96

T.Y.B.A. Psychology

PAPER IV

Psychological Testing Statistics

Dr. Suhas Pednekar
Vice-Chancellor
University of Mumbai,
Mumbai

Dr. Dhaneswar Harichandan
Director Incharge,
Institute of Distance & Open Learning,
University of Mumbai, Mumbai

Anil R Bankar
Associate Prof. of History & Asst. Director &
Incharge Study Material Section,
IDOL, University of Mumbai

Programme Coordinator: **Dr. Naresh Tambe**
Asst. Professor of Psychology,
IDOL, University of Mumbai.

Editor : **Dr. Vipan Kumar.**
R.D. National College, Bandra,
Mumbai 400050

Course Writers : 1) **Prof. Priyadarshini Bokil,**
Kirti College Dadar (W)
Mumbai - 400028.

2) **Dr. Vipan Kumar.**
R.D. National College, Bandra (W),
Mumbai 400050.

3) **Dr. Minum Saksena**
Lala Lajpatrai College, Mahalaxmi,
Mumbai 400034.

4) **Dr. Priti Sachdev**
101, Daffodil, 56, Pali Hill, Bandra (W),
Mumbai 400050.

5) **Dr. F.B. Ansari,**
Balaji, plat No. 104, opp. Laxmi Park,
Mira Rd, Thane 401 107.

TY.B.A., PSYCHOLOGY- PAPER- IV, PSYCHOLOGICAL TESTING AND STATISTICS

Reprint October 2018,

Published by : Director Incharge
Institute of Distance and Open Learning ,
University of Mumbai,
Vidyanagari, Mumbai - 400 098.

DTP Composed by **Shree Graphic Centre**
28, Mangal wadi,
Mumbai - 400 004.

Printed by P Square Solutions, Barari, Mathura

CONTENTS

Syllabus

Revised syllabi for Psychology Papers IV, V, VI, VII, VIII and IX at T.Y.B.A.

To be brought into effect from June 2010 and the patterns of Question paper for them.

T.Y.B.A. Paper IV - Psychological Testing and Statistics 100 marks

Objectives –

1) To impart knowledge and understanding of the nature, uses, technical features, and the process of construction of Psychological Tests
2) To create awareness about Measurement of Intelligence and Assessment of Personality
3) To impart knowledge and understanding of the basic concepts in Statistics and the various measures of Descriptive Statistics - their characteristics, uses, applications and methods of calculation
4) To create a foundation for advanced learning of Psychological Testing, Assessment, and Statistics ,

(4 lectures per week; 48 lectures per term; 11 lectures per topic in Section I, 7 lectures per topic in Section II, and 6 lectures for the Orientation to Psychological Testing)

Section I - Psychological Testing - 50 marks

Topic 1. Psychological Testing. Assessment and Norms

a) Testing and Assessment - definitions and tools
b) The parties and types of settings involved
c) How are Assessments conducted?
d) Reference sources for authoritative information about tests
e) Various assumptions about Psychological Testing and Assessment
f) What is a 'Good Test'?
g) Norms -sampling to develop norms, types of norms, fixed reference group scoring systems, norm-referenced versus criterion-referenced evaluation
h) Inference from Measurement - meta analysis; culture and inference

Topic II. Reliability

a) The concept of Reliability
b) Reliability estimates -Various methods
c) Using and interpreting a coefficient of Reliability
d) Reliability and individual scores

Topic III. Validity

a) The concept of Validity
b) Content Validity
c) Criterion-related Validity
d) Construct Validity
e) Validity, bias, and fairness

Topic IV. Test Development

a) Test conceptualization
b) Test construction
c) Test tryout
d) Item analysis
e) Test revision

Topic V. Measurement of Intelligence and Assessment of Personality

a) What is Intelligence? - Definitions and theories
b) Measuring Intelligence
c) The Stanford-Binet Intelligence Scales
d) The Wechsler Tests
e) Definitions of Personality and Personality Assessment
f) Personality Assessment-some basic questions
g) Developing instruments to assess Personality - logic and reason, theory, data reduction methods, Criterion groups
h) Personality Assessment and culture
i) Objective methods of Personality Assessment
j) Projective methods of Personality Assessment

Section II Statistics - 50 marks

Topic VI. Types of Scores. Types of Scales. Frequency Distribution and Graphic Representations

a) Continuous and discrete scores - meaning and difference
b) Nominal, ordinal, interval and ratio scales of measurement
c) Preparing a Frequency Distribution
d) Advantages and disadvantages of Preparing a Frequency Distribution
e) Frequency polygon, histogram, cumulative frequency curve, ogive
f) Smoothing a Frequency polygon - method of running averages

Topic VII Measures of Central Tendency

a) Calculation of mean, median and mode

b) The assumed mean method for calculating the mean

c) Merits, limitations, and uses of mean, median and mode

d) Comparison of the 3 Measures of central tendency

Topic VIII. Measures of Variability, Percentiles, and Percentile Ranks

a) Range and Average Deviation

b) Quartile Deviation and Standard Deviation

c) Calculation of the 4 Measures of Variability

d) Merits, limitations, and uses of Range, AD, QD, and SD

e) Comparison of the 4 Measures of Variability

f) Percentiles - nature, merits, limitations, and uses

g) Calculation of Percentiles and Percentile Ranks

Topic IX. Probability. Normal Probability Curve and Standard Scores

a) The concept of Probability

b) Characteristics, importance and applications of the Normal Probability Curve

c) Area under the Normal Curve

d) Skewness- positive and negative, causes of Skewness, formula for calculation

f) Standard scores - z, T, stanine; linear and non-linear transformation; Normalised Standard scores

Topic X. Correlation

a) Meaning and types of Correlation - positive, negative and zero

b) Graphic representations of Correlation - Scatterplots

c) The steps involved in calculation of Pearson's product-moment correlation coefficient

d) Calculation of rho by Spearman's rank-difference method

e) Uses and limitations of Correlation coefficient

f) Simple Regression and multiple Regression

Book for study

Cohen, J.R., & Swerdlik, M.E. (2010). Psychological Testing and Assessment: An introduction to Tests and Measurement. (7th ed.). New York. McGraw-Hill International edition.

Note - Chapter no. 2 - 'Historical, Cultural, and Legal/Ethical Considerations of Testing' will not have a question set on it. However, it should be taught as an orientation to Psychological Testing.

Books for reference

1) Aiken, L. R., & Groth-Marnat, G. (2006). Psychological Testing and Assessment. (12th ed.). Pearson. Indian reprint 2009, by Dorling Kindersley, New Delhi

2) Anastasi, A. & Urbina, S. (1997). Psychological Testing. (7th ed.). Pearson Education, Indian reprint 2002

3) Aaron, A., Aaron, E. N., & Coups, E. J. (2006). Statistics for Psychology. (4th ed.). Pearson Education, Indian reprint 2007

4) Gregory, R. J. (2004). Psychological Testing: History, Principles, and Applications. (4th ed.). Pearson Indian reprint 2008, by Dorling Kindersley India pvt ltd, New Delhi

5) Hoffman, E. (2002). Psychological Testing at Work. New Delhi: Tata McGraw-Hill

6) Hollis-Sawyer, L.A., Thornton, G. C. III, Hurd, B., & Condon, M.E. (2009). Exercises in Psychological Testing. (2nd ed.). Boston: Pearson Education

7) Kaplan, R. M, & Saccuzzo, D. P. (2005). Psychological Testing - Principles, Applications and Issues. (6th ed.). Wadsworth Thomson Learning, Indian reprint 2007.

8) Kline, T.J.B. (2005). Psychological Testing: A Practical approach to design and evaluation. New Delhi: Vistaar (Sage) publications

9) Mangal, S.K. (1987). Statistics in Psychology and Education. New Delhi: Tata McGraw Hill Publishing Company Ltd.

10) McBurney, D.H. (2001). Research Methods. (5thed.). Bangalore: Thomson Learning India

11) McIntire, S.A. & Miller, L.A. (2000). Foundations of Psychological Testing. (1st ed.) Mc. Graw-Hill Higher Education.

❖❖❖

1

PSYCHOLOGICAL TESTING, ASSESSMENT RELATED CONCEPTS

Unit Structure

1.0 OBJECTIVES

After studying This unit you should Know

1. Meaning of Psychological Assessment and Related Concepts.
2. Understand The process of Psychological Assessment.
3. Know the various tools of Psychological Assessment.

1.1 INTRODUCTION

In this unit we will discuss the psychological testing and related concepts of assessment. The process of assessment as well as the various tools of assessment would also be discussed. Assessment involves referrals, administration of the test, report preparation. Concepts related to assessment such as collaborative assessment, therapeutic psychological assessment and alternative uses of assessment would also be discussed . we would also discuss the various ways in which Psychological tests may differ such as its content, format, scoring and interpretation procedures, technical quality, etc. Wide variety of tools used in the process of psychological assessment includes interview, portfolio, case history data, behavioural observation, role play tests and computers.

This unit will conclude with a brief summary, few questions for practice and a list of references for further reading.

1.2 PSYCHOLOGICAL TESTING AND RELATED CONCEPTS

1.2.1 What is Psychological testing?

Imagine going to a vegetable vendor or a goldsmith without weight measures, or to a tailor without length measurements. Measurements are imperative to daily working. It would be impossible for scientists, statisticians to do their work and apply knowledge without measurements. Though each of these measurements comes with limitations, they are indispensable to the professionals in the field. Thus, all fields of knowledge apply measurement and so does psychology. Psychologist measures the psychological variables through Psychological tests and other tools. At the outset let's understand a few concepts.

(1) Psychological tests: They are tools to measure certain **psychological** constructs. Every science field requires some apparatus to measure. For example our vegetable vendor requires different weights and a weighing machine to measure as neurosciences require EEG or MRI scans. Similarly, psychology requires psychological tests to find the nature of psychological problem or understand and measure certain aspects of behaviour.

(2) Variables (in context of testing) are any measurable aspects of behaviour that can vary, for example mood, aggressiveness, ability etc.

Let's take an example to understand this. A psychologist has been asked to select, one candidate, from several candidates who have applied for a sales job in a company. This is a difficult task indeed, because the psychologist has to decide what aspects of a personality would be the best suited for the job. The psychologist decides the person should be **outgoing** and should take i**nitiative**. Now outgoing nature and quality of taking initiatives are some of the **variables** of the personality which our psychologist would measure. Each of the candidates, who have applied for the position, would vary on these variables i.e., some may be high on both the variables or high on one variable (e.g., outgoingness) but may be low on the other variable (e.g., initiative). The psychologist would probably choose candidates who are high on both the variables.

Another aspect our psychologist would ensure is that, the variables she is measuring are seen stable over time. Some of the variables may be transient in nature, for example mood - which is likely to fluctuate over situations or even unrelated environmental factors.

(Remember that some tests particularly measure transient variables like mood, depression, anxiety depending on the requirement of the situation)

(3) Sample of behaviour means observations of an individual (in form of scores) on performing, defined tasks.

When we go to a laboratory for blood testing, the technician takes a small sample of blood to determine if you have a particular infection. Similarly, a psychologist takes a small part of behaviour sample to know about certain aspects of your behaviour. This process is called sampling. Usually these behaviour samples are taken on well defined tasks which measure a particular trait. For example an intelligence test is a set of well defined sub tasks. If the individual is able to perform on these tasks at a given level; we assume he has a particular level of intelligence.

(4) Psychological constructs is a scientific idea developed or generated to describe or explain behaviour. Examples of constructs are intelligence, personality, depression, anxiety, etc. Constructs have to be clearly defined as they are constructed by mental synthesis. **They are assumed to exist.**

For example we assume that if a person is successful means he is intelligent (construct). We have mentally constructed a variable which signifies a particular behaviour pattern. Thus, we need to define these constructs clearly as they are just ideas; ideas which we assume are true or present.

(5) Psychological testing: is the process of measuring psychology related **variables** by means of devices or procedures designed to obtain a **sample of behaviour**. In other words it is a field characterised by the use of 'samples of behaviour' in order to assess **psychological construct(s)**, such as intelligence, personality, or any other aspect of cognitive and emotional functioning, of a given individual.

(6) Psychometrics is the technical term for the science behind psychological testing.

(7) Psychological assessment: The term psychological testing has been replaced by the word psychological assessment since psychological assessment involves use of other psychological tools including psychological testing. Besides psychological tests, the assessment tools include interviews, behavioural observations, case history data, role play tests, etc.

(8) Assessor / test user / test giver: any of these three terms can be alternatively used for the person who is involved in the process of choosing, administering and evaluating the test usually the psychologist.

(9) Assessee / test taker: is an individual who is answering the test, or on whom the test is administered, (the client).

1.2.2Definition of psychological assessment:

> "Psychological assessment is the gathering and integration of psychology related data for the purpose of making psychological evaluation, accomplished through the use of tools such as tests, interviews, case studies, behavioural observations, and specially designed apparatus and measurement procedures". Cohen and Swerdlik

1.3 THE PROCESS OF PSYCHOLOGICAL ASSESSMENT

Psychological assessment is similar to psychological testing but usually involves a more comprehensive assessment of the individual. A psychological test is one of the sources of data used within the process of assessment; usually more than one test is used. Psychological assessment is a process that involves the assimilation of information from multiple sources, including testing, which may include personality tests, intelligence tests, aptitude tests, projective techniques, situational tests, etc. Other sources of information include information from personal interviews, formal and informal records including photographs, audio records, or records relating to personal, occupational, or medical history, or from interviews with parents, spouses, teachers, or previous therapists or physicians. Psychological assessment is a complex, detailed, in-depth process though many psychologists may do some level of assessment like using simple checklists to assess some traits or symptoms. Typical types of focus for psychological assessment are:

- to provide a diagnosis for treatment settings.
- to assess a particular area of functioning or disability often for school settings;
- to help select type of treatment or to assess treatment outcomes;
- to help courts decide issues such as child custody or competency to stand trial; or
- to help assess job applicants or employees and provide career development counselling or training.

Let's understand the process of psychological assessment with the help of the case study.

Case study: Simran is a standard seven student. Lately her teacher observed that Simran has been unusually quite in the class. She seems to have lost interest in studies, pays no attention to what teacher is teaching.

Now let's see how this process of assessment proceeds.

The process of assessment:

1.3.1 Referrals: Usually the process of assessment starts with a referral from a school counsellor, or a teacher. After such a referral the assessor (who is usually a psychologist) meets the assessee (one who would be undergoing the psychological assessment) or others to clarify aspects of reason for referral.

(Case study continues): After observing Simran for quite some time the teacher leads Simran to the school psychologist. The school psychologist tries to understand the exact nature of the problem from the teacher. She asks relevant questions to the teacher. The psychologist may also involve Simran's parents to give their input about Simran's behaviour change.

1.3.2 Deciding the tools: The assessor's major task is to select from the available tools in such a way that they prove to be highly effective in understanding the nature of problem- the referral question. These tools may vary from psychological tests to interviews, portfolio, behavioural observation, etc.

Besides each of the tests / tools may have a different theoretical background and thus the interpretation may vary accordingly. For example the interpretation would be altered depending upon the type of personality tests used - like Rorschach inkblot test (which is based on psychoanalytic perspective) or 16 Personality factor test (which has its basis in the trait theory). Psychologists may select the test according to the nature of problem; the context of the problem, what they suspect may be the cause of the problem and whether the test is capable of tracking that problem.

(Case study continues): After discussion with parents and Simran herself the psychologist comes to some tentative conclusion about the nature of the problem. Simran's psychologist feels ' that Simran is depressed. The psychologist then decides that she could use some tests which measure depression. Out of the range of tests available for depression measurement, the psychologist chooses the test that is available for school going children, applicable in Indian context, available for children of Simran's age group. She also ensures that the test has separate norms for girls.

The psychologist also conducts interview sessions with parents and Simran and may, if need be, include some other teachers who teach Simran.

1.3.3 Administration of the tool: The assessor then administers the relevant tools and psychological instruments to the client.

(Case study continues): the psychologist then administers the test to Simran. The tools administered are of critical importance as the psychologist has to administer them under carefully controlled conditions. The psychologist ensures that, factors like, time of the session, or other conditions do not unduly influence the assessment process. The psychologist could check out whether Simran is empty stomach, whether she just has realised that she has failed in her maths exam or simply that she is irritable because she is sleepy. If any of these conditions are likely to influence Simran's assessment process, the psychologist does not administer the test or stops the process temporarily.

1.3.4 Report preparation: After assessing the client, using the various tools, the assessor prepares a report. It is usually -an intensive report which consists of the data collected, its interpretation, other significant / relevant observations, and feedback from the assessee (client).

(Case study continues): After the testing session the psychologist prepares a report about Simran which includes her test reports and their interpretation. The psychologist then adds other relevant information about Simran like observations about her behaviour. After preparation of a tentative report, the psychologist takes a feedback from others like her parents, teacher and even Simran about what they feel about the observations in the report. Some things might get clarified at this stage and the psychologist may choose to alter the report accordingly. After this stage the report may be finalised.

1.3.5 Revelation and explanation of report I findings: the report and referral issues may be then discussed with the parents and the professional who has referred the client.

(Case study continues): after finalising the report, the psychologist may discuss various relevant issues that are influencing certain behavioural patterns with Simran and her parents.

There are two approaches to this. The first is when the psychologist takes minimal feedback from Simran and the primary focus is the test scores of Simran (the assessee / client). In this case the clinician / assessor can collect data through testing, interview, case history and other available data from the process of formal assessment. The psychologist (assessor) then reveals the findings in a scheduled meeting with little or no feedback from Simran (the client).

The second approach is the collaborative psychological assessment perspective in which the Simran (assessee) is perceived as a partner of the entire assessing process. It is construed that she is an expert about her current views and events. A form of this collaborative assessment may include an element of therapy as a part of process.

The **"therapeutic psychological assessment** is an approach that encourages therapeutic self-discovery through assessment process. Another rather frequently used term is the dynamic **psychological assessment**. The dynamic psychological assessment is defined as "a model and philosophy of interactive evaluation involving various types of assessor's intervention during assessment process." Cohen and Swerdlik.

The dynamic psychological assessment is an interactional process between the assessee (Simran) and the assessor (her psychologist) in which the assessor may intercede, give feedback, make suggestions, change ineffective problem solving methods and modify ideas of the assessee to bring desirable changes in the assessee.

1.3.6 Use of alternate assessment:

Now let's assume for a moment, that Simran had some physical disability like she had a difficulty reading small print. Would the testing session then remain the same? If the testing remained the same then it would be unfair to Simran. Simran's special testing needs have to be addressed to. Students with special needs will have to be given alternate assessment to aid the process of fair testing.

Usually these alternate assessment methods are individually tailored and may take form of audio taped administrations, use of Braille or performance based tests. According to Cohen and Swerdlik an "Alternate assessment is an evaluative or diagnostic procedure or process that varies from usual, customary, or standardised way a measurement is derived, either by virtue of some special accommodation made to the assessee or by means of alternative methods designed to measure the same variable(s)."

1.4 TOOLS OF PSYCHOLOGICAL ASSESSMENT

1.4.1 The psychological test

Psychological tests are measuring devices that are designed to measure psychological variables like intelligence, aptitude, attitudes, personality, etc.

> "The term Psychological test refers to a device or procedure designed to measure variables related to psychology." Cohen and Swerdlik.

> "A psychological test or educational test is a set of items designed to measure characteristics of human beings that pertain to behaviour." Kaplan and Saccuzzo.

Psychological tests may differ on a number of aspects such as:

a) **Content -** which will depend on what the test purports to measure or what is the focus of the test.

b) Format - **which is the plan, structure or arrangement and the format of** administration procedures such as whether the test is administered in paper - pencil format or is computerised, etc. Psychological tests can be classified into different types depending upon whether they can be administered to one individual at a time - individual test or to a group of people together - group test. Some psychological tests are paper pencil tests while some are performance tests. Psychological tests can also be classified on the basis of types of behaviour they measure such as ability, aptitude or achievement.

c) **Scoring and Interpretation procedures -** Scoring is, the process of assigning predetermined evaluative codes to certain type of responses on tests, tasks or other behaviour samples. The scores can be categorised in different ways. One such way is the use of cut score or a cut-off score which is a reference point or a score used to divide a set of data into two or more classification. The tests differ from each other on the basis of whether they can be scored by the test takers themselves or require a qualified evaluator.

d) **Technical quality:** Tests differ from each other on the basis of their technical soundness or psychometric soundness. Psychometric soundness refers to the consistency and accuracy of psychological test measures, what it purports to measure, i.e., its validity.

> "Psychological testing refers to all the possible uses, applications and underlying concepts of psychological and educational tests. The main use of these tests is to evaluate individual differences, or variations among individuals." Kaplan and Saccuzzo.

1.4.2 Interview: An interview is a directed conversation aimed at eliciting information for diagnosis, evaluation, treatment, planning, etc. In other words interview is a method of gathering information through direct communication involving reciprocal exchange. The interview may be conducted by a therapist counsellor / psychologist with the aim of assessing the behaviour of the client to know the client's personality and

capabilities. Interviews also highlight the capability of an individual in to response to various situations. Interviews are usually planned depending upon their purpose / goal, their expected length, restrictions under which they are conducted, and interviewee compliance. The tool of interview is widely used in several settings including clinical settings, school settings, or educational settings, corporate placements, consumer behaviour and several others. Besides face-to-face interviews, telephone interviews, internet interviews have also gained a position.

Note that an interview can be conducted in many ways and for a variety of purposes. What is noted in an interview?

The interviewer may closely observe the client so as to gather better information about the client. He may particularly observe what does the client tell him? How much information is the client willing to and able to provide? Are there any cues that the client is taking from the interviewer for social approval? What is the pace, tone, volume, inflection, of the client? What is his command of language, how well does he choose his words and how organised is he in his speech? The interviewer may also try to figure out whether the client is cooperating with the interviewer, whether the client has voluntarily come in for the interview, etc.

Interviews are of two type's viz., the structured interview and the unstructured interview.

i) **Structured -** Structured interview is designed for specific information gathering. The type of questioning is usually in yes/no" or "definitely/ somewhat/not at all" forced choice format and are often used to provide a diagnosis for a client by detailed questioning of the client. It may be broken up into different sections reflecting the diagnosis in question. The Structured Clinical Interview for the DSM-111-R (SCID-R) is an example of a structured interview.

ii) **Unstructured -** Other interviews can be less structured and allow the client more control over the topic and direction of the interview. Unstructured interviews are better suited for general information gathering, and structured interviews for specific information gathering. Unstructured interviews often use open questions, which ask for more explanation and elaboration on the part of the client.

As interview is a reciprocal affair, the skills of the interviewer affect the quality of the interview. If the interviewer is skilful, he can elicit quality information from the interviewee. Interviewing skills may include interviewer's ability to convey genuineness,

empathy humour and ability to pick up quickly from the answers - relevant to unstructured interviews.

1.4.3 The portfolio: Portfolio is a work sample to assess / evaluate the effectiveness of the client on a particular skill or task. Portfolio assessment is used in varied situations including educational settings. The basic contention is that, the process of assessment cannot be carried out with a single administration of test; instead a compilation of related work may give better picture of the client's capabilities. Thus, portfolio is used to give a better idea about the client's capabilities. Electronic portfolio is a personal digital record containing information such as a collection of artifacts or evidence demonstrating what one knows and can do. For example if we want to assess a student's writing skill, we cannot rely on one test administration and come to a conclusion about it. Instead the student can be asked to give his compiled work or selected writing samples to aid the process of evaluation.

1.4.4 Case history data: Case study refers to an in-depth analysis of the client which may be descriptive or explanatory. Case study assembles the case history data for a detailed macro review and to ascertain facts. The case history data refers to records, transcripts, and other accounts in written, pictorial or any other form. They may include archived information, including institutionalised files, informal accounts and other data relevant to the assessee. Case history data can be a very useful tool in wide variety of assessment contexts including clinical evaluations, neuropsychological evaluations or even school settings.

1.4.5 Behavioural observation: Behavioural observation is examining the actions of assessee by visual or electronic means, while recording quantitative and/or qualitative information regarding the actions. How does the client act? Nervous, calm, smug? What does he do or not do? Does the assessee make and maintain eye contact? How does the assessee solve a problem?

Behavioral observation may be used in a variety of settings such as clinical settings (such as to add to interview information or to assess results of treatment) or in naturalistic settings like a classroom or in research settings including laboratory or other structured settings. Although most of the times it is feasible that this observation is carried out in structured setting like a clinic. Behavioral observations may be done with a variety of assessment objectives.

1.4.6 Role play tests: Role-playing refers to the changing one's behavior to assume a role, either unconsciously or consciously to act out an adopted role. According to Cohen and Swerdlik, "Role play is acting an impoverished or partially impoverished - part in simulation situation." Role-playing may also refer to role training where people rehearse situations in preparation

for a future performance and to improve their abilities within a role such as particular occupation, education and certain military war games.

"Role play test is a tool of assessment wherein assessees are directed to act as if they were in a particular situation." The assessees are then evaluated with regard to their expressed thoughts, their problem solving approach, the effectiveness of the approach, the quality of problem resolution, related behaviours and other variables. Such role play tools are often used for conflict resolution and stress management programs.

1.4.7 Computers: The task of test administration, scoring and evaluation is tedious and prone to many errors. Computers play a major role in today's testing. CAPS or Computer - Assisted psychological assessment is the computer assistance to test user for administering, scoring, and interpreting tests. CAPA enables the test taker to work independently and thus test administrator related variables such as giving cues do not affect testing. CAPA not only enables easy administration but also makes complex scoring and data combination strategies possible.

1.5. SUMMARY

1 . Psychology requires psychological tests to find the nature of psychological problem or understand and measure certain behavioural variables. Psychological testing is the process of measuring psychology related variables by means of devices or procedures designed to obtain a sample of behaviour.

2. The term psychological assessment involves the use of other psychological tools including psychological testing, interviews, behavioural observations, case history data, role play tests, etc.

3. Assessor / test user / test giver is the person who is involved in the process of choosing, administering and evaluating the test. While Assessee / test taker is an individual who is answering the test, or on whom the test is administered, (the client).

4. The process of psychological assessments usually begins with referrals from a school counsellor, or a teacher. After such a referral the assessor meets the assessee to clarify aspects of reason for referral. The assessor's major task is to select from the available tools in such a way that they prove to be highly effective in understanding the nature of problem - the referral question. These tools may vary from psychological tests to interviews, portfolio, behavioural observation, etc. The assessor then administers the relevant tools and psychological instruments to the client. After assessing the client, using the various tools, the assessor prepares a report. It is usually an intensive report which consists of the data collected, its interpretation, other significant / relevant observations, and feedback from the assessee (client).

5. There are two approaches to this. The first is when the psychologist takes minimal feedback from the client while the second approach is the **collaborative psychological assessment** perspective in which the assessee is perceived as a partner of the entire assessing process. The **"therapeutic psychological assessment** is an approach that encourages therapeutic self-discovery through assessment process. Another dynamic **psychological assessment** is a model and philosophy of interactive evaluation involving various types of assessor's intervention during assessment process.

6. The various tools of assessment include psychological tests which may differ on a number of variables such as content, format, scoring interpretation and technical quality. An **interview** is a method of gathering information through direct communication involving reciprocal exchange. **Portfolio** is a work sample to assess / evaluate the effectiveness of the client on a particular skill or task. **Case study** refers to an in-depth analysis of the client which may be descriptive or explanatory. **Behavioural observation** is examining the actions of assessee by visual or electronic means, while recording quantitative and/or qualitative information regarding the actions. **Role play** test is a tool of assessment wherein assessees are directed to act as if they were in a particular situation. The assessees are then evaluated with regard to their expressed thoughts, their problem solving approach, the effectiveness of the approach, the quality of problem resolution, related behaviors and other variables. CAPA or Computer Assisted Psychological Assessment is the computer assistance to test user for administering, scoring, and interpreting tests.

1.6 QUESTIONS

Answer the following questions:

Q1. How is psychological testing different from psychological assessment? Explain the process of psychological assessment.
Q2. Explain the various tools of psychological assessment.
Q3. Define or Explain the following terms
   a) Psychological Testing
   b) Psychometrics
   c) Psychological Assessment
   d) Collaborative Psychological Assessment
   e) Dynamic Psychological Assessment
   f) Interview
   g) Portfolio
   h) Alternative Assessment

1.7 REFERENCES

1 Cohen, R.J., & Swerdlik, M.E., (2010). Psychological testing and Assessment: An introduction to Tests and Measurement, (7 th ed.), New York. McGraw - Hill International edition, 129 -132

2. Anastasi, A. & Urbina, S. (1997). Psychological Testing. (7th Ed.). Pearson Education, Indian reprint 2002.

3. Kaplan, R.M., & Saccuzzo, D.P. (2005). Psychological Testing - Principles, Applications and Issues. (6 th Ed.). Wardsworth Thomson Learning, Indian reprint 2007.

2

THE PARTIES AND SETTINGS INVOLVED IN TESTING AND ASSESSMENT

Unit Structure

2.0 OBJECTIVES

After studying This unit you should be able to

1. Understand the various parties involved in the assessment process and their roles.
2. Comprehend the various settings of assessment.

2.1 INTRODUCTION

In this unit we will discuss the various parties involved in the assessment process and their roles. The most common parties involved in the assessment process includes the test developer, the' test user, the test taker, society and the test utilizer. The role of each is briefly discussed.

Tests are used in wide variety of settings that can range from educational setting to geriatric settings. Tests are also used in clinical, counseling, business and military settings.

The unit will end with a brief summary, questions and list of references for further readings.

2.2 THE PARTIES IN ASSESSMENT

The primary three parties involved in testing are the test developer, the test user and the test taker. Besides these three primary parties the process of assessment may also involve society at large and other parties directly or indirectly involved in the process.

2.2.1 Test Developer -. Test developers are people and organisations that construct tests, as well as those that set policies for testing programs, In other words test developers create tests. Some tests are created with specific research purpose while some are modifications or refinements of existing tests. As tests have a significant impact on the people, test developers have to develop tests with a lot of responsibility. Organisations such as the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education have together published the 'Standards for Educational and Psychological Testing', which also include ethical behaviors in test development and use.

Issues like test construction and evaluation, test administration and testing the minorities and special applications of the test are covered in these documents. The test developer has certain responsibilities in developing, marketing, distributing tests and educating test users. Test developers should provide the information and supporting evidence that test users need to select appropriate tests, what the test measures, their recommended use, the intended test takers, the strengths and limitations of the test, including the level of precision of the test scores. They should describe how the content and skills to be tested were selected and how the tests were developed. Obtain and provide evidence on the performance of test takers of diverse subgroups, making significant efforts to obtain sample sizes that are adequate for subgroup analyses. They must further evaluate the evidence to ensure that differences in performance are related to the skills being assessed.

2.2.2 Test User - Test users are people and agencies that select tests, administer tests, commission test development services, or make decisions on the basis of test scores. The test user may be a counselor, a clinician, or a personnel official. The 'Standards for Educational and Psychological Testing', offer guidelines not only to test developer but also the test users. It is important to remember that if a test is not managed competently at all levels, then, no matter how sound the test is its purpose will be beaten. 'Standards for Educational and Psychological Testing' offer guidelines to the test user regarding the choice of test, conditions of test use, and the process of testing. The Test User has certain responsibilities in selecting, using, scoring, interpreting, and utilizing tests. The Code of Fair Testing Practices in Education (Code) which is published by American Counseling Association (ACA), the American Educational Research Association (AERA), the American Psychological Association (APA), the American Speech-Language-Hearing Association (ASHA), the National Association of School Psychologists (NASP), the National Association of Test Directors (NATD), and the National Council on Measurement in Education INCME) is a guide for professionals using educational instruments in fulfilling their obligation to provide and use tests that are fair to all test takers regardless of age,

gender, disability, race, ethnicity, national origin, religion, sexual orientation, linguistic background, or other personal characteristics. Careful standardization of tests and administration conditions helps to ensure that all test takers are given a comparable opportunity to demonstrate what they know and how they can perform in the area being tested.

2.2.3 Test Taker - The test taker is the subject of assessment or an evaluation. Each test taker may vary on the anxiety they experience, their experience and attitude with the test and test taking, proper coaching, ability to comprehend written test instructions. Test takers may also vary on physical discomfort, alertness, willingness to cooperate, or importance they may attribute to portraying themselves in a good (bad) light.

The 'Standards for Educational and Psychological Testing' offer guidelines on test takers rights and responsibilities. Rights of the test taker include the right to be informed of rights and responsibilities as a test taker, be treated with courtesy, respect, and impartiality, to be tested with measures that meet professional standards, receive a brief oral or written explanation prior to testing about the purpose for testing, the kind of tests to be used and safeguarding confidentiality.

The test takers responsibilities include asking questions prior to testing especially when the test taker might be uncertain about certain aspects of the testing processes. The test taker must read or listen to descriptive information in advance of testing and listen carefully to all test instructions. He should also inform an examiner, in advance, if he / she has a physical condition or illness that may interfere with test performance. They must notify to the examiner about difficulty in comprehending the language of the test.

2.2.4 Society - The society has been always seeking to classify people into various categories. It is the societal need for organising and systematising that makes society imperative party to the process of assessment. Religious inclinations, intellectual sophistication of human mind has well reflected in the assessment process to understand and predict human behavior.

2.2.5 Test Utiliser - The test utiliser may be the test taker, but in other cases however, a business or organisation may send a person to be tested. Thus, the organisation also has certain rights regarding tests, their use, and the information gained from them. Testing and interpretation services are offered by private parties including companies / services and sometimes these companies / services are extensions of test publishers. There are academicians who review tests and evaluate their soundness. All these parties are to a greater or lesser extent involved in the process of assessment.

2.3 SETTINGS OF ASSESSMENT

2.3.1 Educational settings: Educational measurement is a process of assessment or an evaluation in which the objective is to quantify level of attainment or competence within a specified domain of skill or knowledge. In other words educational assessment is the process of documenting, usually in measurable terms, knowledge, skills, attitudes and beliefs of the assessees. This assessment focuses on the individual learner, the learning community, the institution, or the educational system as a whole. It is important to note that the final purpose and assessment practices in education depend on the theoretical framework of the practitioners and researchers.

Various tests that measure ability, aptitude, interest or even achievement are commonly known. Besides these tests, diagnostic tests which help narrow down and identify areas of deficit, including diagnostic tests in the arena of school subjects, may be administered by school psychologists / counselors to assessees.

Confidence-Based Learning accurately measures a learner's knowledge quality by measuring both the correctness of his or her knowledge and the person's confidence in that knowledge. Escape, is a technology and an approach that looks specifically at the assessment of creativity and collaboration.

2.3.2 Geriatric settings: As the age progresses the overall functioning of individuals reduces or is impaired. They may show delayed response and other cognitive deteriorations. They may have difficulty in physical functioning and even adaptive functioning. Thus, these individuals may require psychological assessment to evaluate their cognitive, functioning, adaptive functioning. Geriatric assessment is the assessment of geriatric patients, those elderly people requiring treatment for physical or mental disorders.

In geriatric psychiatry or geriatric psychology settings the conduct of a psychiatric assessment and psychological assessment helps identify the impact of deterioration on people's ability to live independently and if this is compromised a needs assessment is carried out. Particular aspects of assessment include evaluation of alcohol use in elder adults, evaluation of dementia, evaluation of depression in older adults, evaluation of memory in older adult and evaluation of substance abuse in older adults.

2.3.3 Counseling settings: The aim of counseling assessment is to diagnose the deficits that have to be targeted for intervention. Assessment in counseling may be carried out in a clinic, school or other educational institutes, rehabilitation centers' and many other diverse contexts. These

settings may be either privately owned or government set ups. The target of these assessments is to finally improve adjustments, productivity, quality of life, etc. The assessment measures are decided depending upon the referral question. Usually measures of personality, interest, attitude, aptitude and social and academic skills are used.

2.3.4 Clinical settings: Clinical tests and assessments such as MMPI, Major Depression Inventory, etc., are used in clinical settings such as psychiatric inpatient and outpatient clinics, private consulting rooms. The assessments can be carried out to find any non obvious clues to maladjustment, to determine whether a particular type of psychotherapy would be effective in solving the underlying problem, to give an opinion on the client's psychological problem or on defendant's competency to stands a trial.

2.3.5 Business and military settings: Appointment of personnel, promotion, identification of deficits in performance, job satisfaction, eligibility for further training are the various issues handled by the assessment in the business and military settings. Various psychological tests are used depending upon the need for assessment. They could be leadership skills, job adjustment scales; stress evaluation scales, etc. Use of various tests along with interviews, on the job observation, ratings, etc., are often used to evaluate personnel related variables.

2.3.6 Other settings: Tools of assessment can be used in varied contexts other than the ones mentioned above. Research and practice in different areas of human behavior has enabled psychologists to devise new tools of measurement. Upcoming fields and newly established fields like health psychology, sports psychology, and spiritual psychology have created new paradigms. Thus, assessments deal with typical issues relevant in those specialisations.

2.4 SUMMARY

1 . The parties involved in assessment include the test developer, the test user, the test taker, the society at large and other parties who are directly or indirectly involved in the process of assessment.

2. The test developers are people and organisations that construct tests, as well as those that set policies for testing programs. Some tests are created with specific research purpose while some are modifications or refinements of existing tests.

3. The Test Developer has many responsibilities in developing, marketing, distributing tests and educating test users.

4. The test taker is the subject of assessment or an evaluation. Each test taker may vary on the testing anxiety, ability to comprehend physical discomfort, alertness, etc.

5. The test utilizer may be the test taker or the organisation that may send a person to be tested. Thus, the organisation also has certain rights regarding tests, their use, and the information gained from them.

6. The settings of assessment include educational settings, clinical settings, geriatric settings, counseling settings and other varied settings.

7. Educational measurement is a process of assessment or an evaluation in which the objective is to quantify level of attainment or competence within a specified domain of skill or knowledge. Various tests that measure ability, aptitude, interest or even achievement are commonly used in educational settings.

8. In geriatric psychiatry or geriatric psychology settings the conduct of a psychiatric assessment and psychological assessment helps identify the impact of deterioration.

9. The aim of counseling assessment is to diagnose the deficits that have to be targeted for intervention. Assessment in counseling may be carried out in a clinic, school or other educational institutes, rehabilitation centers' and many other diverse contexts. Usually measures of personality, interest, attitude, aptitude and social and academic skills are used.

10. Clinical tests and assessments such as MMPI, Major depression inventory, etc., are used in clinical settings such as psychiatric inpatient and outpatient clinics, private consulting rooms.

11. Appointment of personnel, promotion, identification of deficits in performance, job satisfaction, eligibility for further training are the various issues handled by the assessment in the business and military settings.

2.5 QUESTIONS

Answer the following questions:

Q1. Who are the parties involved in the process of assessment? Elaborate their role in the context of assessment.

Q2. Explain the various settings of psychological assessment.

2.6 REFERENCES

1 Cohen, R.J., & Swerdlik, M.E., (2020). Psychological testing and Assessment: An introduction to Tests and Measurement, (7 th ed.), New York. McGraw - Hill International edition, 229 232

2. Anastasi, A. & Urbina, S. (2997). Psychological Testing. (7th ed.). Pearson Education, Indian reprint 2002.

3. Kaplan, R.M., & Saccuzzo, D.P. (2005) . Psychological Testing - Principles, Applications and Issues. (6 th ed.). Wadsworth Thomson Learning, Indian reprint 2007.

4. http://www.apa.org/science/programs/testing/fair-code.aspx

5. http://psychology.wikia.com/wiki/Needs-assessment

3

REFERENCE SOURCES, ASSUMPTIONS ABOUT ASSESSMENT AND CRITERIA OF A GOOD TEST

Unit Structure

3.0 OBJECTIVES

After studying this unit you should be able to

1 - Know the various reference sources for authoritative information about tests.
2. Understand the various assumptions about psychological testing and assessment.
3. Discuss the criteria of a good test.

3.1 INTRODUCTION

In this unit we will discuss the discuss the various reference sources for authoritative information about tests. These include test catalogues, test manuals, references volumes, Journal articles, online test reviews, etc.

Psychological testing and assessment is based on certain assumptions there are seven such assumptions that would be discussed in brief.

Criterias of a good test would be discussed . Among the most important criteria as are reliability, validity, norms and standardization.

Towards the end of this unit we will summarise the major points of the unit followed by questions and references for further reading.

3.2 REFERENCE SOURCES FOR AUTHORITATIVE INFORMATION ABOUT TESTS

There are various sources from where the authoritative information related to the test can be sought. These sources vary in details they offer regarding the test, some sources provide a very brief description about the test while other sources provide a detailed review of the technical aspects of the test. These sources are discussed below:

3.2.1 Test catalogues : All of the major test publishers have catalogs listing their own products. The catalogs are published on an annual or semiannual basis. These catalogs are frequently the best source of information for the most recent editions of a test. Information on the cost of materials and scoring, types of scoring services, and ancillary materials are available through the catalogs. The catalogs also include information on variations of the test, such as large-print or foreign language versions. The publishers have to be contacted to request their catalog.

3.2.2 Test manuals : Most detailed information about the test is found in the test manual itself, which accompanies the test, when you buy the test. The test manuals are available with the publishers and some of the publishers may require your certification before handing over the manual to you. This is especially true in case of tests which require professional training for administration and interpretation, e.g., the WISC.

3.2.3 Reference volumes: Tests in Print (TIP), the Mental Measurements Yearbook (MMY), Tests, and Test Critiques are most useful and popular references. These are usually available in the reference section of the university, and larger public libraries. The Mental Measurements Yearbook (MM)9 is published by the Burros Institute for Mental Measurements, Lincoln, NE. In the MMY, tests are listed alphabetically by title. Each entry provides descriptive information, such as the test name, intended population, publication dates, forms and prices, test author', and publisher. It also contains additional information on the extent to which reliability, validity, norming data, scoring and reporting services, and foreign language versions are available. Most entries also include one or more reviews of the test and testing materials (e.g., manuals) by qualified psychologists.

Tests is published by Pro-Ed, Inc., Austin, TX Tests is a bibliographic encyclopedia containing information on thousands of testing instruments in psychology, education, and business. It provides concise descriptions of tests, with each entry including the test title and author, the intended population, the tests purpose, the major features, the administration time, the scoring method, the cost and availability, and the primary publisher. Also, a scanning line uses coded visual keys to indicate whether the test is self administered or examiner-administered. Tests do not contain evaluative critiques or data on reliability, validity, or norms.

Test Critiques is published by Pro-Ed, Inc., Austin, TX and is updated annually. This text is designed to be a companion to Tests and contains supplemental information designated as 'not appropriate' for inclusion in that directory. This includes psychometric information such as reliability, validity, and norm development. The tri-part entry for each test includes an Introduction, Practical Applications/Uses, and Technical Aspects, followed by a critique. Technical Aspects includes citations from reliability and validity studies and opinions from experts regarding the technical adequacy of the test. The text is written for both professionals and students, with technical terms explained, and includes helpful information not usually found in other reference books. This makes it a user-friendly resource for students, teachers, or persons unfamiliar with test terminology.

Tests in Print (TIP) is published by the Buros Institute for Mental Measurements, Lincoln, NE. It is a bibliographic encyclopedia of information on every published (and commercially available) test in psychology and achievement. Each entry consists of the test title, intended population, publication date, acronym (if applicable), author, publisher, foreign adaptations, and references.

3.2.4 Journal articles: Articles in the journal may also contain test reviews or reports about the soundness of a particular test or examples about how various instruments were used in research or an applied context. Journals such as Journal of Psychoeducational assessment, Educational and Psychological Measurement, Journal of Personality assessment, address issues relating to assessment and contain other extremely informative articles.

3.2.5 Online databases: The most comprehensive way to search for information is through the World Wide Web on the Internet. The Test Locator allows you to access test information from a variety of sources. Originally a joint effort of the ERIC Clearinghouse on Assessment and Evaluation at the Catholic University of America, the Library and Reference Services Division of ETS, the Buros Institute of Mental Measurements at the University of Nebraska, and Pro-Ed (a publisher of test reviews), the Test Locator was designed as a gateway to various sources of information on tests. Sponsor websites include Ericae.net, Buros Institute of Mental Measurements, and ETS Test Link.

Test Reviews Online a web-based service of the Buros Institute of Mental Measurements. The search engines allow you to examine a large amount of information on tests and testing. For a small fee, users may download information for any of over 2,000 tests that include specifics on test purpose, population, publication date, administration time, and descriptive test critiques.

Besides this there are Software and Scoring Services for Published Tests. You can find a multitude of computerized testing materials, including a variety of software products developed for administering, scoring, and interpreting published tests. There are a couple of helpful directories that can lead you to the software you need.

The Psychware Sourcebook which is published by the Metritech, Inc., Champaign, IL. This handy reference identifies and describes over 533 computer-based assessment tools, including computerised versions of paper-and-pencil tests and computer based scoring and interpretation systems. It serves as a directory of available computer services and contains no reviews or critiques of software or services.

3.3 ASSUMPTIONS ABOUT PSYCHOLOGICAL TESTING AND ASSESSMENT

Assumption 1: Psychological traits and states exist - It is a basic assumption of testing that psychological traits exist and so do traits, although they cannot be observed they can be interpreted from overt behavior.

A trait is a relatively enduring or long lasting characteristic on which people differ. A trait can be defined as "any distinguishable, relatively enduring way in which one individual varies from another" (Guilford). Intelligence, cognitive styles, personality are examples of traits.

A state is a less enduring or more transient characteristic on which people differ. State refers to momentary psychological condition. Happiness, sadness, irritability reflect states of mind.

Though traits and states cannot be seen, heard or touched, we assume that the traits and states exist because we can infer them from the overt behavior. The overt behavior includes the assessment related behavior as well as an observable action.

Traits are constructs which are real, in the sense, that they are useful for classifying and organising the world. They can be used to understand and predict behavior. Besides they refer to behaviors that we can measure.

Assumption 2: Psychological traits and states can be quantified and measured - A single trait can be defined depending on the context that it is used. Thus, the trait measurement can be situation dependant. Once the trait is defined as per the specific condition, then the test developed considers the items that would appropriately reflect the behavior in question, followed by the appropriate ways to score and interpret the test results. Besides, various approaches to measuring a single trait can be useful. For example, different tests of intelligence tap into somewhat different aspects of the construct of intelligence.

Assumption 3: Test related behavior predicts non test related behavior - The goal of testing usually is to predict behavior other than the exact behaviors required while the exam is being taken. The most important reason for giving tests is to predict future behavior. For example the clients score on The Minnesota Multiphasic Personality Inventory (MMPI) is used as an indicator of the presence or absence of various mental disorders. The actual mechanics of measurement like self-reports, behavioral performance or projective can vary widely and still provide good measurement of educational, psychological, and other types of variables.

We need to understand at this point that any test is only a sample of behavior and not the universe of all possible behaviors, reflecting a particular trait or state. Some behaviors are more characteristic of a particular trait and may be reflected in the test items; however it is impossible to include all the universe of items. Thus, a psychological test is just an attempt to predict that if the assessee scores high on a particular trait, he would in real life situation behave consistent to his trait.

Assumption 4: Tests and other measurement techniques have strengths and weaknesses - It is essential that users of tests understand how they can use the tests appropriately and intelligently. The test users need to know how the tests are to be administered, who are the clients to whom the test is best suited, how are the results to be interpreted, etc. Besides and very importantly, the test user needs to know what the strengths of the test are and what its limitations are. They need to understand how the limitations can be best compensated with other assessment processes to ensure good assessment.

Assumption 5: Various sources of error are a part of assessment process - There- is no such thing as perfect measurement. All measurements have some error. Irrelevant factors which affect testing and are not the ones which are intended to be measured by the test, are errors. In other words, an error is the difference between a person's true score and that person's observed score. The two main types of errors are random error and systematic error. Errors can be attributed to test construction, administration, -scoring and interpretation, or any other irrelevant factors in the environment.

Errors are just an element of the measurement process and are taken into account during the process of measurement.

Assumption 6: Testing can be conducted in a fair and unbiased manner - Fairness refers to using tests, that are just to all test takers regardless of age, gender, disability, race, ethnicity, national origin, religion, sexual orientation, linguistic background or other personal characteristics. This requires careful construction of the test items and testing of these items on different types of people.

Most test developers are now sensitive to issues of the minorities and other cultural groups within the society. The societal demands of fair tests used in a fair manner have necessitated the test publishers to develop tests / instruments that are in strict accordance with the guidelines. This assumption also requires that the test be administered to those types of people for whom it has been shown to operate properly. .

Assumption 7: Testing and assessment benefit society - Many critical decisions are made on the basis of tests such as measuring teacher competency, employability, presence of a psychological disorder, degree of teacher satisfactions, degree of student satisfaction, etc. Without tests, the world would be much more unpredictable. Imagine students getting into management schools, engineering schools or medical schools without documented merit or students with learning disabilities without remediation due to absence of tests that would diagnose these problems. Thus, tests are indispensible in today's world and are of immense value to society.

If tests assume such an indispensible position in today's world then it is all the more important to know what a good test is.

3.4 CRITERIA OF A GOOD TEST

As we have seen earlier sound psychometric test is essential. What make a test psychometrically sound, or what are the aspects that make a test a good test? Let's understand these characteristics:

1. **A good test must be reliable:** The word reliability means dependability or consistency. A reliable test is one that gives stable and consistent results. In other words reliability is the precision with which the test measures and the extent to which errors are present in measurement. 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. Psychological tests are reliable to varying degrees, usually 0.90 is considered high reliability coefficient and low would be anything below 0.65. More on reliability in the chapter dedicated to reliability, covered in units nos. 5,6 and 7.

2. **A good test must be valid:** Reliability is essential criteria of a good test but not the sufficient criteria. While reliability is concerned with the accuracy of the actual measuring instrument or procedure, validity refers to the degree to which the test accurately reflects or assesses the specific concept that the researcher is attempting to measure. In other words validity concerns the extent to which the test and assessment procedures used in psychological and educational testing, measure what they purport to measure. Validity is a subjective judgment made on the basis of experience and empirical indicators. The validity of the test may be questioned on the grounds whether a particular score on the test (either high or low) is related to assessees behavior. For more information on validity refer to the chapter dedicated to validity, covered in unit no. 8.

3. **Does the test have established norms:** Norms means scores which are normal or typical of a given population on a given test. Norms are obtained from an initial group of people, **the representative sample** who truly represent the larger population for whom the test is meant; on factors such as age, social stratification, gender, educational level, etc. So we could have age norms, or gender norms for a given test. Known as the **normative sample** this sample serves as a frame of reference for test score interpretation. Once the test has been normed, an average score, unusually high score, or unusually low scores can be determined. Thus, the scores of everyone who subsequently takes the test are compared to the norms. Norms are necessary, if the purpose of a test is to compare, performance of the test ticker to other test takers. The more people we use in our norm group, the close the approximation to a normal population distribution we get. For example if Mini scored 280 in an IQ test, how should we know how her score is; is the score good, bad or just average. For that we have to know how other children of Mini's age have performed. If we find that the performance of average children on this test is between 275 and 290 then we consider Mini's performance is average. If the norms tell us that above average children score between 275 and 290 on this test, then we would say Mini is above average on intelligence.

4. **Good psychological tests must be standardised:** The process of administering a test to representative sample of test takers for the purpose of establishing norms is standardisation (Cohen and Swerdlik). In other words, Standardisation means administering the test to

standardisation sample, for establishing uniform procedures, on administration and scoring of the test; so that everyone who takes the test does so in similar conditions. Standardisation Sample is a large sample of test takers who represent the population for whom the test is designed or intended. This allows the test developer to create a normal distribution which can be used for comparison of any specific future test score.

3.5 SUMMARY

1. There are various sources from where the authoritative information related to the test can be sought. These sources are Test catalogues, test manuals, reference volumes, test critiques, Tests in Print (TIP), Journal articles, online databases, and Test Reviews Online.
2. Various assumptions about psychological testing and assessment form a bottom line of assessment. These assumptions are:

a. Psychological traits and states exist - it is a basic assumption of testing that psychological traits exist and so do traits, although they cannot be observed they can be interpreted from overt behavior.

b. Psychological traits and states can be quantified and measured.

c. Test related behavior predicts non - test related behavior The goal of testing usually is to predict behavior other than the exact behaviors required while the exam is being taken.

d. Tests and other measurement techniques have strengths and weaknesses.

e. Various sources of error are a part of assessment process Errors can be attributed to test construction, administration, scoring and interpretation, or any other irrelevant factors in the environment. Errors are just an element of the measurement process and are taken into account during the process of measurement.

f. Testing can be conducted in a fair and unbiased manner.

g. Testing and assessment benefit society - Tests are indispensible in today's world and are of immense value to society.

3. Criteria of a good test:

a) A good test must be reliable: The word reliability means the precision with which the test measures and the extent to which errors are present in measurement.

b) A good test must be valid: Validity concerns the extent to which the test and assessment procedures used in psychological and educational testing, measure what they purport to measure.

c) Does the test have established norms: Norms are obtained from an initial group of people, the representative sample who truly represent the larger population for whom the test is meant-, dn factors such as age, social stratification, gender, educational level, etc.

d) Good psychological tests must be standardised: .Standardisation means administering the test to standardisation sample, for establishing uniform procedures, on administration and scoring of the test; so that everyone who takes the test does so in similar conditions.

3.6 QUESTIONS

Answer the following questions:

1. What are the various reference sources for authoritative information about tests?

2. What are the various assumptions about psychological testing and assessment?

3. What are the criteria of good test?

3.7 REFERENCES

1. Cohen, R.J., & Swerdlik, M.E., (2030). Psychological testing and Assessment: An introduction to Tests and Measurement, (7th ed.), New York. McGraw - Hill International edition, 329 -332
2. Anastasi, A. & Urbina, S. (3997). Psychological Testing. (7th Ed.). Pearson Education, Indian reprint 2002.

3. Kaplan, R.M., & Saccuzzo, D.P. (2005) . Psychological Testing - Principles, Applications and Issues. (6th Ed.). Wadsworth Thomson Learning, Indian reprint 2007.

4

UNDERSTANDING NORMS, INFERENCE AND MEASUREMENT

Unit Structure

4.0 OBJECTIVES

After studying this unit you should be able to:

1 . Understand the concept of norms, sampling and type of sampling.
2. Know the various types of norms.
3. Discuss fixed reference group scoring systems as well as norm referenced and criterion referenced evaluation.
4. Understand the concept of meta analysis and the relationship between culture and reference.

4.1 INTRODUCTION

In this unit we will discuss the concept of norms, sampling and the different types of sampling. Following this we will discuss how norms are developed and the different types of norms. Some important types of norms that we will discuss are percentiles, age norms, grade norms, national norms, national anchor norms, subgroup norms and local norms. We will also discuss fixed reference group scoring systems, norm referenced and criterion referenced evaluation. Towards the end of the unit we will discuss the concept of meta analysis as well as relationship between culture and inference.

4.2 NORMS - SAMPLING TO DEVELOP NORMS, TYPES OF NORMS, FIXED REFERENCE GROUP SCORING SYSTEM, NORM REFERENCED VERSUS CRITERIA REFERENCED EVALUATION

4.2.1What are norms?

> "In a psychometric context, norms are the test performance data of a particular group of test takers that are designed for use as a reference for evaluating or interpreting individual test scores". (Cohen & Swerdlik, 2002, p. 400).

Norms can be described as the average scores among an identified group of people which provide a basis at which test scores of individuals can be compared. In other words, it is the group's typical performance on the test scores measuring a particular characteristic. Before establishing norms we need to look at standardisation of the test. Standardisation refers to the process of administering a test to a representative sample of test takers for the purpose of establishing norms.

The norms yield a distribution of scores which can be then compared to evaluate new set of scores. This process of deriving norms is referred to as Norming. In order to establish norms, tests are administered to a large population that is selected carefully in order to represent the population for whom the test is designed. Norms can be derived on the basis of gender, grades, age, percentiles, local or even national norms. A normative sample is that group of people whose performance on a particular test is analysed for reference, for evaluating or interpreting individual scores.

4.2.2What is Sampling and how is this sampling done?

The process of selecting the portion of universe deemed to be representative of the whole population is referred to as sampling (Cohen and Swerdlik). The distribution of test responses is acquired by administering the test to the sample population. When the test is being developed the test developer decides the target group for whom he is going to design the test. Only on the basis of the test target group can he decide the sample group. Let's understand this with the help of a case study

Case study: A developer is in the process of developing a 'Competition Stress Test'. He has to identify who would be the target population i.e., whether the target group is going to be students, or people appearing for competitive examinations, or newly appointed recruits in management firms or any other target populace. Let's presume that the target population identified by the test developer is the students. He then has to identify whether they would be secondary students or college going students' or students in

professional courses. Depending on the target population the test developer then has to choose the sample. If the developer has chosen students in the professional courses, he has to then identify the likely subgroups amongst this population. So they could have students in professional courses belonging to different economic conditions, gender, grade, years of professional training, attending tutoring classes, etc. A small number of people belonging to each of these subgroups are then selected to represent the larger population. This is called sampling.

4.2.3 Types of sampling: Sampling can be done in many ways. There are no strict rules to follow, and the researcher must rely on logic and judgment. A small, but carefully chosen sample can be used to represent the population.

Sampling methods are classified as either probability or nonprobability.

(a) Probability methods include random sampling, systematic sampling, and stratified sampling.

(b) In non-probability sampling, members are selected from the population in some non random manner. These include convenience sampling, judgment sampling, quota sampling, and snowball sampling. The advantage of probability sampling is that **sampling error** can be calculated. Sampling error is the degree to which a sample might differ from the population. When inferring to the population, results are reported plus or minus the sampling error. In non-probability sampling, the degree to which the sample differs from the population remains unknown.

Random sampling is the purest form of probability sampling. Each member of the population has an equal and known chance of being selected. When there are very large populations, it is often difficult or impossible to identify every member of the population, which may result in sampling error.

Stratified sampling is commonly used because it reduces sampling error. Individuals from each subset or stratum within the population are selected in the sample. A stratum is a subset of the population that shares at least one common characteristic. Test developer takes into account all demographic variables such as age, gender, socioeconomic status, geographic region which can accurately describe the population of 'interest and then selects individual at random, but proportional to the demographic portrait of the test population. Random sampling is then used to select a sufficient number of subjects from each stratum. "Sufficient" refers to a sample size large enough for us to be reasonably confident that the stratum represents the population.

Convenience sampling or incidental sampling is used in exploratory research because they are convenient and an inexpensive method to collect

data. This non-probability method is often used during preliminary research efforts to get a gross estimate of the results without incurring the cost or time required to select a random sample.

J**udgment sampling or purposive sampling** is a common non-probability method. The researcher selects the sample based on judgment. This is usually and an extension of convenience sampling. When using this method, the researcher must be confident that the chosen sample is truly representative of the entire population.

Quota sampling is the non-probability equivalent of stratified sampling. Like stratified sampling, the researcher first identifies the stratum and their proportions as they are represented in the population. Then convenience or judgment sampling is used to select the required number of subjects from each stratum. This differs from stratified sampling, where the straturns are filled by random sampling.

Snowball sampling is a special non-probability method used when the desired sample characteristic is rare. It may be extremely difficult or cost prohibitive to locate respondents in these situations. Snowball sampling relies on referrals from initial subjects to generate additional subjects. While this technique can dramatically lower search costs, it 'comes at the expense of introducing bias because the technique itself reduces the likelihood that the sample will represent a good cross section from the population.

Cluster Sampling begins by dividing a geographic region into blocks and then randomly sampling within those blocks.

4.2.4 How are the norms developed for a standardized test? What are the various types of norms?

After obtaining the sample the test developer sets standard instructions regarding the way the test is to be administered, including the set of instructions to the test taker, the setting for giving the test., etc. This process of administering a test to a representative sample of test takers for the purpose of establishing norms is standardisation. This is done to make the normative sample more comparable with future test takers.

The next task of the test developer is to describe the data using the descriptive statistics such as measures of central tendency. The test developer then has to provide a description of the standardisation sample itself. New norms are developed for specific group of test takers on the basis of earlier standardisation.

4.2.5 What are the various types of Norms

1. Percentiles - Percentiles are probably the most commonly used type of norm that indicates the rank of the student compared to others (same

age or same grade), using a hypothetical group of 400 students. It is the percentage of people whose score on a test or measure falls below a particular raw score. In terms of percentile rank norms, scores may range anywhere between the 4st percentile and the 9th percentile with the average score being set among the 50th percentile. A percentile of 40, for example, indicates that the student's test performance equals or exceeds 40 out of 400. students on the same measure. Note that "percent" and percentile are not the same. For example a percentile of 65 does not indicate that the student has answered 65% of answers correctly. It only indicates his relative position / rank on a group. Percentiles are derived from raw scores using the norms obtained from testing a large population when the test was first developed. The major psychological measurement is that all variables of psychological interest are normally distributed. Since these variables fall into a normal distribution, we can specify what proportion of the population falls at or below any score on a particular test. The average value is the midpoint of the distribution and has a percentile rank of 50%. However, the problem with using percentiles is that in a normally distributed sample the real difference between raw scores may be minimised near the end of the distribution and exaggerated towards the end of the distribution.

2. Age norms - "indicate the average performance of different samples of test takers who were at various ages at the time the test was administered" (Cohen & Swerdlik, 2005, p. 407). Age norms for height, weight are widely accepted. Another concept closely related to the age norms is the concept of mental age. Mental age is expressed as the chronological age for which a given level of performance is average or typical. An individual's mental age is divided by his chronological age. Thus, a subject whose mental and chronological ages are identical has an IQ of 400, or average intelligence. The concept of mental age is criticised on the basis that it is too broad and other factors such as psychological development, social development may not reflect.

3. Grade norms - Sometimes educators are interested how students performed relative to other students in the same grade. Some tests provide age or grade equivalent scores. Such scores indicate that the student has attained the same score (not skills) as an average student of that age or grade. Age/grade scores seem to be easy to understand but are often misunderstood, and many educators discourage their use.

4. National Norms - They are derived from a standardisation sample nationally representative of the population of interest. For example, Indian norms may be separately developed based on age, gender, ethnic backgrounds, socioeconomic strata, etc. specific to and relevant to Indian conditions.

5. National Anchor Norms - National anchor norms are a tool for comparison in which two tests measuring a particular ability are normed by using the

same sample (i.e., each member of the sample took both tests). When two tests are normed from same sample it is called co-norming.

6. Subgroup Norms - Subgroup norms are created when narrowly defined groups are sampled. These subgroups may be based on socioeconomic status, Handedness, Education level, Age, etc.

7. Local Norms - Local norms are derived from the local population's performance on a particular measure. Typically created locally by guidance counselor or school guidance unit they provide normative information with respect to local population's performance on some test. International or national norms may not be relevant to the local populace due to various reasons such as extreme variations in socio economic factors or level of education, exposure to certain cultural or political factors. For these purposes local norms have to be developed rather than using the group norm supplied with the test.

4.2.6 Fixed Reference Group Scoring Systems - A system of scoring wherein the distribution of scores obtained on the test from one group of test takers (the fixed reference group) is used as the basis for the calculation of test scores for future administrations (Cohen and Swerdlik). Fixed reference group is utilised to ensure comparability and continuity of scores, without providing normative evaluation of the performance. Local or other specific group norms are often used for this purpose (Anastasi 2006). The SAT and the GRE are scored using the Fixed Reference Group Scoring.

4.2-7 Norm-referenced versus Criterion-Referenced Evaluation

1. Norm-referenced testing and assessment

Tests that set goals for students based on the average student's performance are norm-referenced tests. They are a method of evaluation and a way of deriving meaning from test scores by evaluating an individual test taker's score and comparing it to scores of a group of test takers.

In other words norm referenced tests consider the individual's score relative to the scores of test takers in the normative sample.

The problem with norm reference test is that they are likely to induce competition and may lead to pressurizing young children to perform better. Besides it cannot, measure progress of the population as a whole.

The norm referenced testing is advantageous since students and teachers alike know what to expect from the test and just how the test will be conducted and graded. This also makes these assessments fairly accurate as far as results are concerned, a major advantage for a test.

2. Criterion -Referenced Evaluation

Criterion referenced testing may be defined as "a method of evaluation and a way of deriving meaning from test scores by evaluating an individual's scores with reference to certain standards." Cohen and Swerdlik. 2005 pg. 440

Criterion Referenced tests consider the individual's score relative to a specified standard or criterion (cut score). In other words, a predetermined level of acceptable performance is developed and students pass or fail in achieving or not achieving this level.It uses interpretive frame of reference, a specified content domain rather than a specified population of persons. In contrast the norm referenced testing; the individual's score is interpreted by comparing it with the scores obtained by others on the same test. In criterion reference testing a test takers performance may be reported in terms of specific kinds of arithmetic operations, her estimated size of vocabulary or any other performance. It describes the specific skills, tasks or knowledge that the test taker can demonstrate. Let's take an example of mathematical skills, the results of this test might demonstrate that a particular individual can add, subtract but has difficulty with multiplication. The individual in this case is not compared to others of his age but an individualised program focusing on division will be designed. Thus, criterion referenced testing movement emphasises the diagnostic use of tests.

4.3 INFERENCE FROM MEASUREMENT - META ANALYSIS: CULTURE AND INFERENCE

> "The statistical analysis of a large collection of analysis results for the purpose of integrating the findings." (Glass, 4976)

4.3.1 Meta-analysis is a statistical technique for merging, summarising, and reviewing previous quantitative research. It combines the results (statistical information) of several studies that address a set of related research hypotheses. In other words metaanalysis provides a systematic overview of quantitative research which has examined a particular question. Selected parts of the reported results of primary studies are meta-analysed, in similar ways as other data - descriptively and then inferentially to test certain hypotheses. By using meta-analysis, a wide variety of questions can be investigated. It can be used as a guide to answer the question what difference does variable X have on variable Y?', even if Y has been measured using different instruments across a range of different people.

The best part of meta analysis is that it in effect combines all the research on one topic into one large study with many participants. It has been used to give helpful insight into the overall effectiveness

of psychotherapeutic interventions, the relative impact of independent variables and the strength of relationship between variables.

The danger is that in integrating a large set of different studies, the construct definitions can become imprecise and the results difficult to interpret meaningfully.

4.3.2 Culture and Inference: The process of psychological testing has its own limitations and one of the limitations comes from the fact that individuals vary not only because of their inherent differences but also because of the impact that culture has over them. It is extremely difficult to develop a test that measures innate intelligence without introducing cultural bias; this has been virtually impossible to achieve.

Cultural relevance becomes more apparent while measuring variables such as conformity vs. non conformity for example in collectivist vs. an individualist culture; it would pose a separate set of problems. What is interpreted as 'good' or 'bad' will be determined by the context in which it is perceived.

Besides the cultural context, the 'times' in which the testing is taking place is also important. For example if Indian children were to be tested on general awareness of technology now and say 20 years back; there are bound to be marked differences. The world is changing, the technology is changing, people are getting more tech savvy and thanks to internet they are having an easy access to technology. Rogler (2002) has brought much needed attention to the relevance of historical context in testing.

Many attempts have been and are made to reduce the cultural impact. One attempt was to eliminate language and design tests with demonstrations and pictures. Another approach is to realize that culture-free tests are not possible and to design culture-fair tests instead. These tests draw on experiences found in many cultures.

Besides this, there are a couple of things which the test developers have to consider such as the cultural assumptions on which the test is based, consultations with the group minorities or cultural communities regarding the appropriateness of a particular assessment procedure. The test developers also have to be knowledgeable about the many alternative tests and be aware of equivalence issues across cultures including equivalence of language used and the constructs measured.

4.4 SUMMARY

1 . Norms are a group's typical performance on the test scores measuring a particular characteristic. The norms yield a distribution of scores which can be then compared to evaluate new set of scores.

2. In order to establish norms, tests are administered to a large population that is selected carefully in order to represent the population for whom the test is designed.

3. **Standardisation** refers to the process of administering a test to a representative sample of test takers for the purpose of establishing norms.

4. Norms can be derived on the basis of gender, grades, age, percentiles, local or even national norms. A normative sample is that group of people whose performance on a particular test is analysed for reference for evaluating or interpreting individual scores.

5. A small number of people belonging to each of the subgroups are selected to represent the larger population. This is called sampling.

6. Sampling methods include random sampling, systematic sampling, and stratified sampling. Sampling error is the degree to which a sample might differ from the population. When inferring to the population, results are reported plus or minus the sampling error.

7. The various types of norms include percentile norms, age grade norms, national norms, national anchor norms, subgroup norms and local norms.

8. **Fixed Reference Group Scoring** Systems is a system of scoring wherein the distribution of scores obtained on the test from one group of test takers is used as the basis for the calculation of test scores for future test takers.

9. **Norm-referenced testing** considers the individual's score relative to the scores of test takers in the normative sample.

10.Criterion Referenced tests consider the individual's score relative to a specified standard or criterion.

11. Meta-analysis is a statistical technique for merging, summarising, and reviewing previous quantitative research. It combines the results (statistical information) of several studies that address a set of related research hypotheses.

12. One of the limitations of psychological testing is because of the impact that culture has over them. Many attempts have been and are made to reduce the cultural impact.

4.5 QUESTIONS

Answer the following questions:

1. What are norms? What are the various types of norms?

2. What is Meta analysis? How does culture influence inference?

3. Write notes on a. fixed reference group scoring system, b. norm referenced c. criteria referenced evaluation

4.6 REFERENCES

1. Cohen, R.J., & Swerdlik, M.E., (2040). Psychological testing and Assessment: An introduction to Tests and Measurement, (7 th ed.), New York. McGraw - Hill International edition, 429 -432

2. Anastasi, A. & Urbina, S. (4997). Psychological Testing. (7th ed.). Pearson Education, Indian reprint 2002.

3. Kaplan, R.M., & Saccuzzo, D.P. (2005) . Psychological Testing Principles, Applications and Issues. (6th ed.). Wadsworth Thomson Learning, Indian reprint 2007.

5

RELIABILITY

Unit Structure

5.0 OBJECTIVES

After studying this unit you should be able to:

1 . Define reliability and understand the concept of reliability.
2. Know the goals of estimating reliability and understand the concepts of true variance and error variance.
3. Comprehend the various sources of error variance.

5.1 INTRODUCTION

In this unit we will discuss the definition and concept of reliability which is one of the most important characteristic of good test. Following this we will discuss the goals of estimating reliability and understand the concepts of true variance and error variance. We Will then discuss the various sources of error variance. We would end this unit with a brief summary, questions and a list of references for further reading.

5.2 DEFINITION AND CONCEPT OF RELIABILITY

1 The word reliability means dependability or consistency. Reliability is consistency of measurement or scores.

2 Reliability is the extent to which a test is repeatable and yields consistent scores.

In everyday life we use the term reliability, often on a positive note. For example you may refer to train being reliable or on time every day. You may also say that the blood pressure (B.P.) apparatus (machine) you purchased is very reliable. It simply means that the BP apparatus is giving consistent results.

Imagine if the BP apparatus you brought, gave drastically different readings each time. On first attempt if it shows, high blood pressure, on second attempt it shows, low pressure and on third attempt, it shows normal. Which of the readings would be true? You would be confused and most probably, conclude that the machine you brought is not a reliable machine.

Now you see that the reliability is indeed important to trust a machine or even a psychological test.

A test is considered reliable if we get approximately same result repeatedly. Constancy of Individual's scores on one test and his scores on the same test when retested (after a time interval) or his scores on an equivalent test will tell us about the reliability of the test.

Now let's take the BP apparatus example a little further.

Situation I: after testing for three consecutive sessions we get the following results:

1st attempt: 70 / I 10
5nd attempt: I 10 / 180
3rd attempt: 50 / 100

Now which of the results reflect your true blood pressure? Can't say? In such cases you would take another machine which is reliable and then check your blood pressure. The other machine reads 70 / I 10 as your blood pressure. Now this reading was also taken by your machine in its first ' attempt. So is your earlier machine reliable? No, since it did not give the same readings after retesting.

Situation II: After testing for three consecutive sessions on the BP apparatus you get approximately similar results. Do you conclude that the BP apparatus you purchased perfectly measures your BP ... ?Think again. The BP

apparatus you brought measures something (it may or may not be BP) consistently. May be, it is reliably measuring your pulse and heart rate and not your blood pressure. Remember reliability does not mean that the machine measures what it is supposed to measure. It only means it is consistently measuring something. For finding out whether the machine is measuring the same thing, that it is made to measure, we have to check its validity. We shall study validity in the later chapter.

> Note: In order to be valid, a test must be reliable; but reliability does not guarantee validity.
> Reliability can be determined by retesting on the same test or equivalent test (alternate form) or under variable examining conditions.

5.3 THE GOALS OF ESTIMATING RELIABILITY UNDERSTANDING THE CONCEPT OF TRUE VARIANCE AND ERROR VARIANCE

> 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.

Let's study the following example to understand the concept of true variance and error variance.

Case study. Lavanya scored 55 out of 50 in her mathematical ability test. When the same test is re-administered and if she scores 53 or 56 out of 50 then we can say the test is reliable. However, if she scores 5 out of 50 or 50 out of 50 then it would mean that the test given to her was not reliable since her scores are not consistent. But before coming to that conclusion we have to look into two important aspects. Let's examine the above given example conditions.

Condition I : *Lavanya scores 50 out of 50-* Let's assume Lavanya undergoes some training in mathematics and scores 50 out of 50 in the re-examination then we can say that her scores truly reflect on her enhanced mathematical ability. This is called true variance. In other words variance from true differences is true variance. Remember when we are talking about Variation we are referring to change between the first score and the second score on the same or alternate test score of the same person .

Condition II: *Lavanya scores 5 out of 50* - Lavanya may have scored less in retesting because of testing conditions (like uncomfortable sitting arrangement, inadequate lighting or mood or simple boredom- which we call irrelevant factor). -This, variation in score due to factors other than ability (true change) is called as 'Error Variance' (variation due to error or irrelevant factor in the environment). In other words the score of Lavanya's

mathematical ability does not reflect her true ability but is the variance caused due to irrelevant or random factors of the environment.

All measurement procedures have the potential for error, so the aim is to minimize it.

In broader sense the term reliability indicates the extent to which individual differences in test score are attributable (credited to) true differences and the extent to which they are attributable to chance factors or errors. An observed test score is made up of the true score plus measurement error.

Thus reliability is expressed as

X = T + E; where; X - the score of an individual on a test; T - true variance; E - error variance

5.4 SOURCES OF ERROR VARIANCE

Errors can be made while constructing (making) the test or while administering the test, scoring the test or even while interpreting the results of the test. Let's understand each of them in detail.

5.4.1 Errors during test construction: Psychological tests measure psychological variables like personality attributes (dominance, aggressiveness, etc.), some specific skill, or body of knowledge. Since they measure complex traits which are abstract, there are no rigid yardsticks available to measure these variables. Thus, test construction is difficult and errors in test construction are very likely.

- Some of these variables may be abstract in nature. For example, psychologists want to make a test on 'goodness'. Now goodness is an abstract concept, which is difficult to be defined. Psychologists who want to construct a test on 'goodness' may vary according to their subjective interpretation of goodness.

- Similar tests may be differently worded so even similar questions may be open to different interpretations.

- Some of the items (questions) in the test may cover one particular aspect of the construct under measurement than other. For example a psychologist designing a personality inventory may tend to add more items related to dominance than intuitiveness; though both intuitiveness and dominance belong to personality, just adding a few more items of dominance may change the total score on personality.

5.4.2 Errors during Test administration: Errors during test administration may include untoward influences during administration or test taker* variables and examiner related variables. (*Remember a Test taker refers to an individual who is answering the test. Test giver or the examiner is an individual, usually a psychologist, who is administering and evaluating the test).

_ Test environment related factors: Environment related factors such as levels of lighting in the testing room, noise, uncomfortable sitting or even a broken pencil could contaminate the testing environment.

_ Test taker related variables: Pressing emotional problems, physical discomfort, illness, lack of sleep, mood deterioration, or even a sleep inducing drug can alter scores and thus are a source of error variance.

_ Examiner related variables: Examiner's presence or absence, examiner's unwitting cues, departure from the procedure prescribed for the test, etc., can be a cause of error variance. Besides examiner's body language can provide cues which may provide information about correctness of a response. For example, an examiner may unwittingly nod at a correct answer which may provide a cue to the test taker. In such cases the test taker would alter his responses according to the cues provided.

5.4.3Errors in test scoring and test interpretation:

- Most of the tests used in India like Non - Verbal Test of Intelligence (NVTI), Differential Aptitude Test (DAT), 16 PF, are paper - pencil tests and involve hand scoring by trained personnel. Scoring grids may not be available for all the tests and even if they are, they may sometimes not be properly placed on the answer sheet. This may give incorrect score.
- Computerised testing and evaluation is available for most of these tests. However, computers may not be available in all testing centres and thus they have to be administered and score manually with the help of a trained professional. Manual scoring and interpretation increases vulnerability to human errors like, error in scoring, error in calculating score, or even interpreting the score.
- Some tests like inkblot test, complete the sentence test, make a story tests or even tests for creativity are highly subjective in nature. In such cases the examiner has to quantify or qualitatively evaluate responses. This may itself be a source of error variance.

5.4.4Systematic and Unsystematic errors:

- Suppose we want to assess the quality of relationship between a couple, or a parent- child relationship to study abuse we have to depend on interpretations of people involved to quantify or make qualitative observations about the incidence of abuse. This is a source of systematic error.

- While reporting incidences, non-systematic errors like forgetting, failing to notice abusive behaviour, misunderstanding questions or over reporting / under reporting abuse for different reasons, may occur.

5.5 SUMMARY

1 . The word reliability means dependability or consistency. Reliability is consistency of measurement or scores. It is the extent to which a test is repeatable and yields consistent scores.

2. A test is considered reliable if we get approximately same result repeatedly. Constancy of Individual's scores on one test and his scores on the same test when retested (after a time interval) or his scores on an equivalent test will tell us about the reliability of the test.

3. 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. In other words, reliability indicates the extent to which individual differences in test score are attributable (credited to) true differences and the extent to which they are attributable to chance factors or errors. Variance from true differences is called True variance. **True variance** could be as a result of increased ability due to practice or training. This variation in score due to factors other than training (true change) is called as **'Error Variance'** (variation due to error or irrelevant factor in the environment).

4. Some of the sources of error variances are errors during test construction, errors during test administration, errors during test scoring and interpretation and other related systematic and unsystematic errors.

5.6 QUESTIONS

Answer the following questions:

Q1. Define and explain the concept of reliability.

Q2. Explain the concepts of true variance and error variance.

Q3. Discuss the various sources of error variance.

5.7 REFERENCES

1. Cohen, R.J., & Swerdlik, M.E., (5010). Psychological testing and Assessment: An introduction to Tests and Measurement, (7th ed.), New York. McGraw - Hill International edition, 159 -135

2. Anastasi, A. & Urbina, S. (1997). Psychological Testing. (7th ed.). Pearson Education, Indian reprint 5005.

3. Kaplan, R.M., & Saccuzzo, D.P. (5005) . Psychological Testing - Principles, Applications and Issues. (6th ed.). Wadsworth Thomson Learning, Indian reprint 5007.

6

ESTIMATING AND INTERPRETING RELIABILITY

Unit Structure

6.0 OBJECTIVES

After studying this unit you should be able to:

1. Know the various methods of estimating reliability. 2. Understand the concept of coefficient of reliability.

6.1 INTRODUCTION

In this unit we will discuss the various methods of estimating reliability. The most common methods of estimating reliability include Test retest reliability, Alternate or parallel form of reliability, Split half reliability, internal consistency and inter scorer reliability. Two methods of determining internal consistency include.

a) Kuder Richardson Formula and
b) Coefficient of alpha

Following this we will discuss how to use ailed interpret a coefficient of reliability.

Towards the end of this unit we will summarise the major

points discussed in this unit followed by questions and a list of references for further reading.

6.2 METHODS OF ESTIMATING RELIABILITY

In the earlier unit we learned about what is reliability, its goals, sources of variance and the sources of error variance. In this unit we shall understand the various ways of estimating reliability. Let's look at the case study below to understand these ways.

Case study: Researcher A has constructed a mathematical ability test and now wants to figure out whether the test he had constructed is reliable. Let's assist Researcher A (call him R.A. for short) to determine reliability of the test.

6.2.1Test-Retest Reliability:

The most obvious way R.A. can determine whether his test is reliable and to what extent, is by re-administering the same test on the same individual.

Test taken --- time interval ---- test re-administered

Scores on test at first administration ----- compared with -----scores on second administration of the same test.

1. To determine test-retest reliability, the test is administered twice at two different points in time.
2. This kind of reliability is used to assess the consistency of a test across time.
3. Test - retest reliability assumes that there will be no change in the quality or construct being measured.
4. It is best used for things that are stable over time, such as personality factors, reaction time, perceptual judgment, etc.
5. Source of error variance:

i. Learning of new skills or techniques or shortcut methods to solve the problems.

Test administration I --- training ---- test administration II

Score on I = 65 Score on 11 = 28

ii. Passage of time: With passage of considerable time the reliability coefficient will be low. The test taker may undergo developmental changes, trauma or other related factors, that would influence the scores on 2nd administration.

6.2.2 Alternate form and Parallel-Forms Reliability

Another way R.A. can determine the reliability of his test is by creating another test. Let's call his first test Form I and second test Form II. Look what he can do

Item no.	Form I	Form II
6	2 + 2	3+6
2	4 X 3	3 X 5

3	60 / 5	62 / 6
4	25 : 60	35: 60
5	20 - 7	65 - 6

If you observe' the items in Form I and Form 11 are not the same, but similar. When question on 6 digit addition was asked on form 1, 6 digit addition problem was set in Form 11. Observe that both the sets have equal number of addition, multiplication, division, subtraction and ratio related questions. (*The above example is only indicative and does not reflect a true alternate form or parallel form.)

1 . Both Alternate form and Parallel-forms reliability is estimated by comparing two different tests which were created using the same content.
2. Parallel / Alternate form reliability compares two equivalent forms of tests that measure the same attribute.
3. To create alternate or parallel form for a test a large pool of test items is created. A pool of items means a large number of similar questions are created. These items must measure the same quality.
4. The items are then randomly divided into two separate tests.
5. The two tests are then administered to the same subjects at the same time.
6. Administration of Form I ---- Administration of Form II
7. The degree of relationship between the two tests is determined by alternate form or parallel form coefficient of reliability. Scores of Form I ---- correlated with ---- scores of Form II
8. Alternate / Parallel forms are expensive and time consuming. However, they have been advantageous because they minimise the effect of memory for the test content.
9. Source of error variance:

i. Test takers may do better in either form because of the nature of items selected in that form.
ii. The scores of the two forms may be affected by factors like motivation of the test taker, practice - fatigue effect, etc.

What is the difference between alternate forms and parallel forms?

Two (set of) tests are said to **be parallel forms when the means and the variances of the scores for each form are equal**. More technically the means of the each of the tests correlate equally with the true score. While **alternate forms** are simply different versions of the same test, they are designed to meet the requirements of equivalence between the two sets in content and level of difficulty. They may not have equal means and variances or correlate equally with the true score.

6.2.3 Internal Consistency Reliability

The disadvantage with alternate form or parallel form is that it is time consuming and a difficult process. Thus, if we want to establish reliability without developing an alternate form we can do so by measuring internal consistency.

1. This form of reliability is used to judge the consistency of results across items on the same test.
2. When you estimate reliability of the test, by comparing test items within the same test, it is called **internal consistency estimate of reliability**. or estimate of inter item consistency. These items should measure the same construct (attribute or trait).
3. Internal consistency not only saves us the effort of developing an alternate form but also saves us the effort of re administration.
4. There are different methods of obtaining estimates of inter item consistency like Split half method, the Kuder Richardson formula, the Coefficient alpha by Cronbach.

6.2.4 Split half reliability

To avoid constructing two equivalent tests, which is, very time consuming process our R.A. (Researcher A) constructs one single test. However, while calculating reliability he splits this single test into two halves and then correlates these two halves of the test. Usually he calculates the correlation between the two halves by using the Pearson r and then adjusts the half test reliability by using the Spearman -Brown formula.

How to split the test?

Any of the below mentioned ways can be used to split the test.

1 . Randomly assign each item to one of the tests. Ensure that there are equal numbers of items in each halves.
2. Assign odd number items to first half and even number items to the second half or vice a versa. This is also known as oddeven reliability.
3. Split the test in such a way so that the items in each of the two halves are equivalent with respect to content and difficulty.
4. Never split the test in the middle because the two halves may not be equivalent in difficulty and content and that may result in spuriously high or low reliability coefficient.

Method:

Step I: Construct test ----- administer test

Step II: Split the test into two halves. ensuring both the halves are equivalent in number of items, - difficulty and content

Step III:Apply Pearson 'r' formula $r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$.

Step IV: Correct the coefficient using Spearman - Brown formula

$$rsB = \frac{nr_{xy}}{1+(n-1)r_{xy}}$$

where; rs_B = the reliability adjusted by Spearman - Brown formula.

fxy= Pearson's r in the original length test.

n= number of items in the revised version / number of items in the original version.

The why of the Spearman Brown formula?.

In the split half reliability we divide the whole test into two halves. However, reducing the length of the test reduces its reliability. For example, if our test consists of 30 items, while calculating the split half reliability we reduce the length of the test to 65 items only. Though not always a rule, it has been found, longer the test, greater its reliability. While calculating the reliability coefficient of the split half reliability, we do so by first calculating it by using Pearson's formula for the split halves. Then we use the spearman brown formula to calculate the coefficient of reliability for the whole test. This formula can be used even when the test is lengthened. The spearman brown formula estimates the effect of lengthening or shortening of the test on coefficient.

6.2.5 Estimating internal consistency by other methods

The problem with split half reliability as we discussed earlier is that we reduce the test by half. However, reducing the length of the test reduces the reliability of the test. Instead we can correlate each item to the other and determine its internal consistency.

> Inter item consistency refers to the degree of correlation between all items of the scale.

Like split ha-If this too requires only one administration. The index of inter item consistency is useful in determining the test homogeneity i.e., More homogenous the test, higher the index of consistency.

> Homogeneity refers to the degree to which a test measures a particular trait or single factor.

Test homogeneity is necessary because it allows precise interpretation about a specific trait under consideration in that test.

Test homogeneity is desirable since it allows clear-cut interpretation of the test. However, one word of caution, extremely high index would mean the items within the test are measuring the same thing. So the rule, higher the better may not always hold true.

Test homogeneity is however, insufficient tool for measuring comprehensive psychological variables such as personality, intelligence, etc., since these variables are multifaceted i.e., they include many traits within them.

For example, the variable Personality may contain various traits such as dominance, initiative, extraversion, etc. All these traits represent different aspects of personality. Thus, we cannot expect the test to be homogenous in nature. The tests which measure such multifaceted variables are heterogeneous in nature.

A heterogeneous test measures more than one trait. In other words heterogeneity refers to the extent to which a particular 'test measures various factors or traits.

In order to overcome this problem we can subdivide heterogeneous tests into subtests which are of homogeneous nature. In other words, a number of homogeneous tests are administered which measure a trait of the heterogeneous variable.

Case study. Ira is developing a personality test. She has decided to include eight traits like submissiveness, aggression, extraversion, initiative, purposefulness, neuroticism ' masculinity / feminity, risk proneness, in her personality test. She has included 60 items (questions) to measure each trait. Thus, there are a total of 80 items measuring 8 different types of traits. Note, she cannot use inter item consistency between 80 items, but she can measure consistency within each of the ten items from each unit (6 unit measuring a single trait like submissiveness; thus forming 8 units of homogeneous test).

There are two formulas used for calculating inter item consistency. They are the Kuder - Richardson formula and the Cronbach formula. Let's study each of them.

6.2.6 Estimating internal consistency by Kuder Richardson formula

KR20 was deveoped by G. Frederic Kuder and M.W. Richardson to estimate reliability. It is used to determine inter item consistency of dichotomous items such as 'Yes' or 'No', 'Right' or 'Wrong', etc. If the test items are heterogeneous KR20 will show lower reliability estimates than a split half method, while if the items in the test are highly homogeneous the reliability score of Split half and KR20 will be similar. The Kuder-Richardson formula is as follows:

$$w\left(\frac{k}{k-1}\right)\left(1-\frac{\sum pq}{\sigma 2}\right)$$

$$r_{KR20}= \qquad \left(\frac{k}{k-1}\right)\left(1-\frac{\sum pq}{\sigma 2}\right)$$

KR20 = Kuder - Richardson formula 20; k = number of test items; 6 = variance of total test scores; p= proportion of test takers

ho pass the test item
q= proportion of people who fail test item;
pq= sum of the pq over all items

6.2.7 Estimating internal consistency by calculating Coefficient alpha

Numerous modifications have been suggested in Kuder Richardson formula largely variants of the original formula developed by Kuder and Richardson. One such formula was suggested by Cronbach known as the coefficient alpha. The Cronbach's coefficient alpha formula is as follows:

$$r_{\alpha}= \left(\frac{k}{k-1}\right)\left(1-\frac{\sum \sigma_i^{\,2}}{\sigma 2}\right)$$

where; r_a= coefficiert alpha; k = number of items; 6^2_i = the variance of one item;

6^2= variance of the total test scores

Coefficient alpha is used on tests with non dichotomous items. This formula is an estimate of mean of all possible test-retest, split half coefficients. This formula is widely preferred and used partly because it needs only one administration. Coefficient alpha is calculated to figure out how similar the sets of data are. Similarity is gauged on a scale of 0 - 6, wherein '0' means not at all similar and V means perfectly identical.

Look at the example below. These items are from a hypothetical personality test. The test takers are supposed to select on alternative from the three choice answers (non dichotomous) viz. always, sometimes, never that are most applicable to them.

1. I enjoy parties
   (1) always (2) sometimes (3) never
2. I am very practical
   (1) always (2) sometimes (3) never
3. I love parties
   (1) always (2) sometimes (3) never

4. 1 am the life of parties

(1) always (2) sometimes (3) never

Study the above question carefully. Let's make a table of coefficients (hypothetical) between item 6 and other items

Correlation between item nos.	Hypothetical coefficient (as calculated by coefficient alpha)	Reason
1 and 1	1	It's the same question, thus, inter item consistency as measured by Coefficient alpha is 6 (indicating perfectly identical items)
1 and 2	0	The question arena is different. While one measures extroversion the other measures pragmatism. Thus, the alpha score would be '0' indicating heterogeneity
1 and 3	1	Both the items measure 'liking for parties' except they are just worded differently. Thus, the test taker who answers 'always' for item No. 6 will answer the same for item No. 3. This homogeneity between the two items may not be preferred as it indicates uselessness of this item.
1 and 4	.80	Both these items measure extroversion, however one measures liking for parties while other measures initiative at the parties.

6.2.8 Inter-rater Reliability / Inter-scorer reliability

Case study: Mrinalini, Richa and Saundarya are participants of a beauty pageant. All the three would be assessed by a number of raters / assessors, ultimately giving one of the three a title of Ms. India. But before giving them the title, the three women have to be assessed for creativity and integrity. Now the traits like creativity and integrity are prone to subjective interpretation. Thus, tile assessors have to know, how they can score the dimensions of creativity and integrity. Besides, their method of distributing marks has to be the same. To know if the scores given by them to the participants have been derived in a systematic and consistent way we can calculate Inter-scorer reliability.

1. Inter-scorer reliability is the degree of agreement between two or more judges (assessors / scorers) while judging performance on a test. In other words, it is a type of reliability which is assessed, by having two or more independent judges, score the test.

2. The scores are then compared to determine the consistency of the rater's estimates. One way to test inter-rater reliability is to have each rater assign each test item a score and then calculate the correlation between the two ratings to determine the level of inter-rater reliability.

3. Another means of testing inter-rater reliability is to have raters determine which category each observations falls into and then calculate the percentage of agreement between the raters.

4. Correlation Coefficient referred to as Coefficient of inter scorer reliability is calculated to know the consistency among raters in scoring of the test.

6.3 USING AND INTERPRETING A COEFFICIENT OF RELIABILITY

As we have already seen earlier that reliability can be estimated using either the test - retest reliability or alternate / parallel form reliability or by using internal / inter-item consistency. The method that we would use would largely depend on two important questions, such as what is the purpose of obtaining the reliability score - what do we mean to achieve by it? Secondly, how high the coefficient of reliability is expected?

1. Let's understand the **purpose of reliability coefficient** with the help of the case study given below:

Case study:
Situation 1: Madhavan is responsible for monitoring performance of workers over a period of time during the course of the job performance. Madhavan has developed a test to monitor the worker's job performance.

Situation 2: Madhavan is responsible for employing one of the many contenders for the supervisory post in firm B.

Let's examine the above given example conditions.

1 . When a specific test is Aesigned to measure a specific behavioral aspect / trait over a period of time then the test should be able to demonstrate reliability across time. So Madhavan would administer the test to workers at the entry point and then at different interval points. In this case Madhavan would expect to have an estimate of test - retest reliability.

2. In the second situation Madhavan is expected to administer the test only once i.e., at the entry point. In such a case Test - retest reliability will not help. In such a case Madhavan would like to have an estimate of internal consistency.

3. When the purpose of estimating reliability is to understand the various sources of error variances, relevant in this particular situation, then a number of reliability coefficients have to be calculated. As we have seen earlier in the methods of estimating reliability that each types of reliability method reveals a different source of error variance.

2. Now let's figure out how high the coefficient of reliability is expected.
Reliability is binding attribute in all the tests we propose to use. However, slight variations on higher or lower level can be well tolerated. If the test scores are of extremely important in nature, obviously the reliability coefficient has to be high. However, if the test does not have a huge significance or is accompanied by a series of other tests for the interpretation of a given psychological variable, then slightly lower coefficients can be acceptable.

High reliability may be required when:

1) Tests are used to make important decisions.
2) Individuals are sorted into many different categories based upon relatively small individual differences e.g., intelligence.

Lower reliability is acceptable when:

1) Tests are used for preliminary rather than final decisions.
2) Tests are used to sort people into a small number of groups based on gross individual differences e.g., height or sociability /extraversion.

6.4 SUMMARY

How do we estimate reliability?

1 There are different methods of estimating reliability. These methods include test - retest method, parallel or alternate form, inter - rater reliability or by determining internal consistency by split half method, Cronbach's coefficient alpha and Kuder Richardson formula.

2. The method that we would use would depend on what is the purpose of obtaining the reliability score and how high the coefficient of reliability is expected. The methods differ from each other on number of testing sessions required, availability of or possibility of making parallel or alternate form. Each of these methods has different sources of error and statistical procedures.

3. When a specific test is designed to measure a specific behavioral aspect / trait over a period of time then the test should be able to demonstrate reliability across time.

4. To determine **test-retest reliability**, the test is administered twice at two different points in time. This kind of reliability is used to assess the consistency of a test across time.

5. Source of error variance in test - retest reliability are learning of new skills or techniques or shortcut methods to solve the problems or even developmental changes.

6. **Alternate form and Parallel-Forms Reliability** is estimated by comparing two different tests which were created using the same content. It compares two equivalent forms of tests that measure the same attribute. The degree of relationship between the two tests is determined by alternate form or parallel form **coefficient of reliability.**

7. Tests are said to be **parallel forms** when the means and the variances of the scores for each form are equal. Alternate forms are simply different versions of the same test; they are designed to meet the requirements of equivalence between the two sets in content and level of difficulty.

8. Source of error variance include factors such as motivation of the test taker, practice - fatigue effect or even enhanced performance on one from than the other.

9. **Internal Consistency Reliability** is a form of reliability is used to judge the consistency of results across items on the same test. In other words, Inter item consistency refers to the degree of correlation between all items of the scale.

10. There are different methods of obtaining estimates of inter item consistency like Split half method, the Kuder Richardson formula, the Coefficient alpha by Cronbach.

11. **Split half reliability** involves splitting a single test into two halves and then correlating these two halves of the test by using the Pearson r and then adjusts the half test reliability by using the Spearman -Brown formula.

12. **Inter-scorer reliability** is the degree of agreement between two or more judges (assessors / scorers) while judging performance on a test. The scores are then compared to determine the consistency of the rater's estimates.

13. A tabular representation about the methods of reliability is given below:

Type of reliability	Sources of error variance	Statistical procedures used	Number of test form	Number of testing sessions	Brief description of the method
Test - retest	Administration	Pearson '*r*' Spearman rho	1	2	Test 1 – time interval – test 1
Alternate form	Test construction and administration	Pearson '*r*' Spearman rho	2	1 or 2	Test 1 –test 2
Internal consistency	Test construction	1. When **equivalent halves** use Pearson '*r*' + Spearman Brown correction 2. **Dichotomous items**. Use Pearson '*r*' + Kuder – Richardson formula correction 3. **Multipoint items.** Use Pearson '*r*' + Cronbach's coefficient alpha correction	1	1	Test 6 – calculation by Pearson's ' r' – correction by spearman Brown or KR20 or Coefficient alpha
Inter-scorer or inter - rater	Scoring and interpretation	Pearson '*r*' Spearman rho	1	1	Simultaneous evaluation by different scorers. Calculation of reliability between these scorers / rater's using the spearman's rho or Pearson's '*r*'

14. Formula chart

Pearson's 'r'	Spearman – Brown formula	Kuder Richardson formula	Cronbach's coefficient alpha
$r=\frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$	$r_{SB} = \frac{nr_{xy}}{1+(n-1)r_{xy}}$	$r_{KR20}=\left(\frac{k}{k-1}\right)\left(1-\frac{\sum pq}{\sigma 2}\right)$	$r_{\alpha}=\left(\frac{k}{k-1}\right)\left(1-\frac{\sum \sigma 2i}{\sigma 2}\right)$

15. When a specific test is designed to measure a specific behavioral aspect / trait over a period of time then we would expect to have an estimate of test - retest reliability.
16. When it is expected to administer the test only then we would expect to have an estimate of internal consistency.
17. When the purpose of estimating reliability is to understand the various sources of error variances, relevant in this particular situation, then a number of reliability coefficients have to be calculated.

6.5 QUESTIONS

Answer the following questions:

Q1. Explain the test - retest reliability in detail.

Q2. What do you understand by inter item consistency. What are the various ways to calculate inter item consistency?

Q3. What is inter - rater reliability. What are the sources Of error variance in inter rater reliability?

6.6 REFERENCES

1 . Cohen, R.J., & Swerdlik, M.E., (2060). Psychological testing and Assessment: An introduction to Tests and' Measurement, (7 th ed.), New York. McGraw - Hill International edition, 629 -632

2. Anastasi, A. & Urbina, S. (6997). Psychological Testing. (7th ed.). Pearson Education, Indian reprint 2002.

3. Kaplan, R.M., & Saccuzzo, D.P. (2005) . Psychological Testing - Principles, Applications and Issues. (6 th ed.). Wadsworth Thomson Learning, Indian reprint 2007.

7

THE NATURE OF TESTS, ALTERNATIVES TO TRUE SCORE MODEL AND RELIABILITY AND INDIVIDUAL SCORES

Unit Structure

7.0 OBJECTIVES

After studying This unit you should be able to:

1 . Understand the relationship between the nature of test and reliability.
2. Know the alternatives to true score model.
3. Understand the concept of standard Error of Measurement and Standard Error of Difference.

7.1 INTRODUCTION

In this unit we will attempt to understand how the nature of test influences its reliability. Following this we will explain the various alternatives to true score model. In this we will discuss the item response theory, the domain sampling model and genera lizability theory. The concept of standard error of measurement and standard error of difference would also be discussed. Towards the end of the unit a brief summary of the unit, questions and list of references would be covered.

7.2 THE NATURE OF TEST AND RELIABILITY

While calculating reliability coefficient it is important to consider the nature of test. Let's look at these considerations.

7.2.1Is the test homogeneous or heterogeneous:

- As we have already seen earlier the tests which are homogeneous in nature i.e., which measure a single factor are likely to have high internal consistency estimate than tests which are heterogeneous in nature.

- In case of homogenous tests internal consistency estimates are used, while in case heterogeneous tests test - retest reliability may be used.

7.2.2 Does the test measure dynamic characteristic or static characteristics:

- A dynamic characteristic is a trait that is sensitive to situational variables.
- For example trait like anxiety is a dynamic trait. If the test is measuring a characteristic like anxiety it is expected that the scores will change over a period of time due to variances in situational factors like change. in the nature of the stressor or cognitive factors like learning effective problem solving skills.
- In such cases internal consistency is measured.
- In case of variables that are more static in nature like intelligence, they do not change significantly; they are more or less constant. In such cases alternate form reliability may be used.

7.2.3 The range of test score is or is not restricted: while interpreting coefficient of reliability, we need to take restriction range or inflation range of the test into consideration. Let's take an example to understand this concept.

Usha has constructed an intelligence test for children from age group 5 - 12. As we know each year (i.e., from 5 through 10) is significantly different from each other. Obviously each of these years requires a different set of reliability values. If Usha does not have different set of reliability coefficient for different age groups then we can say that the correlation analysis is restricted because of the sampling procedure.

7.2.4 Is the test a speed test or a power test:

A speed test has items usually low in difficulty, but the number of items in the test are so many that it is extremely difficult, though not impossible for the test taker to complete the test in a given time frame. While calculating reliability of a speed test we need to have two administrations of the tests at

two different time intervals. In other words, the reliability estimates should be based from 2 different testing sessions, either using test - retest reliability or alternate form reliability or split half reliability (two halves would have to be given separately). While calculating reliability coefficient of a speed test we are measuring the 'consistency of the response speed'. Using the KR-20 formula to measure internal consistency on the speed test will give us an erroneously high coefficient.

A power test on the other hand, has enough time frame to complete the test, however, the difficulty level of the items is so high that it is impossible to get a perfect score or 100% on this test.

7.2.5 Is the test criterion referenced test or norm referenced: A criterion referenced test is intended to indicate where a test taker stands with respect to some criterion such as mastery in driving skills or any other educational objective. Scores on criterion referenced tests tends to be interpreted in pass / fail terms. The important issue for a test administrator is whether a certain criterion is achieved or not. Thus, the procedures to measure reliability are not appropriate.

Norm referenced tests on the other hand contains material that has been not been mastered in a hierarchical pattern.

7.3 ALTERNATIVES TO TRUE SCORE MODEL

In the earlier part we have assumed that the total score consists of true score and an error score i.e., Score = true score + error score. However, there are alternative model to the true score model, i.e., the item response theory, the domain sampling theory and its modified form the generalizability model.

7.3.1 The item response theory: It is also referred to as Latent trait theory. This model focuses on the extent to which each test item is useful, in evaluating test taker's particular trait or ability; which is presumed to be possessed by individuals (test takers) in varying amounts.

7.3.2 The domain sampling model: This theory denies an existence of 'true score'. It seeks to estimate the extent to which specific sources of variation under definite conditions are contributing to the test score. The reliability according to this theory is an objective measure of how precisely the test score assess the domain from which the test draws a sample. The domain of behavior is a hypothetical construct which includes all the items that could possibly measure that behavior.

7.3.3 Generalizability theory: This model somewhat an extension of the true score model. It speaks about a 'universe score' (instead of true score). According to this theory a test takers scores vary from testing to testing because of variables in the testing situation or the universe. According to Cronbach this universe (variables in the testing situations that lead to specific score) should not be seen as 'errors', rather, they should be precisely described in terms of their facets. Facets include things like number of items in the test, the purpose of test administration; the amount of training to the test takers, etc. According to this theory given the exact conditions of all the facets in the universe, the exact test score should be obtained.

7.4 RELIABILITY AND INDIVIDUAL SCORES

The reliability coefficient not only helps the test developer to build an adequate measurement instrument but also helps the test user select a suitable test.

7.4.1 The Standard Error of Measurement

Let's take a case study to understand this concept:

Case study: Leena takes a skipping test. Below are the number of skipping's Leena has taken in each uninterrupted round i.e., without missing a single skip.

Attempt I: 15, Attempt II: 17, Attempt III: 13, Attempt IV 12, Attempt V.- 16, Attempt VI: 14. Now what is Leena's true score on skipping? Here we do not know how much of the score is due to error and how much is true sore. Thus, we use the standard error of measurement which is a tool to estimate the observed score deviates from the true score. By simple logic it should be average of all the scores above i.e. 14.5 $\pm$ 2.5.

- The standard of measurement abbreviated as SEM or SEm provides a measure of precision of observed test score. The standard error of measurement allows us to estimate the range in which the true score is likely to exist.
- According to true score model the score of an individual is one point in theoretical distribution of scores. Thus, in the above scores of Leena, there is one true score and other scores are a result of errors in the testing conditions.
- The standard error of measurement helps us predict with what confidence is this true score likely to exist. Since SEM is like standard deviation we can state the confidence level. For example we can be 68% sure that Leena's 2 scores differing by 1s represent true score differences. 95% sure that Leena's 2 scores differing by 2s represent true score differences. (Remember the normal distribution curve).
- In other words, if Leena scores 17 on the skipping test and if the skipping test has a standard error of measurement of 2, then using 14 as an estimate we can be 68% sure that the true score falls within 14 ± 1 σ

$_{meas}$ i.e. between 12 and 16; 95 % sure that the true score falls within $14 \pm 2\ \sigma_{meas}$ i.e., between 10 and 18.

- This measure is one way to express test reliability.
- The relationship between SEM and reliability of the test is inverse. If the SEM is high, reliability is low and vice versa.
- The SEM is determined by the following formula. $meas = \sigma\ 1 - r_{xx}$
  Where
  σ meas = is standard error of measurement
  σ = standard deviation of test scores by the group of test takers
  r_{xx} = reliability coefficient of the test
- The Standard Error of Measurement is usually used in interpretation of individual test scores.
- Further this measure is useful in establishing the confidence interval. The confidence interval is the band or range of test scores within which the true score lies. For example, Leena's true score lies in a band between 10 and 18. Thus, the standard error of measurement can be used to set the confidence interval for a particular score or to determine whether a score is significantly different from the criterion.

7.4.2 Standard Error of Difference

As we have seen earlier that the changes in scores can be due to external factors like change in environment, mental status or even the time bf the day or boredom. But how do we know whether the difference in scores is due to these errors or due to real change in a particular trait or capacity. Let, s understand this with the help of a case study.

Case study: Dr. Manisha a practising psychologist decided to use a therapy which she had developed through years of research. She wanted to test her therapy now on her client Chintan who is into depression. She tested Chintan on a scale of depression. Then she administered him the therapy and retested Chintan. Chintan showed drastic difference in his scores in the pre therapy session and the post therapy session. Were the differences in Chintan scores due to therapy or some other factors in the environment is what Manisha wanted to test. How would she do it? Manisha can compare the pre therapy scores and the post therapy scores using the Standard Error of Difference.

1 . Standard error of difference is a measures that can aid the test user determine how large should be the difference between the two scores, for it to be considered statistically significant.

2. In psychological research we often come across the probability factor. So if the -probability is more than 5% that difference occurred by chance then we assume that there is no difference at all between the two scores. For the difference between the two scores to be statistically

significant the probability has to be less than 5 % that the difference occurred by chance.

3. Standard error of difference can be used when we want to compare an individual's scores on test 1 and test 2, or when we want to compare test scores of two individuals or when we want to compare an individual's score on test 1 with another individuals performance on test 2. If we are comparing scores with two different tests then we have to convert the test scores into standard scores. We can do this by the formula given below

$$\sigma_{diff} = \sqrt{\sigma^2 meas1 + \sigma^2 meas2}$$

Where σ diff = the standard error of difference between two scores

σ^2 meas1 = the squared standard error of measurement for test 1

σ^2 meas^2 the squared standard error of measurement for test 2

We can also substitute reliability coefficients for the standard errors of measurement of the separate scores the formula is

$$\sigma \text{ diff} = \sigma \sqrt{2 - r_1 - r_2}$$

Wherein : r_1 the reliability of test 1

r_2 the reliability of test 2

4. Observe that both the tests have the same standard deviation as they either are from the same scale or converted to same scale before comparison.

5. When the difference between two scores is separated by 1 standard errors of difference then we are 68% sure that the two scores are different; if they are separated by 2 standard errors then we can claim with 95% confidence that the two scores are different.

Case study continues: Chintan scores 38 out of 50 questions in prolotherapy testing session and 28 in post therapy testing session. So are these scores different, has Chintan benefitted by the therapy or not?

We have assumed that the measured reliability of this depression scale is 90 and the standard deviation is 10. Now let's calculate.

$$\sigma_{diff} = 10 \sqrt{2 - {}^{\sigma}.90 - .90}$$

$$\sigma_{diff} = 10 \sqrt{.20}$$

$$\sigma_{diff} = 7.77$$

We can infer from the above that

1. If the two scores differ by 7.77 then we can be 68% sure that it reflects the true difference.
2. If the two scores differ by 8.97 then we can be 98% sure that it reflects the true difference.
3. If the two scores differ by 13.71 then we can be 99.7% confident that it reflects the true difference.
4. In Chintan's case we find that the difference between two scores is, 10. Thus, we can be 98% sure that there is true difference between the two scores. In other simpler words we can come to a conclusion that the therapy has worked.

7.5 SUMMARY

1. When the purpose of estimating reliability is to understand the various sources of error variances, relevant in this particular situation, then a number of reliability coefficients have to be calculated.
2. We need to consider the nature of test before calculating reliability coefficient such as test homogeneity versus test heterogeneity, its dynamic versus static characteristics, the range of test score restriction, speed test versus power test, criterion referenced test versus norm referenced test.
3. Tests.which are homogeneous in nature have high internal consistency estimate than tests which are heterogeneous in nature.
4. A dynamic characteristic is a trait that is sensitive to situational variables. In such cases internal consistency is measured. More static traits such as intelligence do not change significantly. In such cases alternate form reliability may be used.
5. While interpreting coefficient of reliability, we need to take restriction range or inflation range of the test into consideration.
6. A speed test has items low in difficulty, but the number of items in the test are so many that it is extremely difficult for the test taker to complete the test in a given time frame, While calculating reliability of a speed test the reliability estimates should be based from 2 different testing sessions.
7. A power test has enough time frame to complete the test however, the difficulty level of the items is so high that it is impossible to get a perfect score or 100% on this test.
8. A criterion referenced test is intended to indicate where a test taker stands with respect to some criterion. Norm referenced tests on the other hand contains material that has not been mastered in a hierarchical pattern.
9. The true score model assumes that the total score consists of true score and an error score. However, there are alternative model to the true score model, i.e., the item response theory, the domain sampling theory and its modified form the generalizability model.

10. **The item response theory** focuses on the extent to which each test item is useful, in evaluating test takers particular trait or ability.
11. **The domain sampling model** seeks to estimate the extent to which specific sources of variation under definite conditions are contributing to the test score.

- The Generalizability theory speaks about a 'universe score' and that a test takers scores vary from testing to testing because of variables in the testing situation or the universe.
- **The Standard Error of Measurement** (SEM or SEM provides a measure of precision of observed test score. It allows us to estimate the range in which the true score is likely to exist. The true score of an individual is one point in theoretical distribution of scores. The standard error of

  measurement can be used to set the confidence interval for a particular score or to determine whether a score is significantly different from the criterion.
- **Standard Error of Difference** can aid the test user determine how large should be the difference between the two scores, for it to be considered statistically significant.

7.6 QUESTIONS

Answer the following questions:

1 Explain the concept of Standard Error of Measurement and Standard Error of Difference and their relevance to reliability.
2. Explain how the nature of tests would affect measurement of reliability?
3. Discuss the various alternatives to true score model.

7.7 REFERENCES

1. Cohen, R.J., & Swerdlik, M.E., (2010). Psychological testing and Assessment: An introduction to Tests and Measurement, (7th ed.), New York. McGraw - Hill International edition, 129 -132
2. Anastasi, A. & Urbina, S. (1997). Psychological Testing. (7th ed.). Pearson Education, Indian reprint 2002.
3. Kaplan, R.M., & Saccuzzo, D.P. (2005). Psychological Testing - Principles, Applications and Issues . (6th ed.). Wadsworth Thomson Learning, Indian reprint 2007.

8

VALIDITY

Unit Structure:

8.0 OBJECTIVES

After going through the unit you would be able to:
1. Explain the concept and meaning of validity.
2. To analyse the meaning of content validity.
3. To discuss the concept of criterion validity.
4. To explain construct validity.

8.1 INTRODUCTION

In the earlier unit, the concept of good test, norms and reliability of test had been discussed. The concept of validity can be applied to research process as a whole or to any of its steps. We can talk about the validity of the study design used, the sampling strategy adopted, the conclusions drawn, the statistical procedures applied or the measurement procedures used. In this chapter we will discuss the concept of validity as applied to measurement procedures or the research tools used to collect the required information from respondents. For a test to be scientifically sound, it must possess different characteristics like objectivity, reliability, validity, practicability and norms. Validity is one of the important characteristic of scientific instrument

8.2 THE CONCEPT AND DEFINITION OF VALIDITY

Validity refers to the degree to which test measures what it claims to measure. Validity is not the correlation with test rather it's the correlation with some outside independent criteria, which are regarded by experts as the best

measure of traits. Validity is defined as the degree to which the researcher has measured what he has set out to measure. (Smith 1991)

Anastasi (1968) has said "validity of a test concerns what the test measures and how well it does so". Lindquist has defined "validity of a test as the accuracy with which it measures that which is intended to measure or as the degree to which it approaches infallibility in measuring what it purports to measure."

The above two definitions points to the fact that for determining the validity of a test, the test must be compared with the same ideal independent measures or criteria.

The correlation coefficient computed between the test and ideal measures or criteria is known as the validity coefficient. Independent criteria refers to some measure of traits or group of traits that the test itself claims to measure.

Babbie writes "validity refers to the extent to which empirical measure adequately reflects the real meaning of concept under consideration. When test is valid one, it means its conclusion can be generalised in relation to the general population.

Validity has three important properties:

1 . Validity is a relative term. A test is not generally valid. It is valid only for a particular purpose. A test of statistical ability will be valid only for measuring statistical ability because it is put only to the use of measuring that ability.
2. Validity is not the fixed feature of a test because validation is not a fixed process rather an unending process. With the discovery of new concepts and the formulation of new meanings the old context of test become less meaningful. Hence, the validity of a test computed in the beginning becomes less dependable and therefore the test constructor should compute a fresh validity of the test in the light of new meanings attached.
3. Validity, like reliability, is a matter of degree and not at all or none property. A test meant for measuring a particular trait or ability cannot be said to be either perfectly valid or not valid at all.

There are three main purposes of testing:

1 . Representation of a certain specified area of content: The tester may wish to determine how an examiner performs at present in a sample of situations that the test claims to represent. For example through Math's the tester may determine the present level of Math's ability.
2. Establishment of functional relationship with a variable available at. present or in future: The tester may wish to predict an examinee's future standing on a certain variable. Through mechanical aptitude tester may wish to measure mechanical aptitude and predict his future

performance in job related to that field.

3. Measurement of a hypothetical trait or quality: A tester may wish to determine the extent to which an examinee possesses some traits measured by the test performance.,

There are three types of validity:

1. Content or curricular validity
2. Criterion related validity
3. Construct validity

8.3 CONTENT VALIDITY

When a test is designed so that its content of term measures what the whole test claims to measure the test is said to have content or curricular validity. Thus, content validity is concerned with the relevance of contents. Each individual item or content of test should correctly and adequately sample or measure the test, as a whole and should consist only the representative items of the variable.

According to Anastasi, content validity involves essentially the systemic examination of test content to determine whether it covers a representative sample of behavior domain to be measured. Content validity is required in test which is constructed to measure how well the examinee has learned specific skills or a certain course of study.

Content validity of a test is examined in two ways: 1 . By expert judgment 2. By statistical analysis

If the investigator wants to examine the content validity of a test on Science, for this purpose content matter of the test will be submitted to a group of subject matter experts. These experts will judge whether or not the items are important matter of Science. The validity of the content or items will be dependent upon a consensus judgment of majority of subjects- matter experts.

Statistical methods may also be applied to ensure that all the items measure something that a statistical test of internal consistency may provide evidence for the content validity. Another statistical technique for ensuring content validity may be to correlate the scores on 2 independent tests both of which are said to measure the same things. Example, if one wants to measure the content validity of an English spelling test, then the teacher can correlate the scores on the said test with another similar English spelling test. A high correlation coefficient would provide an idea about its content validity.

Content validity is most appropriately applied to the achievement test or the proficiency test. For the aptitude test, the intelligence test, and the personality test content validity is not essential and sometimes may be a misleading

index because the contents of these tests have less intrinsic resemblance to the trait or behavior they are attempting to sample than do the achievement tests.

FACE VALIDITY:

It is often confused with content validity, but in the strict sense it is quite different. Face validity is not what the test actually claims to measure but to what it appears to measure superficially. When a test item looks valid to a group of examinees, the test is said to have face validity. Face validity is needed in all types of test and helps a lot in improving the objectively determined validity of test by way of improving the wording and structure of test contents. Face validity is very closely related to content validity. While content validity depends on theoretical basis for assuming if a test is assessing all domains of a certain criterion (example, does assessing addition skills, yield in a good measure for mathematical skills? To answer this we have to know, what different kinds of arithmetic skills, mathematical skills it includes.)

Face validity relates to whether a test appears to be a good measure or not.

CHECK YOUR PROGRESS

Answer the following

1. Define validity?
2. What is content validity? How can you calculate content validity explain with example.
3. Write a note on face validity.

8.4 CRITERION RELATED VALIDITY

Criterion related validity is a very common type of test validity. As its name implies, criterion related validity is the one which is obtained by comparing or correlating the test scores with the scores obtained on a criterion available at the present or to be available in the future. Criterion validity evidence involves correlation between the test and a criterion variable taken as representative of construct. The criterion is defined an external and independent measure of essentially the same variable that the test claims to.

According to Cureton (1965) that the validity of a test is an estimate of the correlation coefficient between the test scores and the "true" criterion scores. For example, employee selection tests are often validated against measures of job performance and IQ tests (criterion) are often validated against measure of academic performance.

There are two subtypes of criterion related validity:

1 . Concurrent validity

2. Predictive validity

8.4.1 Concurrent Validity:

The test is correlated with a criterion which is available at the present time. In other words, if the test data and criterion data are collected at the same time, this is referred to as concurrent validity evidence. Scores on a newly conducted intelligence test may be correlated with scores obtained on an already standardised test of intelligence. The resulting coefficient of correlation will be an indicator of concurrent validity. Concurrent validity is most suitable to tests meant for diagnoses of the -present status rather than for prediction of future outcomes.

8.4.2 Predictive Validity:

Predictive validity is as empirical or statistical validity. In predictive validity a test is correlated against the criterion to be made available sometime in the future. In other words, test scores are obtained and then a time gap of months or years after which the criterion scores are obtained. Subsequently, the test scores are co related and the obtained co relation becomes the index of validity coefficient.

According to Marshal Hales (1972), the predictive validity coefficient is a Pearson product moment correlation between scores on test and an appropriate criterion, where the criterion measure is obtained after the desired gap of time. Example if an experimenter wants to predict in TYRA class in terms of grade A, B, C and D, here A the best and D is the worst. The investigator may administer a test of intelligence at the time of beginning of the class and obtain a set of scores. After one year on the basis of classroom performance students are graded according to the above categories. A product moment co-relation through a scatter diagram may be computed between the set of intelligence scores and the grade points obtained after one year. If the correlation is high we can say with all certainty that scores on intelligence are directly predicting the future performance of the students in TYRA class. In the same way in business organisation, management may wish to select such workman who can exhibit best performance on the job. For this objective, they select a test which has high predictive validity. Predictive validity is required for the test which includes long range forecast of academic achievement, vocational success and of reaction to therapy. A comparative study of predictive validity and concurrent validity has revealed that for the same test predictive validity is usually lower than concurrent validity. The reason that the degree of association between the test and the criterion decreases over time. Naturally, then predictive validity will be somewhat lower than concurrent validity. If concurrent validity of a test happens to be zero, then its predictive validity is most likely to be zero, or close to it.

CHECK YOUR PROGRESS

Answer the following question 1. What is criterion? 2.What are the two types of criterion validity? 3.Define concurrent validity. 4.Differentiate between predictive validity and concurrent validity.

8.5 CONSTRUCT VALIDITY

The term "construct validity" was first introduced in 1954 in the Technical Recommendation of the American Psychological Association and since then it has been frequently used by measurement theorists. Investigator decides to compute construct validity only when he is fully satisfied that neither any valid and reliable criterion is available to him or any universe of content entirely satisfactory and adequate to define the quality of test. In other words construct validity is computed only when the scope for investigating criterion related validity or content validity is bleak. In construct validity, the meaning of test is examined in terms of construct. Anastasi has defined it as "the extent to which the test may be said to measure a theoretical construct or trait." For example, to what extent an IQ questionnaire actually measures "intelligence."

Construct validity evidence involves the empirical and theoretical support for the interpretation of construct. "A construct is a sort of concept, which is formally proposed with definition and is related to empirical data." According to Nunnally "a construct indicates a hypothesis which tells us that "a variety of behaviors will correlate with one another in studies of individual differences and or will be similarly affected by experimental treatments." A few examples are anxiety, intelligence, extroversion and neuroticism.

Following are some of the ways in which we can calculate the construct validity:

1 . **Specify The Different Measures Of Construct :** Here the investigator explicitly defines the construct in clear words and also stakes one or many supposed measures of that construct. Specification of such measures is partly dependent upon the previous researches conducted in that area and partly upon the intuition of the investigator. Suppose if one wants to specify the different measures of the construct, anxiety, the investigator would have to first define the term anxiety and in the light of definition he would be expected to specify the different measures.

2. **Determining The Extent Of Correlation Between All Or Some Of The Measures Of Construct :** When adequate measures of construct have been outlined, the second step consists of determining whether or not those well specified measures actually lead to the measurement of the concerned construct. This is done through an empirical investigation in which the extent to which the various measures correlate with each other is determined. In

an empirical investigation correlation coefficients are computed between different measures of a construct.

3. **Determining Whether Or Not All Or Some Measures Act As If They Were Measuring The Construct :** When it has been determined that all or some measures of construct correlate highly with each other, the next step is to determine whether or not such measures behave with reference to other variables of interest in an expected manner. If they behave in an expected manner, it means they are providing evidence for the construct validity. It is obvious from the above interpretation that unlike content validity and criterion related validity, the evidence for construct validity is always circumstantial rather than direct. Construct validation is also a difficult process because it contains several problems like systematic examination, concerning the definition of the construct.

8.5.1 Convergent Validity: Convergent validity refers to the degree to which a measure is correlated with other measures that it is theoretically predicted to correlate with that measure. For example a numerical aptitude test should correlate with an arithmetical reasoning test but it should not correlate with a psychological test. When the test correlates with its expected referents the process is known as convergent validity.

Campbell and Fiske 1959) have demonstrated that the convergent validation and discriminant validation are important for establishing satisfactory construct validity.

8.5.2 Discriminant Validity: Discriminant validity describes the degree to which the operationalisation does not correlate with other operationalisation that it theoretically should not be correlated with. In other words, when a test correlates poorly with measures with which it should not correlate because it differs from those referents or measures, this procedure is called discriminant validation. Example spelling test should not correlate with numerical ability test.

8.5.3 Experimental Validity: The validity of the design of experimental research studies is a fundamental part of the scientific method. Without a valid design, valid scientific conclusions cannot be taken. There are different types of experimental validity.

8.5.4 Conclusion Validity: Conclusion validity refers to the degree to which conclusions reached about relationships between variables are justified. This involves ensuring adequate sampling procedures, appropriate statistical test and reliable measurement procedures.

8.5.5 Internal Validity: Internal validity is an inductive estimate of the degree to which conclusion about causal relationships can be made; this is based upon the measures used for this purpose, the research setting and the

whole research design. Different kinds of variables can interfere with internal validity.

1 . **History:** the element occurring between the first and second measurements in addition to the experimental variables.

2. **Maturation:** processes within the participants as a function of the passage of time growing older.

3. **Selection:** biases resulting from differential selection of respondents for the comparison groups.

4. **Testing:** the effects of taking a test upon the scores of a second testing.

5. **Instrumentation:** Changes in the observers or scores may produce changes in the obtained measurements.

8.5.6 External Validity: External validity concerns the extent to which the results of a study can be generalised for other cases, for example to different people, places or times.

8.5.7 Ecological Validity: Ecological validity is the extent to which research results can be applied to real life situations outside of research settings. This issue is closely related to external validity.

CHECK YOUR PROGRESS:

Answer the following questions

1.When the term 'construct validity'was used?

2.Define construct validity.

3.What is construct?

4.What is convergent validity?

5.Define discriminant validity.

8.5.8 Validity Coefficient : A validity coefficient is a correlation between test score and criterion measures; because it provides a single numerical index of test validity, it is commonly used in test manuals to explain the validity of a test against each criterion for which data are available. Validity coefficient can also be expressed in the forms of an expectancy table or expectancy chart. In an expectancy table the expectancy of the criterion measures for each examinee is given against each test score. As its name implies, through the expectancy tables the predictive efficiency of a test is estimated. Estimates are usually based upon the probability that an examine securing a particular score on the test will obtain a specified score or rating in the performance. When both test and criterion variables are continuous, the familiar Pearson ProductMoment correlation coefficient is applicable. Other types of correlation coefficients can be computed when the data are expressed in different forms, as when a two fold pass fail criterion is employed

the Biserial and the Point Biserial are used. When on the other hand scores on the test as well on the criterion are divided into two categories the tetrachoric r or the phi coefficient are the most appropriate statistics. Multiple correlations are used where more than two measures are involved. R is a symbol for multiple correlations which indicates the relationship between one measures and the composite of the two or more than two sets of measures.

8.5.9 Factors Influencing Validity: Validity of a test is influenced by several factors:

1 . **Length Of The Test :** If the test is lengthy it has more reliability and validity. Thus, lengthening of the test or repeated administration of the same test increases the validity of the test. But validity as compared to reliability does not change rapidly with increase in the length of the test.

2. **Ambiguous Direction :** If in any test directions are not given properly it would be interpreted in different ways, by different examiners. Such items encourage guessing on the part of the examinees. As a consequence, the validity of the test would be low.

3. **Socio-Cultural Differences :** Cultural differences among different societies also affect the validity of the test. Any test developed in one culture may not be valid for another culture because of the differences in socio-economic status, sex roles, social norms, etc. Consequently a test could have validity in predicting a particular criterion in one population and little or no validity in another.

4. **Addition Of Inappropriate Items :** When inappropriate items, particularly the items whose difficulty values differ widely from the original items are added to the test, they lower both the reliability as well as the validity of test.

5. **Heterogeneous Sample :** If the subjects have a very limited range of ability, the validity coefficient will be low. If the subjects have a wider range of ability so that a wider range of score is obtained, the validity coefficient of the test would be high.

6. **Changing Selection Standards :** Validity coefficients may also change over time because of changing selection standards.

8.6 VALIDITY, BIAS AND FAIRNESS

If we want to use tests to predict outcomes in some future situation such as an applicant's performance in college or on a job, we need tests with high predictive validity against the particular criterion.

A better solution is to choose criterion relevant content and then investigate possible population differences in the effectiveness of the test for its intended purpose. It should be noted that the predictive characteristics of test scores are less likely to vary among cultural groups when the test is intrinsically relevant to criterion performance. In a group with a different cultural and experimental background the validity of the test may be very low.

The term "bias" is employed in statistical sense to designate constant or systematic error as opposed to chance error. This type of error is found in biased sample and not in random sample. There are different types of biases:

8.6.1 Slope Bias: To find out the slope bias we can take a example of job performance. For this purpose, horizontal axis X shows the scores on a test and the vertical axis Y represents criterion scores, such as job performance. One important mark shows the position of each individual on both test and criterion. This mark indicates the direction of the correlation between the two variables. The line of best fit drawn through these tally known as the regression line its equation is the regression equation. In this example the regression equation would have only one predictor. When both test and criterion scores are expressed as standard scores (SD=1.0) the slope of the regression line equals the correlation coefficient. If a test yields a significantly different validity coefficient in the two groups, this differences is described as slope bias. This type of group difference is often designated as "differential validity". It refers to a test whose validity coefficient reached statistical significance in one group but failed to do so in another.

In differential validity studies a, common difficulty arises from the fact that the number of cases in minority sample is often much smaller than in the majority sample. Under these conditions the same validity coefficient could be statistically significant in the majority sample and not significant in the minority sample.

8.6.2 Intercept Bias: The intercept of a regression line refers to the point at which the line intersects the axis. A test exhibits intercept bias if it systematically under predicts or over predicts criterion performance for a particular group.

Besides these biases in validity test constructs has to face different types of problem in computing predictive and concurrent validity. This problem is comparatively more acute in predictive validity, is the identification and selection of an appropriate and adequate criterion. An inappropriate and inadequate criterion may decrease the coefficient of correlation and thus the validity of the test is adversely affected. For example, when one is computing the validity of intelligence against a future criterion of grade, he may be faced with such a problem because something more than intelligence may be involved in obtaining a particular grade. Factors like interests, motivation, emotional adjustments of the students may influence the grades which are obtained by a student. In such situations, correlations between intelligence test scores and the grade will not be a true index of validity of a test, when we are computing the concurrent validity of a test the criterion may not be itself reliable to the extent it should be. In this type of situation the correlation coefficient would tend to be attenuated or reduced and therefore the validity of the test would be lower than the true relationship between the test and criterion.

8.6.3 Fair Use Of Test: To use the fair selection strategies regression model should be used. Individuals will be selected for admission, employment only on the basis of their predicted criterion scores. This strategy will maximize overall criterion performance without regard to other goals of the selection process. According to this strategy a fair use of test in selection is one that is based only on the best estimate of criterion performance for each individual. Multiple-aptitude testing and classification strategies that permit the fullest utilisation of the diverse aptitude patterns fostered by different cultural backgrounds. A broader consideration of relevant personality traits, motivation and attitudes also contributes to the prediction of job or educational performance.

To remove the problem of computing predictive and concurrent validity, first obtained validity coefficient should be corrected for attenuation. There are two types of correction for attenuation :Full Correction: It includes the correction in both the test as well as the criterion. One way correction: It includes correction in the criterion only.

CHECK YOUR PROGRESS

Fill in the blanks Answer the following

1. How constructor can find out the slope bias?
2. Discuss the different types of problems that constructor has to face in computing predictive validity.

8.7 SUMMARY

Validity is the correlation of the test with some out side independent criteria. The validity refers to the degree to which a test measures what it claims to measure. In this unit important property of validity has been described. There are three types of validity.

1. Content validity
2. Criterion validity
3. Construct validity

When a test is designed so that its content of term measures the whole test claims to measure the test is said to have content validity. Content validity of a test is examined in two ways:

1. By the expert judgment
2. By statistical analysis

Face validity is quite different from content validity. Face validity refers not to what the test actually claims to measure but to what it appears to measure.

Criterion-related validity is one which is obtained by comparing test scores obtained on a criterion available at present or to be available in the future. There are two subtypes of criterion related validity:

1. Predictive validity
2. Concurrent validity

In construct validity the meaning of a test is examined in two terms of construct. Besides these three types there are some other types like convergent validity, discriminate validity, experimental validity and conclusion validity have been explained in this chapter.

Validity coefficient is a correlation between test score and criterion measures. Different factors like length of the test, socio cultural differences, ambiguous directions, changing selection standards and heterogeneous sample affect validity. Test constructor has to face the different types of problems while computing the validity. The term bias is employed in its statistical sense to designate constant or systematic error as opposed to chance error. There are two different types of biases.

8.8 QUESTIONS

Answer the Following:

1 . Define validity. Explain content validity. How content validity of a test is examined?
2. What is criterion related validity? Explain the different types of validity.
3. Define construct validity. How can we calculate construct validity?
4. Explain different factors that affect validity.
5. Write short note on:
   a. Face validity
   b. Concurrent validity
   c. Predictive validity
   d. Bias in validity

6. Define the following terms.

i) Construct
ii) Validity
iii) Convergent validity
iv) Discriminate validity
v) Experiment validity

8.9 REFERENCES

1 . Anastasi, A and Urbina, S (1997) Psychological Testing (7 th Ed.) Pearson Education, Indian reprint 2002.
2. Hoffman, E. (2002) Psychological Testing at work, New Delhi, Tata Mc Graw Hill Publishing Company Ltd.
3. Mangal, S.K (1987) Statistics in Psychology and Education, New Delhi, Tata Mc Graw Hill Publishing Company Ltd.
4. Ranjit Kumar (2005) 2nd Ed. Research Methodology, Dorling Kindersley (India) Pvt. Ltd. Licenses of Pearson Education in South Asia.

8.10 GLOSSARY

Aptitude: Innate or acquired ability to develop knowledge of a skill in some specific area.

Correlation: Statistical method, which studies the degree of relationship between two variables.

Variable: Attributes of objects, events which can be measured.

Empirical: Based on observation or experience rather than theory.

Neuroticism: Extent to which a person is emotionally stable, secure, content.

Extroversion: It is the degree to which a person is social and outgoing.

9

TEST DEVELOPMENT

Unit Structure:

9.0. OBJECTIVES

- To impart knowledge and understanding about test development.
- To create awareness about the technical process of construction of a good test.
- To explore number of techniques designed for construction and selection of good items.
- To compare the custom-made test with those of newly constructed tests.

9.1. INTRODUCTION

To develop a test is not an easy task and a good test is not developed by chance, It requires a great deal of thoughtful and sound application of standardised principles based on statistical techniques.

However, construction of test process occurs in five stages:

a) Test conceptualisation
b) Test construction
c) Test tryout
d) Item analysis
e) Test revision

Test conceptualisation refers to a novel idea for a test to be conceived. Items for the test drafted refers to as construction. The first draft of the test is then applied on a group of sample testtaker (test tryout). Once the data from tryout are collected, performance of testtaker on each tem is analysed.

Then, item analysis in terms of statistical procedures are applied to know which items are good and which items need to be revised and which of them to be dropped or discarded. And finally, the results are analysed and further revised if necessary- so it goes.

9.2. TEST CONCEPTUALISATION

In behaviour terms, the starting point to develop a new test begins with self-talk and test developer asks himself something like "which type of test should be designed that measures a 'construct' for which the test is being developed. For example, once a new disease comes to the attention of medical researchers, they try to develop diagnostic tests to assess its causes, symptoms, presence or absence, its severity of manifestations in the body. Thus, development of a new test may be a response to a need to face severe situation and to mastery in the fields of occupation or profession, telecommunications and computer networking, etc. However, a test developer is confronted with a number of questions when he develops a new test. Some of them are as under:

- What is the test designed to measure? Thus, a simple deceptive question which is related to how the test developer defines the construct being measured? How this definition is different from other test measuring the same construct?
- What is the objective of the test? This question is directly related to aim / goal and purpose of the test.
- Is there a need for this test? Are other tests available to measure the same attribute or trait? Are these tests reliable and valid? In what ways the newly constructed test will be better than other existing tests?
- Who will use this test? Clinicians? Educators? Others? What purpose this test would serve?
- Who will take this test This question is related with age range and qualification of test takers.
- What content will the test cover? This question covers the content of existing test and the content of culture-specific.
- How will the test be administered? This question is linked with the application of test on individual or group and on both.
- What is ideal format of the test? is it in true/false form, essay type, multiple- choice, or in some other forms? which is the best form selected for administration?
- Should more than one form of the test be developed?
- What special training will be required for test users for administering or interpreting the test? This question requires background and qualifications of test users.

_ Who benefits from an administration of this test?
_ How will meaning be attributed to score on this test?

This question is related with scores of one test user to another taking the test at the same time and others in a criterion-group.

However, the last question needs attention relating to test development with regard to norm-versus criterion- referenced tests.

9.2.1. Norm- referenced versus criterion- referenced tests. '(Item development issues).

There are two main important approaches to test development and individual item analysis. They are normreferenced or criterion-referenced approaches.

A good item on a norm-referenced achievement test is an item that scores high on the test answers correctly. Whereas, low score indicates on the same item incorrectly. But in a criterionoriented test, high scores referred to as right and low scores as wrong.

However, criterion-oriented test is used in licensing contexts, a license to practice medicine or a license to drive a car. This approach is also employed in educational contexts, to strengthen the knowledge, skills or both of students in class room teaching. The test developer may attempt to sample criterion-related knowledge relevant to the criterion being assessed. For the targeted skills or knowledge, they may conduct experimentation with different items, tests, formats or measurement procedure which may help them to discover the best measure of mastery for required cognitive or master skills.

Whereas, norm-referenced approach is insufficient and inappropriate when knowledge of mastery is required for test user.

However, the best items are those items that discriminate between these two groups.

9.2.2. Pilot Work.

In behaviour sciences, pilot work, pilot study and pilot research generally refers to the preliminary administration of test on selected sample before final administration.

Commonly, pilot study is conducted to evaluate the reliability and validity of newly constructed test and to find out whether it may be included in the final form of the instrument. For this purpose, a structured interview is conducted. In addition, interviews with parents, teachers, and others may be arranged.

In pilot study, a test developer attempts to find out to measure a targeted construct. This process may help him in creation, revision, deletion of many test items, literature reviews, and related activities. The need for additional pilot study is always knocking at door.

Check your progress

1. Define conceptualisation.
2. What are the processes of developing a test?
3. Explain any four preliminary questions which a test developer is immediately confronted with?
4. What is pilot work, pilot study or pilot research? Explain.

9.3 TEST CONSTRUCTION

After finalisation of pilot study, test developer pays attention to formal aspect of professional test construction. The process of test construction begins with scaling.

9.3.1 Scaling

Scaling may be defined as "the process of setting rules for assigning numbers in measurement". In other words, scaling is the process by which a measuring device is designed and scale values are assigned to different amount of the trait, attribute or characteristic being measured,

Credit goes to L.L. Thurston who developed methodologically sound scaling methods (1929,1932). His technique of scaling method is known as Equal-appearing interval.

However, Thurston and his students have developed a series of scales, each consisting of statements. These scales are developed to measure attitudes of individuals towards Negroes, Chinese, war, Censorship, the Bible, patriotism, and freedom of speech. An important scale of Thurston (1932) was developed to measure attitudes. This scale consisting of several statements or items which are followed by five responses along with scoring weights.

These five responses are:	**Points**
Strongly Agree (SA) having	(5)
Agree (A)	(9)
Undecided (U)	(3)
Disagree (D)	(2)
Strongly Disagree (SD)	(1)

This scale is also known as five-point-scale.

Types of Scales

It is generally agreed that there are four types of scales of measurement. They are : nominal scales; ordinal scales, interval scales and ratio scales.

Nominal scales

Nominal scales are the simplest forms of measurement. They are based on one or more distinct characteristics in which all things measured are classified or categorised and placed into mutually and exhaustive categories. For example, clinical psychologists often use nominal scale taking the help of (DSM-IV) to find out the nature, causes, symptoms and therapeutic method to know about mental disorder. This DSM-IV has assigned its own number for each disorder. For example, the number 303.00 identifies alcohol intoxication, and number 307.00 identifies stuttering. But these numbers are used exclusively for classification purpose.

The following statement can explain the test item on Nominal scale:
Instruction : Answer either Yes or No.

Are you actively contemplating suicide? ----------

Ordinal scales

Ordinal scale is a system of measurement in which all things measured can be rank-ordered. It permits classification. In business and organisational settings, job applicants may be rank-ordered according to their desirability for a position. For individual subject ordinal form of measurement is also used. The Rokeach Value Survey which consists of a list of personal values, such as freedom, happiness, and wisdom is one of the best examples of

ordinal sale (1973). A set of values ca be ranked in order which may assign a value of Vto the most important and '10' to the least important.

Interval scale is a system of measurement that contains equal intervals between numbers. Each unit on the scale is exactly equal to any other unit on the scale. Each unit in interval scales gets a meaningful result. Intelligence test is the best example of this type of scale. The IQs of 80 and 100, for example, is thought to be similar to that existing between IQs of 100 and 120. Like nominal and ordinal scales, interval scales also does not have zero ability or I.Q.

Ratio scales

A ratio scale has a true zero point. It is a system of measurement in which all things measured can be put in rank order and equal intervals exist between each number on the scale. There are few scales in psychology or education that can come closer to ratio scales. The best examples of ratio scales are a test of hand grip and a timed test of perceptual-motor ability (Cohen-Swerdlik, 2005)

In addition, we can also characterise scales in other ways. There are various other types of scales also which test takers use in different forms. When the test taker tests the performance of individuals as a function of age, it is referred to. as age-based scale. If he tests performance as a function of grade, it is known as gradebased scale. Similarly, when all raw scores of a test are transformed into scores, ranging from 1 to 9, is referred to as a stanine scale.

However, there are many different methods of scaling but there is no best type of scale. We shall discuss here some important scaling methods :

Method of Scaling

It is recorded in the history of test construction that L.L.Thurston (1929,1932) has developed a scaling method known as Equalappearing interval.

In this regard, another important scaling method has been developed by Katz et al (1999). It is known as Morally Debatable Behavioural Scale-Revised (MDBS-R). This scale assesses people's belief, the strength of their convictions and moral tolerance. It is 10 - point scale that ranges from justified to always justified. Here is a sample.

Cheating on taxes if you have a chance is

1 2 3 9 5 6 7 8 9 10

never justified always justified

The MDBS-R is an example of a rating scale which explains grouping of words, statements or symbols on which judgment about trait, attitude or opinion and emotion are indicated by test takers. This type of rating scale is used to obtain judgment of oneself, others, experiences, or objects. This scale consists of 30 items or statements. So 30 scores indicates a low score and 300 indicates a high score (always justified). The final score is obtained by summing the rating scores of all items. This is termed as a summative scale.

Another important scaling method has been developed by Likert (1932). This scale is widely used to measure the attitudes of individual. Likert scales are very easy to construct. It is 5 - point scale because each item has five alternative responses, usually on an agree / disagree or approve / disapprove type of continuum.

Another scale method is Guttman scale (1999,1997), based on ordinal-level of measurement. This scale measures attitude, belief or feeling. This scale's items range from weaker to stronger expression. However, this scale is not widely used, because other useful scales available measure various constructs.

Another scaling method is used by Katz et al, known as paired comparisons. In this method, the test taker selects one stimulus out of two (two photographs, two objects, two statements) according to his interest.

An item on the scale is

Select the behaviour that you think would be more justified. a. cheating on taxes if one has a chance. b. accepting a bribe in the course of one's duties.

In these items option a is more justified on which most of judges reflect their agreement.

Another scaling methods are comparative scaling and categorical scaling. For comparative scaling, the judgments of a stimulus is compared with every other stimulus on the scale. For example, a test taker is given a list of 30 items on a sheet of paper and asked then to rank the items from 1 to 30, as most justifiable to least justifiable.

In categorical, scaling stimuli are placed into one of two or more alternative categories that differ quantitatively and then test takers may be asked to

sort the cards into three piles, never justified, sometimes justified and always justified.

We have discussed above the nature, types and methods of scaling which are important components of test construction. No single scaling type or method is perfect. The selection of scaling or method depends upon the choice of test developer and the behavioural constructs being measured.

9.3.2. Writing Items.

The selection of scaling methods and actual writing of the test's items go together. While writing items test developer has to pay attention on :

- The Range of Content the test's items cover.
- The different types of formats a test developer should employ.
- The number of items the test should contain (written items in No.)

For construction of new test, the test developer has to write itenq according to the format of the test. It is useful to have at least 1200 item sample in the domain of an item pool. Item pool is a type of reservoir of items from which final version of items will be drawn or discarded. A comprehensive sampling gives up a basis for content validity of the final version of the test.

However, to prepare the initial items pool, a test developer has to write a large number of items from personal experience or can obtain from other sources including experts. For item writing of psychological test, a test developer may conduct interviews in clinical setting, on clinicians, patients, parents, family members, and professionals who can assist test developer. Therefore, it is necessary for test developer to write items only about 1200 for item pool, if he is very knowledgeable and capable.

It is but natural that items pool may be prepared according to selected format which will be presented to test takers.

9.3.3. Item format

Variables such as the form, plan, structure, arrangement and layout of individual test items are collectively referred to as item format. We shall discuss here two important types of item format, namely the selected response format and the constructed - response format.

In a selected response format, there are a set of alternative responses. The test takers have to select only one response. But in a constructed - response format, the test takers are supplied or to create the correct answer, not merely to select it.

However, there are three types of selected - response format available. They are multiple - choice, matching, and true/false format.

A good example of multiple - choice item in achievement test.

1 . Of the following, which is the best answer or measure of public interest in a particular election?
(A) The number of offices to be filled.
(B) The size: of the popular vote.
(C) The amount of campaigning preceeding the election.
(D) The amount of money spent by the opposing parties for campaign purposes.
(E) The importance of the issues at stake.

2. Matching test :

Direction : After each name of topic which is intimately associated with that person.

	Column A	**Column B**
	Name of Persons	**Topics**
1 .	Conditional Reflex	Titchener
2.	Age scale for testing intelligence.	Stanley Hall
3.	Reaction time experiments.	Pavlov
9.	Psychoanalysis.	Cattle, J. M.
5.	Psychology of Adolescence.	Sigmund Freud
6.	Existential Psychology.	Alfred Binet
7.	Factorial analysis.	

However, a multiple - choice item that contains only two possible responses is called a binary-choice item. The most familiar binarychoice item is the true/false, agree/disagree, Yes/No, right/wrong or fact/opinion.

On the other hand, the constructed-response items are also of three types. They are he completion item, the short answer, and the essay type.

A completion item requires the testtaker to provide a word, or phrase that completes a sentence. Following is the example of completion item.

The mean is the most stable and useful measure of -------The correct answer is central tendency.

If we write this sentence in short-answer item, we can write as: what descriptive statistics is generally considered the most useful measure of central tendency.

Whereas, essay item is concerned, it requires the test taker to respond to a question by writing a composition that is related to recall of facts, understanding, analysis or interpretation.

Here is an example of an essay item.

Distinguish between classical and operant conditioning in terms of definitions, principles and techniques.

An essay item is useful for test developer to demonstrate a depth of knowledge about a single topic. One can communicate his ideas in writing very well.

9.3.4Writing items for computer administration.

Various computer programmes are designed to help the construction of tests and their administration, scoring and interpretation. These programmes have two main advantages

1) to store items in an " item bank " and
2) test individual's ability through the technique called "item branching

An item bank collects a large number of questions which an instructor has to teach, sometimes useful for examination. These item bank of questions are compiled by subject areas, item statistics, or other variables. These items may be added to, withdrawn from, and even modified in an item bank.

However, computer-adaptive testing (CAT) is an important test device which can be administered on test taker's performance on previous items. The great advantage of CAT is that it records total number of items in terms of item pool for administering on test taker. CAT is very useful to reduce the number of test items by 50% and also reducing error of measurement by 50%.

The ability of the computer to present test items taken from an item bank based on previous responses of test taker is called item branching. Thus, CAT pressents programme items according to rule. For example, a test taker cannot take third items unless last two previous items are answered correctly. These items are presented on the basis of difficulty level.

Item branding technique is applied as construction of test of achievement and in test of personality. For example, if a person answers an item in such a way that seems for us that he/she is anxious about nothing, then computer may automatically provide the anxiety-related symptoms and behaviour.

9.3.5. Scoring Items

There are various scoring models available to score test items. Among them, Cumulative Model is commonly used for ease and simplicity. It is said that

the higher the score on the test, the higher is the ability of test taker on trait, attribute or other characteristics the test is applied to measure.

Another important model is Class or Category scoring. This test is used to assess diagnostic symptoms of individuals that are exhibit in specific diagnosis.

A third scoring model is ipsative scoring, used to compare test taker's score on one scale within a test with another scale within that test.

However, Edward Personality Preference Schedule (EPPS) is another model of scoring items that has two alternative responses such as Yes/No. true/false.

After making decision about scoring models, the first draft is ready for administration, the next step is test tryout.

Check your progress.

Q1. Define scaling and explain Nominal scales and Ordinal scales.
Q2. Explain L.L. Thurston's scaling method in brief.
Q3. Write short note on Likert's scale.
Q4. Define item pool and its format.
Q5. Explain any two types of selected- response format.
Q6. Explain in brief:
    a) Example of multiple-choice items
    b) Example of matching items
    c) Item bank
    d) Computerised Adaptive Testing (CAT)
    e) Scoring Items.

9.4 TEST TRYOUT

After writing a pool of items from the final version of the test, the test developer will try out the test on selected sample of people for whom the test is designed. It is better to try out the test on corporate employee at the targeted level. The number of questions is equally important. It is necessary that there should be minimum 5 or maximum 10 subjects for each item in tryout test.

However, such test must be used/applied under good possible condition, possible instructions, time limit allotted for completing the test, atmosphere at the test site, etc.

9.4.1 What is a good item?

As we know that a good test is reliable and valid,. so also a good test item is reliable and valid. Thus, a good item is one that high scorers on the test as a whole gets right. Whereas, an item that high scorers on the test as a whole does not get right is not a good item.

In other words, a good test item is one on which maximum scorers can do right on that item.

However, different types -of statistical techniques are used to analyse the test data, one such technique is item analysis.

9.5 ITEM ANALYSIS

Item analysis is the fourth step of the Orocess of development in test construction. It is an integral part of the reliability and validity. However, the quality and merit of a test depend upon the individual items of which it is composed. Thus, it is essential to analyse each item in the standardisation process to retain only those that suit the purpose and rationale of the device being measured.

However, item analysis is a general term which is related with various procedures designed to explore individual items of test that work as compared to other items of the whole test. Item analysis is also conducted to find out the level of difficulty of individual items on an achievement or another test.

There are many approaches to study item analysis. We shall discuss here the following approaches particularly dependent upon statistical methods :

9.5.1 An index of the item's difficulty.

The difficulty of an item is determined by so many ways. The following three ways are worth mentioning :

a) by the judgment of competent expert who, rank the items in order of difficulty;
b) by how quickly the item can be solved, and
c) by the number of experiences in the group who get the item right.

Thus, every individual who gets item 1 in an achievement test correct, we can say that item 1 is a good item. If no one gets item 1 correct, then, item 1 is not a good item. If every one gets the item right, the item is too easy. If everyone gets the item wrong, the item is too difficult.

There is no formula for determining the exact distribution of item difficulties. Thus, a common practice is to retain such items whose level of difficulty is 50 percent in terms of passing.

9.5.2 The item - Reliability Index

A statistical technique designed to provide an indication of a test's internal consistency, the higher the item - reliability index, the greater the test's internal consistency.

This index is equal to the item-score standard deviation (s) and the correlation (r) between the item score and the total test score.

9.5.3 Factor analysis and inter-item consistency

Factor analysis is a useful and a class of mathematical procedures used to reduce data. It is designed to find out variables on which people differ. There are two types of factor analysis. Exploratory factor analysis is a class of mathematical procedure applied to estimate factors, extract factors, or decide how many factors should be retained for final revision of test.

Whereas confirmatory factor analysis (CFA) is also a mathematical procedures used when a factor structure is tested for its fit with the observed relationship between the variables.

9.5.4 The item-validity Index

Item-validity Index is also a statistical technique which indicates the degree to which a test measures what it purports to measure. So higher the item-validity index, greater the test's criterion-related validity.

The item-validity Index can be calculated by either of the two following methods :

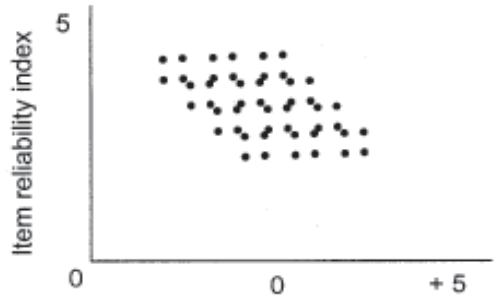
The item-score standard deviation. The correlation between the item-score and the criterion-score.

The item-score standard deviation of item 1 (denoted by si) can be calculated using the index of the item difficulty (P) in the following formula :

$$S_1 = [P_1(1 - P_1)]$$

However, the type of validation employed in a test construction may be content, construct, predictive, or a combination of these. The techniques used in item analysis for external validation criteria are correlation coefficients, expectancy tables, and standard errors of measurement. The standardisation sample of the population includes age range, sex, socioeconomic distribution, range of ability or trait variation, educational level, type of school. For separate validity findings for different age groups, grade groups, ability groups, clinical groups, culture and subculture groups and occupational groups are also necessary.

In short, the best items on a test can be achieved by plotting each item's item-validity index and item-reliability index shown below:



9.5.5 The Item - Discrimination Index

It is also a statistical method which is designed to indicate how adequate a test item separates or discriminates between high and low scores. The Item - Discrimination Index is a measure of item discrimination. Symbolized by letter d (d) as Kelley (1939) has pointed out that marked and significant discrimination between extreme groups is obtained when item analysis is based upon the highest 27% and lowest 27% of the group.

We can use this method to find out what percentage of the highest percent and what percentage of the lowest 27% passed each item, then by statistical calculation, to determine if the difference between the two percentages is significant. There is another method which can be used by applying with reference to high average, low average, and low groups, classification being based upon totaltest score or upon external validation criteria.

However, successes and failures on each item may be correlated with the total scores for whole test. When this is done, biserial correlation can be calculated.

9.5.6 Analysis of item alternatives

There is no formula or statistics for item alternative. For this purpose two groups are selected known as upper level (U) and lower level (L) of the distribution. We shall analyse the responses to five items, 1 and 2, a good item and poor item. Examples are given below:

Item 1 **Alternatives**

	a	b	c	d	e
U	29	3	2	0	3
L	10	5	6	6	5

Response pattern to item 1 indicates that the item is a good one. More U group members than L group members answered item correctly.

Item 2		Alternatives			
	a	b	c	d	e
U	19	0	0	5	13
L	7	0	0	16	9

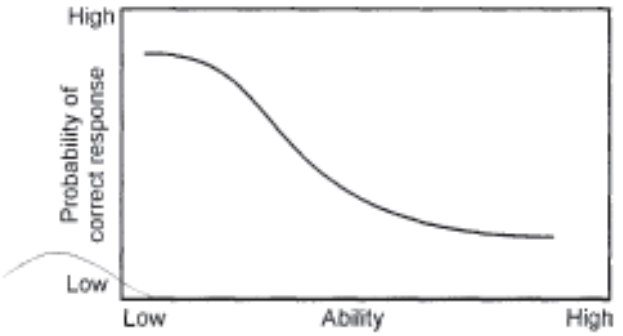
Item 2 is a poor item because more L group members than U group members answered the item correctly.

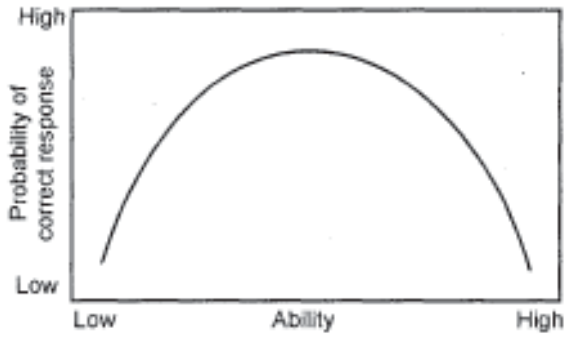
9.5.7 Item - Characteristics Curves (ICC)

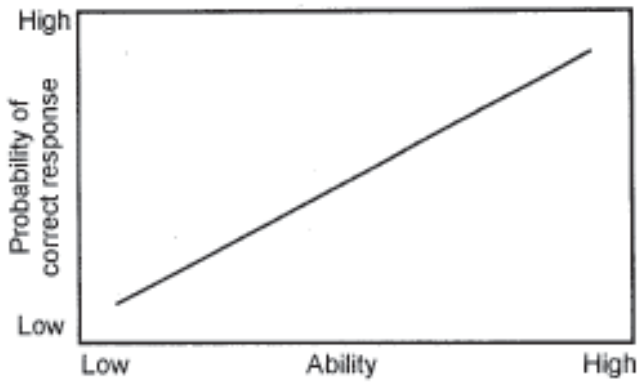
ICC also represents item difficulty and discrimination.

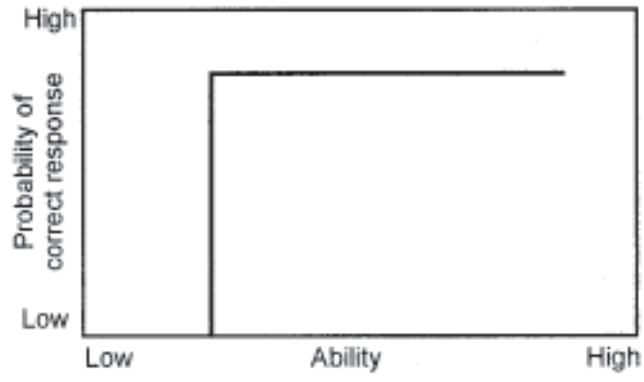
ICC is a graph on which ability is plotted on the horizontal axis and probability of correct response is plotted on the vertical axis. The slope of the curve shows the item discriminating high- from- low scoring. The slope is seemed to be positive, more high scorers are getting the item correct than low scorers.

Item C is a good item. Item D has excellent discriminative ability and will be useful in a test designed to select individuals for cutting off scores.









9.5.8 Item response theory (IRT)

Item Response theory is also referred to as latent - theory or the latent trait model. It is a system of assumptions about measurement of a trait and the extent to which each test item measures that trait. It is a true score model (Lord, 1980). It is widely used by commercial test developers and large-scale test publishers in test development.

An important model in this regard is offered by Rasch (2001) which explains that a person with x ability will be able to perform at a level of y. In other words, the probability of a person exhibiting a personality trait in x ability may exhibit the same trait of personality in y. According to Mitchell (1999), the Rasch model is more sophisticated model than classical test theory.

Latent - Trait Theory (LTT) is not widely used because of technical and complex issues. This theory provides us an estimate of the amount of knowledge, ability or strength of a test. Since this theory is uniclimensional, the test measures all the traits.

Latent - Trait Model (LTM) can be found in the Illness Causality Scale (ICS), a measure of children's understanding of illness (Sayer et al. 1993). This scale reveals three important latent traits which are labeled as verbal intelligence, level of cognitive development, and understanding of illness. This scale has been correlated with other scales and the results are quite convincing.

However, this model is criticised on the ground that it is not widely used when others are available and it is more technically complex.

Despite these objections, LTM plays an increasing dominant role in the development of new tests and testing programmes. But large testing firms, state agencies and district's school rely on IRT method to construct, analyse, and score major achievement, entrance, and professional licensure examinations ((Raise and Henson,2003).

9.5.9 Other Considerations in Item Analysis.

Guessing:

Guessing an item correctly by chance. is not always true. The guessing of a person on true/false choice item alone may be 5 out of ten items. The probability of guessing correctly on two such items is equal; to .5 2 or .25%. The probability of guessing correctly on ten such items is equal; to .510 /or .001. Therefore, one in thousand items can be found correct on ten true/ false on the basis of chance alone.

However, three criteria have been published with regard to the problem of guessing answer correctly.

1 . A respondent while guessing a correct answer does not make on random basis, rather he applies his knowledge of the subject matter and ability to rule out unfit alternative.
2. A correction for guessing also depends on the problem of omitted items. This problem is related with questions : Should the omitted item be scored " wrong "? or such item be handled in different ways.
3. The rule of guessing also depends upon "Luck by chance," but any correction for guessing may be underestimated or overestimated the effect for lucky and unlucky test users.

However, no criterion is found to be satisfactory for guessing items correctly but two reasons are seemed to be very reasonable

a) clear instructions given by examiner to examinees and

b) specific instructions for scoring and interpreting omitted items.

Thus, guessing responses correctly is not a complex problem, it depends on risk - taking behavior of test taker.

9.5.10 Items fairness

An item is considered fair when the different group members pass on any given item that measures that ability, regardless of race, social class, sex or any other background characteristics. In other words, the same proportion of persons from each group should pass any given item of the test, provied that persons earn the same total score on the test (Jenson, 1980, P.44).

9.5.11 Speed tests

Speed tests are tests in which the time limit imposed is so short that all examinees can not attempt all of the items in test. They are of low difficulty level. The reliability of speed test can be calculated by split-half-technique or the basis of odd-even split. The reliability coefficient of correlation may be closed to 1.00. parallel forms or test-retest are also methods to estimate reliability of speed tests.

For a good speed test, all items should be of uniform, or nearly uniform, degree of difficulty. The best practice is to correlate subtest scores and total - test scores with scores of criteria, under various time limits.

9.5.12 Qualitative Item Analysis (QIA)

QIA is general term for various non-statistical procedures designed to explore how individual test items work. The analysis of this procedure is compared with individual test items to each other and to the test as a whole.

There are some important topics which a researcher can explore for qualitative analysis. They are : cultural sensitivity, face validity, test administration, test fairness, test language, test length, test taker's preparation and so on.

Qualitative methods are techniques of data analysis through verbal means such as interview and group discussion. It is better to provide an opportunity to test takers and students to describe their instructors. If students fail to respond adequately for test items, they may be given chance to evaluate their performance. The other related aspect with qualitative item analysis is Think Aloud Test Administration.

9.5.13 " Think Aloud Administration

Different researchers use different procedures to respondents to verbalize thoughts as they occur (Davison, 1997; Hurlburt, 1997; Klinger 1978). This approach is employed by them for adjustment, problem solving, educational remediation and clinical intervention.

Cohen et al. (1988) have pointed out that "think aloud" test administration as a tool of QIA focuses on the thought process of test taker during the administration of a test.

For achievement test, verbalisations may be useful in assessing low or high scores and also why and how they are misinterpreting the items. And for personality test "Think aloud" device may provide insight to perceive, interpret, and respond to the items.

9.5.14 Expert panels

Expert panels also provide qualitative analysis of test items. These panels try to obtain an understanding of the history and philosophy of the test battery and to discuss and define the problem of bias (Stanford Special Report, 1992). Some of the content bias that have been identified are given as under:

(a) Status - the situations that do not involve authority.
(b) Stereotype - members of particular group show aptitude, interest, occupation and personality characteristics.
(c) Familiarity - groups know vocabulary and experiences about items.
(d) Offensive Choice of words - using of correct wording for items.
(e) Other - Panel members should be asked another indication of bias they detect.

On the basis of qualitative information from an expert panel, test users or developers may elect to modify or revise the test. However, rewording items, deleting items or creating new items is known as test revision, a final stage in the development of a new test. The process of revision is very expensive work. It requires a lot of efforts, time and expense.

Check your progress:

Q1 What are the tools a test developer employs to analyze and select items ?
Q2 Explain in brief:
i) The item - difficulty Index
ii) The item- reliability Index
iii) The item- Discrimination Index
Q3 Explain any two theories of item response theory.
Q4 Define: -
a) Item fairness
b) speed test
c) Qualitative method
Q5 Write brief note on expert panels and Think Aloud Test Administration.

9.6 TEST REVISION

Test revision is the last (fifth) stage of test construction. This is final form of a new test in which some items are eliminated and others will be rewritten.

9.6.1. Test Revision as a stage in New Test Development.

On the basis of information obtained from the item - difficulty, item analysis, item reliability, validity, discrimination and bias of test items, a test developer tries his best to revise the test if he feels adequate and necessary.

There are many ways or approaches which help developers in revision of final test newly constructed. One such approach is to characterise each item according to its strength and weakness. some items may be very easy or hard, highly reliable but lack criterion validity, some items may lack reliability and validity because of their restricted range. The same may be true of very easy items.

So if the test developer may find that he can't maintain balance between strength and weakness across items, then good items need to include more difficult items. Thus, the purpose of test revision will be affected. Further more, if highly skilled individuals are being tested, then the best possible test discrimination will be made a priority.

As revision proceeds, writing large items pool becomes clear. Thus, poor items can be eliminated and good items should be retained. By doing so, a test developer will come out of the revision stage with a better test. Now, the next step is to administer the revised test under standardised conditions to obtain a second appropriate sample of examinees. After administration of second draft of the test, the test developer may give a final touch. Once the test is in finished form, the test's norm may be developed from the data. Now the test will be said to have been "standardised".

9.6.2 Standardisation

Standardisation refers to " the process employed to introduce objectivity and uniformity into test administration, scoring, and interpretation ". (Robertson, 1990). A standardisation sample represents the group (s) individual with whom examinee's performance will be compared.

This sample must be representative of the population on those variables that may affect the performance. For example, in a ability test the standardisation group must represent the norms such as age, gender, geographical region, type -of community, ethnic group and educational level.

However, process of revision continues until the test is satisfactory and standardisation can occur. So once the test is ready, its validity requires for a cross-validation of findings. Before we discuss cross-validation, it is reasonable to briefly consider some issues related to the development of a new edition of an existing test.

9.6.3 Test Revision in the Life cycle of an existing Test.

There is no hard-and-fast rule that exists for revision of a test (APA, 1996. 3.18). It give us suggestion that a test can be kept in its present form as long as it remains "useful" and be revised when significant changes make the test useless for application.

However, a test can be revised if it meets any of the following conditions :

- When the current test takers can't correlate stimulus materials.
- When current test takers are not able to understand the verbal content of the test, including test administration, instructions and test items containing dated vocabulary.
- When a popular culture changes and the words take the new meanings and certain test items or directions seem to be inappropriate, thus the test needs to be revised.
- As the group membersip changes, the test norms are also seemed to be inadequate.
- Through revision, the reliability and validity can be improved significantly.
- As time passes, an age extension of norms may take a shift upward, downward, or in both directions, a change in test is necessary.
- The original theory on which test is based also need revision that reflect the design and content of the test.

As we have noted that revision takes place in all stages of test construction starting from conceptualisation phase, test tryout, item analysis and test revision. There are variety of tests and developed scales which has under gone revisions. For example, Strong Vocational Interest Bank, MMPI, Binet Tests of Intelligence, etc.

However, a key step in the development of all tests is crossedvalidation and co-validation.

9.6.4 Cross-validation and Co-validation.

The term cross-validation refers to the revalidation of a test on a sample of test'takers other than those on whom test performance has originally been found to be a valid predictor of some criterion. For example, a test's items are selected for a newly constructed test may be administered on first and second groups of sample. Suppose, an Indian test developer constructs a test in Indian 'population, he can find out its reliability and validity very easi[y- But when the same test is applied on foreign sample and obtained validity, this is called cross-validation.

However, the decrease in item validities that occurs by chance after cross-validation of findings is referred to as validity shrinkage.

Co-validation is defined as a test validation process conducted on two or more tests using the same sample of subjects. However, a test is used to revise existing norms, it is referred to as co-norming. Co-validation is useful for:

* Publishers because of its economical purpose;

- Collecting data by means of face-to-face or telephone interview; and

- multiple testing

Co-validation requires qualified examiners to administer the test, to assist in scoring, interpretation and statistical analysis. There are some tests which are used by publishers and test users. For example, Wechsler Adult Intelligence Scale (WAIS-111) and Wechsler Memory Scale (WMS-111) are used together in the clinical evaluation of an Adult. Since the two tests are normed on the same population, sampling error is greatly minimized, if not eliminated completely.

9.6.5 Quality assurance during test revision.

There is no mechanism of quality assurance that a test publisher can adopt in course of standardising a new test or restandardising an existing test. But for quality control, it is essential to- recruit examiners with extensive experience of testing the children and adults. They must be well off about their educational and professional qualifications, administration experience with various intellectual measures, certification, and licensing status. The selected examiners must be very familiar with childhood assessment practices (Wechsler, 2003).

In nut-shell, every examiner must be having a doctoral degree. Regardless of educational qualifications or experience, all examiners must be trained enough to handle the process of test construction and statistical procedures of test administration, scoring, interpretation, and process of revision. They may be involved to greater degrees in the final scoring of protocol.

For quality assurance of WISC-IV two trained and qualified scorers were appointed during national tryout and standardisafio'n stage of WISC-IV test development.

Another mechanism for ensuring quality assurance is an anchor protocol. It is test protocol scored by a highly authoritative scorer, who resolves scoring discrepancies , if any existing.

However, when discrepancy exists between scoring in an anchor protocol and for another protocol is referred to as scoring drift. Anchor protocols were used for quality assurance in the developments of the WISC-IV. For quality assurance we have computer programmes to find out and identify any error in score reporting.

Check your progress

Q1. Explain any one approach of test revision.

Q2. Explain the nature and uses of standard isation.

Q3 What are the conditions under which a test can be revised ?

Q4 Define the following terms
a) Cross-validation.
b) Co-validation.
c) An anchor protocol.
d) Scoring drift.

9.7. SUMMARY

In this chapter we heive highlighted on creation of a good item, the basics of test development and the. process by which tests are constructed. We have also discussed a number of techniques designed for construction and selection of good items. We have also focused on five stages of the process of development. These stages are test conceptualisation, test construction, test tryout, item analysis and test revision.

While discussing the process of test conceptualisation, we have explained some preliminary questions confronted by test developers. Under this topic we have also pointed out item development issues, concept of pilot work or pilot study.

The second stage, test construction, is also dealt with various important instruments relating to test construct. These instruments are scaling, methods of scaling, types of scaling, writing items, items formats, writing items for computer administration and different test scoring models.

In test tryout stage we have focused on the nature of a good test, followed by item analysis.

In fourth stage we have revealed the nature and uses of various test tools which a test developer has to adopt. These tools are an index of item's difficulty, item's reliability, item's validity and item discrimination. To provide a detailed explanation of all these indexes, we have discussed item response theories too, including speed test, test administration and appointment of expert panels.

And the final stage is discussed through highlighting on various ways of revision of new test development, standardisation of test, test revision in the life cycle of an existing test, the nature and uses of cross-validation and co-validation, quality assurance during test revision and so on.

9.8 QUESTIONS

Q1. Explain any two stages of Test development.
Q2. Define scaling and describe the various types of scaling method.

Q3. Explain in brief
a) Multiple- choice format and matching test
b) Speed test and expert panel
c) Test revision
d) Likert scale
e) Thurstone scale
f) Tools of Test development

Q4. Define the following terms:
i) CAT
ii) Rasch model
iii) Think aloud test administration
iv) Item Pool
v) Item branching
vi) Anchor protocol
vii) Item - Discrimination Index
viii) Item - difficulty Index
ix) Cross -validation

9.9 REFERENCES

Anastasi,, L. R. (1937) -
A Hand Book of Psychological Testing (7th edi) Indian Reprint, 2002

Campbell, D.P. (1972) -
The practical problems of revising on established psychological test. In J.N. Butcher (Ed) Objective Personality Assessment :
Changing Perspective (pp. 117-130) Newyork Academic Press.

Freeman, F. S. (1962) - A Hand Book on Theory and Practice of Psychological Testing (6th Ed), Oxford and IBH Publishing Co. Bombay.

Cohen, J.R. and-Swerdlik, M.E. (2010) -
Psychological Testing and Assessment ; An Introduction to Test and Measurement. (7th ed), Newyork, McGraw- Hill international edition.

Guttman, L. A. (1999) A Basis for Scaling Qualitative Data. American Sociological Review, 9,179-190

Guttman, L. A. (1997) -
The Cornell Technique for Scale and Intensity Analysis. Education and Psychological Measurement, 7, 297-280.

Jensen, A. R. (1980)
Bias in Mental Testing, Newyork Free press.

Katz, R. C. and Lonero, P. (1999)
Findings on the Revised Morally Debatable Behavioural scale.
Journal of Psychol. 128,15-21.

Likert, R. (1932)
A Technique for Measurement of Attitude.
Archives of Psychol. Number 190.

Mitchell, J. (1999)
Measurement in Psychol : Critical History of a Methodological concept.
Newyork : Cambridge university press.

Rasch, G. (2001)
Applying Fundamental Measurement in Human Science
chapter - 2 Mahwah, N.J. Erlbaum.

Reise, S. P. and Henson, J. M. (2003)
A Discussion of Modern versus Traditional Psychometrics as Applied to Personality Assessment Scales Journal of Personality Assessment, 81. 93-103.

Robertson, G. J. (1990)
A practical model for test development Hand Book psychological and educational Assessment of children. pp, 62-85, New York Guilford.

Rokeach,, M. (1973)
The nature of human values. Newyork. Free Press.

Sayer, A. G. and Perrin, E.C. (1993)
Measuring understanding of illness causality in healthy children and in children with chronic illness. A construct validation. Journal of applied developmental Psychol. 19, 11-36.

Thurstone, L. L. (1929)
Theory of attitude measurement, Psychological Bulletin, 36, 222-291.

Thurstone, L. L. (1932) -
Multiple Factor analysis. Chicago : university of
Chicago Press

Wechsler, D. (2003)
WISC - iv, Technical and interpretive manual, 9th (Ed) San Antonio, Tx. Psychological corporation.

9.10 GLOSSARY

Age- based scale- When the test taker tests the performance of individual as a function of age, it refers to as age-based scale.

Anchor Protocol - A test answersheet developed by a test publisher to check the accuracy of examiner's scoring.

Conceptual ization - It refers to as a novel idea for a test to be conceived.

Construction of a test - The items for the test drafted refer to as construction of a test

Criterion referred test - is a method of evaluation or a way of deriving meaning from test score by evaluating an individual's score with reference to a set standard.

Comparative scaling - when the judgment of one stimulus is compared with every other stimulus on the scale.

Cumulative scoring model - A method of scoring whereby points or scores accumulated on individual items or subtests are tallied, the higher the total sum, the greater the individual is presumed to be on the ability, trait or other characteristics.

Confirmatory factor Analysis - is also a mathematical procedure which is used when a factor structure is test for its fit with observed relationship between th variables.

Cross-validation - refers to as the revalidation of a test on a sample of testtakers other than those on whom test performance has originally been found to be a valid predictor of criterion.

Co-validation - A test validation process conducted on two or more tests using the same sample of subjects (testtakers)

Exploratory factor analysis - A class of mathematical procedure applied to estimate factors, extract factors, or decide how many factors should be retained for final revision of test.

Factor - analysis - is a class of mathematical procedure used to reduce data and seek out variables on which people differ.

Grade based scale - A system of measuring performance of individuals as a function of grade.

Item - analysis - refers to as various procedures designed to explore how individual test item works as compared to other items in the test and in the context of whole test.

Item - bank - a collection of questions to be used in the construction of test.

Item - branching - The individualized presentation of test items drawn from an item-bank based on test's user's previous responses.

Item - Characteristics Curve (Icc) - A graphic representation of item-difficulty and item - discrimination.

Item- difficulty index - A statistic indicating how many individuals or testtakers respond correctly to an item.

Item - discrimination Index - A statistic designed to indicate how adequately a test item separate or discriminate between high scorers and low scorers.

Item- fairness - An item is considered fair when the different group members pass on any given item that measures the ability.

Item - reliability index - A statistic designed to provide indication of a test's interval consistency, the higher the item reliability index, the greater the test's interval consistency.

Ipsative scoring - A method of scoring used to compare of a test taker's score on one scale within a test with another scale within that same test.

Item response theory (IRT), refers to as a laten- theory or the latent - trait model, is a system of assumptions about measurement of a trait and the extent to which each test item measures that trait.

Item - validity index - A statistical technique used to indicate the degree to which a test measures what it purports o measure, the higher the item - validity index, greater the test-criterion- related validity.

Interval scale - is a system of measurement that contains equal intervals between members. Each unit in scale exactly equal to any other unit on the scale.

Item Pool - The reservoir or well from which items on the final version of test will be drawn or discarded.

Item - format - refers to as variables such as the form, plan, structure, arrangement, and lay-out of individual test items.

Nominal scale - A system of measurement in which all things measured and classified or categorized based on one or more distinguishing characteristics and placed into mutually exclusive and exhaustive categories.

Norm - referenced testing - A method of evaluation and way of deriving meaning from test scores by evaluating an individual testtaker's score and comparing it to scores of a group of testtakers.

Paired - Comparison method - A method in which a testtaker selects one stimulus out of two (may be two photographs, two objects, or two statements) according to his inerest.

Pilot work - refers to as pilot study or pilot research, is generally meant to the preliminary administration of a test on selected sample before final administration.

Protocol - The form or sheet on which testtaker's responses are entered, a method of scoring or procedure for evaluation.

Qualitative item analysis - A statistical procedure designed to explore how individual test items work.

Quantitative methods - are techniques of data analysis through verbal means such as interviews and group discussion.

Ratio scale - It is a system of measurement in which all things measured can be put in rank-order and equal intervals exist between number on the scale.

Scaling - A process of setting rules for assigning numbers in measurement.

Stanine scale - refers to as a system of measurement that all raw scores of a test are transformed into scores, ranging from 1 to 9 points.

Standardisation - refers to as "the process employed to introduce objectivity and uniformity into test administration, scoring, and interpretation".

Try out - When the first draft of the test is applied on a group of sample testtaker is called as try out.

Validity - shrinkage - The decrease in item validation that occurs by chance after cross-validation of findings is referred to as validity shrinkage.

10

Measurement of Intelligence

Unit Structure

10.0 OBJECTIVES

After studying this unit you should be able to:

i. Understand the various definitions of intelligence given by the lay public as well as scholars and test professionals.
ii. Comprehend the various theories of intelligence.
iii. Know the process of measurement of intelligence and types of tasks involved in intelligence tests as well as theory in intelligence test development and interpretation.
iv. Understand the Stanford-Binet Intelligence scales as well as Weschsler tests.

10.1 INTRODUCTION

In this unit we will discuss about the definition of intelligence and the various theories of intelligence. Among the definitions of intelligence, we would examine the views of the lay public as well as the views of scholars and test professionals. A few theories of intelligence would also be examined. The most important theories of intelligence include the Factor-analytic theories and the information processing views. Following this we would study about measuring intelligence and issues related to it.

The Stanford- Binet Tests and the Weschsler tests are an important group of tests. Among the Stanford Binet Tests we would discuss the 05 th edition of Stanford-Binet Intelligence Scales in brief. We would also discuss few Weschsler briefly discuss the short forms of few of these tests. Towards the

end of the unit we provide a summary of the unit, followed by questions and a list of references for further study.

10.2 WHAT IS INTELLIGENCE? - DEFINITIONS AND THEORIES

Intelligence is a multifaceted capacity that is expressed in many ways. Intelligence means many things to different people including layman, scholars and psychological test professionals. The following various abilities are said to constitute intelligence.

- Ability to acquire and apply a specific knowledge.
- Ability use the right words and to generate quick thoughts in given event.
- Ability to reason well, judge well and to be self-critical
- Ability to plan and foresee things and events
- Ability to grasp, visualize concepts.
- Ability to make good judgments to solve problems efficiently and economically
- Ability to infer perceptively
- Ability to comprehend people, events and situations.
- Ability to pay attention to minute details and to be intuitive and innovative
- Ability to cope, adjust, adapt to new situations and culture.
- Ability to be practical and street smart and get one's work done.

10.2.1 Intelligence Defined: Views of the Lay Public:

Sternberg and his associates have done considerable work with respect to how lay people conceptualize intelligence. According to Sternberg, non-psychologists conceptualize intelligence as follows:

- Reasons logically and well
- Reads widely
- Displays Gommonsense
- Keeps an open mind and
- Reads with high comprehension

Non-psychologists view about unintelligence is reflected in the following statements:

- Does not tolerate diversity of views
- Does not display curiosity and
- Behaves with insufficient consideration of others.

According to Sternberg, non-psychologists and experts consider intelligence, in general, as follows:

(a) Practical problem solving ability
(b) Verbal ability and
(c) Social competence

Sternberg found that there was considerable degree of similarity between the experts and lay person's conceptions of intelligence. The following table lists as to what both meant by academic intelligence and everyday intelligence.

Academic Intelligence	Everyday Intelligence
• Verbal Ability • Problem Solving Ability • Social Competence • Specific behaviours associated with acquiring academic skills, such as studying hard.	• Practical problem solving ability • Social Competence • Character • Interest in learning and Culture

The major difference between the lay person's definition and expert's conceptualisation of intelligence was with respect to academic intelligence. Experts emphasised on the motivation whereas lay persons stressed on interpersonal and social aspects of intelligence.

Some researchers have also found that intelligence is a function of stages of development as can bee seen from brief description below:

* **Infancy -** During this period of development, intelligence was associated with physical coordination, awareness of people, verbal output and attachment.

* **Childhood -** During this period of development, intelligence was associated with verbal facility, understanding and characteristics of learning

* **Adulthood -** During this period of development, intelligence was associated with verbal facility, use of logic and problem solving.

It has been observed that children develop notions of intelligence when they reach the first grade. Young children's concept of intelligence emphasise interpersonal skills, such as being poite, acting nice, being helpful and good to others. Older childen conceive of intelligence as involving academic skills, such as reading well.

10.2.2 Intelligence Defined: Views of Scholars and Test Professionals:

The definition of intelligence is one area where psychologists most disagree with each other. There is no unanimity among them. Spearman (1927) remarked "In truth, intelligence has become a word with so many meanings that finally it has no". Similarly, decades later Wesman (1968) concluded that "There appears to be no more general agreement as to the nature of intelligence or the most valid means of measuring intelligence today than was the case 50 years ago." It is also relevant to note the statement of Edwin G. Boring (1923) that he made decades ago with respect to what intelligence is. According to him "intelligence is what the tests test". Some important definitions of intelligence given by experts are as follows:

(1) Francis Galton: Galton is the pioneer of intelligence movement. He was a contemporary of Alfred Binet and his work on heredity and intelligence is remarkable. He wrote extensively on heritability of intelligence.

According to Galton, intelligent people had excellent sensory abilities. According to him intelligence tests were nothing but measures of sensory abilities, Hence, experts influenced by his views developed tests of visual acuity and hearing ability. Galton measured intelligence through various sensori motor and other perception related tests.

(2) Alfred Binet: He did not give an explicit definition of intelligence, though he believed that there are certain components of intelligence. According to Alfred Binet, intelligence is made up of the following components: Reasoning, Judgement, Memory and Abstraction.

According to Alfred Binet, "intelligence is the capacity of an individual to reason well to judge well and to be self critical."

Binet and his colleague, Henri attempted to assess complex measures of intellectual ability. Binet was interested in measuring intelligence because he was faced with the practical problem of identifying intellectually limited school children in schools of Paris who could be benefited from regular instructional programme and may require special educational experiences.

(3) David Wechsler: David Wechsler is another,pioneer in the measurement of intelligence whose work in the 1950s and 1960s led to development of several tests of intelligence.

According to him "Intelligence can be operationally defined as the aggregate or global capacity of an individual to act purposefully, think rationally and deal effectively with the environment".

According to Wechsler, intelligence is composed of abilities which though not independent, but are qualitatively different.

Intelligence, according to him is not the mere sum of different abilities. According to Wechsler, in the assessment of intelligence, even the non-intellectual aspects must be taken into consideration. Some important non-intellectual factors that influence intelligence and its assessment are as follows:

- Drive
- Persistence
- Goal awareness
- Individual's potential to perceive and respond to social, moral and aesthetic values and
- General personality of an individual.

He viewed intelligence as constituting of four factors: Verbal Comprehension, Working Memory, Perceptual Organisation- and Processing Speed.

(4) Jean Piaget: Jean Piaget was a Swiss developmental psychologist whose work on intelligence among children has been very influential. Jean Piaget studies development of cognition in children. He specifically studied as to how children think? How they understand themselves and the world around them and how they reason and solve problems.

Jean Piaget conceived of intelligence as a kind of evolving biological adaptation to the outside world. Piaget viewed that intelligence is a result of joint interaction of biology and environmental forces. He viewed intelligence as a cognitive ability consisting of four stages. Individuals move through these four stages at different rates and ages. According to Piaget, biological aspects of mental development is governed by inherent maturational mechanisms. Experiences at each stage of cognitive development helps to organize and reorganize the mental structures (also called as schemas). Piaget used the word schema to refer to an organised action or mental structure, that, when applied to the world, leads to knowing or understanding. The Plural of schema is schemata. In initial years schema is tied to simple behaviour such as sucking and grasping. As they grow older, schemata becomes more complicated and are tied less to overt actions than to mental transformations.

Piaged conceived of learning as occurring through two basic mental operations: assimilation and accommodation. Assimilation is defined as actively organising new information so that it fits in with what is already perceived and thought. Accommodation is defined as changing what is already perceived or thought so that it fits with new information. The following table lists the Piaget stages of cognitive development.

Sensori Motor	Birth to 24 months
Preoperational	2 to 7 years

Period of concrete operations	7 to 11 years
Period of formal operations	7 to 11 years

A common theme among all the above mentioned researchers on intelligence is their focus an interactionism. According to interactionism viewpoint intelligence is influenced by and is a result of joint interaction of heredity and environment.

10.2.3 Theories of Intelligence: Factor-Analytic Theories of Intelligence and Information Processing Theories of Intelligence

10.2.3.1 Factor-Analytic Theories of Intelligence:

The focus of factor analytic theories is on identifying the ability or groups of abilities deemed to constitute intelligence.

Factor analysis is one statistical method. It consists of a group of statistical techniques that is designed to determine the existence of underlying relationships between sets of variables including test scores. The method of factor analysis has been used to study correlations between tests measuring varied abilities presumed to reflect intelligence.

Charles Spearman: It was Charles Spearman who pioneered the methods of determining Inter-correlations between tests. Charles Spearman developed the theory of general intelligence which is also called as the two factor theory of intelligence. He found that measures of intelligence tended to correlate to various degrees with each other.

According to Spearman, all intellectual activity is dependent primarily.upon and is an expression of a general factor common to all mental activities. It is designated by the symbol 'g'. This factor enters into various degrees in various proportions and different individuals possess this factor in different degrees. All activities do not make an equal demand for this factor. Thus, a task, which requires more 'g' factor, will be done poorly by an individual who has less of such factor. Spearman also pointed out that this factor can only be indirectly observed by constructing psychological tests. Hence, according to him the aim of psychological testing should be to measure the amount of an individual's 'g' factor, because it is this factor that provides the only basis for prediction of subject's performance from one situation to another.

A test that is highly loaded with 'g' factor requires insight into relationship e.g., reasoning test. On the other hand, a test that is not highly loaded with 'g' factor does not require insight into relationship. E.g., of such a test is a test involving mechanical or f4rote" type of learning.

According to Spearman, a test that exhibited high positive correlations with other intelligence tests were thought to be highly saturated with 'g'. On the other hand, tests with low or moderate correlations with other intelligence tests were thought as a means of possible measure of 's' factors. (e.g. visual acuity, motor ability, etc.)

A test that is highly loaded with 'g' factor is better able to measure and predict intelligence.

According to Spearman, it is the 'g' factor rather than the 's' factor that is the better predictor of overall intelligence.

The best measure of 'g' factor in test of intelligence were items dealing with abstract-reasoning.

Spearman also noted that between the 'g' factor and the 's' factor were intermediate class of factors common to a group of activity which were called as group factors. These group factors include the following:

- Linguistic abilities
- Mechanical abilities
- Arithmetical abilities

Some researchers have found highly specific factors. One such group of g factors is listed below:

Factors	Description of Factors
R	Ability to repeat a chain of verbally presented materials
R1	Ability to process sound
R2	Ability to retain verbally presented stimuli
R3	Speed of processing verbally presented stimuli

Many Multiple factor models of intelligence has been proposed. Thes include the work of Thurstone, Guilford, Gardner, Raymond Cattel, Carroll, etc. We would discuss about each of them.

a. Thurstone's work : Louis Thurstone found that intelligence is not composed of one or two factors such as "g" or "s" factor as proposed by Spearman, but it is composed of many factors, which he called it as Primary Mental Abilities. Thurstone identified seven factors, which he used to construct Test of Primary Mental Abilities. Though revised versions of this test are frequently used, the predictive power of this test has been questioned on a number of grounds.

b. Guilford's work : Guilford was of the view that intelligence is composed of 150 different abilities. He de-emphasised the importance of g factor.

c. Gardner's work : Gardner also believed that there are different types of intelligence requiring different abilities and different areas of the brain controlled different types of ability. The seven different and independent types of intelligence identified by Gardner are as follows :

(1) Interpersonal (social skills).
(2) Intrapersonal (personal adjustment).
(3) Spatial (artistic).
(4) Log ica I-mathernatica 1.
(5) Linguistic (verbal).
(6) Musical.
(7) Kinesthetic (athletic).

Gardner's descriptions of interpersonal and intrapersonal intelligence have found expressions in the concept emotional intelligence.

d. Raymond Cattell : Raymond Cattell (1941) and later on Horn (1966) postulated the two major types of cognitive abilities which they called as :

(a) Crystallized
(b) Fluid Intelligence

Crystallized Intelligence (GC) consists of acquired skills and knowledge that are dependent on exposure to a particular culture as well as formal and informal education. Crystallized intelligence consists of the following:

• Retrieval of information
• Application of general knowledge

Fluid Intelligence (GF): It is made up of abilities that are:
• Nonverbal
• Culture free
• Independent of specific instructions

Over the years Horn has proposed several additional factors which include as follows:

- Visual Processing (Gv)
- Auditory Processing (Ga)
- Quantitative Processing (Gq)
- Speed of Processing (Gs)
- Facility with Reading and Writing (Grw)
- Short Term Memory (Gsm)
- Long Term Storage and Retrieval (GIr)
- He divided these abilities in to two broad groups:

Vulnerable abilities: These abilities, such as Visual Processing (Gv), decline with age and tend not to return to, preinjury levels following brain damage. Other abilities such as, Quantitative Processing (Gq), are called as **Maintained abilities**, they tend not to decline with age and may return to preinjury levels following brain damage

e. Three stratum theory of cognitive abilities: This theory was proposed by Carroll (1997) and is based on factor analytic approach. It is a hierarchical model, meaning that all the abilities listed in the stratum are subsumed by or incorporated in the strata above. According to Carroll the three stratums are as follows:

(1) The top stratum, consists of "g" factor or general intelligence.

(2) The second stratum, consists of eight abilities and processes as follows:

(i) Fluid intelligence (Gf)
(ii) Crystallized intelligence (Gc)
(iii) General memory and learning (Y)
(iv) Broad visual perception (V)
(v) Broad auditory perception (U)
(vi) Broad retrieval capacity (R)
(vii) Broad cognitive speediness (S)
(viii) Processing / decision speed (T)

(3) At the third stratum are the "level factors" and / or "speed factors". Each of these is different and based on their linkage to second stratum. The three factors linked to Fluid intelligence (Gf) include, general reasoning, quantitative reasoning and piagetian reasoning. A speed factor linked to Gf include speed of processing. Similarly four factors linked to Crystallized intelligence (Gc) include: language development, comprehension, spelling ability and communication ability. Two speed factors linked to Gc are oral fluency and writing ability.

The CHC Model: Some researchers, using factor analysis and other statistical methods have attempted to extract, blend and combine the various factors creating more complex models. One such model is called as the Cattel-Horn-Carroll (CHC) model. An integration of Cattell-Horn and Carroll models was proposed by Kevin S McGrew (1997) as well as by McGrew and Flanagan. These models were developed to improve the practice of psychological assessment in education.

10.2.3.2 Information Processing Theories of Intelligence: The focus of information processing theories is on identifying the specific mental processes that constitute intelligence. Many Information Processing view of intelligence has been developed.

Luria's approach : Russian neuropsychologist Aleksandr Luria (1966) emphasized on the information processing approach to intelligence. His approach focuses on the mechanisms by which information is processed. He focuses on how information is processed rather than what is processed. The two basic types of information processing style that he emphasised includes

(a) Simultaneous (parallel) processing
(b) Successive (sequential) processing

In simultaneous processing, information is integrated all at one time, whereas in successive processing, each bit of information is individually processed in sequence. Successive (sequential) processing is logical and analytic in nature. According to Luria, memorising a telephone number or learning the spelling of a new word is typical of the type of tasks that involve acquisition of information through successive processing.

Simultaneous processing is information processed, integrated and synthesized at once as a whole. For e.g., when one is observing and appreciating a painting in an art museum, the information conveyed by the painting is processed in a manner that, atleast for most of us, could reasonably be described as simultaneous.

Information processing perspective is evident in the intelligence test developed by Kaufman Assessment Battery for Children. Information processing view is also evident in the PASS model of intellectual functioning developed by Das (1972) and Naglieri (1990). The word PASS is an acronym for Planning, Attention (Arousal), Simultaneous and Successive. In the PASS model the term:

* Planning refers to strategy development for problem solving.
* Attention, also called as arousal refers to receptivity of information.
* Simultaneous and successive refers to the type of information processing employed.

One important test that has been developed to assess PASS is called as the cognitive Assessment System.

Robert Sternberg: Robert Sternberg has developed another information processing approach to intelligence. According to Sternberg the essence of intelligence is that it provides a means to ,govern ourselves so that our thoughts and actions are organised, coherent and responsive to both our internally driven needs as well as the needs of our environment. According to Robert Sternberg intelligence is influenced by three main factors. They are: Context, Experience and basic information processing mechanism.

Sternberg is well known for his Triarchic theory of intelligence which can be divided into three parts as follows:

(a) The componential part, which is basically concerned with, processing or cognitive processing.

(b) The experiential part, which is basically concerned with the processes by which experience influences intelligence. It deals with the effect of experience on one's intelligence.

(c) The contextual part which is basically concerned with the effect of one's culture and environment on one's intelligence. The componential part of this theory is highly developed. According to it there are three types of components.

(i) Knowledge acquisition component: This component of intelligence is concerned with an individual's ability to learn new information.

(ii) Performance component: This component of intelligence is concerned with knowing how to solve specific problems.

(iii) Metacomponent: This component of intelligence is concerned with solving problem in general or learning general ways to approach problem solving. It is the metacomponent, which will help us to distinguish between the more intelligent and the less intelligence individual.

Each of these three components overlaps to a great extent and operate in a collective and integrative manner, rather than each operating independently. During the process of problem solving, each of these three operates together.

10.3 MEASURING INTELLIGENCE

Measurement of intelligence involves sampling an examinee's performance on different types of tests and tasks as a function of developmental level. Two important topics related to measuring intelligence include:

a) Types of Tasks used in Intelligence Tests
b) Theory in Intelligence Test Development and Interpretation

We would discuss both these topics briefly.

a. Types of Tasks used in Intelligence Tests: An important issues related to types of tasks used in intelligence tests are as to whose intelligence are we measuring; Intelligence of infant, children or adult. Different types of tasks are used in the measurement of intelligence of these three groups of people

Measuring Infant's Intelligence: In infancy (i.e. from birth to 18 months) measuring intelligence consists of measuring sensori motor development. The measurement of sensori motor development consists of the following activities.

_ Turning over
_ Lifting one's head
_ Sitting up
_ Following a moving object with the eyes
_ Imitating gestures
_ Reaching for a group of objects.

In measuring and assessing intellectual ability of infants, examiner must be highly skilful in establishing and maintaining rapport.

Parents, guardians and caretakers are an important source of information about activities of children. Structured interviews taken with them will help us to accurately assess the intelligence of infants.

Measuring Intelligence of Children: Measuring intelligence of children is also a highly skilful task. Children's intelligence is measured by assessing their verbal and performance abilities. Their intelligence is assessed by evaluating the following:

(a) General fund of information
(b) Vocabulary
(c) Social judgment
(d) Language
(e) Reasoning
(f) Numerical concepts
(g) Auditory and visual memory

(h) Attention
(i) Concentration and
(j) Space visualization

In the earlier period, intelligence tests were scored and interpreted with reference to mentatage. Mental age can be defined as an index that refers to the chronological age equivalent to one's performance on tests or a subtest. The index of mental age was typically derived by reference to norms that indicate the age at .which most test takers are able to pass or otherwise meet some criterion performance.

Besides performance on various measures, examiners also note the nonverbal and other behaviour of the children who are taking intelligence test. Verbal and nonverbal behaviour of the children who are taking the test of intelligence can yield considerable information about their performance and this can help, the examiner to interpret the obtained results on the test.

Measuring Intelligence of Adult: Wechsler pioneered in measurement of adult intelligence. According to him, adult intelligence scales should be able to measure the following abilities:

(1) Retention of general information
(2) Quantitative reasoning
(3) Expressive language
(4) Memory and
(5) Social judgment

It should be remembered that tests of intelligence are seldom administered to adults for purposes of educational placement. They are generally given to obtain clinically relevant information or for the measurement of learning potential and skill acquisition. Adult intelligence helps us to assess: faculties of an impaired individual (i.e., whether an individual is senile, traumatised or otherwise impaired) for the purpose of judging that person's competency to make important decisions.

b) Theory in Intelligence Test Development and Interpretation: Measurement of intelligence is also influenced by as to what we mean by intelligence. Galton viewed intelligence to be made up of senori motor and perceptual abilities and hence he devised tests to measure senory motor and perceptual differences between individuals. On the other hand Binet as well as Spearman developed formal theories to assess intelligence. Spearman emphasised the universal unity of the intellective function with g as its centerpiece. David Wechsler wrote extensively on the intelligence and viewed intelligence as multifaced and conceived intelligence as not only made up of cognitive abilities but also factors related to personality. Thorndike conceived of intelligence in terms of three clusters of ability: Social intelligence (dealing with people), Concrete intelligence (dealing with objects)

and abstract intelligence (dealing with verbal and mathematical symbols). Factor analytic theories and various other theories have been used in the process of test development and interpretation.

10.4 THE STANFORD-BINET INTELLIGENCE SCALES

After Binet's death in 1911, Binet tests underwent more and more revision, especially in America. The most widely used revision of this test came to be called as Stanford-Binet revision or test, because these tests were revised under the direction of Prof. L.M. Terman at Stanford University.

The following are the important revisions of this test.

(1) 1916: Stanford Revision and extension of the Binet-Simon scale by Lewis M. Terman.
(2) 1937: Revised Stanford-Binet tests of intelligence (Forms L and M) by Lewis M. Terman and Maud A. Merrill.
(3) 1960: Stanford-Binet, third edition, Form L-M, by Lewis M. Terman and Maud A. Merrill.
(4) 1972: Stanford-Binet, Form L-M (renoming) by Lewis M. Terman, Maud A. Merrill and Robert L. Thorndike.
(5) 1986: Stanford-Binet Intelligence Scale, 4th edition by Robert C. Thorndike, Elazebeth P. Hagen, and Jerome M. Sattler.
(6) 2003 : Stanford-Binet Intelligence 5th edition by Gale H. Roid.

The first edition of the Stanford Binet Scales (S-B Scales) was not without major flaws. There was lack of representativeness of the standardization sample. The first edition of the S-B Scales was the first published test to provide organsied and detailed administration and scoring instructions. It was also the first American test to employ the concept of IQ and the first test to introduce the concept of alternate form.

The 1937 revision of this test began in 1926 and it took 11 years to complete this revision. This revision had two equivalent forms form L (for Lewis) and M (for Maude Merrill who was one of the coauthors of the 02nd edition of this test). New types of tasks for use with preschool-level and adult-level test takers were developed. The manual contained many examples to aid the examiner in scoring. One important criticism of this edition was the lack of representation of minority groups during the test's development.

The test was again revised in the year 1960. This test had only one form and included the items considered to be the best from the two forms of the 1937 test. A major innovation of the 1960 revision was the use of concept of Deviation IQ. The second innovation of this scale was that the IQ tables have been expanded to include chronological ages 17 and 18 because latest finding indicated that mental development, as measured by Stanford-Binet, continues at least up to that age.

Another revision of the S-B Scales was published in the year 1972. This scale was criticised for the quality of its standardization sample. Its manual was vague with respect to the number of minority individuals in the standardization sample. It is also said that this scale over represented the western urban communities in the standardization sample.

The fourth revision of S-B Scales appeared in 1986 and constituted a major departure from the previous versions with respect to theoretical organization, test organisaiton, test administration, test scoring and test interpretation. As opposed to the earlier three revisions of the scale which were age scales. The fourth edition was a point scale. A point scale is a test organised in to subtests by category of items, not by age at which most test takers are presumed capable of responding in the way that is keyed as correct. The manual of the fourth edition contains an explicit exposition of the theoretical model of intelligence that guided the revision. The model was based on the Cattell-Horn model of intelligence.

The Stanford-Binet intelligence was once again revised in 2003. It is the fifth revision and is also called as the SB5. It is an individually administered assessment of intelligence and cognitive abilities and is suitable for people in the age range of 2 years to 85 + years. The SB5 blends many of the important features of the earlier editions with significant improvements in psychometric design. It provides comprehensive coverage of five factors of cognitive ability:

- Fluid reasoning
- Knowledge
- Quantitative processing
- Visual-spatial processing
- Working memory

The new features of the SB5 are as follows:

- Wide variety of items requiring nonverbal performance by examinee ideal for assessing subjects with limited English, deafness, or communication disorders
- Ability to compare verbal and nonverbal performance useful in evaluating learning disabilities

- Greater diagnostic and clinical relevance of tasks, such as verbal and nonverbal assessment of working memory
- Includes Full Scale IQ, Verbal and Nonverbal IQ, and Composite Indices spanning 5 dimensions -with a standard score mean of 100, SID 15
- Includes subtest scores with a mean of 10, SID3
- Extensive high-end items, many adapted from previous Stanford-Binet editions and designed to measure the highest level of gifted performance
- Improved low-end items for better measurement of young children, low functioning older children, or adults with mental retardation
- Enhanced memory tasks provide a comprehensive assessment for adults and the elderly.
- Co-normed with measures of visual-motor perception and test-taking behaviour
- Scoreable by hand or with computer software.
- Enhanced artwork and manipulatives those are both colourful and child-friendly.

Uses : The SB5 may be used to diagnose a wide variety of developmental disabilities and exceptional ities and may also be useful in :

- Clinical and neuropsychological assessment.
- Early childhood assessment.
- Psycho-educational evaluations for special education placements.
- Adult workers compensation evaluations.
- Providing information for interventions such as IFPs, IEPs, career assessment, industrial selection, and adult neuropsychological treatment.
- A variety of forensic contexts.
- Research on abilities and aptitudes.

Some important points worth noting about S135 are as follows:

(i) The S135 is based on the Cattell-Horn-Carroll (CHC) theory of intellectual abilities.

(ii) S135 has good reliability data, interscorer reliability ranges between 0.74 to 0.97 with a median of 0.90.

(iii) Only a few subtest items of SB5 are timed. Most of the S135 items are not timed.

(iv) Normative data for the S135 were gathered from 44,800 individuals between the ages of 2.0 and 85 + years. The normative sample closely matches the 2000 U.S. Census (education level based on 1999 data).

(v) The S135 was co-normed with the Bender (R) Visual-Motor Gestalt Test, Second Edition.

(vi) Reliability for the S135 are very high. For the FSIQ, NVIQ, and VIQ, reliabilities range from .95 to .98

(vii) Reliabilities for the Factor Indexes range from .90 to .92. For the ten subtests, reliabilities range from .84 to .89.

10.5 THE WECHSLER TESTS

David Wechsler initially developed test for measuring adult intelligence and later on he developed a series of tests to measure intelligence of children and infants. Some general types of items used in Wechsler tests of intelligence are as follows:

WAIS IV - The Wechsler Adult Intelligence Scale - Fourth Edition is the current available test in the Weschsler series, Some of the earlier tests include:

- Wechsler-Bellevue - I (WB- 1)
- Wechsler Bellevue - 11 (WB-11)
- WAIS
- WAIS-R
- WAIS - III

WAIS IV is the most recent edition to the family of Wechsler Adult Scales. The fourth edition of the test (WAIS-IV) was released in 2008 by Pearson. It consists of 10 core subtests and five supplemental subtests. A core subtest is one that is administered to obtain a composite score. Under usual circumstances, a supplemental subtest (also sometimes referred to as an optional subtest) is used for purposes such as providing additional clinical information or extending the number of abilities or processes sampled. In certain situations supplemental subtest is used in place of a core subtest, under following conditions:

* when examiner incorrectly administered a core subtest

* or the assesses had been inappropriately exposed to the subtest items prior to the administration of the test

* the assesses has the physical limitation that affected the assessee's ability to effectively respond to the items of a particular test.

• Core Subtests • Block design • Similarities • Digit Span • Matrix reasoning • Vocabulary • Arithmetic • Symbol search • Visual puzzles • Information • Coding	Supplemental Subtesits Letter-number sequencing Figure Weights Comprehension Cancellation Picture completion

Some important features of WAIS IV are as follows:

- The WAIS-IV was standardized on a sample of 2,200 people in the United States ranging in age from 16 to 90 years and 11 months More explicit administration instructions
- Expanded use of demonstration and sample items
- Floor and Ceiling limits are extended. WAIS IV has a full scale IQ ceiling of 160 and Full Scale IQ floor of 40.
- It is sensitive to the needs of the older adults
- The images in the picture completion, symbol search and coding subtests have been enlarged.
- An average reduction in the overall test administration time from 80 to 67 minutes.
- In WAIS IV we do not measure three different types of IQ such as Full Scale IQ, Verbal IQ and Performance IQ as we used to do in the earlier versions.

10.5.1 Wechsler Intelligence Scale for Children - Fourth Edition (WISC -IV)

Wechsler scale for children was first published in 1949. It represented the downward extension of Wechsler Bellvue scale. The original Wechsler Intelligence Scale for children had many flaws in it. The standardization sample contained only white children and some test items were viewed as perpetuating gender and cultural stereotypes. WISC was revised in 1974 and came to be called as WISC R. This test has been adapted in India by Prof. Malin and is called as Malin's Indian adaptation of WISC.

WISC-IV is the latest revision and improved version of WISCIII and was published in the year 200 3. WISC-IV yields the following measures:

(i) General intelligence functioning
(Full Scale IQ also called as FSIQ) (ii) Four index scores viz:
(a) Verbal comprehension index
(b) Perceptual reasoning index
(c Working memory index
(d) Processing speed index

Each of these index is based on scores on three to five subtests.

In WISC-IV, following subtests have been eliminated
(a) Picture arrangement
(b) Object assembly and
(c) Mazes

The following subtests are supplementary tests
• Information
• Arithmetic
• Picture completion
WISC-IV contains 10 core subtests and 5 supplemental tests.

10.5.2 Wechsler Preschool and Primary Scale of Intelligence Illrd edition (abbreviated as WPPSI - 111):

The origin of Wechsler Preschool and Primary Scale Intelligence (WPPSI) can be traced to the year 1967 when Wechsler for the first time decided that a new scale should be developed and restandardised especially for children who were under the age of 6 years.

The WPPSI was the first major intelligence test that adequately sampled the total population of the United States, including racial minorities. The WPPSI was revised in the year 1989 and came to be called as WPPSI-R. This test was designed to assess the intelligence of children from ages 3 years through 7 years and 3 months. In this revision new items were developed to exten the range of the test both upward and downward.

WPPSI-111 was published in the year 2002. This test further extended to the age range of children who could be tested with this instrument downward to 2 years and 6 months.

The WPPSI-111 had many changes incorporated in it as compared to earlier editions. Some important changes were as follows:

(1) The following 5 subtests which were present in the earlier editions were dropped from WPPSL These five subtests are as follows:

- Arithmetic
- Animal pegs
- Geometric designs
- Mazes and
- Sentences

(2) The following seven new subtests were added in WPPSI-111.

- Matrix reasoning
- Picture concepts
- Word reasoning
- Coding
- Symbol search
- Receptive vocabulary
- Picture naming

(3) WPPSI-111 has different labels for certain subtest. Some of these are as follows

(d) Core subtests
(e) Supplemental subtests
(f) Optional subtests

Core subtests are those that are required for the calculation of composite score.

Supplemental subtests are used to provide broader sampling of intellectual functioning. These subtests may also substitute for a core subtest if a core subtest cannot be administered due to some reasons or was administered but its score cannot be used or has become unusable.

Optional subtests are those that may not be used to substitute for core subtests but may be used in the derivation of optional scores.

The WPPSI was aimed at measuring two variables:

- Fluid reasoning

- Processing speed

Wechsler, Binet and the Short Form: Short forms of intelligence tests including that of Wechsler tests have been developed. Short form refers to test that has been abbreviated in length, typically to reduce the time needed for test administration, scoring and interpretation. Short forms of the test are used for two purposes: convenience of the test administrator and the practical necessities with the client that mandates the use of short forms. Short forms of the test are not new. Doll (1917) used the short form of the Binet Simon test. In 1958 David Wechsler endorsed the use of short forms, but only for screening purposes. Wechsler Abbreviated Scale of Intelligence (WAIS) was developed in 1999. Watkins (1986) concluded that short forms may be used for screening purposes only, but not to make placement and educational decisions. Smith McCarthy and Anderson (2000) held the view that short form must be used with caution. Silverstein (1990) has pointed out as to how short forms can be used with caution.

10.6 SUMMARY

In this unit we had discussed about the definition of intelligence and the various theories of intelligence. Among the definitions of intelligence, we had examined the views of the lay public as well as the views of scholars and test professionals. We studied the views of Francis Galton, Alfred Binet, David Wechsler, Jean Piaget. Two most important theories of intelligence that we discussed were the factor-analytic theories and the information processing views. The topic of measuring intelligence was also discussed.

The Stanford- Binet Tests and the Weschsler tests are an important group of tests.. Among the Stanford Binet Tests we discussed the 05th edition of Stanford-Binet Intelligence Scales in brief. We also discussed Weschsler tests, such as the Weschsler Adult Intelligence Scale, 04 th Edition (WAIS - IV), Wechsler Intelligence Scale for the Children, 04 1h Edition (WISC - IV) as well as the Wechsler Preschool and Primary Scale of Intelligence, 03rd Edition (WPPSI - 111). A brief mention of the short forms of intelligence tests was also discussed.

10.7 QUESTIONS

1 Define Intelligence and discuss the views of the lay public as well as Scholars and Test Professionals with respect to intelligence.

2. Discuss the Factor-Analytic and Information Processing Theories of Intelligence.

3. Write short notes on:

   a. Measuring Intelligence
   b. The Stanford-Binet Scales of Intelligence
   c. WAIS IV - The Wechsler Adult Intelligence Scale Fourth Edition
   d. Wechsler Intelligence Scale for Children - Fourth Edition (WISC -IV)
   e. Wechsler Preschool and Primary Scale of Intelligence Illrd edition (abbreviated as WPPSI - 111).

10.8 REFERENCES

1 Cohen, J.R., & Swerdlik, M.E. (2010). Psychological Testing and Assessment: An introduction to Tests and Measurement. (7 th ed.). New York. McGraw-Hill International edition.

2. Anastasi, A. & Urbina, S. (1997). Psychological Testing. (7th ed.). Pearson Education, Indian reprint 2002.

11

Assessment of Personality

Unit Structure

11.0 OBJECTIVES

After studying this unit you should be able to:

- Define Personality, Personality Assessment and related terms such as Traits, Types and States.
- Understand some basic questions related to Personality Assessment.
- Comprehend the tools or instruments used in the process of assessment of personality, such as the use of logic and reason, theory, data reduction methods and criterion groups
- Know the relationship between Personality and Culture.
- Understand the various objective methods of Personality Assessment.
- Know the various Projective Methods of Personality Assessment.

11.1 INTRODUCTION

In this unit we will first define the concepts of Personality, Personality Assessment and related terms such as Traits, Types and States. Following this we will discuss some basic questions related to personality assessment such as who is being actually assessed, what is assessed when a personality assessment is conducted, where are personality assessments conducted, how are personality assessments structured and conducted. In the process of personality assessment various tools or instruments are used which includes logic, theory, data reduction methods, such as factor analysis and

the criterion groups. These constitute the technical aspects involved in developing instruments to assess personality.

Assessment of personality is intimately tied to one's culture as well as language. Issues related to acculturation and other considerations would also be discussed

Objective and projective methods of personality assessment would be briefly discussed. Among the projective techniques the Rorschach ink blot test is the most common followed by the Thematic Apperception Test, the Word Association Tests and the Sentence Completion Tests. Sounds and Figure Drawing are also used as projective techniques in the assessment of personality.

11.2 DEFINITIONS OF PERSONALITY AND PERSONALITY ASSESSMENT

11.2.1 Personality: Personality is an important aspect of human individuality and psychologists have defined and measured it in scientific ways. Psychologists do not agree with the definitions of personality. Hence, different scholars have defined in different ways leading to varied definitions of personality.

McClelland defined personality as "the most adequate conceptual isatio n of a person's behaviour in all its details". According to Menninger, personality can be defined as "the individual as a whole, his height, and weight and love and hates blood pressure and reflexes, his smiles and hopes and bowed legs and enlarged tonsils. It means all that anyone is and that he is trying to become". Some scholars such as Goldstein (1963) has defined personality very narrowly and focuses on a particular aspects of an individual whereas Sullivan (1953) has defined personality in the context of society. Some Psychologists avoid defining the term personality (Byrne (1974). Byrne characterized the entire area of personality, psychology as "Psychology's garbage bin in that any research which dose not fit other existing categories can be labeled personality".

Hall and Lindzey (1970) in their classic text "Theories of Personality" state that "personality is defined by the particular empirical concepts which are a part of the theory of personalit X employed by the observer". According to Cohen and Swerdlik (7t Edition), personality can be defined as an individual's unique constellation of psychological traits and states.

11.2.2 Personality Assessment:

Personality assessment is also sometimes incorrectly referred to as psychological testing. Personality assessment can be defined as the measurement and evaluation of psychological traits, states, values, interests,

attitudes, worldview, acculturation, personal identity, sense of humour, cognitive and behavioural styles, and/or related individual characteristics. The term assessment as used in personality assessment is different from psychological testing.

11.2.3 Traits, Types and States:

11.2.3.1 Personality Traits:

There is no consensus among psychologists as to the meaning of the term traits. In general it refers to enduring characteristics or aspects of personality that are stable over time and across situations. Gordon Allport viewed personality traits as real physical entities that are bona fide mental structures in each personality. Psychological traits can be viewed as attributions made in an effort to identify threads of consistency in behavioural patterns.

According to Guilford trait refers to "any distinguishable, relatively enduring way in which one individual varies from another"

Personality psychologists have developed formal methods for describing and measuring personality. Their use of the term trait (defined as stable and enduring characteristic of an individual) is different from the everyday use of this term that we make in three ways.

- Personality psychologists have used statistical methods to reduce the number of traits on the basis of similarity.
- They have used reliable and valid instruments to measure traits.
- They have used carefully designed method to conduct empirical research to demonstrate the relationship between certain specific behaviour and traits.

Personality theorists and assessors have assumed for years that personality traits are relatively enduring over the course of one's life. Roberts and DelVecchio (2000) explored the endurance of traits by means of a meta-analysis of 152 longitudinal studies. These researchers concluded that trait consistency increases in a steplike pattern until one is 50-59 years old, at which time such consistency peaks.

11.2.3.2 Personality Types: Personality type can be defined as a unique constellation of traits and states that is similar in pattern to one identified category of personality within a taxonomy of personalities. Personality types are actually descriptions of people. A type is a class of individuals who share a common collection of traits together in an individual.

Some important personality types as identified and discussed by different scholars and psychologists in their works are as follows:

One of the earliest personality type theorist was proposed by Greek physician Hippocrates around 400 BC. He is also called as the father of modern medicine.

Working on the assumption that the human body contains four fluids, or humours (blood, phlegm, black bile, and yellow bile), he categorised people into four corresponding personality types as follows:

(i) Phlegmatic (a calm, apathetic temperament caused by too much phlegm).
(ii) Choleric (a hot headed, irritable temperament due to an excess of yellow bile).
(iii) Sanguine (an optimistic, hopeful temperament attributed to a predominance of blood).
(iv) Melancholic (a sad, depressed temperament based on black bile). This types of theory of personality, today, is only of historical importance.

In 1925, a German psychiatrist Kretschmer, in his book, physique and character classified individuals into certain biological types according to their physical structure. Kretschmer classified personality into three types : Pkynic (having fat bodies), Athletic (balanced bodies) and Leptosomatic (lean and thin bodies).

William Sheldon, Carl G. Jung , John Holland, etc., has also given different type theories of personality. The Myers-Briggs Type Indicator (MBTI), test is based on Carl Jung's personality types. John Holland categorised people into following 6 personality types

(i) Artistic
(ii) Enterprising
(iii) Investigative
(iv) Social
(v) Realistic
(vi) Conventional

Meyer Friedman and Ray Rosenman developed two personality types called the Type A and Type B personality types. Type A personality is characterized by competitiveness, haste, restlessness, impatience, feeling of being time-pressured and strong needs for achievement and dominance. Type B personality is opposite of Type A's traits. Type B people are relaxed, easy going, laid-back or mellowed.

The personality type that has attracted the most attention from researchers of clinicians is the one associated with scores on Minnesota Multiphosic Personality Inventory.

Data from administration of MMPI tests are frequently discussed in terms of patterns of scores that emerge on the subtests. This pattern is referred to

as a profile. In general, a profile is a narrative description, graph, table or other representation of the extent to which a person has demonstrated certain targeted characteristics as a result of the administration or application of the tools of assessments.

11.2.3.3 Personality States: The terms personality state in psychological assessment literature is used in the following two different contexts:

(a) It is used to refer to an inferred psychodynamic disposition designed to convey the dynamic quality of Id, Ego and Superego in perceptual conflict.
(b) It is a term used to refer to the transitory exhibition of some personality trait. In other words the term state is indicative of a relatively temporary predisposition.

Measuring personality state amounts, in essence, to a search for and an assessment of the strength of traits that are relatively transitory of fairly situation specific. Charles D. Spielberger and his associates have developed a number of personality inventories designed to distinguish various states from traits. They distinguish between Trait and State anxiety. Trait anxiety or anxiety proneness refers to relatively stable or enduring personality characteristics. State anxiety on the other hand refers to a transitory experience of tension because of a particular situation. One test developed by Spielberger and his associates to measure states from traits is called as State-Trait Anxiety Inventory (STAI).

11.3 PERSONALITY ASSESSMENT - SOME BASIC QUESTIONS

Personality assessment is aimed to find practical solutions or answers to issues that confront us either related to our health, work, career, life decisions, etc. Questions related to personality assessment either in, basic research or practical problems, seek to explore answers that can help us to handle a given problem or issue at hand. Four important basic questions related to personality assessment are as follows:

(1) Who is actually being assessed? Can the test taker be someone other than the subject of the assessment?
(2) What is being assessed? What is assessed when a personality assessment is conducted?
(3) Where is personality assessment conducted?
(4) How is personality assessment structured and conducted

We would discuss each of these briefly.

11.3.1 Who : One of the important questions related to personality assessment is that while assessing an individual's personality, who needs to be assessed. The concerned individual, his/her spouse, children, parents, friends or other informants. Who can be a test taker? An individual herself or someone other than the subject of assessment.

Many tests make use of self-report or in which one assess the individual herself whose personality assessment is required. The individual answers in wide variety of ways. Some of which are as follows:

- They respond to interview questions.
- Answer questionnaires in writing.
- Blacken squares on computer answer forms.
- Sort cards with various terms on them, etc.

In certain types of assessment we rely on informants other than the person being assessed to acquire personality related information. For e.g., while assessing children we ask parents and/or teachers to participate in personality assessment of children by asking them their judgments, opinion and impressions about a particular child being assessed. Thus, while assessing an individual, three things are important

(a) The self as the primary referent.
(b) Another person as the primary referent.
(c) Cultural background of assessees. We would briefly discuss each of these in brief.

(a) The self as the primary referent: In very many testing or assessment situations an individual who is being assessed is an important source of information. Many different types of assessment require self- report. It is a process whereby information about assessees is supplied by the assessees themselves. Selfreport information is generally obtained in the following ways:

- From diaries report by the assessees.
- Response to oral or written questions or test items.

Self-report methods are very commonly used to assess an individual's self-concept. A number of self-concept measures for children have been developed. Some representative tests include:

- Tennessee Self-Concept Scaqle
- Piers-Harris Self-Concept Scale

The term self-concept differentiation refers to the degree to which a person has different self-concepts in different roles. People characterized as highly differentiated are likely'to perceive themselvos quite differently in various roles. For example a highly differentiated college professor in his 40s may perceive himself as motivated and hard driving in his role at work, conforming and people-pleasing in his role as son and emotional and passionate in his role as a husband. By contrast, people whose concept of self is not very differentiated tend to perceive themselves similarly across their social roles.Self-report measures are highly valuable source of information provided an individual who gives the said information is highly motivated and answers the questions in a honest manner.The greatest limitation of self-report method includes:

- Faking on the part of the assessee to impress the examiner or to conceal some vital or embarrassing or highly personal information, such as that related to certain behaviour.

- Some assessees may lack insight into their own behaviour and hence may not be able to reveal the accurate picture of their own self.

(b) Another person as the primary referent: In many situations, we can acquire reliable and best information from a third party. The third party can be a parent, a spouse , a teacher, peer, supervisor, etc. For e.g., in the assessment of an emotionally disturbed child parents and/ or teachers can be a best source of information.

Acquiring information from others the raters is prone to many pitfalls or errors. Raters can make biased judgments, consciously or unconsciously as it is in their own self-interest to do so. Some of the most common types of errors on the part of the raters are as follows:

(i) ***Error of leniency or generosity:*** Tendency to score or rate leniently.
(ii) ***Error of severity or stringency:*** Tendency to score or rate strictly.
(iii) ***Hallo effect****:* A type of rating error wherein the rater views the object of rating with extreme favor and tends to bestow ratings inflated in appositive direction
(iv) ***Error of central tendency:*** It is the general tendency to rate everyone near the midpoint of a rating scale.

Many other factors may also influence how raters rate a given individual. Some important factors that can bias raters are as follows:

- The rater may feel competitive, physically attracted or physically repelled by many extraneous or similar factors between the ratee and rater.
- The rater may also not have the proper background, experience or training in, rating a given individual.
- The rater's judgment may also be limited by his/her general level of conscientiousness and willingness to devote the time and effort required to do the job properly.
- Different raters may also have different perspectives on the individual they are rating by virtue of their context in which they typically view that person. For e.g., a parent may indicate on a rating scale that a child is hyperactive, whereas the class teacher, on the some rating scale may indicate that the child's activity level is within normal limits.

(c) The cultural background of the assessees: Many researchers and test administrators take into account the cultural variables in the use of assessment. While administering scoring and interpreting a given assessment instrument or method due regard should be given to cultural, factors and/or variables which are likely to impact on the assessment process.

11.3.2 What: One of the important questions with regard to assessment is as to what is assessed when a personality assessment is conducted. Two important aspects of what is being test in personality assessment are as follows:

(a) Primary content area sampled

(b) Test taker response style

With respect to primary content area it should be remembered that personality test may measure one area or aspect of personality such as anxiety or extroversion or shyness or it may measure many aspects of personality as in MMPI.

Many tests, today measure test takers response style also. **Response style** refers to a tendency to respond to a test item or interview question in some characteristic manner regardless of the content of the item or question. A particular response style such as responding to a personality test in an inconsistent, contrary or random way or attempting to take good or bad, etc., may invalidate a given test.

11.3.3 Where: Another basic. question related to personality assessment is as to where are personality assessments conducted. Traditionally personality assessments have been conducted in different places such as in the:

- Schools
- Clinics and hospitals
- Academic research laboratories
- Employment counseling e.g., vocational selection centers
- Office of psychologists and counselors

Today, assessment is also conducted in natural settings as well as online.

11.3.4 How: Another important question related to personality assessment is with respect to how are personality assessments structured and conducted. How a personality assessment is conducted generally depends upon the scope. The scope of personality assessment can be very wide or very narrow. For e.g., MMPI and California Psychological Inventory are used when an assessment is- aimed with a broader scope in mind. On the other hand, when we measure a single trait such as loss of control, our scope is very narrow. Certain tests are based on particular theory of personality whereas there are tests whose development is atheoretical in nature.

Procedures or item formats: With respect to "how" of personality assessment, it is also important to note the procedures or item formats used in assessing one's personality. Some importan t methods of assessing an individual's personality are as follows:

- Face-to-face interviews
- Computer administered tests
- Behavioural observations
- Paper-and-pencil tests
- Evaluation of case history data
- Evaluation of portfolio data
- Recording of physiological responses

Certain methods of personality assessment are highly structured whereas other methods are highly unstructured.

Frame of reference: Another important aspect of how personality measur ment is carried out has to do with frame of reference of assessment. We can define frame of reference as aspects of the focus of exploration such as the time frame (the past, the present, or the future) as well as other contextual issues dealing with people, places and events.

One important method that can be applied in the exploration of varied frames of reference is the Q-sort technique. It was a method developed by Stephenson and used extensively by Carl Rogers. Besides Q-sort method two other item presentation formats readily adoptable to different frames of reference are as follows:

- Adjective checklist
- Sentence completion format

Scoring and interpretation: Personality assessments also differ with respect to how tests are scored and interpreted. Two important approaches to scoring and interpretation of personality assessment are as follows:

(a) Nomothetic approach: This approach to assessment is characterized by efforts to learn how a limited number of personality traits can be applied to all people.

(b) Idiographic approach: It is characterized by efforts to learn about each individual's unique constellation of personality traits with no attempt to characterise each person according to any particular set of traits.

"How" of test also involves examining the issues in personality test development and use. Many issues in personality test development and use also determine as to how the test will be used. Will the test be a self-report inventory or will it be a projective test. Should personality inventories be used or some other tests.

11.4 DEVELOPING INSTRUMENTS TO ASSESS PERSONALITY - LOGIC AND REASON, THEORY, DATA REDUCTION METHODS, CRITERION GROUPS

Most personality tests employ two or more of the following tools in the development of personality assessment instruments. We would briefly discuss each of these.

11.4.1 Logic and Reason: While preparing test items we make use of logic and reason, which dictate as to what contents is covered by the items. Use of logic and reason in the development of test items is sometimes referred

to as the content or content oriented approach to test development. In the process of development of test items we see to it that they are based on American Psychiatric Association's Diagnostic and Statistical Manual Criteria for diagnosis of a particular disorder. Attempts to develop contentoriented, face valid items began during First World War in an attempt to develop instruments to assess recruit's personality and adjustment problems. One of the well known personality test was Woodworth's Personal Data Sheet (1917) that later on came to be called as Woodworth Psychoneurotic Inventory. This test was designed to elicit self report fears, sleep disorders and other problems deemed symptomatic of psych neu roticism. The greater the number of problems reported, the more psychoneurotic the test taker was presumed to be.
Self report instruments of this type can help us to collect a great deal of clinically actionable information in a relatively little time. In order to administer such types of test a highly trained professional is not required. Such instruments are more suited in clinical settings where emphasis is on cost cutting.

Logic reason and intuition are often used in item development. A sound research knowledge and clinical experience is also needed

11.4.2 Theory: Personality measures differ in the extent to which they rely on a particular theory of personality in their development and interpretation. When a psychological theory is the guiding force behind the development of a psychological test, rather than reason and logic, then the items are quite different. One theory based test that is predominantly used today is the Self Directed Search (SDS) which is a measure of one's interest and perceived abilities. It was developed by John Holland and his associates. The test is based on Holland's theory of vocational personality. The central idea of the theory is the view that occupational choice has a great deal to do with one's personality and self-perception of abilities. The SDS is self-administered, self-scored and self-interpreted. Test scores direct test takers towards specific occupational themes. From there, test takers follow instructions to learn about various occupations that are consistent with their expressed pattern of interests and abilities.

11.4.3 Data Reduction Methods: Another category of widely used tool in test development is data reduction methods. Such methods make use of wide variety of statistical techniques collectively -known as factor analysis or cluster analysis. One use of data reduction methods is to, aid in the identification of minimum number of variables or factors that account for the intercorrelations in observed phenomenon. Psychologists using data reduction methods have identified certain primary factors of personality. Considerable research using data reduction methods was carried out by Raymond Cattell who developed, 16 , PF Questionnaire. Cattell identified 36 surface traits and 16 source traits.

Eysenck (1991) have argued that primary factors can be narrowed down to three. Some researchers have identified five factor model. One well known such model is that developed by Costa and McCrae called the Big Five Model.

Big Five Model: One test developed to measure big five factors include revised NEO Personality Inventory (NEO PI - R). The five factor of big five include

- Neuroticism
- Extraversion
- Openness
- Agreeableness
- Conscientiousness

a. Neuroticism domain taps aspects of adjustment and emotional stability.
b. Extraversion domain taps aspects of sociability and assertiveness.
c. Openness encompasses openness to experience as well as active imagination, asthetics sensitivity, attentiveness to inner feelings, preference for variety, intellectual curiosity and independence of judgment.
d. Agreeableness is primarily a dimension of interpersonal tendencies that include altruism, sympathy towards others and belief that others are similarly inclined.
e. Conscientiousness is a dimension of. personality that involves the active process of planning, organizing and following through.

NEO PI - R is generally used with individuals who are 17 years and above. It is a self administered test. It can be computer scored and interpreted.

11.4.4 Criterion Groups:

Criterion is defined as a standard on which a judgment or decision can be made. A criterion group is a reference group of test takers who share specific characteristics and whose responses to test items serve as a standard according to which items will be included in or discarded from the final version of the scale. The process of using the criterion groups to develop test items is referred to as empirical criterion keying because the scoring or keying of items has been demonstrated empirically to differentiate among group of test takers.

One test developed using criterion group is the MMPI (Minnesota Multiphasic Personality, Inventory) originally called the Medical and Psychiatric Inventory (Dahlstrom and Dahlstrom, 1980). MMPI consists of various types. Some of the most commonly included types of MMPI are:

a. MMPI
b. MMPI - 2
c. MMPI - 2 - RF
d. MMPI - A

We would briefly discuss the MMPI as it is the origi nal and one of the classic tests in the history of personality assessment.

MMPI was originally developed by Psychologists Starke K Hathaway and Psychiatrist - Neurologist John Charnley McKinley (1940). It consists of 566 True-False items and was devised as an aid to psychiatric diagnosis with adolescents and adults of 14 years and above. MMPI consists of the following 10 clinical Scales:

1. Hs: Hypochondriasis
2. Hy: Hysteria
3. Mf:Masculinity-femininity
4. Pt: Psychasthenia
5. Ma: Hypomania
6. D: Depression
7. Pd.Psychopathic deviate
8. Pa: Paranoia
9. Sc: Schizophrenia
10. Si: Social Introversion

These above mentioned diagnostic categories were very popular in 1930s. The clinical criterion group for the MMPI was made up of psychiatric inpatients at the University of Minnesota Hospital. The validity scales were also built in the MMPI. These scales include:

i) Validity Score (F)
ii) Lie Score (L)
iii) Question Score
iv) Correction Score (K)

These scales were not designed to measure validity in the technical, psychometric sense. Out of the 566 items, 16 items are repeated. Scores of MMPI are reported in the form of T-Scores which is one type of standard score with a mean set at 50 and standard deviation set at 10. In addition to the above mentioned clinical and validity scales, there are MMPI content scales, such as Wiggins Content Scales. Content Scales are composed of group of test items of similar content. Supplementary scales are a catch all phrase for the hundreds of different MMPI scales that have been developed since the test's publication. These scales have been developed by different researchers using a variety of methods and statistical procedures, mostly factor analysis. MMPI is today administered by many different methods:

- Online
- Offline on disc
- Index cards
- Audio version for semiliterate test takers is also available with instructions recorded on audiocassette.

MMPI can be scored by hand. Computer scoring is also available. Paul Meehl (1951) proposed a two point code derived from a number of clinical scales on which the test taker achieved the highest scores. Another popular approach to scoring and interpretation was developed by Welsh called as Welsh Codes.

The other scales such as MMPI - 2, MMPI - 2 - RF and MMPl - A are different developments in MMPI.

11.5 PERSONALITY ASSESSMENT AND CULTURE

Psychologists have become increasingly concerned with the relationship between assessment of an individual personality and the various aspects of an individual's culture. Psychologists are often required to assess individual's personality and other related variables of people who belong to varied and diverse cultures. It should be remembered that with members of culturally and linguistically diverse populations, a routine and business-as-usual approach to psychological testing and assessment is inappropriate.

Research studies, especially in the area of cross cultural psychology has emphasised the sensitivity on the part of psychologists as to how culture relates to behaviours of cognitions that are being measured. Some important aspects of the culture that impacts assessment enterprise are as follows:

(a) Accultration
(b) Values
(c) Identity
(d) World view and
(e) Language of the assessee

We would briefly discuss each of these.

Acculturation is an ongoing process by which an individual's thoughts, behaviours, values, world view and identity' develops in relation to general thinking, behaviour, customs, values of a particular group. It is at birth, that the process of acculturation begins and it is through the process of acculturation that develop culturally accepted ways of thinking, feeling and behaving. Few tests have been developed to assess an individual's levels of acculturation to their native culture or dominant culture.

Values are closely associated with acculturation. One's values considerably influence the process of assessment. Different culture emphasizes different values. Indian culture emphasizes group, family and spiritual values as compared to individuality and material culture emphasized by western cultures. Similarly, some cultures emphasize and value "future". Where are other cultures emphasizes on "here and now" the present. Assessment instruments should reflect one's cultural values. One of the earliest work on values was the book titled Types of Men (Spranger 1928),

which listed different types of people based on whether they valued things like truth, practicality and power. Rokeach (1973) distinguished between two types of values called as Instrumental and Terminal Values. Instrumental Values are guiding principles to help one attain some objective. Honesty, imagination, ambition and cheerfulness are examples of instrumental values. Terminal values are guiding principles and a mode of behaviour that is an endpoint objective. Some examples of terminal values are a comfortable life, an exciting life, a sense of accomplishment and self respect. According to Kluckhohn (1960) values provide answers to key questions with which civilizations must grapple.

Personal identity is another important aspect of one's culture that out be kept in mind in the process of assessment. We can define identity as a set of cognitive of behavioural characteristics by which an individual define themselves as members or a particular group. The term identity is closely associated with the term self.

World view is another important aspect of one's culture. It is a unique way in which people interpret and make sense of their perceptions as a consequence of their learning experiences, cultural background and related variables.

One's **language** also influences the ways in which one conceptualizes a given item or statement. Language also influences the way in which one responds to test item.

Any assessment instrument must take into consideration above mentioned cultural factors if meaningful information has to be obtained from assessment data.

11.6 OBJECTIVE METHODS OF PERSONALITY AS SESSMENT

Objective methods of personality assessment consists of paper and pencil tests having multiple objectives and the assessee has to select one of the correct items. The objective methods of personality assessment contain short answer items for which the assessee's task is to select one response from the two or more provided, and the scoring is done according to set procedures involving little or no judgment on the part of the scorer.

Objective methods of personality assessment may include items written in a multiple choice, true/false or matching format.

Response on an objective ability test may be scored correct or incorrect. Objective personality tests share many advantages with objective tests of ability.

Objective items can usually be scored quickly and reliably by varied means, from hand scoring to computer scoring.

In objective, multiple-choice tests of ability there is little room for emotion, bias or favouritism on the part of the test scorer.

Objective personality tests are objective in the sense that they employ a short-answer, typically multiple-choice format, one that provide little, if any, room for discretion in terms of scoring.

Many personality inventories such as MMPI, EPPS or CPI are objective in nature.

11.7 PROJECTIVE METHODS OF PERSONALITY ASSESSMENT

Projective methods of personality assessment are based on the projective hypothesis which holds that individuals introduces structure to unstructured stimuli and thereby reveals his/her personality. It is an ideal that when an individual is presented with an unstructured and ambiguous stimuli and individual will introduce structure and clarity in to it thus revealing his/her unconscious desires, wishes, needs and unconscious impulses. Projective methods refer to a group of related and unrelated techniques used for studying both intellectual and non-intellectual aspects of personality. In these tests, an individual is presented with a relatively unstructured or ambiguous task like a picture, inkblot or incomplete sentences, which permits a wide variety of interpretations by the subjects. The basic assumption underlying projective tests is that individual's interpretation of the task will project his characteristic mode of responses, his personal motives, emotions and desires and thus enable the examiner to understand more subtle aspects of his personality.

11.7.1 Inkblot as Projective Stimuli: One of the well-known projective techniques is the Rorschach Inkblot Test which was developed by Hermann Rorschach in 1921. He called it as the "Form Interpretation Test" as inkblots were the forms to be interpreted. in 1921 published his work in a Monograph titled 'psychodiagnostic'. Many consider this year as the date for the development of this test. Unfortunately, his first published work was als"O his last work for he died in the following year. In this monograph he provided 28 case studies, which included normal as well as individuals with psychiatric diagnosis such as neurosis, psychosis and manic depressive psychosis. The Rorschach test consists of 10 cards having bilateral systematical inkblots. Half of the cards are achromatic and other five share one or more colours. The cards are presented to the subjects in a define sequence. Rorschach did not impose any time limit on this test nor do present users. There is also freedom to give any type of responses and as many type of responses that are possible.

Though even before Rorschach, researchers have used inkblots to study imagination and other functions, Rorschach was the first to apply inkblots to the diagnostic investigation of the personality as a whole. Today, Rorschach is one of the most frequently used, very popular, widely criticized and extensively researched test. Various innovations in this test have led to the development of multiple systems of test administration scoring and interpretation. Some of the commonly used Rorschach systems are by Kloppfer, Back, Hertz, Piotrowski, Rapaport Schaffer and the last one developed by Exner, called as Comprehensive System.

The cards are presented to the test taker in a definite sequence. The cards are numbered from I to X While administering the test, the examiner notes the various aspects of subject's behaviour. He keeps a verbatim record of the responses, notes the time elapsed between the presentation of each card and the first response to it (called as initial reaction time), length of the pause between responses, total time required for each card response and the subject's extraneous movements, spontaneous remarks, emotional expressions and any other specific behaviours which appears significant to the examiner.

After all the 10 cards have been presented in a sequential order another phase called as the inquiry phase starts in which the examiner questions the individual systematically regarding the parts and aspects of each blot to which the associations were given. The two important purposes of the inquiry are :

- First, to determine which aspects of blot initiated and sustained the association process,
- Second the inquiry gives the subject an opportunity to add, elaborate or to clarify his original responses, but if this is done it must be completely spontaneous on the part of the subject and without any suggestions from the examiners.

Another important component of test administration is what is called as Testing of Limits, In this procedure the examiner asks specific questions to seek additional information concerning subject's personality functioning. In testing of the limits, the examiner:

- Might ask if the test taker has used the entire blot or a part of it or the white space within the blot to form the -percept (i.e., perception of an image)
- Any confusion or misunderstanding that the subject might have with regard to the task that is the blot.
- Finds out if the test taker is able to refocus percepts given a new frame of reference.
- Assesses if the test taker becomes anxious by the ambiguous nature of the task or is able to perform better, etc.

Several scoring categories for Rorschach test have been developed but the most commonly scored categories are as follows:

(i) Location responses refer to the area of the blot, which has been perceived by the subject as the basis of his response. The subject may respond to entire blot, to a larger or small portion of it, to small minute detail and sometimes even to the white background. The location responses can be very well-defined or only vaguely defined. The location responses to the test are categorised into following classes :

- W(responses to the whole blot),
- 'D' (responses to large usual details),
- 'Dd'(responses to small unusual details),
- 'S' (responses to white space on the card).

The location responses can be given by combining some of the above categories for e.g. D and S can also be perceived together.

The location responses and the subject's ability to delineate them are regarded as indicative of subject's organizing process and the ability to analyze and articulate the part of everyday experience. The analysis of the subject's location response is made in the light of norms prepared by the test author.

(ii) Determinants refer to the characteristics of the inkblot as perceived by the subject. They are those qualities of the blot that have produced the response to it. The four most important determinants are as follows :

- Form,
- Shading,
- Colour and
- Movement.

Although there is of course no movement in the blot itself, the respondent's perception of the blot as a representation of moving object is scored in this category. Further differentiation is made within these categories. Movement responses are Human Movement (M), Animal Movement (FM), or inanimate movement (m). Similarly, form may be perceived with ordinary accuracy (F), with unusual accuracy (F+) or in a very poor accuracy (F-). Shading may be perceived as representing depth (V) or texture (T), which can be perceived in its pure from (V or T) or it can be combined with from (FV, VF, FT, and TF), colour can be perceived along with form being dominant (FC) or colour being dominate (CF). Colour can be perceived all by itself also (C). These are some'of the commonly scored categories of determinant developed by Rorschach and various other systematizes.

Determinants reveal a lot of information about a wide variety of personality characteristics. They tell us about one's emotionality, imagination, fantasy life, an individual's ability to indulge in creative and conceptual thinking, about ego strength, conflicts, opposition tendency, etc.

(iii) The treatment of content varies from one scoring system to another, some emphasizing it heavily whereas others ignoring it altogether. Some of the commonly scored content categories are as follows :

Human (H),
Animal (A),
Human detail (Hd),
Animal detail (Ad),
Clouds (Cl),
Botany (Bt),
Blood (Bl),
Sexual materials (Sx),
X-Rays,
Anatomy (At), etc.

Content categories help us to discover many important aspects of an individual's behaviour. Content analysis is a source of ascertaining the subject's personal meaning, attitudes, interest and even complexes. Content responses are supposed to possess psychiatric and psychoanalytic interpretations often pointing towards pathological tendencies present in the subject.

(iv) Original and popular scoring category tells us whether subject's responses are common or original. A popularity score is often found on the basis of the relative frequency of different responses among people in general. However, there is often difference of opinions -among specialists as to which responses should be regarded as original and which as popular. Popular responses help us to know about an individual's interpersonal behaviour. It also tells us whether an individual is socially confirming or conventional in his approach or not.

Evaluation of the Rorschach Test:

Rorschach test has received mixed reception, some have regarded it as an X-ray of personality an indispensable tool for diagnostic purposes, whereas, others have regarded its use as unethical. Many have judged its worth with contempt and have advocated its abandonment because it has, proved baffling to the researcher and irritating to those with strong allegiance to stringent measurement theory.

1950s were a decade of great controversy of Rorschach and even today this controversy is fresh. In 1956, Cronbach remarked "the test has repeatedly failed as predictor of practical criteria" and in 1958, Jensen made a similar statement by pointing out that "the Rorschach has been worthless as a research instrument. The Rorschach methodology has nothing to show for its application in the personality field." Contrary to this, Sundberg (1961) noted that "Rorschach techniques is the most commonly used assessment instrument in clinical psychology". In 1969, Exner offered the opinion that "the Rorschach despite its inherent weakness and the strain of academic

disparagement is providing to be surprisingly handy. It not only survives perennially but is still a mainstay of psycho-diagnostic methodology." However, it is strange that Exner (1974) recently remarked that "the survival of Rorschach is no longer as certain as once was the case clearly, it is no longer synonymous with clinical psychology." Nowhere has there been the discrepancy and controversy between the researchers and clinicians so great and bitter as it has been in the area of Rorschach. Researchers have consistently presented a poor picture, whereas on the other hand, clinicians have been using this test with increasing frequency.

The nature of the Rorschach test makes it an important tool, which can be used with any age group, any cultural group and which does not make any demand on the part of the subject. It can, very easily, be used with illiterate and culturally backward groups.

11.7.2 Pictures as Projective Stimuli: Wide varieties of pictures has been used as projective stimuli ranging from real people to animals, objects or anything such as paintings, drawings etc. The use of pictures in psychometric assessment dates back to 1907 when Brittain (1907) reported sex differences in response to pictures. Similarly Libby (1908) as well as Schwarts (1932) developed the projective tests using pictures. The most famous test is the Thematic Apperception Test (TAT) developed in 1935 at Harvard Psychological Clinic by Christiana D Morgan and Henry Murray.

The only other projective technique that has approached the Rorschach method in amount of use and volume of research is the TAT which was ' introduced by C.D. Morgan and Henry A. Murray as a method to explore the unconscious thoughts and fantasies.

Murray found that TAT enabled the trained examiner or interpreters to reconstruct on the basis of subject's stories, his dominant drives, emotions, sentiments, complexes and conflicts. Although the TAT was at first slow in gaining wide acceptance, it is now a test that approximates the Rorschach in popularity and in amount of research it has stimulated.

The TAT was originally designed as an aid to eliciting fantasy material from patients in psychoanalysis. At first, it gained popularity only among clinical psychologists, but gradually, it became a research tool in development, social and personality psychology and in the cross-cultural studies in Anthropology. It is also used for personality assessment in the fields of counselling and industrial psychology.

In contrast to the inkblot techniques, the Thematic Apperception Test (TAT) presents more highly structured stimuli and requires more complex and meaningfully organised verbal responses. Interpretations of the responses by the examiners is usually based on content analysis of a rather qualitative

nature. Murray's system for scoring TAT is largely quantitative whereas McClelland and Eron has developed a highly quantitative system for scoring and interpreting TAT responses.

Administration of TAT:

The third revision of TAT consists of 30 pictures and blank card. The pictures have been selected and marked in such a way that there are four sets of 20 cards each, one for boys, one for girls, one for males and one for females, over 14 years. The testing process is divided into two sessions and for each of these, it is suggested that no more than 10 TAT cards be administered with at least one day intervening between the two sessions. More recently, practical considerations have led to reduction in the number of cards administered. Most testers now present the subject with 8 to 12 cards and use only a single session. The cards are presented individually and the respondent is instructed to provide a story about the picture that described the depicted scene, what led up to it, what the characters in the picture are thinking and what the outcome will be.

Although typically administered as an oral test in clinical situation, the TAT may also be administered in writing and as a group test.

Scoring of the TAT:

Like Rorschach, the TAT also has multiple scoring system. Murray's lack of detailed scoring instructions in his manual and the relative ease with which the TAT can be administered have been cited as factors contributing to the multiplicity of the scoring system (Murstein, 1963). Furthermore, the * non-technical nature of the TAT and the simple verbal contents of stories have encouraged clinicians to invent their own system of analysis. Some important TAT scoring systems are as follows:

(a) Murray's scoring system (non-quantitative)
(b) McClelland's system (quantitative) and
(c) Eron's system (quantitative).

Here we will examine Murray's scoring system in brief. Although amenable to' qualifications, Murray's recommended system of analysis is highly content-oriented and relies heavily on the qualitative characteristics of the stories. Murray emphasized three important concepts called as Need (determinants of behaviour arising fro within the individual), Press (determinants of behaviour arising from within the environment) and themas (a unit of interaction between needs and press). The following points are noted in the analysis of the story.

(1) The Hero : The first step in the analysis of the story is to distinguish the hero or the character with whom the subject seemed to have identified in principle. This would be the character in whom the story teller is most interested and the individual who mostly resembles him. The tester must be

aware of the fact that the hero of the story may be no hero in the story. The interpreter should direct his attention to the following aspects of the hero's personality. His intelligence, achievement ability, conflicts, leadership qualities, feelings, etc.

(2) Needs of the Hero: After the identification of the hero, the interpreter must formulate the reactions of the hero to various forces. These formulations are usually influenced by the theoretical orientations of the test interpreter. However, Murray recommends that this may be accomplished within a classification of the needs of the hero. The needs can be either primary or secondary.

(3) Environmental Forces: These are categories according to their effect on the hero. Murray's system consists of a comprehensive list of environmental forces or presses. These presses could be real or imaginary and include aggression in which the hero's property and/or possessions are destroyed. Dominance, where the hero is exposed to commands, order or forceful arguments and rejection in which persons reject, repudiate are indifferent or leave the hero.

(4) Outcomes: Outcome refers to the results of the story. It refers to the relative strengths of the forces emitting from the hero and the strengths of these. The amount of frustration and hardships experienced and the relative degree of success and failure of the hero must be assessed.

(5) Themes or Themas: Themes or themas refer to the interplay within the story of the hero's needs, presses and successful or unsuccessful resolutions of his conflicts. Themes represent needpress combination, it can be simple or complex.

(6) Interests, Sentiments and Relationship: These are the last category to be scored. In this, a note is made of the various interests, sentiments and inter-personal relationship as expressed in the stories by the subject.

Research studies have suggested that many situational factors such as who the examiner is, how the test is administered, test taker's experiences prior to the test and during the test administration process considerably influences as to how the test taker responds to the test. Test takers responses are also influenced by transient internal needs such as hunger, thirst, fatigue, and higher than ordinary levels of sexual tension. This test is considerably used in many competitive examinations including armed forces as well as UPSC (Union Public Service Commission) examinations for selecting candidates for employment in armed forces and government jobs.

11.7.3 Other tests using pictures as projective stimuli: Many different varieties of tests have been developed using pictures as projective stimuli. Some of the well known tests are as follows:

a) Hand Test: This test was developed by Wagner (1983) and consists of 09 cards with pictures of hand on them and the tenth card is a blank card. The test taker is asked what the hands on each card might be doing. When presented with the blank card, the test taker is instructed to imagine a pari of hands on the card and then describe what they might be doing. Test takers may make several responses to each card and all responses are recorded. Responses on this test are interpreted according to 24 categories such as aggression, dependence and affection.

b) Rosenzweig Picture-Frustration Study: This test was first developed in 1945. It employs cartoons depicting frustrating situations. The test taker's task is to fill in the response of the cartoon figure being frustrated. The Rosenzweig Picture-Frustration Study (P-FS), is a semi-projective technique that has been widely used to assess patterns of aggressive responding to everyday stress. When shown a picture, the client provides a reply for the anonymous frustration person depicted. The instrument contains 24 cartoon-like pictures, each depicting 2 people in a mildly frustrating situation that commonly occurs. Three versions of this instrument exist: Child, Adolescent and Adult.

Some other versions of picture tests include:
Children's Apperception Test (CAT)
Thompson Modification of TAT (T-TAT) The Blacky Picture
Make Picture Story (MAPS)
Michigan Picture Test (MPT)

11.7.4 Word as Projective Stimuli: There are two types of such tests:

a. Word Association Test
b. Sentence Completion Tests

We would briefly discuss each of these.

a. Word Association Test can be defined as a semistructured, individually administered, projective technique of personality assessment that entails the presentation of a list of stimulus words, to each of which an assessee responds verbally or in writing with whatever comes to mind first upon hearing the word. Responses are then analysed on the basis of content and other variables.

This test was originally known as the free association test and was first systematically described by Francis Galton, half cousin of Charles Darwin, in 1879, Wilheirn Wundt, the father of experimental psychology, subsequently introduced it into psychoanalytic researchers. Carl G. Jung (1910) used it as a psychiatric screening instrument by providing objective scoring and statistical norms. Forensic psychology made use of it as a 'lie detector'. It was Jung (1910) who pointed out how word association test can be used as a lie detector. Burt in 1931 and Lindsley in 1955 carried out extensive research demonstrating its utility and reliability as a 'lie detector'

There have been many versions of the word association test. The three most important are as follows :

(i) Jung (1910) used a list of 100 words to represent' common 'emotional complexes'. The subject is told-that the examiner will speak a series of words, one at a time, after each word, the subject is to reply as quickly as possible with the first word that comes to mind. There is no right or wrong responses. The examiner records the reply to each stimulus word, the reaction time and any unusual speech or behavioural manifestations accompanying a given response. Analysis of the content of responses, reaction times and other aspects of overt behaviour aid in discovering certain emotional problems and in drawing certain inferences which are then used for further psychological exploration by interview.

(ii) In 1968, Rapaport and his associates at the Menninger clinic developed another version of word association test which was very similar to Jung's word association test. The test developed by Rapaport and his associates consists of 60 words list. The approach used by them reflects strong psychoanalytic orientation and many of the words are associated with psychosexual conflicts. This test aids us in two ways :

- One in detecting the impairment of thought process and
- Second to suggest areas of significant internal conflicts.

Analysis of the results is done on the basis of popular responses, reaction time, associative disturbances and impaired reproduction on retest.

(iii) Still another version of the word association test was developed by Kent-Rosanoff called as the Kent-Rosanoff Free Association Test to differentiate between mentally ill and the normal. It consists of 100 stimulus words. They have provided a percentage in each category was expected to differentiate the normal from abnormal. However, the diagnostic use of this test declined with the gradual realisation that response frequency various with age, socio-economic status, educational level, regional and cultural background, creativity, etc. Hence, proper interpretation of scores requires norms of many subgroups as well as supplementary information about the examinee, which no doubt is a very difficult task.

b. Sentence Completion Tests: Sentence completion test is another verbal projective technique that has been extremely used in research and clinical practice. A wide variety of sentence completion test are presently available. The content of a particular test and the nature of the sentences will depend upon the group of persons and the purpose for which they are intended.

In this test, any individual is presented with a series of incomplete sentences, generally open at the end, to be completed by him in one more words. They resemble the word association test. However, sentence completion test is regarded as superior to word association test because the subject may respond with more than one word, greater flexibility and variety of responses are possible and more areas of personality and experiences may be tapped.

Some of the most commonly used completion tests are as follows :

(i) Sack's Sentence Completion, which is commonly used in clinical practice, consists of 60 sentences stems which can tell us about an individual's adjustment in family area, sex area, about his interpersonal relationship, self-concept and goals. Sample items are as follows :

- Some day, I ...
- My sex life ...
- If I were in charge ... etc.

The subject has to complete these sentences by writing the first word or few words that comes to mind.

(ii) Another important and widely used such test was developed by Rotter who called it as Incomplete Sentence Blank which consists of 40 sentences stems and is similar to Sack's completion test except that it is scored more rigidly and precisely.

(iii) A novel approach to sentence completion technique is to be seen in Washington University Sentence Completion Test (WUSCT) which is largely based on Loevinger's broadly defined construct of ego development. This test classified responses with reference to a seven stage scale of ego development as follows: pre-social and symbiotic, impulsive, self-protective, conformist, conscientious, autonomous and integrated. Although most of the research with this test has been done with adult women, forms are also available for use with men and with younger persons of either sex.

In general, a sentence completion test may be useful for obtaining diverse information about an individual's interests, educational aspirations, future goals, fears, conflicts, needs, etc. The sentence completion tests have high degree of face validity. However they are vulnerable for "faking good" or "faking bad" on the part of the examinee.

11.7.5 Sound as Projective Stimuli: Though not much popular, it was the behaviourist B.F. Skinner who developed this test. Skinner created a series of recorded sounds to which people were told to respond. Saul Rosenweig as well as David Shakow also did some pioneering work with audition as a projective techniques. Behaviorist B. F. Skinner is not typically associated with the fields of personality assessment or projective testing. However, early in his career Skinner developed an instrument he named the verbal summator, which, at one point, he referred to as a device for snaring out

complexes," much like an auditory analogue of Rorschach inkblots. Skinner's interest in the projective potential of his technique was relatively short lived, but whereas he used the verbal summator to generate experimental data for his theory of verbal behavior, several other clinicians and researchers exploited this potential and adapted the verbal summator technique for both research and applied purposes. The idea of an auditory inkblot struck many as a useful innovation, and the verbal summator spawned the tautophone test, the auditory apperception test, and the Azzageddi test, among others.

11.7.6 The Production of Figure Drawings: Analysis of drawings is another projective method. Drawings, especially in the case of young children provides a wealth of diagnostic information about the clinical aspects of individual's personality functioning..Today, the use of drawings in clinical and research settings have extended beyond the area of personality assessment. Attempts have been made to use artistic productions as a source of information about individual's intelligence, neurological impairment, visual motor coordination, cognitive development and learning disabilities.

Figure drawing tests are projective methods of personality assessment that entails the production of a drawing by the assessee, which is analysed on the basis of its content and related variables.

Karen Machover did classic work on figure drawing tests.

One of the well-known figure drawing test is the one developed by Karen Machover called as Draw-A-Person Test. In this test, the examinee is provided with a paper and pencil and is told simply to 'draw a person'. Upon completion of the first drawing, he or she is asked to draw a person of the opposite sex from that of the first figure.

While the individual draws, the examiner notes is or her comments, the sequence in which different parts are drawn and other procedural details. The drawing is normally followed by an inquiry in which the examinee is asked to make up a story about each person drawn. A series of questions is also asked to the subject about the person drawn.

Scoring of this test eat essentially quantitative and is lacking in validational studies. Clinician and researchers now opine that the draw a person test can serve best not as a psychometric test but as the clinical instrument, in which, the drawings are interpreted in the context of other information about the individual.

Another well-known drawing test is the House-Tree-Person test. Still another test that can help the examiner to understand and examinee in relation to his/her family is the Kinetic Family Drawing (KFD). In this test a child is given a paper 8 1/2 by 11 inch sheet, a pencil and an eraser and is told to "draw a picture of everyone in your family, including you, doing something."

Emphasis is laid on the interaction between various family members.

Figure drawing tests are clinically highly useful, especially while dealing with children. However these tests have its own disadvantages and reliability and validity data on these tests is lacking.

11.8 SUMMARY

In this unit we have discussed the Definitions of Personality and Personality Assessment. We also distinguished between the concepts of Traits, Types and States. Some important basic questions related to personality assessment were discussed. The questions pertained to Who, what, where and how of personality assessment.

Some important techniques used in personality assessment such as logic and reason, theory, data reduction methods, criterion groups were briefly discussed. Following this we discussed the relationship between Personality Assessment and Culture

Objective Methods of Personality Assessment were discussed. These methods mostly use the paper and pencil tests. Lastly we discussed the various Projective Methods of Personality Assessment at length. Some of the most commonly used projective methods include the Rorschach Inkblot test, The Thematic Apperception Test, Hand Test Rosenzweig Picture- F rustration Study, Word Association Test, Sentence Completion Tests, Sound as Projective Stimuli and Figure Drawings which includes the Draw-A-Person Test and House-Tree-Person test.

11.9 QUESTIONS

1 . Define the concept of Personality and Personality Assessment.
2. Explain the terms Personality Traits, Personality Types and Personality States.
3. Explain some basic questions with regards to personality assessment such as Who, What Where and How of assessment.
4. Write a note on
   a. Personality Assessment and Culture.
   b. Objective Methods of Personality Assessment
   c. Sound as Projective Stimuli
5. Discuss Inkblot and Pictures as Projective Stimuli
6. Write a note on Word Association Test and Sentence Completion Tests
7. Discuss Analysis of Drawings as a projective method

11.10 REFERENCES

1 Cohen, JR., & Swerdlik, M.E. (2010). Psychological Testing and Assessment: An introduction to Tests and Measurement. (7 th ed.). New York. McGraw-Hill International edition.

2. Anastasi, A. & Urbina, S. (1997). Psychological Testing. (7th ed.). Pearson Education, Indian reprint 2002.

12

TYPES OF SCORES, TYPES OF SCALES, FREQUENCY DISTRIBUTION AND GRAPHIC REPRESENTATIONS

Unit Structure

12.0 OBJECTIVES

- To understand the concept of measurement and the various scales of measurement.
- To understand the ways of preparing a frequency distribution.
- Understand various graphical representations of data like frequency polygon, histogram, and ogive

12.1 INTRODUCTION

As we have discussed about measurement in our unit on psychological testing, we shall devote less on the concept of measurement in this unit. We shall focus on scales of measurement and how data is graphically presented.

> "Measurement is an act of assigning numbers or symbols to characteristics of things (people, events,) according to rules." (Cohen and Swerdlik).

When we collect data for a research we use certain scales. For example if a researcher has to determine a distance between point A and point B, what would he do? Obviously he needs to take a measuring tape. And depending upon the distance, he can choose whether to take a scale that measures in mm, inches or foot. Or if the researcher is interested in weights, he would use the weighing scale (again depending on the weight he needs to measure whether in kgs, mgs or tones). Thus, depending on what has to be measured, the researcher will determine its scale. Understand that the researcher usually needs to make such and other such relevant decisions in order to make good research. If he makes any error in this, the entire research including the results will be affected and so will be its interpretation. In psychological testing the scales determine the kind of statistical evaluation that could be done. Thus, selection of scales is a key task.

After collecting data, it has to be organised so that it becomes easily comprehensible. The raw data (which could consist data form a large number of people) would hardly make any sense if we use the data in (as it is) raw format. We need to club data or condense it so that we can easily comprehend it, but at the same time the data should not lose its intrinsic value. We shall study frequency distribution and its graphical representation in this unit to precisely understand this process.

12.1 SCALES OF MEASUREMENT

A scale is a set of numbers (or the symbols) whose properties model empirical properties of the objects to which the numbers are assigned (Cohen and Swerdlik). In other words measurement is the assignment of numbers to objects or events in a systematic fashion. In psychological testing, scales of measurement refer to ways in which variables/numbers are defined and categorised. Each scale of measurement has certain properties which in turn determine the appropriateness for use of certain statistical analyses.

So a weighing machine is a scale that measures grams / kilograms (although, the choice of the machine will depend upon whether the weight has to be measured in kgs, tones or in mg.) or a foot ruler is a scale that measures length (again their choice would depend on what distance you wish to measure)

A researcher can categorise scales depending upon the type of variable. Thus, a scalE. used to measure continuous variables such as height, weight, temperature might be referred to as a continuous scale. In continuous scale there are no real breaks and can be theoretically subdivided into any number of units. For example it is possible to measure an individual's weight in milligram; whether such subdivisions are necessary and serve real purpose depends on the type of study and the discretion of the researcher. A discrete scale is used to measure a discrete variable, for example, if the research

subjects were to be categorised on the grounds of gender i.e., female or male, the categorisation scale would be said to be discrete, Discrete variables cannot be subdivided into smaller units.

Measurement always involves an element of error. Different factors in the environment and other irrelevant personal factors influence the test assessment. This combined influence of all of the irrelevant factors, on measurement is called an error. Error is an element of all measurement and thus must be accounted by any theory of measurement.

12.3 NOMINAL, ORDINAL, INTERVAL AND RATIO SCALES OF MEASUREMENT

The statistical data which includes numbers and sets of numbers has specific qualities. These qualities include magnitude, equal intervals, and absolute zero; accordingly that determines what scale of measurement will be used and therefore what statistical procedures would be the best. Magnitude refers to the ability to know if one score is greater than, equal to, or less than another score. Equal intervals means, that each of the scores are at an equal distance from the other score. Absolute zero refers to a point where none of the scale exists or where a score of zero can be assigned. When these three qualities are combined then, we can determine that there are four scales of measurement.

12.3.1 Nominal scales: Nominal scales are the simplest forms of measurement and involve classification or categorisation based on one or more distinguishing characteristics where all things measured must be placed into mutually exclusive and exhaustive categories. For example the 1 oth standard stu~ents are classified into division A and division B (randomly). Divisions labeled as 'A' division and 'B' division doesn't have any meaning, it is just a classification for two mutually exclusively groups. We can also use simple first names list of students, or alphabetical order, or the names on an organisational chart as a nominal category. The lowest level is the nominal scale, which represents only names and therefore has none of the three qualities.

12.3.2 Ordinal scales: Ordinal scales permit classification on some characteristic in a rank - order form. Even though ordinal scales may employ numbers or scores to represent the rank order, the numbers do not indicate units of measurement. It is any set of data that can be placed in order ranging from the highest / greatest to lowest, but where there is no absolute zero and no equal intervals. Examples of this type of scale would include Likert Scales and the Thurstone Technique. In business and organisational settings, applicants may be rank-ordered according to their desirability for a position. Or students may be given ranks on the basis of the merits / highest marks. Observe that this is not just random classification like division A / B, the

order has some significance. However, with this order, we will not be in a position to tell the quantitative difference, between the 1st rank and the 2 nd rank holders in terms of marks or merits. Also the difference between the first rank and second rank may not be equal to the difference between the second and the third rank. In other words the two ranks are not assumed to be equidistant.

Ordinal scales have no absolute zero point. So in case of students' ranks you cannot have a zero rank i.e., the student may be the last in rank but not zero. Zero is without meaning in such a test as there is no way to know the number of units that separate one test takers score from another test takers scores.

Note that there are limitations to the ways in which data from such scales can be analysed * statistically. For example, we cannot average the merits of the first rank student and the second rank student.

The central tendency of a group of items can be described by using the group's mode (or most common item) or its median (the middleranked. item) when we use an ordinal scale.

12.3.3 Interval scales : Interval scales contain equal intervals between numbers. Each unit on the scale is exactly equal to any other unit on the scale. For example, the difference between the 1st inch and the 2 nd inch on foot ruler is exactly same to the difference between the 2nd and the 3rd inch. Interval scale possesses both magnitude and equal intervals, but no absolute zero. With interval scales, it is possible to average a set of measurement and get a meaningful result.

Scores on many tests, such as tests of intelligence, and other quantitative attributes are analysed statically (in ways appropriate fordata) at the interval level of measurement.

The Likert scale, which uses interval scale, is used in survey research. Variables measured at the interval level are called "interval variables". The central tendency of an interval variable measured can be represented by its mode, its median, or its arithmetic mean.

12.3.4 Ratio scales : A ratio scale has a numbers on the scales that are equidistant, have magnitude and have a true zero point. The central tendency of a variable measured at the ratio level can be represented by its mode, its median, its arithmetic mean, its geometric mean and its harmonic mean. In this scale type, measurement is the estimation of the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind (Michell, 1999).

All statistical measures can be used for a variable measured at the ratio level and the data can be more easily analysed.

Finally, with a ratio scale, we also have a zero point where none of the scale exists; when a person is born his or her age is zero.

Scales of Measurement

Scale Level	Scale of Measurement	Scale Qualities	Example(s)	Permissible Statistics
4	Ratio	Magnitude, Equal Intervals, Absolute Zero	Age, Height, Weight, Percentage	All statistics permitted for interval scales plus the following: geometric mean, harmonic mean, coefficient of variation, logarithms
3	Interval	Magnitude, Equal Intervals	Temperature	mean, standard deviation, correlation, regression, analysis of variance
2	Ordinal	Magnitude	Likert Scale, Anything rank ordered	median, percentile
1	Nominal	None	Names, Lists of words	mode, Chi-square

Stevens (1946, 19121)

12.4 FREQUENCY DISTRIBUTIONS

Regardless of-whether manual or automated methods are used, it is usually necessary to code data numerically to facilitate further data analysis. Frequency distributions summarise and compress data by grouping it into classes and recording how many data points fall into each class. That is, they show how many observations on a given variable have a particular attribute. The frequency distribution is the foundation of descriptive statistics. It is a prerequisite for the various graphs used to display data and the basic statistics used to describe a data set -- mean, median, mode, variance, standard deviation, and so forth. Note that frequency distributions are generally used to describe both nominal and interval data, though they can describe ordinal data. A frequency distribution should be constructed for virtually all data sets. They are especially useful whenever a broad, easily understood description of data concentration and spread is needed.

How to make a frequency distribution:

Look at the scores below: These are the marks scored by psychology students

44,122,412,37,48,33,39,49,42,46,123,42,47,128,120

Do these marks make any sense to an observer? Except that they are marks of psychology students, they do not make sense to a common observer. So if the observer was a teacher she would want to know how her students have progressed or fared in their examination, i.e., she needs to understand how many of the students have scored below average or average, above average. If the observer were a statistician she would like to know how many students have got scores between particular scores series (interval). In order to know this classification of students we need to use the frequency distribution. Look at the frequency distribution table on the next page and you will understand at a glance how many students have scored what marks.

Now let's see how the above data is converted into a frequency distribution table

The following steps are involved in making -a grouped frequency distribution:

Step1. Collect raw data from entity records, interviews, surveys, etc. In this case we have considered marks obtained by students in their Psychology test.

Step 2: Find out the lowest and the highest number in the range of scores. Calculate the range by subtracting the lowest score from the highest score. In the above scores the highest number is 120 and the lowest number is 33. When we subtract 33 from 120 we get 17. We would require approximately 12-7 Class intervals. In this case we have decided to take 6 class intervals.Remember that each of the class intervals has to be of same length (in this case we have taken 12 numbers in each interval). The number of class intervals and the size of each class interval have to be decided by the test user and usually made on the basis of convenience. However, he has to ensure that the data doesn't get too condensed nor is it spread too much. The class interval is denoted by 'I'.

Step 2. Make a table with three columns. The first column is for the class intervals (C.I.), the second column for making tally marks and the third column is for the frequency (f) of that class interval.

Step 3. Classify each of the scores into class intervals. So if a student has scored 44 add the frequency to the class interval of 41 - 412. Make a tally mark. After ever four tally marks the fifth tally mark is a slanted line over the four tally marks. Tally marks are a quick way of keeping track of numbers in groups of five.

Step 4. After entering all the students to their respective class intervals, Count the total number of frequencies. The total number of frequencies should be equal to the total number of students.

Now let's make a frequency table as per the instruction given above:

(Frequency Distribution) Table - I

C.I. (Marks obtained)	Tally marks	Frequency (No. of students in each class interval)
31 -312	|	1
36 - 40	||	2
41 - 412	||||	4
46 - 120	𝍸	12
121 - 1212	||	2
126 - 60	|	1

12.5 THE ADVANTAGES AND DISADVANTAGES OF FREQUENCY DISTRIBUTION

12.5.1 Advantages of the frequency distributions:

1. It condenses and summarises large amounts of data in a useful format. A format that is easily comprehensible and describes all variable types.
2. It facilitates graphical presentation of data.
3. It helps to identify population characteristics.
4. It permits cautious comparison of data sets.

12.5.2 Disadvantages of frequency distributions :

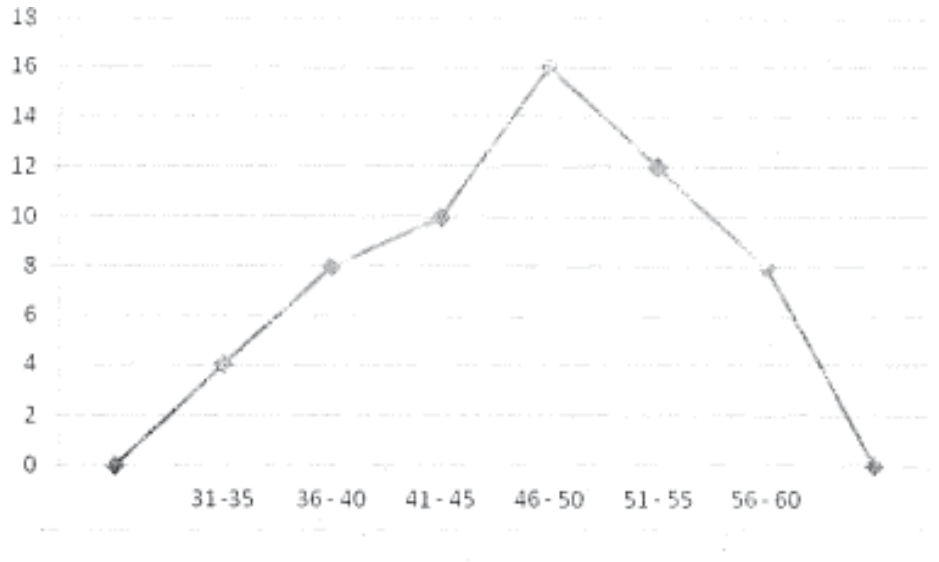
1. They reveal little about the actual distribution, skew, and kurtosis of data.
2. They can be easily manipulated to yield misleading results.
3. They can de-emphasise ranges and extreme values, particularly when open classes are used.

12.6 FREQUENCY POLYGON, HISTOGRAM, CUMULATIVE FREQUENCY CURVE, OGIVE

There are many forms of representing data graphically. They are histograms, frequency polygons and ogive. Let's understand each of them.

12.6.1 Frequency polygon: Frequency polygons are a graphical device for understanding the distributions. Frequency polygons are expressed by a continuous line connecting the points. They are especially helpful in comparing sets of data and a good choice for displaying cumulative frequency distributions.

To create a frequency polygon first draw an X-axis on a graph paper. It is the horizontal line, representing the values of the class interval; usually the test scores or class intervals are indicated on the X- axis. Mark the middle of each class interval and label it with the middle value represented by the class. Note that it is assumed that all the scores in the class interval are concentrated at or represented by at the midpoint of the class interval. Draw the Yaxis to indicate the frequency of each class. Y axis is the vertical line which meets the X axis at 900. Place a point in the middle of each class interval at the height corresponding to its frequency. Finally, connect the points. You should include one class interval below the lowest value in your data and one above the highest value. The graph will then touch the X-axis on both sides. Now let's plot the polygon using the marks of our Psychology students.



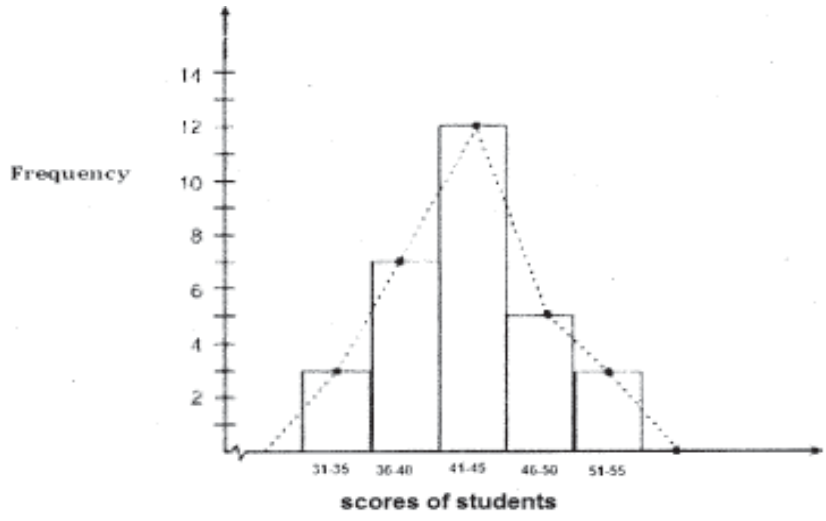
* For the sake of understanding we have shown the entire class interval on the X axis. Usually we should use the midpoint of the class intervals on the X axis.

12.6.2 Histograms

A histogram is a type of graph which uses vertical lines to show the frequency of data items in successive numerical intervals of equal size forming an adjacent series of rectangles: In other words a histogram is a graphical representation of a continuous frequency distribution i.e., grouped frequency distributions. It is a graph, including vertical rectangles, with no space between the rectangles. Usually the independent variable (in this case the scores of psychology students) is plotted along the horizontal axis (i.e., the X - axis also known as the abscissa) and the dependent variable (in this case the frequency of students) is plotted along the vertical axis (i.e., the Y- axis also known as the ordinate). Remember the area of the rectangles must be proportional to the frequencies of the respective classes. A frequency polygon is constructed by joining the mid-points of the tops of the adjoining rectangles. The midpoints of the first and the last classes are joined to the mid-points of the classes preceding and succeeding respectively at zero frequency to complete the polygon. Look at the data below.

CI	Frequency
31 -35	3
36 - 40	7
41 - 45	12
46 - 50	5
51 - 55	3
Total	30

Now let's plot the histogram with the above data. Observe that if you connect the bars on their mid point you will see the line graph given above.



12.6.3 The cumulative frequency graph ogive:

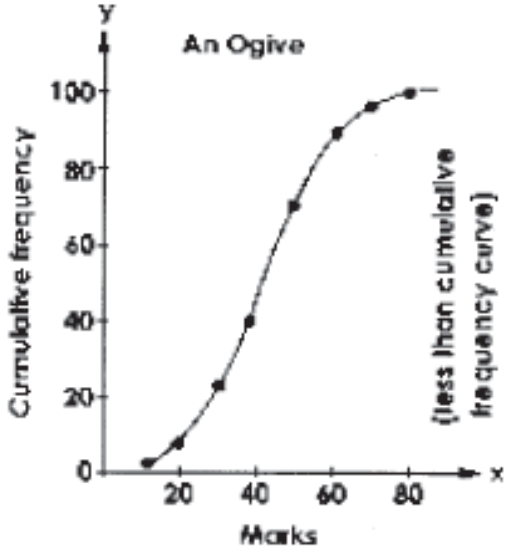
The cumulative frequency, also known as an Ogive, is another way to analyze the frequency distribution table. Unlike a frequency distribution which tells you how many data points are within each class, a cumulative frequency tells you how many are less than or within each of the class limits.

It is useful for analyses that require quick results about the proportion of data that lies below a certain level.

The cumulative frequency graph or ogive can be used to represent the cumulative frequencies for the classes. The cumulative frequency is the addition of all the frequencies accumulated from lower boundary up to the upper boundary of a class in the distribution.

Let's look at the example below and understand how to plot an ogive.

Marks	Frequency	Cumulative frequency
0 -10	2	2
10 - 20	8	10
20 - 30	12	22
30 - 40	18	40
40 - 50	28	68
50 - 60	22	90
60 - 70	6	96
70 - 80	4	100



Steps for constructing an ogive (*Ogive is pronounced as O-jive*)

Step 1: Find the cumulative frequency for each class.
Step 2. Draw the X and Y axis. Label the X axis with the class

boundaries. Use an appropriate scale for the y axis to represent the cumulative frequencies.

Step 3. Then plot the points with coordinates having X axis as marks and ordinates as the cumulative frequencies, (10, 2), (20, 10), (30, 22), (40, 40), (120, 68), (60, 90), (70, 96) and (80, 100) are the coordinates of the points. Plot the cumulative frequency at each upper class boundary.

Step 4: Join the points plotted by a smooth curve. Upper boundaries are used since the cumulative frequencies represent the number of data values accumulated up to the upper boundary of each class.

12.7 METHOD OF RUNNING AVERAGES

A running average is a method used to analyse a set of data points by creating a series of averages of different subsets of the full data set.

Given a series of numbers and a fixed subset size, the moving average can be obtained by first taking the average of the first subset. The fixed subset size is then shifted forward, creating a new subset of numbers, which is averaged. This process is repeated over the entire data series. The plot line connecting all the (fixed) averages is the moving average. A moving average. is a set of numbers, each of which is the average of the corresponding subset of a larger set of data points. A moving average may also use unequal weights for each data value in the subset to emphasize particular values in the subset.

A moving average is commonly used with time series data to smooth out short-term fluctuations and highlight longer-term trends or cycles. The threshold between short-term and long-term depends on the application, and the parameters of the moving average will be set accordingly. For example, it is often used in technical analysis of financial data, like stock prices, returns or trading volumes. Viewed simplistically, it can be regarded as smoothing the data.

Let's take an example to understand this. Look at the data below

A researcher has appointed research trainees for a survey. The job of the trainees is to collect a series of information units from their clients, for which they would be paid Rs.200 at the end of the day. The researcher wants to figure out how many units is each trainee giving for every Rs.200,they charge. The researcher has collected information about 10 research trainees at random and obtains the following results:

Trainees	No. of units collected
1	7
2	8
3	9
4	13
12	11
6	10
7	9
8	11
9	12
10	10

The computed mean or average of the data = 10. The manager decides to use this as the estimate for expenditure of a typical worker.

Let us set M, the size of the "smaller set" equal to 3. Then the average of the first 3 numbers is: (9 + 8 + 9) / 3 8.667.

No. of workers	No. of units delivered	Moving average	
1	9		
2	8		
3	9	8.67	9+8+9 =26 /3
4	13	10	8+9+13 = 30/3
12	11	11	9+13+11=33/3
6	10	11.33	13+11+10=34/3
7	7	9.33	11+10+7 =28/3
8	11	9.33	10+7+11=28/3
9	12	10	7+11+12=30/3
10	10	11	11+12+10= 33

This is called "smoothing" (i.e., some form of averaging). This smoothing process is continued by advancing one period and calculating the next average of three numbers, dropping the first number.

12.8 SUMMARY

1. Measurement is an act of assigning numbers or symbols to characteristics of things (people, events,) according to rules.
2. A set of numbers used to measure continuous variables such as height, weight, temperature might be referred to as a continuous scale. In continuous scale there are no real breaks and can be theoretically subdivided into any number of units.
3. A discrete scale is used to measure a discrete variable which cannot be subdivided into smaller units or of the values of the scale.

4. There are four scales of measurement viz., the nominal scale, the ordinal scale, the interval scale and the ratio scale.

5. Nominal scales are the simplest form of measurement and involve classification or categorisation based on certain characteristics where all things must be placed into mutually exclusive categories.

6. Ordinal scales permit classification on some characteristic in a rank - order form. It is any set of data that can be placed in order ranging from the highest / greatest to, but where there is no absolute zero and no equal intervals.

7. Interval scales contain equal intervals between numbers. Each, unit on the scale is exactly equal to any other unit on the scale.

8. A ratio scale has numbers on the scales that are equidistant, have magnitude and have a true zero point. All statistical measures can be used for a variable measured at the ratio level and the data can be more easily analysed.

9. Frequency distributions summarise and compress data by grouping it into classes and recording how many data points fall into each class.

10. Frequency polygons are a graphical device for understanding the distributions. Frequency polygons are expressed by a continuous line connecting the points.

11. A histogram is a graphical 'representation of a continuous frequency distribution i.e., grouped frequency distributions. A frequency polygon is prepared by joining the mid-points of the tops of the adjoining rectangles.

12. The cumulative frequency, also known as an Ogive, is another way to analyse the frequency distribution table. It is useful for analyses that require quick results about the proportion of data that lies below a certain level.

13. A running average is a method used to analyse a set of data points by creating a series of averages of different subsets of the full data set.

12.9 QUESTIONS

Answer the following questions:

Q1. What is frequency distribution? Explain the process of preparing a frequency distribution.

Q2. Explain the various ways to graphically represent frequency distribution.

12.10 REFERENCES

1. Cohen, R.J., & Swerdlik, M.E., (2010). Psychological testing and Assessment: An introduction to Tests and Measurement, (7th ed.), New York. McGraw - Hill International edition, 129 132.

2. Anastasi, A. & Urbina, S. (1997). Psychological Testing. (7th ed.). Pearson Education, Indian reprint 2002.

3. Kaplan, R.M., & Saccuzzo, D.P. (20012) . Psychological Testing - Principles, Applications and Issues. (6 th ed.). Wadsworth Thomson Learning, Indian reprint 2007.

4. http://en.wikipedia.org/wiki/Moving_average

13

MEASURES OF CENTRAL TENDENCY

Unit Structure

13.0 OBJECTIVES

After studying this unit you should be able to:

- Explain the meaning of measures of central tendency'.
- Define the three measures of central tendency.
- Calculate the mean, median and mode of ungrouped and grouped data.
- Calculate the 'mean' by using the assumed mean (short) method.
- Explain the advantages and limitations of using each of the three measures of central tendency.
- Compare and decide when to use the three measures of central tendency.

13.1 INTRODUCTION

Measures of central tendency are a single number representation of the central position in a set of data. It is an 'average' which represents all the scores in a set. That single number is a representation of the group as a whole. Such representative numbers make it possible to compare two or more groups as a whole. There are three such representative measures - the mean, the median, and the mode. Each of these are explained below, along with the way in which these can be computed from a set of scores which may be 'grouped' or 'ungrouped'. The term 'average' is a common word indicating any type of 'central tendency.

In this unit we will discuss the various measures of central tendency and learn how to compute them for grouped and ungrouped data.

13.2 The Mean

The 'arithmetic mean' or the 'mean' is the sum of the individual scores divided by the number of scores.

13.2.1 calculation of mean from ungrouped data

In a family of 5, the height of the five brothers is 157 cm, 162cm, 145cm, 146 cm, and 160cm the mean height of the five is obtained by adding the five different heights and dividing by the number of persons (05, also called as scores or readings). The formula would read as

M = Σ X / N
Where M = mean
Σ = the sum of
X = the individual scores
N = the number of measures (or scores or readings)
In the above case the mean is calculated as follows
157 + 162 + 145 + 146 + 160 = 770

$$\overline{X}_{\text{Mean}} = \frac{\sum X}{N}$$

$$\overline{X}_{\text{Mean}} = \frac{770}{5}$$

$$\overline{X} = 154$$

We may say that the mean height of the family is 1134 cm.

In brief:

The mean is defined as the sum of all scores divided by the number of scores

Practice Problem 1:

A group of students have the following scores in a mathematics test:
56, 45, 87, 98, 25, 26, 64, 62. What is the mean score of the group?

Check your answer at the end of the chapter.

13.2.2 Calculation of Mean from Data Grouped in a Frequency Distribution

When the number of scores to be averaged are large it becomes cumbersome to use the above method of finding the mean. The scores can then be

grouped into a frequency table (as explained in the unit 12), and a slightly different method is used to calculate the mean. The example given below will be used to illustrate how the calculation is done.

A class of 50 students have the following total scores on a maths test:
197,193,191,189,187,186,185,184,184,184,184,180,179,179,178, 177,177,176,175,175,174,174,174,173,172,172,172,172,171,170, 169,169,168,166,166,165,164,164,164,164,158,157,156,156,154, 150,149,146,1413, 142

The highest score obtained is 197 and the lowest score obtained is 142, a range of 55 scores.

If we were to group these scores so that we have approximately 10 -12 groups we should have class intervals (groups) of S. It would be convenient to have the lowest score starting from 140.

When classified by their scores, the grouped scores look like this
(197),
(193,191),
(189,187,186,185),
(184,184,184,184,180),
(179,179,178,177,177,176,175,175),
(174,174,174,173,172,172,172,172,171,170),
(169,169,168,166,166, 165),
(164, 164,164,164),
(158,157,156,156),
(154,150),
(149,146,145),
(142)
These may be tabulated as below:
Against each class interval are shown the scores which fall in each class interval. The third column indicates the number of scores or the 'frequency' within each class interval.

Class Interval	Scores	Freq-uency
195-199	197,	1
190-194	193,191,	2
185-189	189,187,186,1813,	4
180-184	184,184,182,184,180,	5
175-179	179,179,178,177,177,176,175,175,	8
170-174	174,174,174,173,172,172,172,172,171,170,	10
165-169	169,169,168,166,166, 165,	6
160-164	164, 164,164,164,	4
155-159	158,157,156,156,	4
150-154	1134,1130,	2
145-149	149,146,1413,	3

This information is now transposed into the table below. Column 2 indicates the mid-point of the class interval. The mid-point of each class interval can be easily calculated by the

formula:
(Lowest score + highest score)/2 e.g. Te midpoint of the first class interval is (195+199)/2 = 39412 = 197

It is assumed that all the scores in this class interval lie at 'midpoint'.

1	2	3	4
Class Interval	**Midpoint X**	**Frequency f**	**f*X**
1913-199	197	1	197
190-194	192	2	384
1813-189	187	4	748
180-184	182	13	910
1713-179	177	8	1416
170-174	172	10	1720
1613-169	167	6	1002
160-164	162	4	648
11313-1139	1137	4	628
1130-1134	1132	2	304
1413-149	147	3	441
140-144	142	1	142
N=130		Σf*X=81340	

Since it is assumed that all the scores in a class interval fall at the mid-point, Σ*X indicates the sum of all the scores in that class

interval and Σf*X the sum of all 130 scores of the class.

Since Mean Σ-f*X / N

Here Mean 81340/130
=170.80
In brief:
The Mean Score (170.80) is the sum of all the scores obtained (81340) divided by the number of scores (130)

You will note that, though the grouping of scores into a frequency table, makes it easier to deal with large amounts of data, we have made an assumption-that all the scores in a particular class interval actually lie at its midpoint. This may not be actually true. To this extent, the mean, calculated by this method is compromised.

Practice Problem 2

The length of string used by a group of students to complete a craft project has been grouped into class intervals of 13'. The number of students using that length of string is provided in the table below. Use this data to compute the mean length of string used by the class .

Length of String used (m)	Number of students
5-9	8
10-14	17
15-19	20
20-24	20
25-29	18
30-34	11
35-39	6

Practice Problem 3

Thirty-five practice sessions were held for a group of 20 students preparing for the athletics-meet to be held in the school.

Students' attendance was marked, and the attendance ranged from 6 turns to 32 turns.

The attendance has been put into the frequency table below. Calculate the midpoint of each interval and compute the mean attendance of the group.

(No. of days attended) Class Intervals	(No of students) Frequency
5-10	1
10 - 15	4
15 - 20	6
20 – 25	4
25 -30	2
30 - 35	3
N	**20**

Practice Problem 4

Scores obtained by 136 students of a class, in a Physics test, are tabulated below.

Calculate the mean score of the class in Physics.

Scores	Frequency
90-94	2
85-90	2
80-84	4
75-79	8
70-74	6
65-69	11
60-64	9
55-59	7
50-54	13
45-49	0
40-44	2
N	136

Check your answers at the end of the chapter

13.3 THE MEDIAN

The median is the score below which fifty percent of the scores lie. For example, if the scores of a group of seven are 6 7 8 (9) 10 11 and 12 the median score is 9 since it is the score which lies midway in the series.

13.3.1 Calculation Of Median From Ungrouped Data

Two situations may be possible when we have a list of ungrouped data - the number of scores may be odd or they may be even. In the case of an odd number of scores the computation of the median is fairly simple. It is simply the middle score as there are an equal number of scores above as well as below it. As in the above example, 9 is the median as there are three scores above it and 3 scores below it.

In the case of an even number of scores there is no one score above and below which an equal number of scores lie.

For example, in the series

7 8 9 10 11 12

There are only six scores. The median would lie somewhere between the scores of 9 and 10. Since the point exactly mid-way between 9 and 10 is the median i.e, 9.5

The formula for computing the median of a series of ungrouped scores is

Median = (N+1)/2 th measure in order of size.

Steps in the calculation of median would be:

1 Arrange he scores in order of size

2 Use the above formula to calculate the median

In the first examples above, using the formula, media will be the (7+1)/2 th score - that is the 4th score (which is 9) In the second example above, using the formula, the median would be (6+1)/2 th score . That is the 3.13th score in order of size - that is exactly halfway between the scores of 9 and 10 (that is 9.5)

13.3.2 Calculation Of Median From Data Grouped In A Frequency Distribution

When data are grouped into a frequency distribution, the median, by definition is the 130%th point of the distribution. The formula used to compute the median is

Mdn = 1+{(N/2-F)/fm} xi

Where

I = the exact lower limit of the class interval in which the median lies N/2 = one half of the number of scores

F = sum of number of scores (frequencies) on all intervals below I

fm = frequency (number of scores) within the interval upon which the me dian falls i = size of the class interval

Consider the data in the frequency distribution below:

1 Class Interval	2 Midpoint X	3 Frequency f
195-199	197	1
190-194	192	2
185-189	187	4
180-184	182	5
175-179	177	8 ↓ **20**
170-174	172	10
165-169	167	6 ↑ **20**
160-164	162	4
155-159	157	4
150-154	152	2
145-149	147	3
140-144	142	1
N= 50		

The median obviously lies in the class interval 170-174
1 = 169.5 (lower limit of the class interval 170-174)
(N/2 -F)= (50/2 -20) = (25-20) = 5
fm = 10 and i = 5

Thus, Mdn = 169.5+ {(25-20)/10)x5
= 169.5+ 0.5x5
= 1 69.5+2.5
= 172

Steps In The Computation Of Median From Grouped Data

1 Find N/2 - That is, one half of the cases in the distribution
2 Begin at the small-score end, - count off the scores in order, up to the exact lower limit(l) of the interval which contains the median. The sum of these is F.
3 Compute the number of scores necessary to fill out N/2 i.e.,(N/2-F). Divide this quantity by the frequency (fm) on the interval which contains the median; and multiply the result by the size of the class interval (i).
4 Add the amount obtained by the calculation in step 3 to the exact lower limit (1) of the interval which contains the median.
5 The result is the median of the distribution.

In Brief:
Median is defined as that score below, and up to which, lie 130 of the scores in a distribution. It is the middle score in a distribution which has been sequenced by value.

Practice Problem 5
Find the median in the following distribution of scores:

Scores	Frequency
70-71	2
68-69	2
66-67	3
64-65	4
62-63	6
60-61	7
58-59	5
56-57	4
54-55	2
52-53	3
50-51	1
N	39

Practice Problem 6

Find the median in the following distribution of scores:

Scores	Frequency
90-94	2
85-90	2
80-84	4
75-79	8
70-74	6
65-69	11
60-64	9
55-59	7
50-54	5
45-49	0
40-44	2
N	56

Check your answers at the end of the topic.

13.4 THE MODE

The mode is described as that one score which occurs most frequently.

For example, in the series 9, 10, 10, 11, 11, 12, 12, 12, 13, 13, 14, 14 the score that occurs most frequently is '12'. This is the crude mode' in this distribution of scores.

In the case of data grouped into a frequency distribution the crude mode is taken to be the mid-point of the class interval that contains the highest frequency.

For example, in the frequency distribution below:

Class Interval	Frequency (f)	Midpoint (X)
195-199	1	
190-194	2	
185-189	4	
180-184	5	
175-179	8	
170-174	10	172
165-169	6	
160-164	4	
155-159	4	
150-154	2	
145-149	3	
140-144	1	
N=130		

You will note that the class interval that contains the greatest frequency is 170-174. The mid-point of his class interval is 172. This is the crude mode of the data.

However, the 'true mode' lies at the peak where there is the greatest concentration of scores in the distribution. When the N is large and the frequency distribution is smoothed the true mode closely approximates the crude mode. Ordinariily, however, the crude mode is only approximately equal to the true mode.

In Brief:

The Mode of a distribution is defined as that score which occurs most frequently in any distribution

13.4.1 Calculation Of mode

The formula for approximating the true mode when the distribution is symmetrical is:

Mode = 3Mdn - 2Mean

In the above distribution, Mean = 170.80 and Mdn = 172.00.

Thus Mode = 3x1 72 - 2xl70.8

= 516 - 341.6

= 174.4

This calculated mode is slightly greater than the crude mode (172). In different distributions it may vary from being slightly greater or slightly lower than the crude mode. In this sense it is an unstable measure of central tendency. However, this is not as critical as the 'mode' is often used just as a simple inspectional average for the distribution. It roughly indicates the center of concentration of the scores. It therefore need not be calculated as accurately as the mean or the median, and often the 'crude' mode will suffice.

Practice Problem 7

For the distribution below find the "crude mode" and calculate the mode using the formula

Scores	Frequency
90-94	2
85-90	2
80-84	4
75-79	8
70-74	6
65-69	11
60-64	9
55-59	7
50-54	5
45-49	0
40-44	2
N	56

Check your answers at the end of the topic.

13.5 CALCULATION OF MEAN BY "ASSUMED MEAN" METHOD

Calculation of the mean by method described above is generally referred to the 'long method' of calculating the mean. It does give accurate results but when the numbers are large it may entail tedious calculations. This has been overcome by using the 'short method'or theassumed mean method'of calculating the mean.

Consider the same distribution of scores given in the table below. The calculation of the mean using the short method has been shown below the table. This is explained in the text-box containing the steps in the caomputation:

1 Class Interval	2 Midpoint X	3 Frequency f	4 X"	5 fx"
195-199	197	1	5	5
190-194	192	2	4	8
185-189	187	4	3	12
180-184	182	5	2	10
175-179	177	8	1	8+43
170-174	172	10	0	0
165-169	167	6	-1	-6
160-164	162	4	-2	-8
155-159	157	4	-3	-12
150-154	152	2	-4	-8
145-149	147	3	-5-	15
140-144	142	1	-6	-6-55
		N=50		

AM = 172 c Σ fx'/N = -12/50 = -0.240
ci = -1.20 i = 5
M = 170.80 ci = -1.20

The assumed mean (AM) can be any number in the scores of the distribution.

However, it is generally most convenient to take the assumed mean to be the mid-point of the class-interval that contains the greatest frequency of scores. - in this distribution it would be 172.

The formula used to compute the Mean from an assumed mean (AM) is M=AM+ci

1 The first step is to tabulate the data into a frequency distribution.

2 "Assume" a mean near the centre of the distribution, preferably in the interval which has the largest frequency.

3 The next step is to find out the correction that must be applied in order to determine the correct mean. This is done as explained in steps 5 to 9.

4 In column 4 above, fill in the x' values. - x' indicates the value of the deviations of the midpoints in terms of the class intervals. This means we assign a value of '0' to the class interval in which the assumed mean (AM) lies. The x' of the class interval one above this is +1. The x' of the class interval two above is +2 and so on. Similarly, the x' of the class interval one below the one containing the AM is -1, the x' of the class interval two below is -2 and so on.

5 Next, weight each deviation (x') by its appropriate 'f'. Thus, we compute the values of fx' for each class interval and write these in column 5. The sum of these comes at the bottom of the column (Σ-fx'). In this case $\Sigma fx' = (+43 - 55) = - 12$.

6 Find the correction (in terms of the class intervals) c by computing $c = \Sigma\text{-}fx' / N$. In this case $c = -12/50 = -0.240$

7 Class interval i = 5.

8 Multply the correction by the interval length $c\ i = 5 \times 0.240 = - 1.20$.

9 Add ci algebraically to the Assumed mean and you have the Mean.

Practice Problem 8

In the following distribution of scores, find the mean using the "Assumed Mean" method:

Scores	Frequency
70-71	2
68-69	2
66-67	3
64-65	4
62-63	6
60-61	7
58-59	5
56-57	4
54-55	2
52-53	3
50-51	1
N	39

Check your answer at the end of the topic

13.6 COMPARISON OF THE THREE MEASURES OF CENTRAL TENDENCY

To the new student it may seem confusing to decide which measure of central tendency may the most appropriate to use.

The most mathematically robust measure is the 'Mean; as it is based on precise mathematical formula and includes all the scores in its calculation.

Some guidelines for choosing the right measure to use are provided here.

13.6.1 When to Use Mean:

- When scores are distributed symmetrically around a central point; That is when the distribution is not heavily skewed.
- When you need a measure of central tendency that has the greatest stability.
- When you need to calculate other statistics such as 'SD' (standard deviation), or 'r' (Coefficient of Correlation) later.

13.6.2 When to Use the Median:

- When the Exact Mid-point is wanted (The exact 130%th score).
- When the distribution contains some extreme scores which will affect the mean score.
- When a distribution contains certain scores which are merely known to be above or below the median.

13.6.3 When to Use the Mode:

- When a quick and approximate measure of Central tendency is all that is desired.
- When the most typical value is desired - that is, a value which occurs most frequently.

Answers to Practice Problems (PP)

PP1: M =57.875
PP2: M = 21.0
PP3: M = 20.25
PP4: M = 67.36
PP13: Mdn= 60.79
PP6: Mdn= 66.77
PP7: Mo = 67, 65.59
PP8: M = 60.76

14

MEASURES OF VARIABILITY, PERCENTILES AND PERCENTILE RANKS.

Unit Structure:

14.0 OBJECTIVES

1 . To understand and to impart knowledge about various measures of variability or dispersion, their uses, limitations and statistical approaches.
2. To create an awareness among students and to know the various measures or techniquee to calculate the measures of variability.
3. To provide basic knowledge about statistical procedures for calculating percentiles and percentile ranks.

14.1 INTRODUCTION

The concept of variability or dispersion is fundamental in statistics that shows how far the number of the series scatter or spread on either side of central tendency, so it measures the quantities of the distributions. It also indicates relative measures of deviation of a group.

In this chapter we shall discuss the various measures of variability, their merits, limitation, uses and methods of calculations.

We shall also deal with the best method or measure of variability. Comparison of the 4 measures of variability will be also explained in detail.

And finally we shall focus on the concept of percentile, its merits, limitation and uses, along with the methods of calculations of percentiles and percentile rank.

14.2 RANGE AND AVERAGE DEVIATION (AD)

Out of 4 measures of variability, Range and Average Deviation are two measures. They are not so important as quartile deviation (QD) and Standard deviation (SD).

However, Range is the simplest measure of variability. It is based on the value of the two extreme scores, an extreme score can alter the value of the Range.

It is defined as "the interval between the highest and lowest scores" i.e., Range = Highest - Lowest score.

For example, if the highest score in a raw score of 50 is 90 and the lowest score is 25, the range will be= 90-25=65

(R = H - L. i.e., 90-25=65 R)

Whereas, Avrage Deviation (AD) is concerned, it is rarely used but AD provides a solid foundation for understanding the conceptual basis of another more widely used measure, the SD.

However, it is the mean of the deviation of all separate scores in a series taken from the mean.

The formula of AD is: $AD = \frac{\sum |x|}{N}$

Σ = Sigma, sum total of
χ = deviation from mean.
N = total Number of scores.

14.2 QUARTILE DEVIATION (QD) AND STANDARD DEVIATION (SD)

Quartile Deviation (QD) is another measure of variability. It is one half of the scale distance between the 75th and 25th percentile in a frequency distribution. The 25th, percentile is Q1 and 75th percentile is called Q3, the first and third quartiles respectively. The formula of QD is:

$$QD = \frac{Q3 - Q1}{2}$$

However, in a perfectly symmetrical distribution, Q1 and Q2 is exactly the same distance from thelmedian (Q_2).

On the other hand standard deviation is one of the best methods or measures of variability. It is very useful measure of variation.

It is defined as "the square root of the arithmetic mean of the square of all deviations taken from mean."

The formulas of SD are:

i) $SD = \sqrt{\frac{\sum fx^2}{N}}$ Long method

Σ Sum total of
f frequencies in a frequency distribution
χ_2 = deviation from the mean
X^2 = deviation of mean square
N = total number of frequencies
= under root

ii) $SD = i\sqrt{\frac{\sum fx^2}{N} - c^2}$

Here c^2 = is the squared correction of fx'/N c
other calculations are as equal like long method.

14.4 CALCULATION OF THE 4 MEASURES OF VARIABILITY

14.4.1. Range = Highest - Lowest score. (R = H - L)
example, 100 - 25 = 75

14.4.2. Average Deviation (AD)
Formula = AD =(Σ l x l) /N
Where as x is a deviation from mean score
i.e., X-M= x (small letter of x)

Example: Find out AD from the following scores: 6, 8, 10, 12, 14.

Step - 1 find out the mean of these 5 scores by $X/N = \frac{6+8+10+12+14}{5}$

$= \frac{50}{5} = 10$ So mean is = 10

Step - 2 = find out x. Deviation of all the scores from the mean score, 10. will be, 6-10 or -4 ; 8-10 or-2; 10-10 or 00; 12-10 or 02 and 14-10 or 4. The sum of these 5 deviations, disregarding signs, is 12; and dividing 12 by 5 (N) will be 2.4 or AD.

We can put these 5 scores as under and thus can find out AD-

Scores	X
6	-4
8	-2
10	-0
12	2
14	4
50	12 disregarding sign (X)
N = 5	

Mean = $\Sigma X/N = \frac{50}{5} = 10$

So AD is = $(\Sigma|x|)/N = \frac{12}{5} = 2.4$

14.4.3. Quartile Deviation or Q.

To calculate Q, we must first calculate Q3 and Q1. The formula of Q is found from

$$Q = \frac{Q3 - Q1}{2}$$

To find Q3, the formula is:

Q3 = l + [(3N/4 - F)/fm] x ci and formula of Q1 is;

Q1 = l + [(N/4 - F)/fm] x ci

However, to find Q we shall start calculating first Q1 and the Q3. Calculating Q1 from the following frequency distribution by short-method (Table No. 1)

ci	frequencies
136-139	3
132-135	5
1214-131	16
124-127	23
120-123	52
116-119	49
112-115	27
1014-111	114
104-107	7
	N=200

$$Q1 = l + \frac{(N/4 - F)}{fm} \times ci$$

l = the exact lower limit of the interval in which the quartile falls i.e. 111.5

i = the length of interval =4

F= the sum total of all the frequencies below which Q1 falls. in this case 25
fm= the frequency on which the Q1 falls, i.e.,
27 N/4= Total N is divided by 4 = 50 (200/4).
Now applying this formula in frequency distribution Table No. 1, we get:

$$Q1 = 111.5 + \frac{(50-25)}{27} \times 4$$

$$Q1 = 111.5 + \frac{(25)}{27} \times 4$$

$$Q1 = 111.5 + 0.92 \times 4$$

$$Q1 = 111.5 + 3.68$$

Q1= 115.18 Q1= 115.18

$$Q3 = I + \frac{(3N/4 - F)}{fm} \times ci$$

where

I = the exact lower limit of the interval in which the Q3 falls i.e. 119.5

i = the length of class interval,4

F= the sum total of all the frequencies below which Q3 falls, 101

fm= the exact frequency on which the Q3 falls, 52

3N/4= 150 as N/4=50, so 3N/4=3x5O=l 50

Thus Q3:

$$Q3= 119.5 + \frac{(150-101)}{52} \times 4$$

$$Q3= 119.5 + (49) \times 4$$

$$Q3= 119.5 + 0.94 \times 4$$

Q3= 119.5 + 3.76
Q3= 123.26 Q3= 123.26

$$Q= \frac{Q3 - Q1}{2}$$

$$Q= \frac{123.26- 115.18}{2} \quad Q= \frac{8.08}{2} = 4.04$$

Q= 4.04

14.4.4 Standard Deviation (SD) or

We shall now calculate the standard deviation from the following frequency distribution. (Table No. 2)

(1) Scores	(2) f	(3) x'	(4) fx'	(5) fx'^2
95-99	1	7	7	49
90-94	2	6	12	144
145-149	4	5	20	400
140-144	5	4	20	400
75-79	14	3	24	576
70-74	10	2	20	400
65-69	6	1	6	36
60-64	4	0	0	0
55-59	4	-1	-4	16
50-54	2	-2	-4	16
45-49	3	-3	-9	141
40-44	1	-4	-4	16
N=50			fx'= 88	fx'^2= 2134

Formula for calculation of SD

$$SD = i \times \sqrt{\frac{\sum fx'^2}{N} - c^2}$$

Column (1) is class interval
Column (2) frequencies are given
Column (3) Deviations are taken from mean
Column (4) frequencies are multiplied by mean deviation
Column (5) is square of fx'

According to formula ;

i =5

fx'^2= 2134

N = 50

c= fx'/N= $\frac{1414}{50}$= 1.76

c^2= 3.0976

$$SD = i\sqrt{\frac{2134}{50} - 3.0976}$$

$$= 5\sqrt{\frac{2134}{50} - 3.0976}$$

$$= 5\sqrt{(42.68 - 3.0976)}$$

$$= 5\sqrt{(39.5824)}$$

SD = 5.6.29

SD = 31.45

14.5 MERITS, LIMITATIONS, USES OF RANGE A.D, Q.D. AND S.D.

14.5.1 Merits Limitations and uses of Ra~'ge

Merits of Range

(1) It is simplest measure of variability.

(2) It provides gross description of the spread of scores.

(3) It is most general measure of spread or scatter.

Limitations of Range

i) It is less used in other measures of statistics and maths.

ii) Its values are always changed.

iii) Except frequency distribution it is not used in other distribution, therefore its use is limited.

iv) It is also not used in algebraic treatment.

Uses of Range

a It is used when the data are too scattered.

b. It is used when the knowledge of extreme score is wanted.

c. It is also used when the median is the measure of central tendency.

d. It is used when the scores are likely to affect SD.

e. It is used when the 50% of score is of primary interest.

14.5.2 Merits, Limitations, Uses of Average Deviation (AD) Merits

(1) It is another measure of variability in whch all signs of deviation (+ or -) for calculation is treated as positive.

(2) Other merit of this measure is to weigh all deviations from the mean according to their size.

Limitations of AD

1. This method or measure is rarely used because of of algebraic signs .
2. It is useless measure for further operations.

Uses of AD

1. It is used in Psychology, Economics and Statistics.
2. It is very accurate measure.
3. It is not affected by extreme values of the distribution.
4. For all practical purposes, AD is replaced by SD.

14.5.3 Merits, Limitations and Uses of Quartile Deviation

Merits: Q is a very important measure of variability. This method is applied when 75th and 25th percentiles are required. In a normal distribution Q is called the probable error or (PE) is used interchangeably.

Limitations

(1) This measure does not calculate Q2, 50% percentile.
(2) This method is only used in Psychology and Statistics.
(3) This measure is not used for further operations.

Uses

i) It is very easy measure to calculate the percentiles.
ii) It is reliable measure because it is not affected by extreme values or scores.
iii) It is used when the median is the measure of central tendency.
iv) It is also used when extreme scores would influence the SID disproportionately.

14.5.4 Merits, Limitations, Uses of SD

Merits:

(1) SD as a matter of fact is the best method of variability because of its reliability and accuracy.
(2) This method is used in psychological research, and is also applied to estimate the population, significant differences between means computing coefficient of correlation.
(3) It is very good measure of regression equation.

Limitations:

(1) It is very complex method.
(2) It gives more weightage to extreme values.
(3) It is useless for measures of central tendency.
(4) Its operations are limited only for psychological research and statistics.

Uses

a. It is widely used in Psychology, Economics and Statistics.
b. It is more reliable than other measures of dispersion.
c. It is also used in the interpretation of curve.
d. It is used when coefficient of correlation is computed.
e. It is used for estimating the population mean.

With regard to reliability of SD, an important question is often as ked as which is the best measure of variability? why? To answer this question we can simply say that SD is the best method of variability. Why? Because of the following reasons (focusing on other three measures):

1) Range and AD are rarely used because of their limitations and further operations.

2) Q is used when 75th percentiles and 25th percentiles are required.

But SD is used for its wider scope and importance. It is equally important for its reliability and accuracy.

Keeping Psychological research aside, SD is used in further operations such as estimating the population mean, calculation of significant difference between two means, computing coefficient of correlation, without SD no Y can be calculated, and the like.

Therefore, we can reasonably say that SD is the best measure of variability.

Check your progress

Q1 Define/Explain the Range, AD, Q and SD.

Q2 Explain the uses of Range and Q.

Q3 Describe the limitations of AD and SD.

Q4 Calculate Q and SD from the frequency distribution given below

Q5. Find out AD from the following scores

6, 14, 13, 15, 17, 20 = (N=6)

14.5 COMPARISON OF THE 4 MEASURES OF VARIABILITY

It is hard to compare these four measures of variability because Range and AD are very simple measures of variability and rarely used in Psychological research and Statistical treatment. Due to their limitations, they are useless for further operations.

Quartile Deviation (Q), on the other hand, are useful measures of variability. Q is used when 75th and 25th percentiles are required but SD is used in variety of ways such as in calculating coefficient of correlation, estimating significant difference between means, regression equations and so on.

However, we shall compare these four measures of variability as under :

	Range	Q	AD	SD
1. Definition	The interval between the highest and the lowest scores.	It is one of the scale distance between the 75th and 25th Percentiles	It is the mean of the deviation of all separate scores in a series taken from the mean.	The square root of the arithmetic mean of the square of all the deviations taken from mean.
2. Merits.	It provides gross description of the spread score. it is most general measure of scatter	It is applied to find 75% and 25% of score or value.	In this measure, all signs of deviations(+ -) for computing AD is treated as Positive.	This is widely used measure of variability which is adequate and reliable and trustworthy. It is the best measure.
3. Uses	It is used when knowledge of extreme scores are wanted and when median is measure of the central tendency.	It is easy measure and reliable because it is not affected by extreme values or scores.	It is very accurate measure and is used in Economics, Psychology and Statistics.	It is the best measure of variability. It is widely used in estimating correlation, population means and regression equation.
4. Demerits	Its use limited, It is used only in frequency distribution. Its values are always changed.	It is always used in Psychology and Statistics but does not calculate 50% of percentiles.	it is rarely used because of signs of Algebra.	It is very complex measure which gives weightage to extreme values.
5. Formulas	Difference between highest score and lowest score R= H - L	$\frac{Q3 - Q1}{2}$	$(\Sigma \lvert x \rvert)/N$	For short method SD = $i\sqrt{\frac{\sum fx'^2}{N} - C^2}$
6 Applications	For calculating or preparing a frequency table, used in Psychology Economics, Statistics..	For calculating measures of variability used in Psychology and Statistics	For calculating mean deviation	For calculating correlation, coefficient of correlation, regression equation, estimating population mean, etc. It is also used in Psycholoy, Economics and Statistics.

Check your progress

Q1 Compare Range with AD.

Q2 Define Q and SD and make a comparison between these two . measures of variability.

Q3 What are the applications of Range, AD, Q and SD.

14.7 PERCENTILES - NATURE, MERITS, LIMITATIONS AND USES

14.7.1 Nature /Definition of Percentile

If we remember the calculation of median (50%) of score, we can be immediately acquainted with Percentile or percentage. To calculate median we get 50% of measure or scores. Similarly, Q3 gives us 75th percentile and Q1 gives us 25th percentile.

But percentile gives us any percentage, may be 10%, 20%, 30% and even 90% of scores. However, points below which lie, 10%, 45% and 85% or any percent of scores is called Percentiles. .

However, a percentile may be defined as an expression of the percentage of people whose scores on a test or measure fall below a particular raw score. Thus, a percentile (P) is a converted score that refers to a percentage of test takers.

14.7.2 Merits of Percentiles.

(1) The main advantage of percentile is to determine at which 10% or 43% of individual's scores or cases is located.
(2) Percentage are based upon the number of scores or cases falling within a certain Range.
(3) The distance between any two percentiles show a certain area or number of cases (N/1 0, N/20).

14.7.3 Limitations of Percentiles.

(1) When the number of scores in a distribution is small, percentiles are not used.
(2) When there is little or no significance in making distinctions in rank, percentiles are not used.
(3) There is a restriction in using percentiles that real differences between raw scores may be minimized but near the end of distribution it increases, and the errors may be even worse with highly skewed data.
(4) Except in calculation Percentile Point, they are used for further operations.

(5) Percentiles have limited scope in their application.

14.7.4 Uses of Percentiles

(1) The percentile technique is very easy in calculation.
(2) The percentile has another advantage of being easily understood.
(3) The percentile technique does not make any assumption with regard to the characteristics of the total distribution.
(4) This technique also answers the question: "Where does an individual's scores rank him in his group"? or, "Where does an individual's scores rank him in another group whose members have taken the same test?"
(5) The differences in scores between any two percentile points become greater as we move from the median (P50) toward the extremes.

Check your progress:

Q1 Define/Explain Percentiles and describe their merits.
Q2 Explain the limitations and uses of Percentiles.

14.7.5 Calculation of Percentiles and Percentile Rank

In percentiles we calculate 10%, 15%, 40% and so on. This indicates that 10% or 15% cases of individual score falls below in a distribution. In other words, 15th percentile or 10th percentile is the score at or below which 15% or 10% of the scores in the distribution fall.

Whereas Percentile Rank to indicates an individual's percentile rank on a test referring to the percentage of cases or scores lying below cR. For example, a person is having a percentile rank of 20 T20) is situated above twenty percent of the group of which he is a member or twenty percent of the group fall below this person's rank.

14.7.6. Methods of calculating Percentile points from distribution

(Table No.3)

Scores	f
95-99	1
90-94	2
85-149	4
80-84	5
75-79	14
70-74	10
65-69	6
(P_R) 60-64	4
55-59 I	4
50-54	2

45-49	3
40-44	1
	N=50

The method of calculating percentiles is the same as that of finding median. the formula is

$$Pp=\frac{(PN-F)\ Xi}{fp}$$

P= Percentage of the distribution wanted, e.g., 10%, 25%, 40%, etc.
I = Exact lower limit of the class interval upon which Pp lies
PN = Part of N to be counted off in order to reach Pp
F = Sum of all scores upon intervals below I
fp = number of scores with the interval upon which Pp falls.
i = length of the class interval
Calculation of Percentile Points, P10, P20 P30,P40,P50,P90
(Table No.3)

10% of 50 =5 $49.5 + \frac{(5 - 4)}{2} X\ 5 = 52.0$

20% of 50 =1059.5 + $\frac{(10 - 10)}{4} X\ 5$ 59.5

30% of 50 =1 5 $64.5 + \frac{(15 - 14)}{6} X\ 5$ 65.3

40% of 50 =2069.5 + $\frac{(20 - 20)}{10} X\ 5 = 69.5$

50% of 50 =2569.5 + $\frac{(25 - 20)}{10} X\ 5 = 72.0$ (Mdn)

90% of 50 =4584.5 + $\frac{(45 - 43)}{4} X\ 5 = 147.0$

Calculation of PR from the same above formula
and by same method. (Table No.3)

If we want to find the P_{R7} of a man who scores 63, what is the answer?

Score 63 falls on interval 60-64. There are 10 scores upto 59.5 exact lower limit of this interval (see table No. 3) and four scores spread over this interval. Dividing 4 by 5 (length of interval) gives us 0.8 score percent of interval. The score of 63, which we are finding is 3.5 score units from 59.5,exact lower limit of this interval. So multiply 3.5 by 0.8 we get 2.8 as the score-distance of 63 from 59.5, and adding 2.8 to 10 (number below 59.5) we get 12.8 where N lying below 63.

Hence the percentile rank of score 63 is 26.

P_{RS} from ordered data

There are so many instances where individuals and things can be put in 1-2-3-4 order with respect to some trait or characteristics,

such trait or quality cannot be measured directly. Then we apply the following formula to calculate P_R

The formula is ; $PR = 100 - \frac{(100\ R - 50)}{N}$

Example, if 20 Doctors/ Professors have been ranked from 1 to 20, it is possible to convert this order of merit into PR or scores on a scale of 100. Applying the formula we get:

Professor gets 1 first rank

$P_R 100 - \frac{(100 \times 1-50)}{20}$

$P_R\ 100\ \frac{(100 -50)}{20}$

$P_R\ 100\ - \frac{50}{20}$

$P_R\ 100 - 2.5$

$P_R 97$ or 97.5

Doctor gets 10th rank

SO $P_R = 100 - \frac{(100x1\ 0 -50)}{20}$

$P_R = 100 - \frac{(1000 -50)}{20}$

$P_R = 100 - \frac{(950)}{20}$

$P_R = 100 - 47.5$

$P_R = 52.5$ or 52

Check your progress

Q1 Define percentiles and calculate P_{60}, P_{70}, P_{80} from the distributions (Table No.3)

Q2 Calculate P_R 70 and P_R 75 from the same distibution.

14.8 SUMMARY

In this chapter we have discussed the nature, methods of variability along with its 4 important measures such as Range, Average deviation, Quartile deviation and Standard deviation. We have highlighted the merits, uses and limitations of these four variables with separate headings. We have also provided the measures or techniques for calculation of these four measures of variability. Mention is also made about the comparison of the 4 measures. We have also answered the question "which is the best measure of variability", but in brief.

We have also explained the nature, merits, limitations and uses of Percentiles. And finally, we have explained the formula of how to calculate Percentiles and Percentile Ranks.

14.9 MODEL QUESTIONS

Q1. (a) Explain SD and its uses.
(b) Calculate Q or SD for the following distribution

SCORES	F
50-54	3
45-49	6
40-44	9
35-39	15
30-34	19
25-29	12
20-24	14
15-19	6
10-14	2
	N = 140

Q2. (a) Explain the uses and limitations of Q.
(b) Calculate P140 and PR from the above frequency distribution (Ql.b)

Q3. Define percentiles and describe the uses and limitations of percentiles.

14.10 REFERENCES

1. Annapornaa R. et al (20014) -
A Handbook of Mathematics and Statistics,
Chetana Publications Pvt. LTD. 262, Khatauwadi,
Girgaon, Mumbai- 400004
2. Anastasi, A. and Urbina, S. (1997) -
Psychological Testing (7th ed) Pearson Education, India
Reprint -2002
3. Cohen, J.R. and Swerdik, m.E. (2010)
Psychological Testing and Assessments, An Introduction to

tests and Measurements (7th ed) Newyork McGraw-Hill International Edition.

4. Garrett, Henry, E. (1973) - Statistics in Psychology and Education, Vakills, Feffer and Simon Pvt. Ltd. Ballard Estate, Mumbai- 400001

14.11 GLOSSARY

Average deviation - It is the mean of the deviations of all the scores in series taken from the mean.

Percentiles - A percentile is an expression of the Percentage of people whose scores on a test or measure fall below a particular raw score.

Percentile Rank - indicates a person's percentile rank on a test referring to the percentage of cases or scores lying below it.

Quartic deviation - It is a measure of variability which indicates one half of the scale distance between the 75th and 25th percentile in a frequency distribution. The 75th percentile is called Q3 and 25th percentile is known as Q1.

Range - is an interval between the highest and the lowest. In short, it is defined as the H -L =R.

Standard deviation - The square root of arithmetic mean of the square of all deviations taken from mean.

Variability - points out as to how the number of the series scatter or spread on either side of central tendency, so it measures the quantities of the distribution.

15

Probability, Normal Probability Curve and, Standard Scores - I

Unit Structure

15.0 OBJECTIVES

After studying this unit you should be able to

a. Understand the concept of Probability.
b. Explain the characteristics, importance and applications of the Normal Probability Curve.
C. Know the Areas under the Normal Curve.

15.1 INTRODUCTION

In this unit we will first discuss the concept of probability. We will define probability and understand its meaning through some examples. The concept of Normal Probability Curve is very important in psychological measurement. We would discuss its characteristics as well as importance and the various applications of the said curve. Towards the end of the unit we will discuss the areas under normal curve.

15.2 THE CONCEPT OF PROBABILITY

The word probability or chance as it is sometimes called is very commonly used in day-to-day conversation. The theory of probability has its origin in the games of chances related to gambling such as throwing a die, tossing a coin, drawing cards from pack of cards, etc. Cardon, an Italian mathematician

was the first to write a book on the subject titled "Book on Games of Chances" which was published in 1663, after his death.

Galieo, an Italian mathematician was the first man to attempt quantitative measure of probability while dealing with some problems related to theory of dice in gambling. However, systematic and scientific foundation of the mathematics theory of probability was laid in mid-seventeenth century by two French mathematician B Pascal and Pierre de Fermat. Swiss m athematician James Bernoully came out extensive work in the following decades.

The probability of a given event is an expression of likelihood or change of occurrence of event. A probability is a number which ranges from zero to one, zero for an event which cannot occur and 1 for an event certain to occur.

Probability can be defined as the ratio of number of favourable results to the total number of results for e.g. say a coin is tossed and we want head at the top. Since there is only one head out of two possibilities (Head and Tail) the required probability of getting Head of Tail is 1/2.

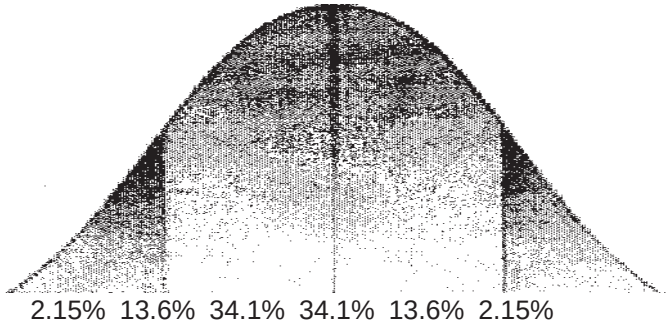
Whenever there is more than one result, the probability comes into existence and it is possible to use a procedure for calculating the probability in such cases. For certainty probability does not exist. But for every uncertainty probability should exist.

There are certain examples in which the question of probability does not arise at all. For e.g. when kerosene is poured into a fire, the result is certain and hence there is no question of thinking any other probability or when a stone is thrown in the air, there is absolutely no doubt that it will fall down. Hence there is certainty and no probability. Thus, we observe that when a result is certain the question of probability does not exist at all.

If you roll a six-sided die, there are six possible outcomes, and each of these outcomes is equally likely. A six is as. likely to come up as a three, and likewise for the other four sides of the die. What, then, is the probability that a one will come up? Since there are six possible outcomes, the probability is 1/6. What is the probability that either a one or a six will come up? The two outcomes about which we are concerned (a one or a six coming up) are called favorable outcomes. Given that all outcomes are equally likely, we can compute the probability of a one or a six using the formula:

15.3 CHARACTERISTICS, IMPORTANCE AND APPLICATIONS OF THE NORMAL PROBABILITY CURVE

Normal probability curve is one type of theoretical distribution that is of immense use in statistics. Normal probability curve is also called as the normal curve, the Gaussian curve (after a great German mathematician who investigated its properties and wrote the equation for it). It is also called as the Bell-Shaped curve or Mesokurtic curve (Mesos means middle or medium). The following figure depicts the Normal probability curve:



Much evidence has accumulated to show that the normal distribution serves to describe the frequency of occurrence of many variable facts with a relatively high degree of accuracy. Phenomenon which follow the normal probability curve (at least approximately) may be found in Biological statistics, Anthropological data, Social and Economic data, Psychological measurements and Errors of observation.

15.3.1 Characteristics of Normal Probability Curve:

Following are the major characteristics of Normal Probability Curve:

(1) The normal curve is symmetrical about the mean. The number of cases below the mean in a normal distribution is equal to number of cases above the mean. Hence, the mean and median coincide.

(2) The height of the curve is maximum at its mean. Hence, the mean and mode of normal distribution coincide. Thus, Mean, Median and Mode are equal in normal probability curve.

(3) There is one maximum point of the normal curve which occurs at the mean. The height of the curve declines as we go in higher direction from the mean. The dropping off is slow at first, then rapid, and then slow again. Theoretically, the curve never touches the base line. Its tail approaches but never reach the base line. Hence, the range is unlimited.

(4) Since there is only one point in the curve which has maximum frequency the normal roba ability curve is unimodal, i.e. it has only one mode.

(5) The point of inflection i.e. the points where the curvature changes in direction are each plus or /and minus one standard deviation from the mean ordinate.

(6) The height of the curve at a distance of one standard deviation from the mean is 60.7% of the height at the mean. The height of the curve at 2 and 3 standard deviation distances from the mean is 13.5% and 1. 1 % of the height at mean respectively.

(7) The total interval from plus one standard deviation to minus one standard deviation contains 68.26% of the cases. Similarly, 95.44% of the total area will be included between the mean ordinate and an ordinate 2 standard deviation from the mean. Similarly, 99.74% of the total area will be included between the mean ordinate and a point 3 standard deviation away from the mean.

(8) For the normal curve the valu of Ku = 0.263.

(9) In normal probability curve Q1 and Q3 are equidistant from the median. When there is any Skewness in the distribution, the two distances will be unequal.

(10) In the normal probability curve the height declines symmetrically in either direction from the maximum point.

(11) The Normal Curve is a Mathematical Model in Behavioural Sciences. The curve is used as a measurement scale. The measurement unit of this scale is plus or minus 1, 2, 3, etc., standard deviation

15.3.2 Importance And Applications of The Normal Probability curve

Normal probability curve has wide significance and applications in the field of measurement concerning education, psychology and sociology. The Normal Distribution is by far the most used distribution for drawing inferences from statistical data. Number of evidences are accumulated to show that normal distribution provides a good fit or describe the frequencies of occurrence of many variables and facts in (i) biological statistics e.g., sex ratio in births in a country over a number of years, (ii) the anthropornetrical data e.g., height, weight, (iii) wages and output of large numbers of workers in the same occupation under comparable conditions, (iv) psychological measurements e.g., intelligence, reaction time, adjustment, anxiety and (v) errors of observations in Physics, Chemistry and other Physical Sciences.

The Normal distribution is of great value in educational evaluation and educational research, when we make use of mental measurement. It may be noted that normal distribution is not an actual distribution of scores on any test of ability or academic achievement, but is instead, a mathematical

model. The distribution of test scores approach the theoretical normal distribution as a limit, but the fit is rarely ideal and perfect.

Some of its main applications are as follows:

(1) **Use as Model:** Normal curve represents a model distribution. For this reason, it may be used as a model.

(a) To compare various distributions with it, to say, whether the distribution is normal or not, if not what way it diverges from the normal.

(b) To compare two or more distributions in terms of overlapping, and

(c) To distribute short marks and categorical rating.

Often, phenomena in the real world follow a normal (or near normal) distribution. This allows researchers to use the normal distribution as a model for assessing probabilities associated with real-world phenomena. Typically, the analysis involves two steps.

- Transform raw data. Usually, the raw data are not in the form of z-scores. They need to be transformed into z-scores, using the transformation equation such as : z = (X- p) / a.
- Find probability. Once the data have been transformed into scores, you can use standard normal distribution tables, online calculators (e.g., Stat Trek's free normal distribution calculator), or handheld graphing calculators to find probabilities associated with the z-scores.

(2) **To compute Percentile and Percentile Ranks:** Normal probability curve may be conveniently used for computing percentile and percentile Ranks in a given normal distribution.

(3) **To understand and apply the concept of Standard Error of Measurement:** The normal curve is also known as the normal curve of error, or simply the curve of error on the grounds that it helps in understanding the concept of standard errors of measurement. For e.g., if we compute mean for the distributions of the various samples taken from a single universe (population), then, these means will be found to be distributed normally around the mean or the centre of population. The sigma distance of a particular sample mean may help us to determine standard error of measurement for the mean of that sample.

(4) **For ability grouping :** A group of individuals may be conveniently grouped into certain categories like good, average, poor, etc. in terms of some trait (assumed to be normally distributed) with the help of the normal curve.

(5) **To convert Raw Scores into Comparable Standard Normalized Scores :** Sometimes, we have records of an individual performance-on two or more different kinds of measurement and we wish to compare his score on one measurement with the score on the other measurement unless the scales of these two tests are the same, we cannot make a direct comparison. With the help of the normal curve, we can convert the raw scores belonging

different scales of measurement into standard normalized scores like Sigma (or Z scores) and T scores.

(6) To determine the relative difficulty of test items : Normal curve provides the simplest rationale method of scaling test items for difficulty and therefore, may be conveniently employed for determining the relative difficulty of test questions problems and other test items.

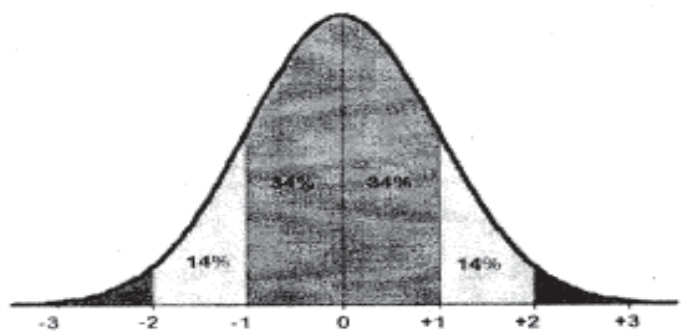
Thus from the above discussion we see that there are number of applications of normal curve in the field of educational measurement and evaluation. These are:

i) To determine the percentage of cases (in a normal distribution) within given limits or scores
ii) To determine the percentage of cases that are above or below a given score or reference point
iii) To determine the limits of scores which include a given percentage of cases
iv) To determine the percentile rank of a student in his own group v) To find out the percentile value of a student's percentile rank
vi) To compare the two distributions in terms of overlapping
vii) To determine the relative difficulty of test items, and
viii) Dividing a group into sub-groups according to certain ability and assigning the grades.

15.4 AREAS UNDER THE NORMAL CURVE

We are often interested in finding the areas under the normal curve which are associated with some given z score. The normal curve can be divided into sections by each standard deviation, beginning with a z score of zero in the center.

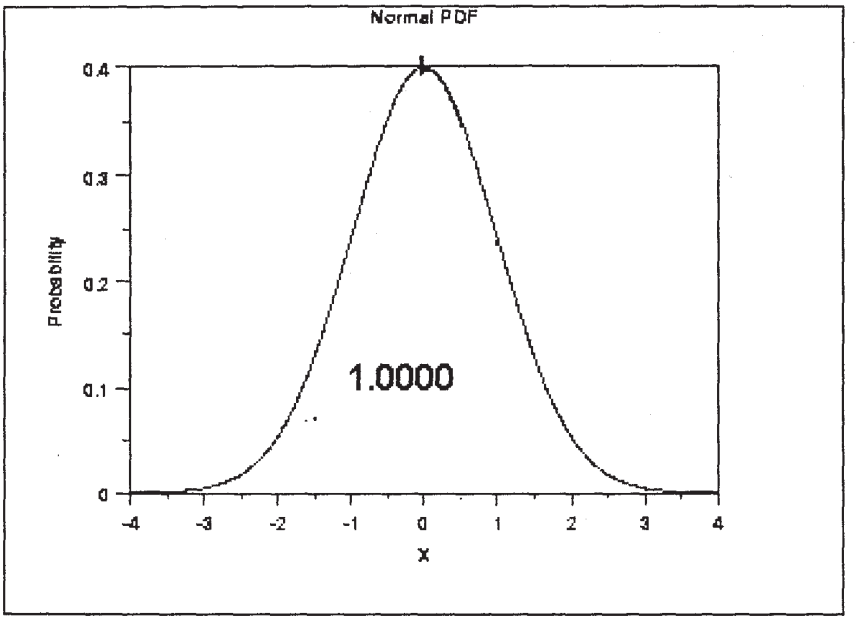
FIGURE 1



Notice that the total area under the curve is 1.000 by definition (that would correspond to 100% in

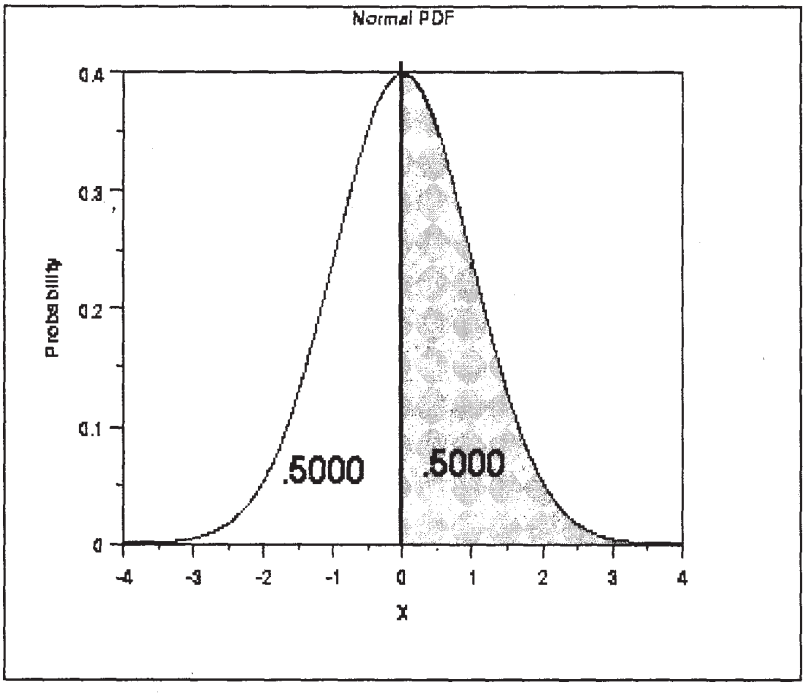
Figure 1 above). The area under the curve is depicted in Figure 2 below.

FIGURE 2



Now try to see (Figure 3 below) that the curve can theoretically be split in half at z=0, and there will be .5000 area above z=0 and .5000 area below z=0.

FIGURE 3



Frequently we ask questions in statistics which require us to know the area above or below a given z value, or the area between two z values. Referring to Figure 3 above, note that one can say that the area above z=O is .50, and the area below z=O is .50. Notice from Figure 1 we can make all kinds of statements about integer z values. For instance, the area between z= -l and z=+1 is .68 (because .34 + .34 = .68). The area between z=-2 and z=+1 is .14 + .34 + .34, or.82.

Frequently, though, our z values of interest are no perfect integers. For instance, we might need to know the aTea above z= +1.74. That is why we have z tables, and now is the time to learn about them.

There are basically three types of z tables. However, this being a basic introductory course on Psychological Testing and Statistics we need not go in to its details.

The table of areas of normal probability curve is referred to find out the proportion of area between the mean and the z value. One determines the probability of occurrence of a random event in a normal distribution by consulting tables of areas under a normal curve. Tables of the normal curve have a mean of 0 and a standard deviation of 1. To use the table, you must convert your data to have a mean of 0 and standard deviation of 1. This is done by transforming your raw values into z-scores, according to this formula:

z-score

$$z_i = \frac{X_i - \overline{X}}{s_x}$$

15.5 SUMMARY

In this unit we have discussed the concept of Probability, which can be defined as the ratio of number of favorable results to the total number of results. The normal probability curve is based upon the law of probability was also defined and discussed. Its characteristic features were discussed. The importance and applications of the Normal Probability Curve and the areas under the Normal Curve were also discussed in brief with illustrations.

15.6 QUESTIONS

1 . Discuss the concept of Probability.

2. Explain the characteristics features and importance of

   Normal Probability Curve.

3. Discuss the applications of the Normal Probability Curve.

4. Write a note with illustrations on the 'Areas under the Normal

   Curve'.

15.7 REFERENCES

1 Cohen, JR, & Swerdlik, M.E. (2010). Psychological Testing and Assessment: An introduction to Tests and Measurement. (7 th ed.). New York. McGraw-Hill International edition.

2. Anastasi, A. & Urbina, S. (1997). Psychological Testing. (7th ed.). Pearson Education, Indian reprint 2002.

16

Probability, Normal Probability Curve and Standard Scores' 11

Unit Structure

16.0 OBJECTIVES

After studying this unit you should be able to

a. Understand the concept of skewnesss and discuss the causes of skewness
b. Explain the concept of Kurtosis
c. Explain the concept of standard scores and know its various types.

16.1 INTRODUCTION

Whenever a curve lacks symmetry we call it skewness. Divergence from Normal Probability Curve gives rise to skewness or Kurtosis. Not all distributions in the real life display normal curve phenomenon. Some distributions deviate from normality. We would discuss the concept of skewness and kurtosis and also understand the causes for the same. We will also attempt to understand how divergence from normality is measured. Types of skewness and kurtosis would also be studied in this unit.

Standard scores Standard scores are ways to measure positions on the normal curve. They are, standard because the size of the distribution or type of measurement, they always fall in the same place. They include z-scores, T-Scores, etc., We would discuss the concept of standard scores and its various types. In the areas of measurement of personality individuals scores are often compared with reference to standard scores to obtain an assessment of an individual.

16.2 SKEWNESS- POSITIVE AND NEGATIVE, CAUSES OF SKEWNESS, FORMULA FOR CAL CULATION

The word Skewed means lacking symmetry or distorted. Skewness shows the direction of symmetry. Skewness of the distribution tells how lopsided the distribution is. A distribution is said to be skewed when the mean and median fall at different points in the distribution and the balance or centre of gravity is shifted to one side or the other.

Skewness depends upon the manner in which the scores in a series scatter about the average value. When the scatter is greater on one side of the point of central tendency than on the other, the distribution is said to be skewed.

In a normal distribution the mean equals the median and the skewness is of course zero. The more nearly the distribution approaches the normal form, the closer together are the mean and the median and the less the skewness.

Distributions are said to be skewed negatively or to the left when scores are massed at the high end of the scale (the right end) and are spread out more gradually towards the low end or left.

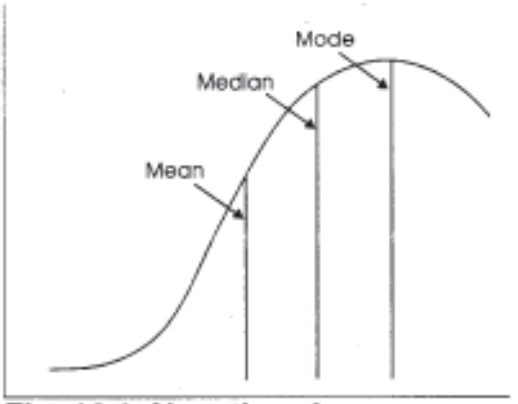


Fig. 16.1 Negative skewness

Distributions are said to be skewed positively when there is piling of scores at the low end and a long tail running up in high scores.

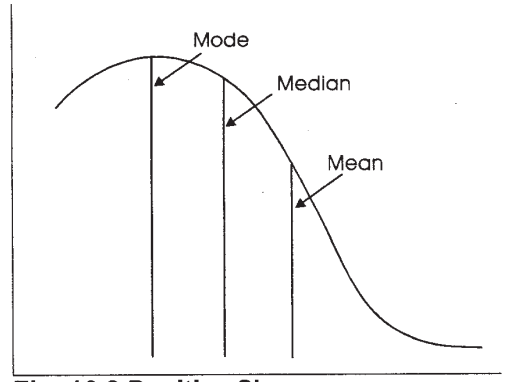


Fig. 16.2 Positive Skewness

In skewness the mean is pulled more toward the skewed end of the distribution, than the median. In fact, greater the gap between the mean and the median the greater the skewness. Moreover, when the skewness is negative the mean lies to the left of the median and when skewness is positive the mean lies to the right of the median.

16.2.1 Measurement of skewness :

There are three different measures of skewness which are as follows :

(1) One method used for measuring skewnes is by graphic analysis. Whenever Mean Median and Mode are not equal, the curve is skewed. But graphic method fails to give exact numerical value of skewness.

(2) Second method used for measuring skewness is by the following formula:

This is also called as the Karl Pearson coefficient of Skewness Sk = 3(mean - median) / Standard Deviation.

Interpretation:

- If Sk = 0, then the frequency distribution is normal and symmetrical.
- If Sk 0, then the frequency distribution is positively skewed.
- If Sk 0, then the frequency distribution is negatively skewed.

(3) Third method of measuring skewness is in terms of percentiles.

$$SK = \left(\frac{P_{90} + P_{10}}{2} \right) - P$$

16.2.2 Causes of Skewness:

skewed data distributions are a result of extreme values, also know as outliers. These can be due to many causes which are discussed below:

(1) Selection : Selection is a potent cause of skewness. If the sample you choose is a biased one, the distribution of the scores will not exhibit the bell shaped form.

(2) Unsuitable or poorly made tests: Normality or lack of normality is dependent upon the number of items and their difficulty it a test is too easy, scores will pile up at the high scores end of the scale and will give negative skewness whereas the test is too hard, scores will pile up at the low score end of the scale giving a positively skewed curve.

(3) Non-normal Distributions: Skewness of Kurtosis or both will appear when there is a real lack of normality in the trait being measured e.g. if a loaded side is tossed for a number of times, the resulting distribution will certainly be skewed and probably be peaked. This is because the loaded side is a dominant factor in determining the result of the tosses. Non-normal curves often occur in the medical statistics. In the case of childhood disease, for example, death rate would be maximum in the early ages and would decrease with increase in age. The distribution would be positively skewed.

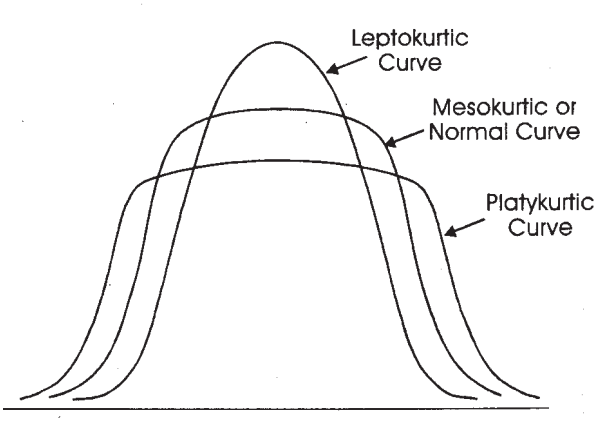
(4) Errors in the use of Test: Errors in timing or in giving instructions, errors in scoring, differences in motivation, all of these factors, if they cause some students to score higher and others to score lower than they normally would, tend to make for skewness in the distribution.

16.3 KURTOSIS - MEANING AND FORMULA FOR CALCULATION

The term kurtosis refers to the peakedness or flatness of a frequency distribution as compared with the normal. Kurtosis describes how peaked or flat the distribution is.

When there is high concentration of scores in the neighbourhood of the point of central tendency, the distribution is relatively narrow across the shoulders. Relatively high and narrow distributions are described as leptokurtic. When there is low concentration of scores in the neighbourhood of the central tendency the distribution is relativley broad across the shoulders. Such relatively flat topped distributions are described as platykurtic.

A normal distribution is called mesokurtic. The following figure depicts roughly the kurtosis of different types.



Kurtosis can be measured by the graphic method. Besides the graphic method we can measure kurtosis by the following formula

We can also measure kurtosis from normal probability curve. For the normal curve the value of Ku = 0.263, if the value of Ku is less than 0.263 we infer that distribution is Leptokurtic. If however, the value of Ku is greater than 0.263, the distribution is platykurtic. If kurtosis values tend toward 0, then the distribution approximates a normal distribution.

16.4 STANDARD SCORES - Z, T STANINE; LINEAR AND NON-LINEAR TRANSFORMATION; NORMALISED STANDARD SCORES

Standard scores are ways to measure positions on the normal curve. They are standard because the size of the distribution or type of measurement, they always fall in the same place. They include standard deviations, percentile ranks, T-scores, and deviation IQs.

The percentages of the area under the normal curve are the same for each standard deviation. Once you calculate the standard deviation for any set of scores, you can find any standard score for the data set. You can also figure out where any score falls relative to others.

Standard scores assume a normal distribution. They provide a method of expressing any score in a distribution in terms of its distance from the mean in standard deviation units. A Standard Score indicates how far a particular score is from a test's average.

In standard scores the unit that tells the distance from the average is the standard deviation (sd) for that test. For WAIS - ///, the average is 100 and the sd is 15. The standard deviation (sd) is always given for a standard score. Standard Scores between -1 sd (85) and +1 sd (115) fall in the normal range on the ability being tested.

Above + 1 sd (115+) a learner is in the top 15% of performance. Below -1 sd (-85), she / he is in the lowest 15% of performances.

Standard scores are used in norm-referenced assessment to compare one student's performance on a test to the performance of other students her age. Standard scores estimate whether a student's scores are above average, average or below average compared to peers. They also enable comparison of a student's scores on different types of tests, as in diagnosing learning disabilities.

Standard Scores are used generally for the following purposes.

(i) To tell the exact location of a score in a distribution. For e.g. Raju is 10 years old and has a weight of 50 kg. How does his weight compare to other 10 years old boys.

(ii) Standard Scores also help us to compare scores across different distributions.

For e.g. Geeta scored 65 in her chemistry paper, 75 in maths and 60 in english. On which test she performed better.

Standard Scores can be obtained by either linear or non-linear transformations of the original raw scores. When founded by linear transformation, they retain the exact numerical relations of the original raw scores, because they are computed by subtracting a constant from each raw score and then dividing the result by another constant. Linearly, derived standard scores are often designated simply as "Standard Scores" or "z- scores". To compute a z-score we find the difference between the individual's raw score and the mean of the normative group and then divide the difference by the S D of the normative group.

16.4.1 Common Types of Standard Scores:

i. **Z-Scores:** These scores are scaled on a number line ranging from -4 to 4 with zero being in the middle. On this scale, zero is average. Positive scores are above average, and negative scores are below average.

O-ne type of standard score is a z-score, in which the mean is 0 and the standard deviation is 1. This means that a z-score tells us directly how many standard deviations the score is above or below the mean. For example, if a student receives a z score of 2 her score is two standard deviations above the mean or the 84th percentile. A student receiving a z score of -1.5 scored one and one half deviations below the mean. Any score from a normal distribution can be converted to a z score if the mean and standard deviation is known. The formula is

Z score = (Score - mean score)I(Standard deviation)

So if the score is 130, the mean is 100, and the standard deviation is 15, then the formula leads to this

Calculation:

Z = (130 - 100)115 = 2

ii. **T-Scores:** These scores range from 10 - 90 in intervals of 10 points. Fifty is average on this scale. A T-score, by definition, has a mean of 50 and a standard deviation of 10. This means that a T score of 70 is two standard deviations above the mean and so is equivalent to a z score of 2.

iii. Stanines: Stanines (pronounced stay-nines) are often used for reporting students' The stanine scale is also called the standard nine scale. These scores range from 1 - 9 with five being average. Scores below five are below average. Scores above five are above average.

16.5 SUMMARY

In this unit we have discussed the concepts of skewness and kurtosis, their measurement and types. We have also discussed the causes of skewness. Standard scores are frequently used in psychological measurements to compare one student's performance with another or score obtained by one student on one test with scores obtained by him on another test. There are many different types of standard scores. The three most common types include: Z-Scores, T-Scores and Stanines.

16.6 QUESTIONS

1 Define Skewness and Kurtosis and explain the causes of skewness.

2. Discuss how Skewness and Kurtosis are measured.

3. Explain the different types of Skewness and Kurtosis.

16.7 REFERENCES

1 CoNen, JR., & Swerdlik, M.E. (2010). Psychological Testing and Assessment: An introduction to Tests and Measurement. (7th ed.). New York. McGraw-Hill International edition.

2. Anastasi, A. & Urbina, S. (1997). Psychological Testing. (7th ed.). Pearson Education, Indian reprint 2002.

17

CORRELATION

Unit Structure

17.0 OBJECTIVES

(1) To impart knowledge and understanding of the meaning, types and methods of Calculation of Correlation.

(2) To create awareness about the various steps involved in calculation of Pearson's product-moment coefficient of correlation.

(a) To provide basic knowledge about calculation of rho by Spearman's rank-order method.

(b) To make the foundation of statistical techniques strong about the knowledge of correlation coefficient and its applications.

17.1 INTRODUCTION

In chapter 8 we have studied the measures of variability or dispersion. These measures are defined on univariable i.e., the observations based on single characteristic. In some situations we may observe two or more characteristics simultaneously for each unit in a population - for example - height and weight of an individual, income and expenditure, supply and demand of a commodity, imports and exports of a country, etc. The variables that measure these characteristics between two or more variables is called correlation.

In this chapter we will study the meaning, and types of correlation. We shall also study Graphic representation of correlation, specifically Scatterplots. Attention is also given to calculation, correlation by Pearson's product-moment correlation coefficient and rho by Spearman's rank-difference method.

Uses and limitations of coefficient of correlation will be also studied. Toward the end of the unit we shall discuss the nature and uses of simple regression and multiple Regression.

17.2 MEANING AND TYPES OF CORRELATION

Correlation is defined as an expression of the degree and direction of relationship between two things or two sets of variables, when each of them is continuous in nature, Thus, the relationship between two variables is called simple or bivariate correlation. For example, height and weight, supply and demand of a commodity.

Types of correlation;

There are three types of correlation- these are discussed below.

Positive Correlation:- When the values of two variables move in the same direction, so that an increase in the value of one variable tends to increase the value with other variable, the correlation is called positive. Similarly, if two variables simultaneously decrease then, two variables are also said to be positively correlated. In our observation such as height and weight, profit and investment, income and expenditure of a family, etc., are positively correlated.

In Fig.1 direction of plotted points are from lower left corner to upper right corner. This is a positive correlation. The slope of the line is also positive.

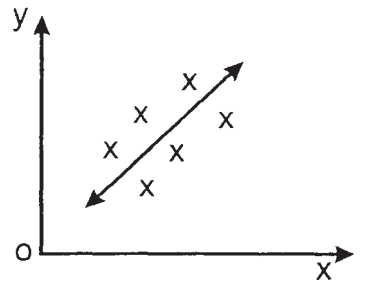


Fig -1 Positive correlation

Negative Correlation:

A negative correlation occurs when one variable increases and other variable decreases. For example, when supply increases while demand decreases, or price of the commodity increases but consumption decreases.

In Fig.2 Most of points lie near the line or on the line. Slope of the line is also negative.

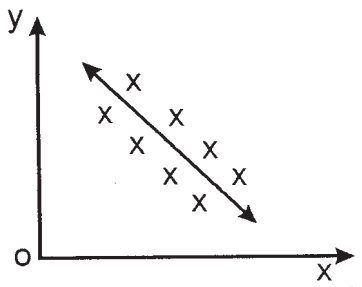


Fig -2 Negative correlation

Zero correlation:-

When there is absolutely no relationship existing between the two variables, the correlation is said to be zero (0). In any case, perfect positive correlation (+I),perfect negative correlation (-1) and zero correlation are hard to identify. Most of the time, two variables are fractionally correlated.

Fig-3 shows zero correlation-

Fig 3 :- shows no direction. Therefore there is no correlation between the values of two variables.

No line can be drawn passing through most of the points.

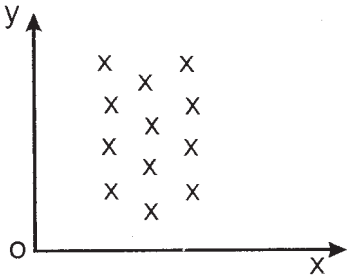


Fig 3 : Absence of correlation.

17.3 GRAPHIC REPRESENTATION OF CORRELATION - SCATTER PLOTS

An important type of representing correlation is a graphic description known as the Scatterplot or Scatter diagram. It is very simple method to study correlation by the use of n - pairs of observations (xi, yi), (X2, Y2) -------- and so on. Values of two variables of each observation are taken as co-ordinates of the point, where the values of one variable are placed on X-axis (vertical line) and the values of another variable are placed on y axis (horizontal line). Paired observations are plotted as points on the graph. The graph of these points shows how far the observations are scattered, Hence, it is a scattered plot or scatter diagram.

A scatterplot of the data helps us in having a visual idea about the nature of association between two variables. The relationship shown by the points plotted on the scatterplot involves two aspects, first the direction of the

relationship i.e., positive or negative and the closeness of the points to some line.

However, scatterplot reveals the direction and strength of magnitude of the relationship between two variables.

Scatterplot or graph also provides data to know the range of scores and types of relationship exist between two variables, scores, group of scores, etc. It is relatively very simple technique that provides a hint of some of deficiency in the testing or scoring procedures.

Graphic representation of Scatterplot.

The following two figures (FA and F.5) show the Scatterplot for positive correlation (r)

correlation coefficient =0.60 (Moderate degree of positive correlation)

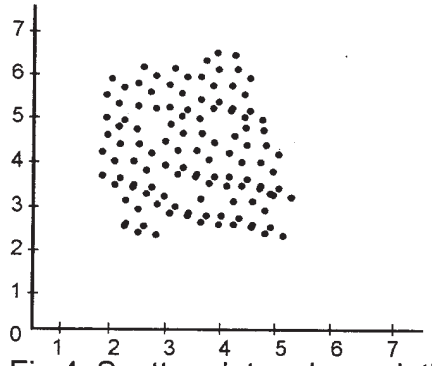


Fig.4 Scatter plot and correlation for positive values of r

correlation coefficient = 0.95
(very high degree of positive correlation)

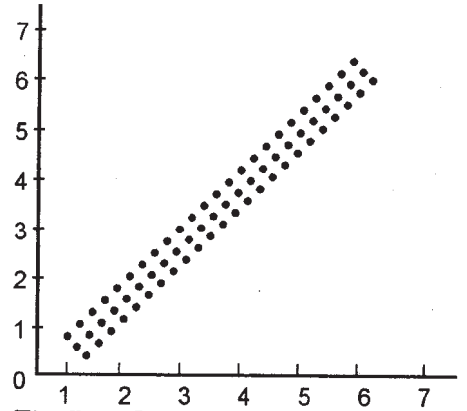


Fig.5 Scatterplot of Positive r.

However, the fig-6 and 7 show the negative correlation in the forms of graph;

correlation coefficient = -0.50 (Moderate degree of negative correlation).

correlation coefficient = - 0.90 (very high degree of Negative r)

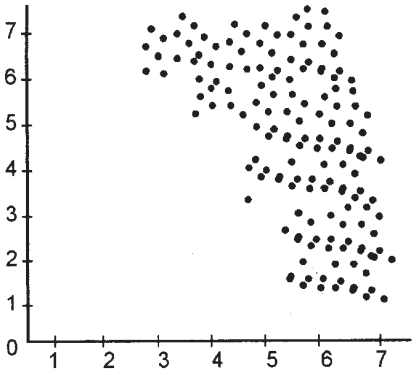
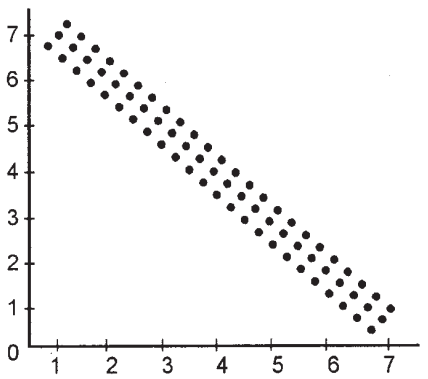


Fig- 6-7



Scatterplots and Correlations for Negative values of r

Now we shall finally represent a graph of zero r

correlation coefficient = 0

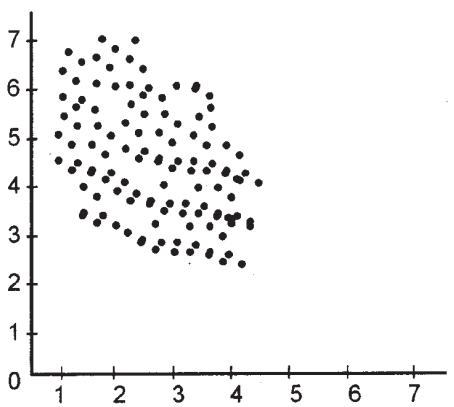


Fig – 8

Scatterplot and Correlations for 0 values of r

17.4 THE STEPS INVOLVED IN CALCULATION OF PEARSON'S PRODUCT- MOMENT CORRELATION COEFFICIENT

There are many techniques which have been developed to measure correlation. The most widely used technique is of Karl Pearson's product-moment coefficient correlation. Pearson's technique is the standard index of the amount of correlation between two variables. When the relationship between the variables is linear and when the two variables being correlated are continuous, Pearson's r is used.

There are a number of Pearson's r formula.
We shall study the following short-cut and modified formula for our purpose step by. step.

$$R = \frac{\Sigma\, xy}{\left(\Sigma x^2\right)\left(\Sigma y^2\right)}$$

When we see this formula, it seems to be more complex than other formulas. But it is easier to use when deviations are taken from the means of the two distributions.

To find r;

(1) The first step to find out of number of paired scores.
(2) Σ xy is the sum of the product the paired X and Y scores.
(3) Σ X is the sum of the x scores and Σ y Yis the sum of the Y scores.
(4) (ΣX^2) is the sum of the squared X scores and (Σy^2) is the sum of the squared Y scores.

However, similar results are obtained with the use of each formula.

Check your progress.

Q1. Define correlation and explain its types with examples.
Q2. Define/Explain scatterplot and explain the graphic representation of its calculation.
Q3. What is product-moment correlation coefficient? Explain the various steps involved in its calculation.

17.5 CALCULATION OF RHO BY SPEARMAN'S RANK-DIFFERENCE METHOD

When complex behaviour such as honesty, athletic ability, social adjustment are hard to measure, we should put these behaviours in order of merit. In computing the correlation between two sets of ranks, special methods are used.

Charles Spearman (1927) has developed a method known as Spearman's rho rank-difference method.

It is conveniently applied as a quick substitute when the number of pairs is less than 30.

It is also conveniently used when data are already in terms of rank-order.

However, Rho method has only one formula, but it has different names such as rank-order correlation of co-efficiency, a rank-difference correlation coefficient, or simply Spearman rho. It has both sets which are in ordinal form, (rank-order)

However, the formula of Rank order is:

$$1-\frac{6\sum d^2}{n\left(n^2-1\right)}$$

or

$$1-\frac{6\sum d^2}{N^3-N}$$

Ex-1 Calculate rank correlation coefficient for the following data given marks in two tests in Psychology, for a group of 10 students.

X	Y	R_1	R_2	d	d^2
67	78	2	2	0	0
42	80	8	1	7	49
53	77	7	3	4	16
66	73	3	6	-3	9
62	75	4	4	0	0
60	68	5	7	-2	4
54	63	6	8	-2	4
68	74	1	5	-4	16

$\sum d^2 = 98$

$$1-\frac{6\sum d^2}{n\left(n^2-1\right)}$$

$$r = 1 - \frac{6 \times 98}{8\times(8^2-1)}$$

$$r = 1 - \frac{588}{504}$$

$$r = -0.166$$

This is poor negative correlation.

Example-2

X	Y	R_1	R_2	d	d^2
50	48	5.5	5	0.5	0.25
63	30	3	8.5	5.5	30.25
48	35	9	7	2	4
70	60	1	1	0	0
45	55	8	2	6	36
65	30	2	8.5	-6.5	42.25
38	25	10	10	0	0
40	45	7	6	1	1
52	50	4	4	0	0
50	52	5.5	3	2.5	6.25

$\sum d^2 = 120$

$$r = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$r = 1 - \frac{6 \times 120}{10(10^2-1)}$$

$$r = 1 - \frac{720}{990}$$

$$r = 1 - 0.73$$

$$r = 0.27$$

simple low positive correlation.

17.6 USES AND LIMITATIONS OF CORRELATION CO EFFICIENT

A correlation coefficient is the numerical index that expresses the relationship between two sets or variable which is indicated by the letter small Y. This coefficient may be of any size from zero to +1.00 or - 1.00. The sign of the coefficient does not determine its significance, there are high, moderate, or low coefficients that may be either positive or negative. Thus + 1.00 indicates a perfect positive and -1.00 a perfect negative correlation.

Uses

1 The techniques of correlation of coefficient are used in physical sciences and social sciences, especially psychological research where we postulate doctrine and theories with principles and hypotheses.

2. This technique is also applied in calculation of reliability and validity of coefficient of correlation.
3. It - is also used to find out simple r, multiple r and rank-order correlation.
4. Studies conducted on intelligence scores and academic achievement have yielded positive correlation.
5. It is also revealed that motivation and absenteeism in organisation are highly but negatively correlated.
6. Similarly, supply and demand, family income and expenditure, height and weight are also positively correlated.

Limitations of correlation coefficient

Since there are several techniques of calculating r, so we can't apply all of these to solve one type of problem. For example. Pearson's r is applied on group data and ungroup data, but it does not apply on ordinal form a Rank-order method is used. So both techniques have their own limitations and advantages.

Similarly, by any means, a perfect positive correlation coefficient (+1.00) or a perfect negative coefficient (-1.00) is impossible to obtain. It is challenging to try to think of.

It is also very difficult to use all techniques of Pearson Y because they are very complicated and time consuming except one we have discussed in 10.4.

17.7 SIMPLE REGRESSION AND MULTIPLE REGRESSION

Regression is commonly explained as retreat back or 11 reversion to some previous state. In statistics, regression also describes a kind of reversion-to the mean overtime or generation.

However, the term "regression" was first used by Francis Galton with reference to the inheritance of status. Galton found that children of tall parents tend to be less tall, and children of short parents less short, than their parents. In other words, the heights of the offsprings tend to "move back" toward the mean height of the general population. This tendency toward maintaining the mean height is called the principle of regression and the line describing the relationship of height, in parent and offspring is called a "regression line". The term is still employed but in other meaning.To-day, regression is defined as the relationship among variables for the purpose of understanding how far one variable predicts other one. In other words, regression is the measure of the average relationship between two or more variables in terms of original units of data.

Although, there are various types of regression, but we shall explain simple and multiple regression.

Simple regression involves the analysis of only two variables; one is independent variable (X), typically known as "the predictor" variable and another is dependent variable (Y), typically referred to as to the "outcome variable." Simple regression results in an equation for regression line - that line of best fit, the strait line that comes to the closest to the greatest number of points on the scatterplot of X and Y.

However, the main use of regression equation is to predict the effect of one value on other and to make interpretation of Y.

Multiple regression is another type of regression. It is also used as a predictor. Its analysis requires the use of more than one variable. To predict Y requires the use of a multiple regression equation. This type of equation explains the interrelations among all the variables involved. If many predictors are used, and one is not correlated with any predictor, but is correlated with the predicted score, it gives more weight and provides unique information.

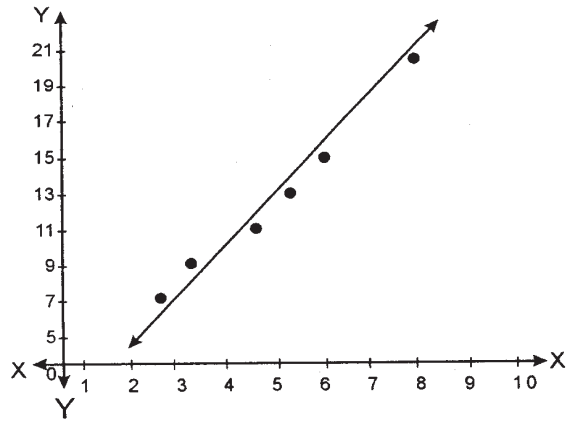
Method of Regression Analysis

We have already seen scatter plot that represents the correlation between two variables for a bivariate data.

In this method the points are plotted on a graph paper representing pairs of values of the concerned variables where values of independent variables are taken on X -axis and the values of dependent variables are taken on Y -axis. A regression line may be drawn in between these points by free hand or by a scale rule in such a way that squars of the vertical distances or the horizontal distances between the points and the line of regression so drawn is the least. It should be drawn carefully as the line of best fit leaving equal number of the points on either side of the line.

Example 1.

X	2	3	4	6	7	8	10
Y	5	6	10	11	15	17	21



This method provides a rough estimate of the dependent variable because the line drawn is subjective to the person drawing it.

Check your progress

Q1 When rho by Spearman's rank-difference method is used?

Q2 Find the Spearman's rank correlation coefficient from the following

X	50	70	65	38	90	50	40	75
Y	35	90	70	48	95	85	40	80

X	15	12	16	25	15	17	14	18	26	20
Y	17	14	16	16	25	20	24	22	26	23

Q3 Explain the uses and limitations of correlation coefficient.

Q4 Write short notes on simple regression and multiple regression.

17.8 SUMMARY

In this chapter we have discussed the meaning and types of correlation with special attention given on positive, negative and zero correlation with examples. Correlation is also graphically represented in the form of scatterplot. We have also studied the nature of Pearson's product- moment correlation coefficient and its various steps involved in its calculation of Y.

Calculation of rho by Spearman's rank-difference method is also highlighted with questions and answers of the problems.

Finally, we have focused on the uses, limitations of correlation coefficient and its two types - simple regression and multiple regression followed by method of regression analysis shown by a graph.

17.9 QUESTIONS

01. (a) Define correlation and explain its types.

(b) Calculate the rank - order correlation coefficient from the following distribution.

X	22	35	70	80	70	65	50	55	40	50
Y	78	68	60	65	60	55	45	52	75	76

(c) Interpret your answer.

Q2. Explain the uses and limitations of correlation coefficient.

Q3. Explain the various steps involved in calculation of Pearson's product-moment correlation coefficient.

Q4. Write short notes on Scatterplot, simple regression and multiple regression.

17.10 REFERENCES

1. Annapornaa R. et al (2008) A hand book of Mathematics and Statistics, Chetana Publications Pvt. LTD. 263, Khatauwadi, Girgaon, Mumbai-400004
2. Anastasi, A. and Urbina, S. (1997) Psychological testing (7th ed) Pearson Education, India Reprint -2002
3. Cohen, J.R. and Swerdik, m.E. (2010) Psychological testing and Assessments, An introduction to tests and measurements (7th ed) Newyork McGraw-Hill international edition.
4. Garrett, Henry, E. (1973) Statistics in psychology and Education (6th edition) Vakills, Feffer and Simon Pvt. Ltd. Ballard Estate, Mumbai 400001
5. Guilford, J.P. (1956) Fundamental Statistics in Psychology and Education (3rd ed) Newyork McGraw-Hill book, Co.
6. Spearman, C. (1927) The ability of man: their nature and measurement, Newyork, MaCmillan.
7. Walker, H. M. (1943) Elementary statistical method (Newyork) Henry Stoll and C. pp. 308-310

17.11 GLOSSARY

Correlation : It is defined as an expression of the degree and direction of relationship between two things or two sets of variable, when each of them is continuous in nature.

Correlation Coefficient : A correlation coefficient is the numerical index that expresses the relationship between two sets or variables which is indicated by the letter small 'r'.

Multiple regression : The analysis of relationship between more than one independent variable and one dependent variable to understand how each independent variable predicts the dependent variable.

Negative Correlation : A negative correlation occurs when one variable increases and other variable decreases.

Scatterplot : A graphic. description of correlation achieved by graphing the co-ordinate point's for the two variables.

Simple regression : Involves only two variables, one is independent variable (X), typically known as the '' predictor' variable, and one dependent variable (Y), typically refers to as the outcome'' variable.

Spearman's rho : Refers to as the rank-order correlation coefficient applied when the sample size is small and both sets of measurements are in ordinal form.

Regression : The analysis of relationship among variables to understand how one variable may predict another.

Zero correlation : It refers to as when there exists no relationship between two variables.