1

PSYCHOLOGICAL TESTING, ASSESSMENT AND NORMS - I

Unit Structure

1.0 OBJECTIVES

After studying this unit, you would 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 DEFINITION OF TESTING AND ASSESSMENT

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 come 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.

Psychological assessment can be considered as a problem-solving process that can take many different forms. Assessment is done to

answer a referral question or arrive at a decision through the use of tools of evaluation.

Psychological testing on the other hand, is done to measure some ability or attribute, usually in numerical form. Psychological testing can be done on individual or in a group. Assessment is typically individualized.

In psychological testing, the tester is not very important to the process and one tester can be substituted by another tester. The tester needs to have technician type skills in administering and scoring a test. On the other hand, assessment generally focuses more on how an individual process rather than simply the results of that processing. The assessor is important to the process of selecting tests and/or other tools of evaluation as well as in drawing conclusions from the entire evaluation.

(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).

1.2.2 Definition 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 bserved 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:

Referrals:

Usually the process of assessment starts with a referral. The referral can be from a teacher, school psychologist, counsellor, judge, clinician, or corporate human resources specialist. 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.

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. Typically, which tool or method to apply to assess will be influenced by the assessor’s own past experience, education, and training.

(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.

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.

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) and even interested third parties such as child's parents, supervisor or referral of the assessee, etc. The assessor will write a report of his findings that is aimed at answering the referral questions.

(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.

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.

This approach can be used in other setting too, such as, correctional, corporate, neuropsychological, clinical, etc.

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 THE 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. However, even two tests measuring the same variable such as personality can have different content, depending upon which theoretical orientation of personality is considered important by the developer.

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) **Administration Procedures**: Different tests have different administration procedures. Some require active and knowledgeable test administrator who may be trained observers. Other tests, such as group tests, may not require the administrator to be even present when testtakers are taking the test.

d) **Scoring and Interpretation procedures - Score** can be defined as a code or summary statement, that is usually but not necessarily is numerical and reflects an evaluation of performance on a test, task, interview, or some other sample of behavior.

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 cut score is sometimes is determined by scientific method and other times by an arbitrary procedure 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.

e) **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 The 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.

Apart from face-to-face format, interviews can be conducted in various other formats too. For example, telephonic interview, in sign language, panel interview, through electronic mediums such as online interviews, e-mail interviews, text messaging and through video conferencing, etc.

When the purpose of an Interview is not only to collect information but also to bring a targeted change in the thinking and behavior of the interviewee, it is called motivational interview. It can be used successfully through telephone, Internet chat or text messaging too.

The success of this tool depends on the skills of the interviewer. The skills of the interviewers may include their pacing of interviews, their rapport with interviewees, and their ability to convey genuineness, empathy, humor, etc.

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. For example, it may include letters, photos and family albums, newspaper and magazine clippings, home videos, movies, audiotapes, work samples, artwork, doodling, postings on social media and other data pertaining to interests and hobbies and items relevant to an 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.

On the basis of case history data and other relevant data, case study or case history is developed. A case history can be defined as a report or illustrative account concerning a person or an event that was compiled on the basis of case history data.

Case history can give information about an individual's past and current adjustment as well as on the events and circumstances that may have contributed to any changes in adjustment.

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?

Behavioural 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. Behavioural 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.

The format of the role play could range from “live scenarios” with live actors, or computer- generated simulations. Outcome measures for such an assessment might include ratings related to various aspects of the individual’s ability to resolve the conflict, such as effectiveness of approach, quality of resolution, and number of minutes to resolution.

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.

CAT refers to computer adaptive testing. Computer can be programmed to adopt in such a way that it gives the testtaker with score feedback as the test proceeds.

Other Tools

Videos are also used for assessment in various settings. Apart from videos, psychologists may use many of the tools traditionally associated with medical health, such as thermometers, biofeedback, etc.

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 behaviours 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

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

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

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

❄❄❄❄❄❄❄

2

PSYCHOLOGICAL TESTING, ASSESSMENT AND NORMS - II

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.
3. Understand the concept of norms, sampling and type of sampling

4. Know the various types of norms
5. Discuss fixed reference group scoring systems as well as norm referenced and criterion referenced evaluation
6. Understand the relationship between culture and reference.

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 organizations 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. Organizations 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. They should 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 person who is being assessed or evaluated. This is a very broad definition and can include even a deceased person as an assessee. For example, in ‘psychological autopsy’ procedure, psychologists reconstruct the psychological profile of a deceased individual on the basis of archival records, artefacts, and interviews previously conducted with the deceased assessee or people who knew him.

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 organizing and systematizing that makes society a crucial party to the process of assessment. Religious inclinations, intellectual sophistication of human mind has been well reflected in the assessment process to understand and predict human behavior.

2.2.5 Test Utilizer

The test utilizer may be the test taker, but in other cases however, a business or organization may send a person to be tested. Thus, the organization 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 more or less extent involved in the process of assessment.

2.3 TYPES OF SETTINGS INVOLVED

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 specializations.

2.4 WHAT IS 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.

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.

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.

2.4.1 Norms: Sampling to Develop Norms

What are norms?

Norms can be described as the average scores in 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 standardization of the test. Standardization 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 analyzed for reference, for evaluating or interpreting individual scores.

What 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 be decided as 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.

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 nonrandom 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.

Judgment 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.

2.4.2 Types of Norms

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 standardization. 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 standardized sample itself. New norms are developed for specific group of test takers on the basis of earlier standardization.

What are the various types of Norms

1. **Percentiles –** A percentile is an expression of the percentage of people whose score on a test or measure falls below a particular raw score. If we divide a distribution of scores into 100 equal parts—100 percentiles. For example, in such a distribution, the 15th percentile is the score at or below which 15% of the scores in the distribution fall. In other words, if 15% of a particular standardization sample answered less than 47 questions on a test correctly, then we could say that a raw score of 47 corresponds to the 15th percentile on this test. Thus, we can say that a percentile is a ranking that gives information about the relative position of a score within a distribution of scores.

   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 100 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 1 and 99, with the 50 being the average score. 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 minimized 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). For example, age norms for height, weight of children is widely accepted. As the age increases, the height and weight also increases, up to middle or late teens. 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 100, or average intelligence. The concept of mental age is criticized on the basis that it is too broad and other factors such as psychological development, social development may not be 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. Another drawback of grade norm is that they are useful only with respect to years and months of schooling completed. They have little or no applicability to children who are not yet in school or to children who are out of school.

4. **National Norms -** They are derived from a standardization 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 takes both tests). When two tests are normed from same sample it is called co-norming. Generally, to make national anchor norms, we begin with first developing percentile norms for each of the tests that we are going to compare. Then equivalency of scores on different tests is calculated with reference to corresponding percentile scores. For example, if the 70th percentile corresponds to a score of 69 on a Test A and if the 70th percentile corresponds to a score of 20 on Test B, then we can say that Test A score of 69 and score of 20 on Test B must have been obtained on the same sample, as each member of the sample took both tests and the equivalency tables or national anchor norms were calculated on the basis of these data. However, one cannot treat equivalence as precise equalities (Angoff, 1964, 1966, 1971).

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.

8. **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 utilized 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. Test items common to each new version of the SAT and each previous version of it are employed in a procedure (termed anchoring) that permits the conversion of raw scores on the new version of the test into fixed reference group scores

2.4.3 Fixed Reference Group Scoring Systems

In Fixed reference group scoring 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. Very often local or other specific group norms are used for this purpose (Anastasi 2006). The SAT and the GRE are scored using the Fixed Reference Group Scoring.

2.4.4 Norm-Referenced Versus Criterion-Referenced Evaluation

a. 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 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.

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.

b. 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 to 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 individualized program focusing on division will be designed. Thus, criterion referenced testing movement emphasizes the diagnostic use of tests.

2.4.5 Culture and Inference

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 having the traces of cultural bias. Avoiding cultural bias is virtually impossible.

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. The 'times' here means the era in which the testing is done. 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 made in the past and still being 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, so one needs to design culture-fair tests instead. Culture fair 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 many alternative tests and be aware of equivalence issues across cultures including equivalence of language used and the constructs measured.

2.5 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 organizations 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 organization that may send a person to be tested. Thus, the organization 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.

12. 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, and 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.

13. 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.

    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.

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

    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 analyzed for reference for evaluating or interpreting individual scores.

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

14. 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.

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

16. 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.

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

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

    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.

2.6 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.

Q.3 What is the criteria for good test?

Q.4 What are norms? What are the various types of norms?

Q.5 Describe various types of sampling.

Q.6 Write notes on a fixed reference group scoring system, b. norm referenced c. criteria referenced evaluation

2.7 REFERENCES

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

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

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

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

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

❄❄❄❄❄❄❄

3

RELIABILITY - I

Unit Structure

3.0 OBJECTIVES

After studying this unit, you should be able to:

1 . Define reliability and understand the concept of reliability.

2. Comprehend the various sources of error variance.

3. Know the various methods of estimating reliability

3.1 INTRODUCTION

In this unit we will discuss the definition and concept of reliability which is one of the most important characteristics of a good test. Following this we will discuss the various sources of error variance.

we will also discuss 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.

We would end this unit with a brief summary, questions and a list of references for further reading.

3.2 THE CONCEPT OF RELIABILITY

In everyday life we use the term reliability, often on a positive note. For example, if a train comes regularly at the same time to a particular station, you may say that this train is very reliable. 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 that you bought a BP apparatus and decided to check your BP. It gave drastically three different readings on three consecutive measure. The first reading showed high blood pressure, the second reading taken immediately after that showed low pressure and third reading taken immediately after the second reading showed normal blood pressure. 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 reliability is indeed important to trust a machine or even a psychological test.

A test is considered reliable if we get approximately the same result repeatedly. The constancy of an 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.

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

1st attempt: 70 / I 10

5th 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 that is reliable and then check your blood pressure. The other machine reads 70 / I10 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. Maybe, 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.

Reliability is of various types. A psychological test may be reliable in one context but not reliable in another context. Similarly, reliability is not an all - or - none matter. It exists in different degrees. Some tests are highly reliable while others may be low in reliability.

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

In a broader sense, the term reliability indicates the extent to which individual differences in test scores 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

The variability in test score is often described in terms of a statistical term called 'variance'. Reliability (consistency) is estimated 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. The greater the proportion of the total variance attributed to true variance, the more reliable the test. Because true differences are assumed to be stable, they are presumed to yield consistent scores on repeated administrations of the same test as well as on equivalent forms of tests. Because error variance may increase or decrease a test score by varying amounts, the consistency of the test score—and thus the reliability—can be affected.

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 factors). -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.

3.3 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.

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.

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

Errors in test scoring and test interpretation:

- Most of the tests used in India like the 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 an incorrect score.

- Computerized testing and evaluation are available for most of these tests. However, computers may not be available in all testing centers and thus they have to be administered and score manually with the help of a trained professional. Manual scoring and interpretation increase vulnerability to human errors like errors in scoring, errors in calculating scores, or even interpreting the score.

- Some tests like an 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.

Systematic and Unsystematic errors:

- Suppose we want to assess the quality of the 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 behavior, misunderstanding questions or overreporting/underreporting abuse for different reasons, may occur.

Other sources of error:

In some cases of survey research, the error in the research may be due to sampling error. It means the sample taken is not representative of the population that it is supposed to represent. For example, the researcher may be wanting to study the stress levels of students attending online classes. He

may have got the age factor wrong and instead of taking a sample of secondary school students, he may have taken a sample of college students. In that case, the sample is not a true representative of the population.

- An error can take place if a researcher has not taken the appropriate sample size.

3.4 RELIABILITY ESTIMATES

Earlier we learned about what is reliability, sources of variance, and the sources of error variance. Now let us understand 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 the reliability of the test.

3.4.1 Test-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 same 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 relatively stable over time, such as personality factors, reaction time, perceptual judgment, etc.
5. Source of error variance:
    a. 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

    b. Passage of time: With the 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. Even when measuring relatively stable variables and even when the time period between the two administrations of the test is short, many other factors, such as, experience, practice, memory, fatigue, and motivation can confound reliability.

3.4.2 Parallel-Forms and Alternate form Reliability

Another way R.A. can determine the reliability of his test is by creating another test that simply a different version of the first test and measures the same attribute. 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 the question on 6 digit addition was asked on form 1, the 6 digit addition problem was set in Form 11. Observe that both the sets have an 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 an alternate or parallel forms 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 attribute.
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 the 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 minimize the effect of memory on 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 firms may be affected by factors like the 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 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.

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 the reliability of the test, by comparing test items within the same test, it is called an **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 the Split half method, the Kuder Richardson formula, and the Coefficient alpha by Cronbach.

3.4.3 Split half reliability

To avoid constructing two equivalent tests, which is, a 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 half.

2. Assign odd number items to the first half and even number items to the second half or vice versa. This is also known as odd-even 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 a number of items, - difficulty and content $r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$.

Step III: Apply Pearson 'r' formula

Step IV: Correct the coefficient using the Spearman-Brown formula

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

where; rsB = the reliability adjusted by the Spearman-Brown formula. fxy= Pearson's r in the original length test.

n = number of items in the revised version *l* 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, the longer the test, the 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 the coefficient.

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.

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

Test homogeneity is necessary because it allows precise interpretation of 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, an extremely high index would mean the items within the test are measuring the same thing. So, the rule, the higher the better may not always hold true.

Test homogeneity is, however, an 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 that are of homogeneous nature. In other words, a number of homogeneous tests are administered that 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, and 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 units measuring a single trait like submissiveness; thus forming 8 units of the 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.

3.4.4 Internal Consistency By Kuder Richardson Formula

KR20 was developed 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:

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

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

KR20 = Kuder - Richardson formula 20; k = number of test items; 6 = variance of total test scores; p= proportion of test-takers who pass the test item q= proportion of people who fail test item; pq= sum of the pq over all items

3.4.5 Estimating internal consistency by calculating Cronbach's 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:

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

6^2= variance of the total test scores

Coefficient alpha is used on tests with non-dichotomous items. This formula is an estimate of the 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. The 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 an 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.

3.4.6 Inter-rater Reliability / Inter-scorer reliability

Case study: Mrinalini, Richa, and Saundarya are participants in a beauty pageant. All the three would be assessed by a number of raters/assessors, ultimately giving one of the three the title of Ms. India. But before giving them the title, the three women have to be assessed for creativity and integrity. Now 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 that 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 observation falls into and then calculate the percentage of agreement between the raters.

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

3.5 SUMMARY

1. The word reliability means dependability or consistency. Reliability is the 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 the same result repeatedly. The constancy of an 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 scores 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. **The true variance** could be a result of increased ability due to practice or training. This variation in score due to factors other than training (true change) is called **'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. There are different methods of estimating reliability. These methods include the test-retest method, parallel or alternate form, inter-rater reliability, or determining internal consistency by split-half method, Cronbach's coefficient alpha, and Kuder Richardson formula.

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 in the number of testing sessions required, availability of, or the possibility of making parallel or alternate forms. Each of these methods has different sources of error and statistical procedures.

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.

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.

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.

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 the alternate form or parallel form **coefficient of reliability.**

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.

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.

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.

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

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 adjusting the half test reliability by using the Spearman-Brown formula.

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.

A tabular representation of 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'*

Formula chart

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

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.

When it is expected to administer the test only then we would expect to have an estimate of internal consistency.

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.

3.6 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.

3.7 REFERENCES

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

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

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

❄❄❄❄❄❄❄

4

RELIABILITY - II

Unit Structure

4.0 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

4.1 INTRODUCTION

In this unit we will attempt to understand how the nature of test influences its reliability. Following this we will explain various alternatives to true score model. In this we will discuss the item response theory, the domain sampling model and generalizability 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.

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

4.2 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?

4.2.1 Purpose Of Reliability Coefficient

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 assigned 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.

4. 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.

4.2.2 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.

Is 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 of heterogeneous tests, test - retest reliability may be used.

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.

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.

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. Since measure of the reliability of a speed test should reflect the consistency of response speed, the reliability of a speed test should not be calculated from a single administration of the test with a single time limit.

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.

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.

4.2.3 The True Score Model of Measurement and Alternatives to It

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. True score model of measurement is also referred to as the Classical test theory.

True score model of measurement:

True score model is most preferred model in psychometry. It is a simple and easy to understand. It propagates the idea that everyone has ‘true score’ on a test that he/she has taken. True score in this model is defined as a value that genuinely reflects an individual’s ability (or trait) level as measured by a particular test. However, the true score of a person can differ from one test to another test of the same attribute. Though classical test theory model can

be easily applied to measure different situations, it has its own share of shortcomings. For example, its assumption that all items contribute equally to the score total has been criticized. Another problem is that test takers and test developers prefer to have shorter tests, but classical test theory favours longer tests.

Therefore, 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.

The item response theory: It is an alternative to true score model . 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. It is also referred to as Latent trait theory as many of the constructs that are measured are psychological or educational and cannot be seen physically. Actually item response theory is not a single theory or method. In stead it refers to a large family of theories and methods. There are more than a hundred varieties of IRT and each model is designed to tackle data with certain assumptions and data characteristics

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. Measure of internal consistency is the most preferred estimate of reliability in domain sampling model.

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.

4.3 RELIABILITY AND INDIVIDUAL SCORES: SEM AND SE-DIFFERENCE

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.

4.3.1 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 ± 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 ± 2 σ_{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 = σ 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.

4.3.2 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 practicing 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 measure 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

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

$$\sigma \sqrt{2 - r_1 - r_2}$$

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?

$$\sigma_{diff} = 10\sqrt{2 - .90 - .90}$$
$$\sigma_{diff} = 10\sqrt{.20}$$
$$\sigma_{diff} = 7.77$$

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,

5. 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.

4 .4 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.

12. 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.

13. **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.

14. **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.

4.5 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.

4.6 REFERENCES

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

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

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

❄❄❄❄❄❄❄

5

VALIDITY AND MEASURES OF CENTRAL TENDENCY - I

Unit Structure

5.0 OBJECTIVES

After going through the unit, you would be able to:

1. Explain the concept and meaning of validity.
2. To analyze the meaning of face and content validity.
3. To discuss the concept of criterion validity.
4. To explain construct validity.

5.1 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 characteristics of scientific instrument.

5.2 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 refer 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 generalized 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

5.2.1 FACE AND 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.

5.2.2 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 as 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

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 standardized 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.

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 TYBA 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 TYBA class. In the same way in business organization, 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.

5.2.3 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 states 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 coefficient is 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.

These various techniques of construct validation may provide evidence, for example, that

- the test is homogeneous, measuring a single construct;
- test scores increase or decrease as a function of age, the passage of time, or an experimental manipulation as theoretically predicted;
- test scores obtained after some event or the mere passage of time (or, post test scores) differ from pertest scores as theoretically predicted;
- test scores obtained by people from distinct groups vary as predicted by the theory;
- test scores correlate with scores on other tests in accordance with what would be predicted from a theory that covers the manifestation of the construct in question.

A test developer can improve the homogeneity of a test containing items that are scored dichotomously (such as a true-false test) is by eliminating items that do not show significant correlation coefficients with total test scores.

The homogeneity of a test in which items are scored on a multipoint scale can also be improved. If all test items show significant, positive correlations with total test scores, then each item is most likely measuring the same construct that the test as a whole is measuring (and is thereby contributing to the test's homogeneity). Coefficient alpha may also be used in estimating the homogeneity of a test having multiple-choice items.

If a test score purports to be a measure of a construct that could be expected to change over time, then the test score, too, should show the same progressive changes with age to be considered a valid measure of the construct.

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 personality 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.

2. **Discriminant Validity:** Discriminant validity describes the degree to which the operationalization does not correlate with other operationalization that it theoretically should not be correlated with. In other words, when a test correlates poorly with measures with which theoretically 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.

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.

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.

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.
    6. **External Validity:** External validity concerns the extent to which the results of a study can be generalized for other cases, for example to different people, places or times.
    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.
    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 Product Moment correlation coefficient is applicable.

Other types of correlation coefficients can be computed when the data are expressed in different forms, as when a twofold 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 measure and the composite of the two or more than two sets of measures.

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 popu lation 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.

5.3 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:

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 tallies 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 de- scribed 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.

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. The 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.

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 utilization 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.

5.4 SUMMARY

Validity is the correlation of the test with some outside 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.

5.5 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

5.6 REFERENCES

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

Hoffman, E. (2002) Psychological Testing at work, New Delhi, Tata Mc Graw Hill Publishing Company Ltd.

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

Ranjit Kumar (2005) 2nd Ed. Research Methodology, Dorling Kindersley (India) Pvt. Ltd. Licenses of Pearson Education in South Asia.

❄❄❄❄❄❄❄

6

VALIDITY AND MEASURES OF CENTRAL TENDENCY - II

Unit Structure

6.0 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.

6.1 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. We will also look at their advantages and limitation as well as their uses.

6.2 CALCULATION OF MEAN, MEDIAN AND MODE OF A FREQUENCY DISTRIBUTION

6.2.1 The Mean

The 'arithmetic mean' or the 'mean' is the sum of the individual scores divided by the number of scores in the given distribution.

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.**

Calculation of Mean from Data Grouped in a Frequency Distri bution

When the number of scores to be averaged are large it becomes cumber-some to use the above method of finding the mean. The scores can then be grouped into a frequency table (as explained in the unit 9), 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 math's 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. The 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.

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

Calculate the mean score of the class in Physics.

Check your answers at the end of the chapter

6.2.2 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.

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)

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.

6.2.3 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. Ordinarily, 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

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

6.3 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 the assumed 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 computation:

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 center 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 Σfx' = (+43 - 55) = - 12.

6. Find the correction (in terms of the class intervals) c by computing c =

   Σ-fx' / N. In this case c -12/50 = -0.240

7. Class interval i = 5.

8. Multiply the correction by the interval length c i = 5 x 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

6.4 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.

6.4.1 Merits and Demerits of Mean, Median and Mode

Merits of Mean:

- It is easy to compute and has a unique value.
- It is based on all the observations.
- It is well defined.
- It is least affected by sampling fluctuations.
- It can be used for further statistical analysis.

Demerits of Mean:

- The mean is unduly affected by the extreme items (outliers).
- It cannot be determined for the qualitative data such as beauty, honesty etc.
- It cannot be located by observations on the graphic method.

Merits of Median

- It is easy to compute. It can be calculated by mere inspection and by the graphical method

- It is not affected by extreme values.
- It can be easily located even if the class intervals in the series are unequal

Demerits of Median

- It is not amenable to further algebraic treatment
- It is a positional average and is based on the middle item
- It does not consider the actual values of the items in the series

Merits of Mode:

- It is comparatively easy to understand.
- It can be found graphically.
- It is easy to locate in some cases by inspection.
- It is not affected by extreme values.
- It is the simplest descriptive measure of average

Demerits of Mode:

- It is not suitable for further mathematical treatment.
- It is an unstable measure as it is affected more by sampling fluctuations.
- Mode for the series with unequal class intervals cannot be calculated.
- In a bimodal distribution, there are two modal classes and it is difficult to determine the values of the mode.

6.4.2. Uses of Mean, Median and Mode

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.

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.

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

6.5 SUMMARY

In this unit, the concept of mean, median and mode is discussed. How to compute mean, median and mode is explained with examples. The advantages and disadvantages of all three measures of central tendency, and their uses are discussed.

6.6 QUESTIONS

1. What is meant by measurement of central tendency? With illustration explain the steps to compute mean.
2. What is meant by median and mode. Explain the computation of median and mode with illustration
3. Compare mean, median and mode as a measure of central tendency.

6.7 REFERENCES

Cohen, J. R., Swerdlik, M. E., & Kumthekar, M. M. (2014). *Psychological Testing and Assessment: An introduction to Tests and Measurement.* (7th ed.). New Delhi: McGraw-Hill Education (India) Pvt Ltd., Indian adaptation

Garrett,H.E (1929). *Statistics in Psychology and education.*

❄❄❄❄❄❄❄

7

TYPES OF SCORES, TYPES OF SCALES, FREQUENCY DISTRIBUTION, GRAPHICAL REPRESENTATIONS - I

Unit Structure

7.0 OBJECTIVES

- To understand the concept of measurement and the various scales of measurement.
- To understand the ways of preparing a frequency distribution.
- To understand various graphical representations of data like frequency polygon, histogram and ogive

7.1 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 to prepare frequency distribution. We will also discuss the advantages and disadvantages of frequency distribution and how to smoothen the frequencies.

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 organized 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.

7.2 CONTINUOUS AND DISCRETE SCORES - MEANING AND DIFFERENCE

Generally, quantitative data is made up of either continuous scores or discrete scores. Discrete scores are the whole numbers. They can be counted and plotted in a frequency table. For example, male/ female, colors, children, etc. If you ask a person how many children do you have? The person can not answer two and a half. There is no such thing as half child. It will be either two children or three children, but not two and a half.

Continuous scores are those that can be a measure in fractions too. For example, height, weight, distance, etc. You can score in extremely fine details. For example, the height of a person can be five feet and ¾ inch or five feet and half inch. Take another example, you can report age in terms of years, days, hours and even minutes in continuous scores but in case of discrete scores it will be only in terms of years.

The difference between continuous and discrete score is that continuous scores can be expressed as fractions too while discrete score can express only as whole numbers. Discrete scores are counted while continuous scores are measurements.

Continuous scores are generally in ratio scale while discrete scores are part of nominal and ordinal scales .

7.2.1 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 categorized. 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 categorize 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 categorized on the grounds of gender i.e., female or male, the categorization 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.

7.2.2 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 mean, 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.

1. **Nominal scales:** Nominal scales are the simplest forms of measurement and involve classification or categorization based on one or more distinguishing characteristics where all things measured must be placed into mutually exclusive and exhaustive categories.

2. For example, the 10^{th} standard students 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 organizational chart as a nominal category. The lowest level is the nominal scale, which represents only names and therefore has none of the three qualities.

3. **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 organizational 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 2nd 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 analyzed 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 middle ranked. item) when we use an ordinal scale.

4. **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 analyzed statically (in ways appropriate for data) 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.

5. **Ratio scales:** A ratio scale has 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 analyzed.

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
			Likert Scale,	median, percentile
2	Ordinal	Magnitude	Anything rank ordered	
			Names, Lists of words	mode, Chi-square
1	Nominal	None		

Stevens (1946, 19121)

7.3 PREPARING A FREQUENCY DISTRIBUTION

Regardless of whether manual or automated methods are used, it is usually necessary to code data numerically to facilitate further data analysis. Frequency distributions summarize 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 concentrated 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

7.3.1 The Advantages And Disadvantages Of Frequency Distribution

1. Advantages of the frequency distributions:

1. It condenses and summarizes 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.

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 deemphasize ranges and extreme values, particularly when open classes are used.

7.3.2 Smoothed frequencies: method of running averages

A running average is a method used to analyze 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

7.4 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 categorization 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 analyzed.
9. Frequency distributions summaries and compress data by grouping it into classes and recording how many data points fall into each class.
10. A moving average is commonly used with time series data to smooth out short-term fluctuations and highlight longer-term trends or cycles.

7.5 QUESTIONS

Answer the following questions:

Q1. Explain various types of scales of measurement.

Q.2. What is frequency distribution? Explain the process of preparing a frequency distribution.

Q.3. What is meant by smoothened frequencies? How is it prepared?

7.6 REFERENCES

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.

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

Kaplan, R.M., & Saccuzzo, D.P. (20012) . Psychological Testing - Principles, Applications and Issues. (6 th ed.). Wadsworth Thomson Learning, Indian reprint 2007.

http://en.wikipedia.org/wiki/Moving_average

❄❄❄❄❄❄❄

8

TYPES OF SCORES, TYPES OF SCALES, FREQUENCY DISTRIBUTION, GRAPHICAL REPRESENTATIONS - II

Unit Structure

8.0 OBJECTIVES

- To understand various graphical representations of data like frequency polygon, histogram, and ogive

8.1 INTRODUCTIO: GRAPHICAL REPRESENTATIONS

As we have already discussed how to make frequency tables, in this unit we will look at what are the different types of graphs and how to smoothen the frequency polygon.

8.1.1 Frequency Polygon:

There are many forms of representing data graphically. They are histograms, frequency polygons and ogive. Let's understand each of them.

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 Y axis 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.

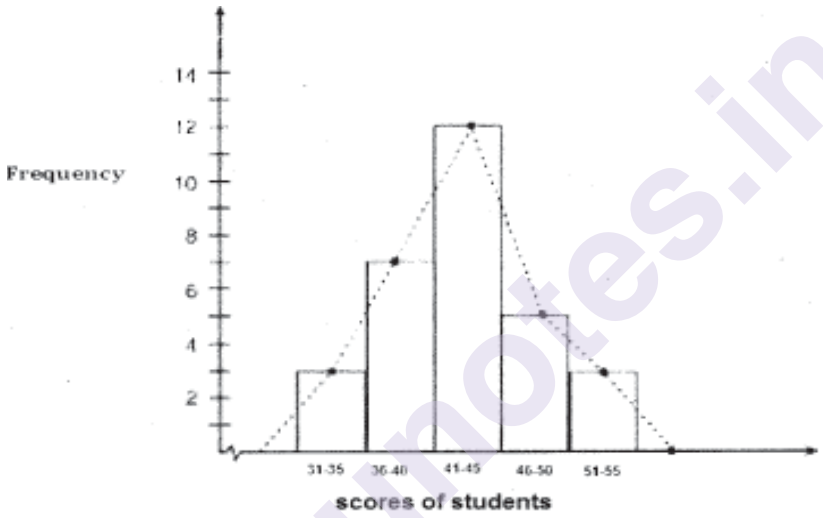
* For the sake of understanding we have shown the entire class interval on the X axis. Usually we should use the midpoint of the class intervals on the X axis.

8.1.2 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 below.



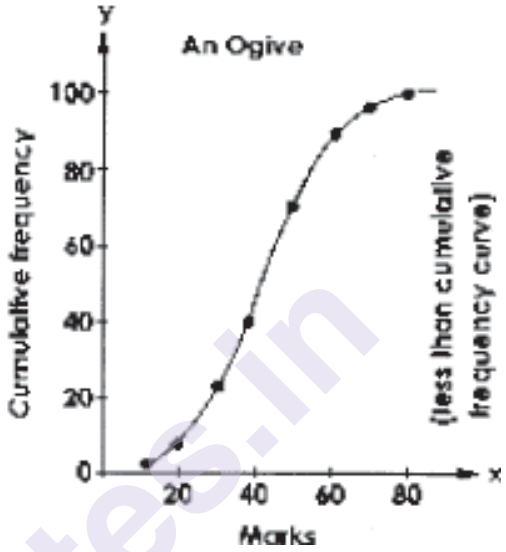
The cumulative frequency, also known as an Ogive, is another way to analyze the frequency distribution table. Unlike a frequency distribution which tells you how many data points are within each class, a cumulative frequency tells you how many are less than or within each of the class limits.

It is useful for analyses that require quick results about the proportion of data that lies below a certain level. Cumulative distributions are useful in telling quickly how many in a group have scored above or below a certain point on the scale

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.

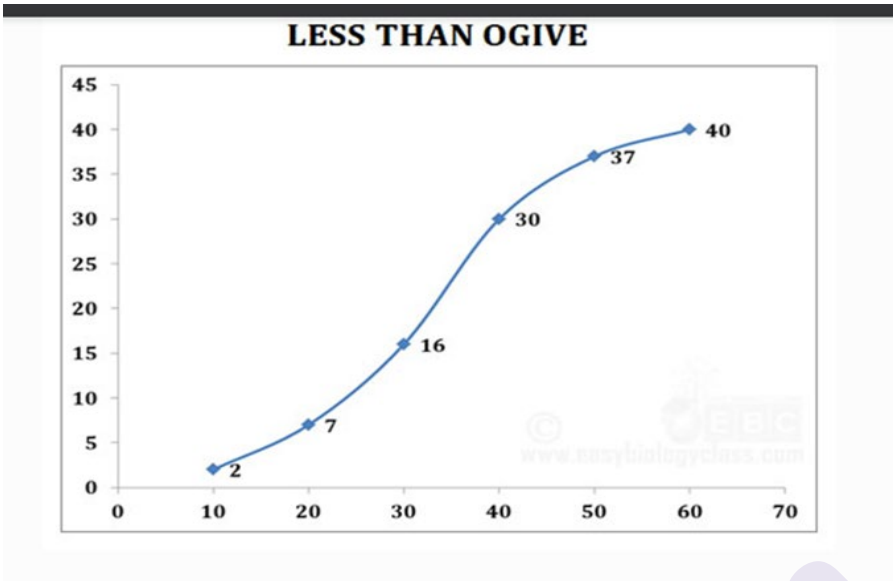
Step 2. Draw the X and Y axis. Label the X axis with the class boundaries. Use an appropriate scale for the y axis to represent the cumulative frequencies.

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.

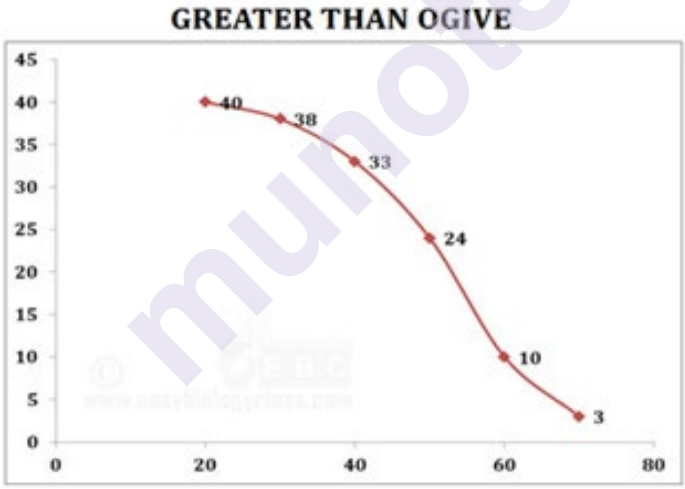
There are two types of ogives

i) **'Less than' Ogive:** The less than cumulative frequencies are plotted against the upper class boundaries of the respective classes and less than cumulative frequencies on Y-axis. It is an increasing curve having slopes upwards from left to right.



ii) **'More than' Ogive :** The more than cumulative frequencies are plotted against the lower class boundaries of the respective classes. It is decreasing curve and slopes downwards from left to right.

For example-

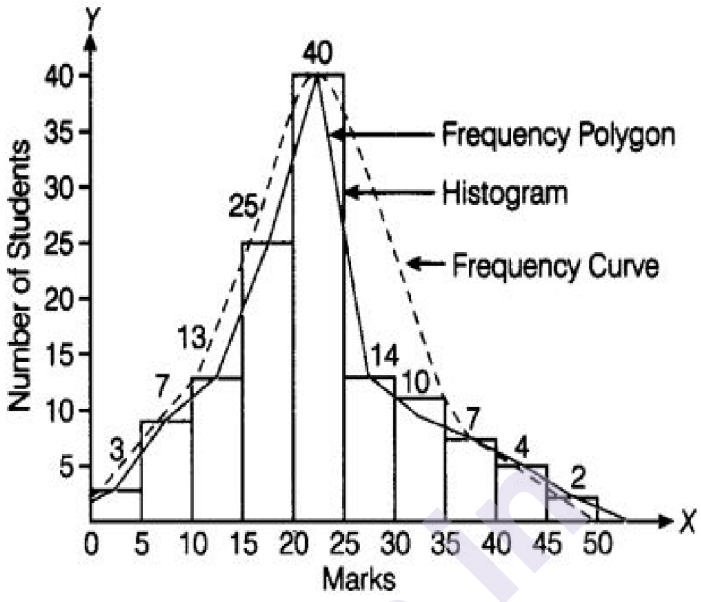


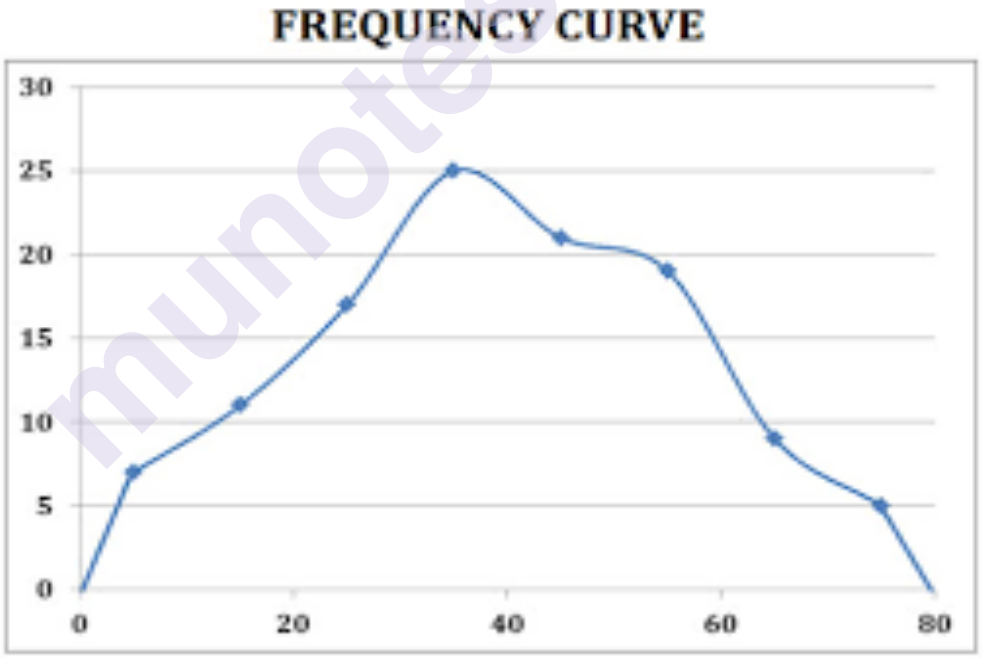
8.1.4 Polygon of Smoothed Frequencies

A smoothed polygon of frequencies can be drawn through the various points of the polygon. The curve is drawn freehand in such a manner that the area included under the curve is approximately the same as that of the polygon. The object of drawing a smoothed frequency curve is to eliminate as far as possible accidental variations that might be present in the data. The curve should look as like normal distribution curve as possible and sudden turns should be avoided. The extent of smoothing would, depend upon the nature of the data.

To draw a smoothed frequency curve or polygon, you need to first draw a histogram than a polygon and then smoothed frequency curve.

See Fig 8.1





While smoothing a frequency curve, please keep in mind certain guidelines, such as -

1. Only frequency distribution based on samples should be smoothed.
2. Only continuous series should be smoothed.
3. The total area under the curve should be equal to the area under the original histogram or polygon.

8.2 SUMMARY

1. Frequency polygons are a graphical device for understanding the distributions. Frequency polygons are expressed by a continuous line connecting the points.

2. 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.

3. The cumulative frequency, also known as an Ogive, is another way to analyze the frequency distribution table. It is useful for analyses that require quick results about the proportion of data that lies below a certain level.

8.3 QUESTIONS

Q.1. What is the difference between histogram and polygon. How will you draw it.

Q.2. What is meant by cumulative frequencies ?

8.4 REFERENCES

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.

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

Kaplan, R.M., & Saccuzzo, D.P. (20012) . Psychological Testing - Principles, Applications and Issues. (6 th ed.). Wadsworth Thomson Learning, Indian reprint 2007.

❄❄❄❄❄❄❄