and Boyd Ltd., Edinburgh.

Appendix V

Percentile Values (^2_p)for p the Chi-Square Distribution with *v* **Degrees of** Freedom (shaded **area** = p)



	$x^2_{.995}$	$X^2_{.99}$	$X^2_{.975}$	$x^2_{.95}$	$x^2_{.90}$	$X^2_{.75}$	$x^2_{.50}$	$X^2_{.25}$	$x^2_{.10}$	$x^2_{.05}$	$x^2_{.025}$	$x^2_{.01}$	$x^2_{.005}$
1	7.88	6.63	5.02	3.84	2.71	1.32	.455	.102	.0158	.0039	.0010	.0002	.0000
2	10.6	9.21	7.38	5.99	4.61	2.77	1.39	.575	.211	.103	.0506	.0201	.0100
3	12.8	11.3	9.35	7.81	6.25	4.11	2.37	1,21	.584	.352	.216	.115	.072
4	14.9	13.3	11,1	9.49	7.78	5.39	3.36	1.92	1.06	.711	.484	.297	.207
5	16.7	15.1	12.8	11.1	9.24	6.63	4.35	2.67	1.61	1.15	.831	.554	.412
6	18.5	16.8	14.4	12.6	10.6	7.84	5.35	3,45	2.20	1.64	1.24	.872	.676
7	20.3	18.5	16.0	14.1	12.0	9.04	6.35	4.25	2.83	2.17	1.69	1.24	.989
8	22.0	20.1	17.5	15.5	13.4	10.2	7.34	5.07	3.49	2.73	2.18	1.65	1.34
9	23.6	21.7	19.0	16.9	14.7	11.4	8.34	5.90	4.17	3.33	2.70	2.09	1.73
10	25.2	23.2	20.5	18.3	16.0	12.5	9.34	6.74	4.87	3.94	3.25	2.56	2.16
11	26.8	24.7	21.9 ,	19.7	17.3	13.7	10.3	7.58	5.58	4.57	3.82	3.05	2.60
12	28.3	26.2,	23.3	21.0	18.5	14.8	11.3	8.44	6.30	5.23	4.40	3.57	3.07
13	29.8	27.7	24.7	22.4	19.8	16.0	12.3	9.30	7.04	5.89	5.01	4.11	3.57
14	31.3	29.1	26.1	23.7	21.1	17.1	13.3	10.2	7.79	6.57	5.63	4.66	4.07
15	32.8	30.6	27.5	25.0	22.3	18.2	14.3	11.0	8.55	7.26	6.26	5.23	4.60
16	34.3	32.0	28.8	26.3	23.5	19.4	15.3	11.9	9.3i	7.96	6.91	5.81	5.14
17	35.7	33.4	30.2	27.6	24.8	20.5	16.3	12.8	10.1	8.67	7.56	6.41	5.70
18	37.2	34.8	31.5	28.9	26.0	21.6	17.3	13.7	10.9	9.39	8.23	7.01	6.26
19	38.6	36.2	32.9	30.1	27.2	22.7	18.3	14.6	1L7	10.1	8.91	7.63	6.84
20	40.0	37.6	34.2	31.4	28.4	23.8	19.3	15.5	12.4	10.9	9.59	8.26	7.43
21	41.4	38.9	35.5	32.7	29.6	24.9	20.3	16.3	13.2	11.6	10.3	8.90	8.03
22	42.8	40.3	36.8	33.9	30.8	26.0	21.3	17.2	14.0	12.3	11.0	9.54	8.64
23	44.2	41.6	38.1	35.2	32.0	27.1	22.3	18.1	14.8	13.1	11.7	10.2	9.26
24	45.6	43.0	39.4	36.4	33.2	28.2	23.3	19.0	15.7	13.8	12.4	10.9	9.89
25	46.9	44.3	40.6	37.7	34.4	29.3	24.3	19.9	16.5	14.6	13.1	11.5	10.5
26	48.3	45.6	41.9	38.9	35.6	304	25.3	20.8	17.3	15.4	13.8	12.2	11.2
27	49.6	47.0	43.2	40.1	36.7	31.5	26.3	21.7	18.1	16.2	14.6	12.9	11,8
28	51.0	48.3	44.5	41.3	37.9	32.6	27.3	22.7	18.9	16.9	15.3	13.6	12.5
29	52.3	49.6	45.7	42.6	39.1	33.7	28.3	23.6	19.8	17.7	16.0	14.3	13.1
30	53.7	50.9	47.0	43.8	40.3	34.8	29.3	24.5	20.6	18.5	16.8	15.0	13.8
40	66.8	63.7	59.3	55.8	51.8	46.6	39.3	33.7	29.1	36.5	24.4	22.2	20.7
50	79.5	76.2	71.4	67.5	63.2	56.3	49.3	42.9	37.7	34.8	32.4	29.7	28.0
60	92.0	88.4	83.3	79.1	74.4	67.0	59.3	52.3	46.5	43.2	40.5	37.5	35.5
70	104.2	100.4	95.0	90.5	85.5	77.6	69.3	61.7	55.3	51.7	48.8	45.4	43.3
80	116.3	112.3	106.6	101.9	96.6	88.1	79.3	71.1	64.3	60.4	57.2	53.5	51.2
90	128.3	124.1	118.1	113.1	107.6	98.6	89.3	80.6	73.3	69.1	65.6	61.8	59.2
100	140.2	135.8	129.6	124.3	118.5	109.1	99.3	90.1	82.4	77.9	74.2	70.1	67.3

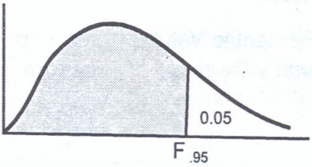
Source: Catherine M. Thompson, *Table of percentage points of the x^2 distribution,* Biometrika, Vol. 32(1941).

Appendix VI

95th Percentile Values for the F Distribution

(v_1 degrees of freedom in numerator)

(v_2 degrees of freedom in denominator)

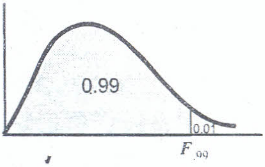


V_1 V_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	879	8.74	8.70	8.66	8.64	8.62	8.59	857	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5,86	5.80	5.77	575	5.75	5.69	5.66	5.63
5	6.61	5.79	5,41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4. 35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2,49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4,54	3.68	3 29	3.06	2.90	2.79	2.71	2.64	2.59	2 54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3 05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	-3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.0 1	1.96	1 91	1.86	1. 81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2,31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.73	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2,37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Source: *E*. S. Pearson and H. O. Hartley*Biometrika Tables for Statisticians,* **Vol. 2 (1972),Table 5, page 178**

Appendix VII

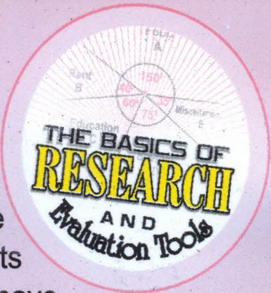
99th Percentiie Values for the *F* Distribution
(v$_1$ degrees of freedom in numerator)
(*v*$_2$ degrees of freedom in denominator)



V_1 V_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	5000	5403	5625	5764	5859	5928	5981	6023	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.5	99.O	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5
3	34.1	30.0	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26. 5	26.4	26.3	26.2	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	139	13.8	137	13.7	13.6	13.5
5	16.3	13.3	12.1	11.4	11.0	107	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	711	7.23	7.14	7.06	6.97	6.88
7	12.2	9.55	8,45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	591	5.82	5.74	5.65
8	11,3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10. 6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4 .56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4. 63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3,80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2. 75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	337	3 23	108	100	2.92	784	975	2,66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	371	3.54	3.41	3. 30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3. 17	3. 03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4 .18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2 .27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3. 18	3.09	2.96	2.82	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2,29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	336	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	255	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4-13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Source: E. S. Pearson and H. O. Hartley, *Biumetrika Tables for Statisticians*. Vol. 2 (1972), **Table** 5, pagc 180.

ABOUT THE Book



'The Basics of Research and Evaluation Tools' is a book focusing aspects of research studies where college, undergraduate and postgraduate students as well as independent and group researchers have been having difficulties. There is no doubt about the importance of research in our days. It is not enough to carry out a research studies but carrying it out using appropriate sample, research instruments, statistical tools and giving correct interpretations of the results. This book, therefore, provides detailed explanations, adequate enough to guide students and researchers on salient issues in research.

ABOUT THE Author

Dr. Joshua Oluwatoyin Adeleke is a Research Fellow in the Institute of Education, University of Ibadan. He has BSc. Ed. (Mathematics and Statistics) of University of Ilorin, M.Ed. (Guidance and Counselling) and Ph.D. (Educational Evaluation), both of the University of Ibadan. He is a seasoned Mathematics Educator and renown quantitative analyst. He has taught both at secondary and University levels. He is a distinguished author who has published in many reputable local and foreign journals. He has served as a resource person for many national projects from where he gathered pool of experiences applied while writing this book. He has also taught courses such as Introduction to data Processing, Inferential statistics and Statistical Methods at Post graduate level. The experience has influenced his writing of this book. He is happily married to Janet and blessed with children.

ISBN: 978-978-49876-4-6