1

THEORETICAL-CONCEPTUAL ORIENTATION - I

Unit Structure

1.0 OBJECTIVES

After learning this chapter, students will understand the following concepts:

- Importance of the measurement in Psychology
- Types of statistics used in Psychology
- Theories of measurement

1.1 INTRODUCTION

Psychological Test:

A psychological test is a systematic procedure for obtaining samples of behaviour, relevant to cognitive or affective functioning, and for scoring and evaluating those samples according to standards. Psychological tests are often described as standardized for two reasons, both of which address the need for objectivity in the testing process. The first has to do with uniformity of procedure in all important aspects of the administration, scoring, and interpretation of tests. The time and place when a test is administered, as well as the circumstances under which it is administered and the examiner who administers it, affect test results. However, the purpose of standardizing test procedures is to make all the variables that are under the control of the examiner as uniform as possible so that everyone who takes the test will be taking it in the same way.

The second meaning of standardization concerns the use of standards for evaluating test results. These standards are most often norms derived from a group of individuals—known as the normative or standardization sample—in the process of developing the test. The collective performance of the standardization group or groups, both in terms of averages and variability, is tabulated and becomes the standard against which the performance of other individuals who take the test after it is standardized will be gauged.

The term test should be used only for those procedures in which test takers' responses are evaluated based on their correctness or quality. Such instruments always involve the appraisal of some aspect of a person's cognitive functioning, knowledge, skills, or abilities.

On the other hand, instruments whose responses are neither evaluated nor scored as right-wrong or pass-fail are called inventories, questionnaires, surveys, checklists, schedules, or projective techniques, and are usually grouped as personality tests. These are tools designed to elicit information about a person's motivations, preferences, attitudes, interests, opinions, emotional makeup, and characteristic reactions to people, situations, and other stimuli. Typically, they use questions of the multiple-choice or true-false type, except for projective techniques, which are open-ended. They can also involve making forced choices between statements representing contrasting alternatives or rating the degree to which one agrees or disagrees with various statements.

Most of the time, personality inventories, questionnaires, and other such instruments are of the self-report variety, but some are also designed to elicit reports from individuals other than the person being evaluated (e.g., a parent, spouse, or teacher). For the sake of expediency and following common usage, the term test will be used throughout this book to refer to all instruments, regardless of type, that fit the definition of a psychological test. Tests that sample knowledge, skills, or cognitive functions will be designated as ability tests, whereas all others will be referred to as personality tests.

1.2 MEASUREMENT IN PSYCHOLOGY AND IN THE NATURAL SCIENCES

1.2.1 Measurement:

The concept of measurement is at the heart of psychological testing as a scientific enterprise for the study of human behaviour. Measurement involves the use of certain devices or rules for assigning numbers to objects or events (Stevens, 1946). If we apply this process systematically, then to a large extent, a phenomenon that is measured is made more easily subject to confirmation and analysis, and thus becomes more objective as well. In other words, by systematically analyzing, categorizing, and quantifying observable phenomena we place them in the scientific arena. Central to the definition of psychological tests is the fact that they consist of carefully chosen samples of behaviour to which a numerical or category system is applied according to some pre-established standards. Psychological testing is largely coextensive with the field of psychometrics, or psychological measurement, and is one of the primary tools for the science and practice of psychology. The use of numbers in testing requires the use of statistics.

1.2.2 Variables and Constants

One of the most basic distinctions we can make in any science is between variables and constants. A variable is anything that varies, whereas a constant is anything that does not. There are many variables in our world and few constants. One example of a constant is π (pi), the ratio of the circumference of a circle to its diameter, a number that is usually rounded to 3.1416.

Variables, on the other hand, are everywhere, and they can be classified in a multitude of ways. For example, some variables are visible (e.g., sex, the color of eyes) and others invisible (e.g., personality, intelligence); some are defined so as to pertain to very small sets and others to very large sets (e.g., the number of children in a family or the average income of individuals in a country); and some are continuous, others discrete.

Discrete variables are those with a finite range of values—or a potentially infinite, but countable, range of values. Dichotomous variables, for instance, are discrete variables that can assume only two values, such as sex or the outcome of coin tosses. Polytomous variables are discrete variables that can assume more than two values, such as marital status, race, and so on. Other discrete variables can assume a wider range of values but can still be counted as separate units; examples of these are family size, vehicular traffic counts, and baseball scores. Although in practice it is possible to make errors in counting, in principle, discrete variables can be tallied precisely and without error.

Continuous variables such as time, distance, and temperature, on the other hand, have infinite ranges and really cannot be counted. They are measured with scales that could theoretically be subdivided into infinity and have no breaks in between their points, such as the scales in analogue clocks, yardsticks, and glass thermometers.

In psychological testing, we are almost always interested in variables that are continuous (e.g., degrees of integrity, extraversion, or anxiety), yet we measure with tools, such as tests or inventories, that are not nearly as precise as those in the physical and biological sciences.

Therefore, we must be aware of potential sources of error and look for pertinent estimates of error whenever we are presented with the results of any measurement process. For example, if polls taken from samples of potential voters are used to estimate the outcome of an election, the estimated margins of error have to be displayed alongside the results of the polls.

In summary, when we look at the results of any measurement process, we need to keep in mind the fact that they are inexact. With regard to psychological testing in particular, whenever scores on a test are reported, the fact that they are estimates should be made clear; furthermore, the limits within which the scores might range as well as the confidence levels for those limits need to be given, along with interpretive information.

1.2.3 The Meaning of Numbers:

Because numbers can be used in a multitude of ways, Stevens (1946) devised a system for classifying different levels of measurement on the basis of the relationships between numbers and the objects or events to which the numbers are applied. These levels of measurement or scales—outlined in Table 1 —specify some of the major differences in the ways numbers may be used as well as the types of statistical operations that are logically feasible depending on how numbers are used.

1. Nominal Scales:

At the simplest level of his classification, Stevens placed what he called nominal scales. As this implies, in such scales, numbers are used solely as labels to identify an individual or a class. Some examples of a nominal scale are the Aadhar numbers that identify most people who live in India; these numbers are useful because each is assigned to only one person and can therefore serve to identify persons more specifically than their first and last names, which can be shared by many people. Numbers can also be used to label categorical data, which is data related to variables such as gender, political affiliation, color, and so forth—that is, data that derives from assigning people, objects, or events to particular categories or classes.

2. Ordinal Scales:

The numbers used in ordinal scales convey one more bit of meaning than those in nominal scales, albeit a significant one. In these scales, in addition to identity, there is the property of rank order, which means that the elements in a set can be lined up in a series—from lowest to highest or vice versa—arranged on the basis of a single variable, such as birth order or level of academic performance within a given graduating class. Although rank - order numbers convey a precise meaning in terms of position, they carry no information with regard to the distance between positions. Thus, the students in a class can be ranked in terms of their performance, but this ranking will not reflect the amount of difference between them, which could be great or small. Similarly, in any hierarchical organization, say, the U.S. Navy, ranks (e.g., ensign, lieutenant, commander, captain, admiral) denote different positions, from lowest to highest, but the differences between them in terms of accomplishments or prestige are not the same.

3. Interval Scales:

In interval scales, numbers acquire yet one more important property. In these scales, the difference between any two consecutive numbers reflects an equal empirical or demonstrable difference between the objects or events that the numbers represent. An example of this is the use of days to mark the passage of calendar time. One day consists of 24 hours, each hour of 60 minutes, and each minute of 60 seconds; if two dates are 12 days apart, they are exactly three times as far apart as two dates that are only 4 days apart. Note, however, that calendar time in months is not an equal-unit scale because some months are longer than others. Furthermore, calendar time also typifies a characteristic of interval scales that limits the meaning of the numbers used in them, namely, that there is no true zero point. In the case of calendar time, there is no agreed-upon starting point for the beginning of time. Different cultures have devised arbitrary starting points, such as the year Christ was presumed to have been born, to mark the passage of years.

On interval scales, the distances between numbers are meaningful. Thus, we can apply most arithmetical operations to those numbers and get results that make sense.

4. Ratio Scales:

Within ratio scales, numbers achieve the property of additivity, which means they can be added—as well as subtracted, multiplied, and divided—and the result expressed as a ratio, all with meaningful results. Ratio scales have a true or absolute zero point that stands for “none of” whatever is being measured. In the physical sciences, the use of this type of measurement scale is common; times, distances, weights, and volumes can be expressed as ratios in a meaningful and logically consistent way. For instance, an object that weighs 16 pounds is twice as heavy as one that weighs 8 pounds (16/8 = 2), just as an 80-pound object is twice as heavy as a 40-pound object (80/40 = 2). In addition, the zero point in the scale of weights indicates absolute weightlessness.

In psychology, ratio scales are used primarily when we measure in terms of frequency counts or of time intervals, both of which allow for the possibility of true zeros.

Table 1.1: Four types of scales

Scale	True Zero	Equal Intervals	Order	Category	Example
Nominal	No	No	No	**Yes**	Marital Status, Sex, Gender, Ethnicity
Ordinal	No	No	**Yes**	**Yes**	Student Letter Grade, NFL Team Rankings
Interval	No	**Yes**	**Yes**	**Yes**	Temperature in

					Fahrenheit, SAT Scores, IQ, Year
Ratio	**Yes**	**Yes**	**Yes**	**Yes**	Age, Height, Weight

Source: https://thebiologynotes.com/nominal-ordinal-interval-and-ratio-data/

1.2.4 Relevance of Numbers in Psychological Testing:

Though it is not universally favored, Stevens's system for classifying scales of measurement helps to keep the relativity in the meaning of numbers in proper perspective. The results of most psychological tests are expressed in scores, which are numbers that have specific meanings. Unless the limitations in the meaning of scores are understood, inaccurate inferences are likely to be made on the basis of those scores. Unfortunately, this is too often the case, as can be seen in the following example:

Example: The problem of ratio IQs. The original intelligence quotients devised for use with the Stanford-Binet Intelligence Scale (S-B) were ratio IQs. That is to say, they were real quotients, derived by dividing the mental age (MA) score a child had obtained on the S-B test by the child's chronological age (CA) and multiplying the result by 100 to eliminate the decimals. The idea was that average children would have similar mental and chronological ages and IQs of approximately 100.

Children functioning below the average would have lower mental than chronological ages and IQs below 100, while those functioning above the average would have higher mental than chronological ages and IQs above 100. This notion worked fairly well for children in the early and middle school years, during which there tends to be a somewhat steady pace of intellectual growth from year to year.

However, the MA/CA ratio simply did not work for adolescents and adults because their intellectual development is far less uniform—and changes are often imperceptible—from year to year. The fact that the maximum chronological age used in calculating the ratio IQ of the original S-B was 16 years, regardless of the actual age of the person tested, created additional problems of interpretation.

Furthermore, the mental age and chronological age scales are not at the same level of measurement. Mental age, as assessed through the first intelligence tests, was basically an ordinal-level measurement, whereas chronological age can be measured on a ratio scale. For these reasons, dividing one number by the other to obtain a quotient simply did not lead to logically consistent and meaningful results.

The following table shows numerical examples highlighting some of the problems that have caused ratio IQs to be abandoned.

Table 1.2 Examples of Ratio IQ Computation

Subject	Mental Age (MA)	Chronological Age (CA)	Difference (MA - CA)	Ratio IQ
Ramesh	6 Years	5 Years	1 Year	(6/5)×100 = 120
Suresh	12 Years	10 Years	2 Years	(12/10)× 100 =120
Rupesh	18 Years	15 Years	3 Years	(18/15)×100 =120

Problem 1:

The mental age score required to obtain any given IQ keeps rising for each successive chronological age, so the ratio of IQs at different chronological ages are not equivalent.

Problem 2: Whereas chronological age rises steadily, mental age does not. Since the highest mental age achievable on a given intelligence test cannot be limitless, even when a limit is placed on the maximum chronological age used to compute IQs—as was done for a long time in the S-B scale—the IQs that most adults can attain are artificially constrained compared to those of children and adolescents.

Solution: Because of this and other problems with ratio IQs as well as with the concept of mental ages, the use of the ratio IQ has been abandoned. The term IQ is now used for a score that is not a ratio IQ and is not even a quotient. This score, known as the deviation IQ, was pioneered by David Wechsler.

1.2.5 What Can We Conclude About the Meaning of Numbers in Psychological Measurements?

In psychology, it is essential to keep in mind that most of our measurement scales are of an ordinal nature. The equality of units is approximated by the scales used in many types of test scores, but such equality is never as permanent or as complete as it is in the physical sciences because the units themselves are relative to the performance of the samples from which they are derived. The use of ratio scales in psychology is limited to measures of frequencies, reaction times, or variables that can be meaningfully expressed in physical units. For example, if we were using assembly-line output per hour as a measure of the speed of performance in an assembly line job, we could say that Worker A, who produces 15 units per hour, is 3 times as fast as Worker B, who produces only 5 units per hour. Note, however, that we could not say that Worker A is 3 times as good an employee as Worker B because speed is probably not the only index of job performance even in an assembly line operation. The overall level of performance is a more complex variable that most likely can be assessed only with a qualitative, ordinal scale.

1.3 TYPES OF STATISTICS

Since the use of numbers to represent objects and events is so common in psychological testing, the field involves the substantial application of statistics, a branch of mathematics dedicated to organizing, depicting, summarizing, analyzing, and otherwise dealing with numerical data. Numbers and graphs used to describe, condense, or represent data belong in the realm of descriptive statistics. On the other hand, when data are used to estimate population values based on sample values or to test hypotheses, inferential statistics—a more ample set of procedures based on probability theory—are applied. Fortunately, although both descriptive and inferential statistics are extensively used in the development of tests, most of the quantitative aspects of test score interpretation require only a good grasp of descriptive statistics and a relatively small number of techniques of the inferential type. Moreover, even though a background in higher-level math is desirable in order to understand thoroughly the statistics involved in testing, it is possible to understand them at a basic level with a good dose of logic and a relatively limited knowledge of math.

The words statistic and statistics are also used to refer to measures derived from sample data—as opposed to those derived from populations, which are called parameters.

Means, standard deviations, correlation coefficients, and other such numbers calculated from sample data are all statistics derived in order to estimate what is of real interest, namely, the respective population parameters. Parameters are mathematically exact numbers (or constants, such as π) that are not usually attainable unless a population is so fixed and circumscribed that all of its members can be accounted for, such as all the members of a college class in a given semester. In fact, one of the main purposes of inferential statistics is to estimate population parameters on the basis of sample data and probability theory.

1. Descriptive Statistics:

Raw data usually consists of a bunch of numbers that do not convey much meaning, even after close examination. With descriptive statistics, we can summarize the data so they are easier to understand. One way to summarize data is to represent them graphically; another way is to condense them into statistics that represent the information in a data set numerically.

2. Measures of Central Tendency:

One of the first things one wants to know when inspecting a data set is where the bulk of the data can be located, as well as the data's most representative or central value. The principal measures of central tendency—the mode, median, and mean—tell us these things. As with any other statistic, each of these measures has particular advantages and disadvantages depending on the types of data and distributions one wishes to describe.

3. Measures of Variability:

These statistics describe how much dispersion, or scatter, there is in a set of data. When added to information about central tendency, measures of variability help us to place any given value within a distribution and enhance the description of a data set. Although there are many measures of variability, the main indexes used in psychological testing are the range, the semi-interquartile range, the variance, and the standard deviation.

- The range is the distance between two extreme points—the highest and lowest values—in a distribution. Even though the range is easily computed, it is a very unstable measure as it can change drastically due to the presence of one or two extreme scores.

- The variance is the sum of the squared differences or deviations between each value (X) in a distribution and the mean of that distribution (M), divided by N. Simply, the variance is the average of the sum of squares (SS). The sum of squares is an abbreviation for the sum of the squared deviation values or deviation scores, Σ (X – M)2. Deviation scores have to be squared before being added in order to eliminate negative numbers. If these numbers were not squared, the positive and negative deviation scores around the mean would cancel each other out, and their sum would be zero. The sum of squares represents the total amount of variability in a score distribution, and the variance (SS/N) represents its average variability. Due to the squaring of the deviation scores, however, the variance is not in the same units as the original distribution.

- The standard deviation is the square root of the variance. Along with the variance, it provides a single value that is representative of the individual differences or deviations in a data set—computed from a common reference point, namely, the mean. The standard deviation is a gauge of the average variability in a set of scores, expressed in the same units as the scores. It is the quintessential measure of variability for testing as well as many other purposes and is useful in a variety of statistical manipulations.

1.4 MEASUREMENT THEORIES: CLASSICAL TEST THEORY, MODERN TEST THEORY

Test developers are basically concerned about the quality of test items and how examinees respond to them when constructing tests. A psychometrician generally uses psychometric techniques to determine validity and reliability. The psychometric theory offers two approaches in analyzing test data: Classical test theory (CTT) and item response theory (IRT). Both theories enable us to predict the outcomes of psychological tests by identifying parameters of item difficulty and the abilities of test takers. Both are concerned with improving the reliability and validity of psychological tests. Both of these approaches provide measures of validity and reliability. There are some identified issues in the classical test theory that concern calibration of item difficulty, sample dependence of coefficient measures, and estimates of measurement error, which are addressed by the item response theory.

1.4.1 Classical Test Theory:

Classical test theory is regarded as the "true score theory." The theory starts from the assumption that systematic effects between responses of examinees are due only to variation in their ability to show interest. All other potential sources of variation existing in the testing materials, such as external conditions or internal conditions of examinees, are assumed either to be constant through rigorous standardization or to have an effect that is non-systematic or random by nature (Van der Linden & Hambleton, 2004). The central model of the classical test theory is that observed test scores (TO) are composed of a true score (T) and an error score (E), where the true and the error scores are independent. The variables were established by Spearman (1904) and Novick (1966) and are best illustrated in the formula: TO = T + E. The classical theory assumes that each individual has a true score, which would be obtained if there were no errors in measurement. However, because measuring instruments are imperfect, the score observed for each person may differ from that individual's true ability. The difference between the true score and the observed test score results from measurement error. Using a variety of justifications, error is often assumed to be a random variable with a normal distribution. The implication of the classical test theory for test takers is that tests are fallible, imprecise tools. The score achieved by an individual is rarely the individual's true score. This means that the true score for an individual will not change with repeated administrations of the same test. This observed score is almost always the true score, influenced by some degree of error. This error influences the observed value to be higher or lower. Theoretically, the standard deviation of the distribution of random errors for each individual reveals the magnitude of measurement error. It is usually assumed that the distribution of random errors will be the same for all individuals.

Classical test theory uses the standard deviation of errors as the basic measure of error. Usually, this is called the standard error of measurement. In practise, the standard deviation of the observed score and the reliability of the test are used to estimate the standard error of measurement (Kaplan & Saccuzzo, 1997). The larger the standard error of measurement, the less certain the accuracy with which an attribute is measured. Conversely, a small standard error of measurement tells that an individual score is probably close to the true score.

Traditionally, methods of analysis based on classical test theory have been used to evaluate tests. The focus of the analysis is on the total test score; the frequency of correct responses (to indicate question difficulty); the frequency of responses (to examine distracters); the reliability of the test, and item-total correlation (to evaluate discrimination at the item level)

(Impara & Plake, 1997). Although these statistics have been widely used, one limitation is that they relate to the sample under scrutiny, and thus all the statistics that describe items and questions are sample-dependent (Hambelton, 2000). This critique may not be particularly relevant where successive samples are reasonably representative and do not vary across time, but this will need to be confirmed, and complex strategies have been proposed to overcome this limitation.

1.4.2 Modern Test Theory or Item Response Theory:

Another branch of psychometric theory is item response theory (IRT). IRT may be regarded as roughly synonymous with latent trait theory. It is sometimes referred to as the strong true score theory or modern mental test theory because IRT is a more recent body of theory and makes stronger assumptions as compared to classical test theory. This approach to testing based on item analysis considers the chance of getting particular items right or wrong. In this approach, each item on a test has its own item characteristic curve that describes the probability of getting each particular item right or wrong given the ability of the test takers (Kaplan & Saccuzzo, 1997). The Rasch model, as an example of IRT, is appropriate for modelling dichotomous responses and models the probability of an individual's correct response on a dichotomous item. The logistic item characteristic curve, a function of ability, forms the boundary between the probability areas of answering an item incorrectly and answering the item correctly. This one-parameter logistic model assumes that the discriminations of all items are assumed to be equal to one (Maier, 2001). Another fundamental feature of this theory is that item performance is related to the estimated amount of the respondent's latent trait (Anastasi & Urbina, 2002). A latent trait is symbolized as theta (□), which refers to a statistical construct. In cognitive tests, latent traits are called the ability measured by the test. The total score on a test is taken as an estimate of that ability. A person's specified ability (□) succeeds on an item of specified difficulty.

There are various approaches to the construction of tests using item response theory. Some approaches use the two dimensions to plot item discriminations and item difficulties. Other approaches use a three-dimension for the probability of test-takers with very low levels of ability getting a correct response. Other approaches use only the difficulty parameter (one dimension), such as the Rasch Model. All these approaches characterize the item in relation to the probability that those who do well or poorly on the exam will have different levels of performance.

1.5 SUMMARY

A psychological test is a systematic procedure for obtaining samples of behaviour, relevant to cognitive or affective functioning, and for scoring and evaluating those samples according to standards. Psychological tests are often described as standardized for two reasons, both of which address the need for objectivity in the testing process. The first has to do with uniformity of procedure in all important aspects of the administration, scoring, and interpretation of tests. Naturally, the time and place when a test is administered, as well as the circumstances under which it is administered and the examiner who administers it, affect test results. The second meaning of standardization concerns the use of standards for evaluating test results. These standards are most often norms derived from a group of individuals—known as the normative or standardization sample—in the process of developing the test.

The concept of measurement is at the heart of psychological testing as a scientific enterprise for the study of human behaviour. Measurement involves the use of certain devices or rules for assigning numbers to objects or events (Stevens, 1946). Psychological testing is largely coextensive with the field of psychometrics, or psychological measurement, and is one of the primary tools for the science and practice of psychology.

There are four types of measurement scales: nominal, ordinal, interval, and ratio scales.

Since the use of numbers to represent objects and events is so common in psychological testing, the field involves the substantial application of statistics, a branch of mathematics dedicated to organizing, depicting, summarizing, analyzing, and otherwise dealing with numerical data. Numbers and graphs used to describe, condense, or represent data belong in the realm of descriptive statistics. On the other hand, when data are used to estimate population values based on sample values or to test hypotheses, inferential statistics—a more ample set of procedures based on probability theory—are applied.

The psychometric theory offers two approaches to analyzing test data: Classical test theory (CTT) and item response theory (IRT). Both theories enable us to predict the outcomes of psychological tests by identifying parameters of item difficulty and the ability of test takers. Both are concerned to improve the reliability and validity of psychological tests. Both of these approaches provide measures of validity and reliability.

1.6 QUESTIONS

1. Discuss various concepts related to the measurement of psychological tests.

2. Critically evaluate classical test theory and modern/item response theory.

1.7 REFERENCES

- Essentials of Psychological Testing by Susana Urbina (2004). Published by John Wiley & Sons, Inc., Hoboken, New Jersey

- Magno, C. (2009). Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data. The

International Journal of Educational and Psychological Assessment, Vol. 1, Issue 1, 1-11.

- Psychological Testing by Anne Anastasi (1976). Fourth Ed. Published by MacMillan Publishing co., Inc. New York
- Psychological Testing and Assessment: An Introduction to Tests and Measurement (2018) by Ronal Jay Cohen and Mark E. Swerdlik. Ninth Ed.

*****

2

THEORETICAL-CONCEPTUAL ORIENTATION – II

Unit Structure

2.0 OBJECTIVES

After learning this chapter, the reader will understand the following:

- How a psychological test is developed
- The concepts related to Item analysis
- Importance of validity and reliability in test development
- Ethical issues related to Psychological Testing

2.1 INTRODUCTION

All tests are not created equal. The creation of a good test is not a matter of chance. It is the product of the thoughtful and sound application of established principles of test construction.

In this chapter, we introduce the basics of test development and examine in detail the processes by which tests are constructed. We explore, for example, a number of techniques designed for the construction and selection of good items.

The process of developing a test occurs in five stages: 1) Test conceptualization, 2) Test construction, 3) Test tryouts, 4) Item analysis, 5) Test revision.

Once the idea for a test is conceived (test conceptualization), the items for the test are drafted (test construction). The first draft of the test is then tried out on a group of sample test-takers (test tryouts). Once the data from the tryout are collected, test-taker's performance on the test as a whole and on each item is analyzed. Statistical procedures, referred to as item analysis, are employed to assist in making judgements about which items are good as they are, which items need to be revised, and which items should be discarded. The analysis of the test items may include analyses of item reliability, item validity, and item discrimination.

2.2 TEST CONCEPTUALIZATION

The beginnings of any published test can probably be traced to thoughts -- self-talk, in behavioral terms. The test developer says to himself or herself something like: "There ought to be a test designed to measure (fill in the blank) in such (as such) way." The stimulus for such a thought could be almost anything. A review of the available literature on an existing test designed to measure a particular construct might indicate that such tests leave much to be desired in terms of psychometric soundness. An emerging social phenomenon or pattern of behavior might serve as the stimulus for the development of a new test. If, for example, celibacy were to become a widely practised lifestyle, then we would witness the development of a variety of tests related to celibacy. These tests might measure variables such as reasons for adopting a celibate lifestyle, commitment to a celibate lifestyle, and degree of celibacy through specific behaviors.

The development of a new test may be in response to a need to assess mastery in an emerging occupation or profession. For example, new tests may be developed to assess mastery in fields such as high-definition electronics, environmental engineering, and wireless communication.

Some preliminary questions:

Regardless of the stimulus for developing the new test, a number of questions are immediately confronted by the prospective test developer:

1) What is the test designed to measure?

2) What is the objective of the test? In the service of what goal will the test be employed? In what way or ways is the objective of the test the same as or different from other tests with similar goals? What real-world behaviors would be anticipated to correlate with test-takers responses?

3) Is there a need for this test? Are there any other tests purporting to measure the same thing? In what ways will the new test be better than or different from the existing ones? Will it be more comprehensive?

Will it take less time to administer? In what ways would this test not be better than existing tests?

4) Who will use these tests? Clinician? Educators? Others? For what purposes would this test be used?

5) Who will take this test? Who is this test for? Who needs to take it? Who would find it desirable to take it? For what age range of test-takers is the test designed? What reading levels are required of a test-taker? What cultural factors might affect the test-taker's response?

6) What content will the test cover? Why should it cover this content? Is this coverage different from the content coverage of existing tests with the same or similar objectives? How and why is the content area different?

7) How will the test be administered? Individually or in a group? Is it amenable to both group and individual administration?

8) What is the ideal format for the test? Should it be true-false, essay, multiple-choice, or in some other format? Why is the format selected for this test the best format?

9) Should more than one form of the test be developed? On the basis of a cost-benefit analysis, should alternate or parallel forms of this test be created?

10) What special training will be required of test users for administering or interpreting the test? What background and qualifications will a prospective user of data derived from an administration of this test need to have? What restrictions, if any, should be placed on distributors of the test and on the test's usage?

11) What types of responses will be required of test-takers? What adaptations or accommodations are recommended for persons with disabilities?

12) Who benefits from the administration of this test?

13) Is there any potential for harm as a result of the administration of this test? What safeguards are built into the recommended testing procedure to prevent any sort of harm to any of the parties involved in the use of this test?

14) How will meaning be attributed to a score on this test? Will a test-taker's score be compared to others taking the test at the same time? To others in a criterion group? Will the test evaluate mastery of a particular content area?

2.3 ITEM ANALYSIS

Statistical procedures used to analyse items may become quite complex. In this section, we will briefly survey some procedures typically used by test developers in their efforts to select the best item from a pool of try-out items. Among the tools test developers might employ to analyse and select items are:

- An index of the item's difficulty
- An index of the item's reliability
- An index of the item's validity
- An index of the item's discrimination

Let us imagine for a moment that we have a test of 100 items for a ninth-grade-level American History Test (AHT). Let's further assume that this 100 - item (draft) test has been administered to 100 ninth-graders. Hoping, in the long run to standardize the test and have it distributed by a commercial test publisher, you have a more immediate, short-term goal: to select the 50 best of the 100 items you originally created. How might that short-term goal be achieved? As we will see, the answer lies in item analysis procedures.

2.3.1 The Item-Difficulty Index:

Suppose every examinee answered item 1 of the AHT correctly. Can we say item 1 is a good item? What if no one answered item 1 correctly? In either case, item 1 is not a good item. If everyone gets the item right, then the item is too easy; if everyone gets it wrong, the item is too difficult. Just as the test as a whole is designed to provide an index of degree of knowledge about American history, each individual item on the test is passed (scored as correct) or failed (scored as incorrect) on the basis of test taker's differential knowledge of American history.

An index of an item's difficulty is obtained by calculating the proportion of the total number of test takers who answered the items correctly. A lowercase "p" (p) is used to denote item difficulty, and a subscript refers to the item number (so p1 is read as item difficulty index for item 1"). The value of an item-difficulty index can theoretically range from 0 (if no one got the item right) to 1 (if everyone got the item right). If 50 of the 100 examinees answered item 2 correctly, then the item-difficulty index for this item would be equal to 50 divided by 100, or .5 (p2=.5). If 75 of the examinees got item 3 right, then p3 would be equal to .75 and we could say that item 3 was easier than item 2. Note that the larger the item-difficulty index, the easier the item. Because p refers to the percent of people passing an item, the higher the p for an item, the easier the item. The statistic referred to as an item-difficulty index in the context of achievement testing may be an item-endorsement index in other contexts, such as personality testing. Here, the statistic provides not a major portion of the percent of people passing the item but a major portion of the percent of people who said yes to, agreed with, or otherwise endorsed the item.

An index of the difficulty of the average test item for a particular test can be calculated by averaging the item-difficulty indices for all the test's items. This is accomplished by summing the item-difficulty indices for all test items and dividing by the total number of items on the test. For maximum discrimination among the abilities of the test takers, the optimal average item difficulty is approximately .5, with individual items on the test ranging in difficulty from .3 to .8. Note, however, that the possible effect of guessing must be taken into account when considering items of the selected-response variety. With this type of item, the optimal average item difficulty is usually the midpoint between 1.00 and the chance success proportion, defined as the probability of answering correctly by random guessing. In a true-false item, the probability of guessing correctly on the basis of chance alone is 1/2, or .50.

Therefore, the optimal item difficulty is halfway between .50 and 1.00, or .75. In general, the midpoint representing the optimal item difficulty is obtained by summing the chance success proportion and 1.00 and then dividing the sum by 2, or

.50+1.00=1.5

1.5/2 =.75

For a five-option multiple-choice item, the probability of guessing correctly on any one item on the basis of chance alone is equal to 1/5, or 20. The optimal item difficulty is therefore .

60: .20+1.00=1.20

1.20/2=.60

2.3.2 The Item-Reliability Index:

The item-reliability index provides an indication of the internal consistency of a test; the higher this index, the greater the test's internal consistency. The index is equal to the product of the item-score standard deviation (s) and the correlation (r) between the item score and the total test score.

Factor analysis and inter-item consistency: A statistical tool useful in determining whether items on a test appear to be measuring the same thing(s) is factor analysis. Through the judicious use of factor analysis, items that do not "load on" the factor that they were written to tap (that is, items that do not appear to be measuring what they were designed to measure) can be revised or eliminated. If too many items appear to be tapping a particular area, the weakest of such items can be eliminated. Additionally, factor analysis can be useful in the test interpretation process, especially when comparing the constellation of responses to the items from two or more groups. Thus, for example, if a particular personality test is administered to two groups of hospitalised psychiatric patients, each group with a different diagnosis, then the same items may be found to load on different factors in the two groups. Such information will compel the responsible test developer to revise or eliminate certain items from the test or to describe the differential findings in the test manual.

2.3.3 The Item-Validity Index:

The item-validity index is a statistic designed to provide an indication of the degree to which a test is measuring what it purports to measure. The higher the item-validity index, the greater the test's criterion-related validity. The item-validity index can be calculated once the following two statistics are known:

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

Calculating the item-validity index will be important when the test developer's goal is to maximize the criterion-related validity of the test. A visual representation of the best items on the test (if the objective is to maximize criterion-related validity) can be achieved by plotting each item-validity index and item-reliability index.

2.3.4 The Item-Discrimination Index:

Measures of item discrimination indicate how adequately an item separates or discriminates between high scorers and low scorers on an entire test. In this context, a multiple-choice item on an achievement test is a good item if most of the high scorers answer correctly and most of the low scorers answer incorrectly. If most of the high scorers fail a particular item, these test takers may be making an alternative interpretation of a response intended to serve as a distracter. In such a case, the test developers should interview the examinees to better understand the basis for the choice and then approximately revise (or eliminate) the item. Common sense dictates that an item on an achievement test is not doing its job if it is answered correctly by respondents who least understand the subject matter. Similarly, an item on a test purporting a particular personality track is not doing its job if responses indicate that people who score very low on the test as a whole (indicating that they are very high on the trait in question, contrary to what the test as a whole indicates).

2.4 VALIDITY

In everyday language, we say that something is valid if it is sound, meaningful, or well-grounded on principles of evidence. For example, we speak of a valid theory, a valid argument, or a valid reason. In legal terminology, lawyers say that something is valid if it is "executed with the proper formalities" (Black, 1979), such as a valid contract or a valid will. In each of these instances, people make judgements based on evidence of the meaningfulness or veracity of something. Similarly, in the language of psychological assessment, validity is a term used in conjunction with the meaningfulness of a test score – What the test score truly means.

A test is considered valid for a particular purpose if it does, in fact, measure what it purports to measure. Questions regarding a test’s validity may focus on the items that collectively make up the test. Do the items adequately sample the range of areas that must be sampled to adequately measure the construct? Individual items will also come under scrutiny in an investigation of a test’s validity. How do individual items contribute to or detract from the test’s validity? The validity of a test may also be questioned on grounds related to the interpretation of the resulting test scores. What do these scores really tell us about the targeted construct? How are high scores on the test related to the test-taker’s behavior? How are low scores on the test related to the test-taker’s behavior? How do scores on this test relate to scores on other tests purporting to measure the same construct? How do scores on this test relate to scores on other tests purporting to measure opposite types of constructs?

We might expect one person’s score on a valid test of introversion to be inversely related to that same person’s score on a valid test of extraversion; that is, the higher the introversion test score, the lower the extraversion test score, and vice versa. Questions concerning the validity of a particular test may be raised at every stage in the life of the test. From its initial development through the life of its use with members of different populations, assessment professionals may raise questions regarding the extent to which a test is measuring what it purports to measure.

The concept of validity:

Validity, as applied to a test, is judgment or estimate of how well a test measures what it purports to measure in a particular context. More specifically, it is a judgement based on evidence about the appropriateness of inferences drawn from test scores.

An inference is a logical result or deduction. Characterizations of the validity of tests and test scores are frequently phrased in terms such as “acceptable” or “weak”.These terms reflect a judgement about how adequately the test measures what it purports to measure.

Researchers have traditionally conceptualized validity according to three categories:

1. Content validity

2. Criterion-related validity

3. Construct validity

Face validity:

Face validity relates more to what a test appears to measure to the person being tested than to what the test actually measures. Face validity is a judgement concerning how relevant the test item appears to be. Stated another way, if a test definitely appears to measure what it purports to measure “on the face of it”, then it could be said to have high face validity.

A paper and pencil personality test labelled the Introversion /Extroversion Test, with items that ask respondents whether they have acted in an introverted or extraverted way in particular situations, may be perceived by respondents as a highly face valid test. On the other hand, a personality test in which respondents are asked to report what they see in inkblots may be perceived as a test with low face validity. Many respondents would be left wondering how what they said they saw in inkblots really had anything at all to do with personality.

Content validity:

Content validity describes a judgement of how adequately a test samples behaviours representative of the universe of behaviours that the test was designed to sample. For example, the universe of behaviours referred to as assertive is very wide-ranging. A content-valid paper and pencil test of assertiveness would be one that is adequately representative of this wide range. We might expect that such a test would contain items sampling from hypothetical situations at home (such as whether the respondent has difficulty making her or his views known to fellow family members), on the job (such as whether the respondent has difficulty asking subordinates to do what is required of them), and in situations (such as whether the respondent would send back a steak not done to order in a fancy restaurant).

Criterion-related validity:

Criterion-related validity is a judgment of how adequately a test score can be used to infer an individual's most probable standing on some measure of interest- the measure of interest being the criterion. Two types of validity evidence are subsumed under the heading criterion- related validity.

Concurrent validity is an index of the degree to which the test score is related to some criterion measure obtained at the same time (concurrently). Predictive validity is an index of the degree to which a test score predicts some criterion measure. Before we discuss each of the validity evidence in detail, it seems appropriate to raise an important question.

Concurrent validity:

If test scores are obtained at about the same time that the criterion measures are obtained, the measure of the relationship between the test scores and the criterion provides evidence of concurrent validity. Statements of concurrent validity indicate the extent to which test scores may be used to estimate an individual's present standing on a criterion. If, for example, scores (or classifications) made on the basis of a psycho diagnostic test were to be validated against a criterion of already diagnosed psychiatric patients, then the process would be one of concurrent validation. In general, once the validity of the inference from the test scores is established, the test may provide a faster, less expensive way to offer a diagnosis or a classification decision. A test with satisfactorily demonstrated concurrent validity may therefore be appealing to prospective users because it holds out the potential for savings in both money and professional time.

Predictive validity:

Test scores may be obtained at one time, and the criterion measures obtained at a later time, usually after some intervening event has taken place. The intervening event may take varied forms, such as training, experience, therapy, medication, or simply the passage of time. Measures of the relationship between the test scores and a criterion measure obtained at a later time provide an indication of the predictive validity of the test; that is, how accurately scores on the test predict some criterion measure. For example, measures of the relationship between college admissions tests and freshman grade point averages, provide evidence of the predictive validity of the admissions tests.

Construct validity:

Construct Validity is a judgement about the appropriateness of inferences drawn from test scores regarding individual standings on a variable called a construct. A construct is an informed, scientific idea developed or hypothesised to describe or explain behavior. Intelligence is a construct that may be invoked to describe why a student performs well in school. Anxiety is a construct that may be invoked to describe why a psychiatric patient paces the floor. Other examples of constructs are job satisfaction, personality, bigotry, clerical aptitude, depression, motivation, self-esteem, emotional adjustment, potential dangerousness, creativity, and mechanical comprehension, to name but a few.

Constructs are unobservable, presupposed (underlying) traits that a test developer may invoke to describe test behavior or criterion performance. The researcher investigating test construct validity must formulate hypotheses about the expected behavior of high scores and low scores on the test. The hypotheses give rise to a tentative theory about the nature of the construct the test was designed to measure. If the test is a valid measure of construct, then high scores and low scores will behave as predicted in the theory. If high scores and low scores do not behave as predicted by the theory, the investigator will need to re-examine the nature of the construct itself or the hypotheses made about it. One possible reason for obtaining results contrary to those predicted by the theory is that the test simply does not measure the construct.

2.5 RELIABILITY

A good test or, more generally, a good measuring tool or procedure is reliable. The criterion of reliability involves the consistency of the measuring tool: the precision with which the test measures and the extent to which error is present in measurements. In theory, the perfectly reliable measuring tool consistently measures the same way.

Whenever we are administering any test, we want to be reasonably certain that the measuring tool or test that we are using is consistent; that is, we want to know that it yields the same numerical measurement every time it measures the same thing under the same conditions. Psychological tests, like other tests and instruments, are reliable to varying degrees. As you might expect, however, reliability is a necessary but not sufficient element of a good test. In addition to being reliable, tests must be reasonably accurate. In the language of psychometrics, a test must be valid.

Reliability refers to consistency in measurement. It is a synonym for dependability/consistency. Due to variation and subjectivity of scoring, individual scores and average group scores always reflect x amount of measurement error.

There are two important features of reliability:

- Consistency
- Dependability

Observed score, True score, Error:

A score on an ability test is not only the test taker’s true score on the ability to be measured but also the test taker's error. The error refers to the component of the observed test score that has nothing to do with the test taker’s ability. If we use X to represent an observed score, it equals the true score plus error. It may be expressed as follows:

•$X=T+E$

Sources of Error:

- Test construction
- Test administration
- Test scoring and interpretation
- Other sources of error

Reliability estimates:

Test-Retest Reliability Estimates:

Test-retest reliability evaluates reliability across time. Reliability can vary with the many factors that affect how a person responds to the test, including their mood, interruptions, time of day, etc. A good test will largely cope with such factors and give relatively little variation. An unreliable test is highly sensitive to such factors and will give widely varying results, even if the person re-takes the same test half an hour later.

The longer the delay between tests, the greater the likely variation. Better tests will give less retest variation with longer delays. The problem with test-retest is that people may have learned something and that the second test is likely to give different results. This method is particularly used in experiments.

Parallel and Alternate Forms:

Parallel-form reliability evaluates different questions and question sets that seek to assess the same construct. Parallel forms of a test exist, and for each form of the test, the means are equal. In theory, the means of scores obtained on parallel forms correlate equally with the true score. More practically, scores obtained on parallel tests correlate equally with other measures.

Alternate forms:

Alternate forms are simply different versions of a test that have been constructed so as to be parallel. Although they do not meet the requirements for the legitimate designation "parallel", alternate forms of a test are typically designed to be equivalent with respect to variables such as content and level of difficulty. Some examples are GED, GRE, SAT.

Split-Half Reliability Estimates:

An estimate of split-half reliability is obtained by correlating two pairs of scores obtained from equivalent halves of a single test administered once. It is a useful measure of reliability when it is impractical or undesirable to assess reliability with two tests or to administer a test twice (because of factors such as time or expense). The computation of a coefficient of split-half reliability generally entails three steps:

Step 1. Divide the test into equivalent halves.

Step 2. Calculate a Pearson r between scores on the two halves of the test.

Step 3. Adjust the half-test reliability using the Spearman-Brown formula.

2.6 NORMS

Norm-referenced testing and assessment can be defined as a method of evaluation and a way of deriving meaning from test scores by evaluating an individual test-takers' score and comparing it to the scores of a group of test-takers. In this approach, the meaning of an individual test score is understood relative to other scores on the same test. A common goal of a norm-referenced test is to yield information on a test-taker's standing or ranking relative to some comparison group of other test-takers.

Norm is the singular and is used in the scholarly literature to refer to behavior that is usual, average, normal, standard, expected, or typical. Reference to a particular variety of norms may be specified by means of modifiers such as age, as in the term age norm. Norms is the plural form of norm, as in the term gender norms. In a psychometric context, norms are the test performance data of a particular group of test-takers that are designed for use as a reference when evaluating or interpreting individual test scores. As used in this definition, the “particular group of test-takers” may be defined broadly (for example, “a sample representative of the adult population of India”) or narrowly (for example, “female inpatients at the Sion Hospital with a primary diagnosis of depression”). A normative sample is a group of people whose performance on a particular test is analysed for reference in evaluating the performance of individual test-takers.

Usually, members of the normative sample will all be typical with respect to some characteristic(s) of the people for whom the particular test was designed. A test administered to this representative sample of test-takers yields a distribution (or distributions) of scores. These data constitute the norms for the test and are typically used as a reference source for evaluating and placing into context test scores obtained by individual test-takers. The data may be in the form of raw scores or converted scores.

The word norm, as well as related terms such as norming, refer to the process of deriving norms. Norming may be modified to describe a particular type of norm derivation. For example, race norming is the controversial practise of norming on the basis of race or ethnic background.

Norming a test, especially with the participation of a nationally representative normative sample, can be a very expensive proposition. For this reason, some test manuals provide what are variously known as user norms or program norms, which “consist of descriptive statistics based on a group of test-takers in a given period of time rather than norms obtained by formal sampling methods” (Nelson, 1994).

Types of Norms:

- Percentile
- Age norms
- Grade norms
- National norms
- National anchor norms
- Sub-group norms
- Local norms

Norm-referenced versus criterion-referenced valuation:

One way to derive meaning from a test score is to evaluate the test in relation to other scores on the same test. This approach to evaluation is referred to as norm-refer. Another way to derive meaning from a test score is to evaluate it on the basis of whether or not some criterion has been met. A criterion may be defined as a standard on which a judgement or decision may be based. Criterion-referenced testing and assessment may be defined as a method of evaluation and a way of deriving meaning from test scores by evaluating an individual's score with reference to a set standard. Some examples:

- To be eligible for a high school diploma, students must demonstrate at least a sixth-grade reading level.
- To earn the privilege of driving an automobile, would-be drivers must take a road test and demonstrate their driving skills to the satisfaction of the state-appointed examiner.
- To be licenced as a psychologist, the applicant must achieve a score that meets or exceeds the score mandated by the state on the licencing test.

2.7 ETHICAL ISSUES IN PSYCHOLOGICAL TESTING

Concern about the use of psychological tests first became widespread in the aftermath of World War I, when various professionals (as well as non-professionals) sought to adapt group tests developed by the military for civilian use in schools and industry. Reflecting growing public discomfort with a lot of the assessment industry were popular magazine articles featuring stories with titles such as "The Abuse of Tests" (Haney, 1981). Less well-known were voices of reason that offered constructive ways to correct what was wrong with assessment practises.

In the USA, Congress passed the National Defense Education Act, which provided federal money to local schools for the purpose of testing ability and aptitude to identify gifted and academically talented students. This event triggered a large-scale testing programs in the schools. At the same time, the use of ability tests and personality tests for personnel selection increased in government, the military, and business.

Laws are rules that individuals must obey for the good of society as a whole. Whereas ethics is a body of principles of right, proper, or good conduct, for example, a principle of ethical research is that the researcher should never fudge data; all data must be reported accurately.

A code of professional ethics is recognised and accepted by members of a profession; it defines the standard of care expected of members of that profession. In this context, we may define standard of care as the level at which an average, reasonable, and prudent professional would provide diagnostic or therapeutic services under the same or similar conditions.

Let us discuss a few points in this regard:

Test-user qualifications:

Should just anyone be allowed to purchase and use psychological test materials? If not, then who should be permitted to use psychological tests?

As early as 1950, the APA Committee on Ethical Standards for Psychology published a report called Ethical Standards for the Distribution of Psychological Tests and Diagnostic Aids. This report defined certain levels of tests in terms of the degree to which the tests use required knowledge of testing and psychology.

Testing people with disabilities:

Challenges analogous to those concerning test-takers from linguistic and cultural minorities are present when testing people with disabling conditions. Specifically, these challenges may include (1) transforming the test into a form that can be taken by the test-taker, (2) transforming the responses of the test-taker so that they are scoreable, and (3) meaningfully interpreting the test data.

Another complex issue—this one ethically charged—has to do with a request by a terminally ill individual for assistance in quickening the process of dying. In Oregon, the first state to enact “Death with Dignity” legislation, a request for assistance in dying may be granted only contingent on the findings of a psychological evaluation; life or death literally hangs in the balance of such assessments.

The right of informed consent:

Test-takers have a right to know why they are being evaluated, how the test data will be used, and what (if any) information will be released to whom. With full knowledge of such information, test-takers give their informed consent to being tested. The disclosure of the information needed for consent must be in a language the test-taker can understand. Test-takers have a right to be informed, in a language they can understand, of the nature of the findings with respect to a test they have taken. They are also entitled to know what recommendations are being made as a consequence of the test data. If the test results, findings, or recommendations made on the basis of test data are voided for any reason (such as irregularities in the test administration), test-takers have a right to know that.

The right to privacy and confidentiality:

State statutes have extended the concept of privileged information and confidentiality to parties who communicate with each other in the context of certain relationships, including the lawyer-client relationship, the doctor-patient relationship, the priest–penitent relationship, and the husband-wife relationship. In most states, privilege is also accorded to the psychologist–client relationship. Professionals such as psychologists who are parties to such special relationships have a legal and ethical duty to keep their clients’ communications confidential.

In short, you should remember the following points in relation to ethical issues related to psychological testing:

- Informed consent

- Confidentiality
- Competence
- Integrity
- Professional and scientific responsibility
- Respect for people's rights and dignity
- Concern for others' welfare
- Social responsibility

2.8 SUMMARY

The creation of a good test is not a matter of chance. It is the product of the thoughtful and sound application of established principles of test construction. In this chapter, you studied the basics of test development and examined in detail the processes by which tests are constructed. A number of techniques designed for the construction and selection of good items were explored.

The process of developing a test goes through five stages: test conceptualization, test construction, test tryouts, item analysis, and test revision.

Once the idea for a test is conceived (test conceptualization), items for the test are drafted (test construction). The first draft of the test is then tried out on a group of sample test-takers (test tryouts). Once the data from the tryout are collected, the test-takers performance on the test as a whole and on each item is analyzed. Statistical procedures, referred to as item analysis, are employed to assist in making judgements about which items are good as they are, which items need to be revised, and which items should be discarded. The analysis of the tests items may include analyses of item reliability, item validity, and item discrimination.

After completing the above steps, the researcher has to check the test for validity and reliability. In psychological assessment, validity is a term used in conjunction with the meaningfulness of a test score – what the test score truly means. Validity is of three types: content validity, construct validity, and criterion-related validity. A good test or, more generally, a good measuring tool or procedure is reliable. The criterion of reliability involves the consistency of the measuring tool: the precision with which the test measures and the extent to which error is present in measurements. In theory, the perfectly reliable measuring tool consistently measures the same way. There are methods to check for test reliability, such as test-retest reliability, split-half reliability, parallel form reliability, etc.

Norm-referenced testing and assessment can be defined as a method of evaluation and a way of deriving meaning from test scores by evaluating an individual test-takers' score and comparing it to the scores of a group of test-takers. In this approach, the meaning of an individual test score is understood relative to other scores on the same test. A common goal of a norm-referenced test is to yield information on a test-taker's standing or ranking relative to some comparison group of other testakers.

Concern about the use of psychological tests first became widespread in the aftermath of World War I, when various professionals (as well as non-professionals) sought to adapt group tests developed by the military for civilian use in schools and industry. Reflecting growing public discomfort with a lot of the assessment industry were popular magazine articles featuring stories with titles such as "The Abuse of Tests" (Haney, 1981).

The committee on Ethical Standards for Psychology published a report called Ethical Standards for the Distribution of Psychological Tests and Diagnostic Aids. This report defined a few levels of tests in terms of the degree to which the tests use required knowledge of testing and psychology.

2.9 QUESTIONS

1. Write a note on test conceptualization and some related questions related to test conceptualization.
2. Discuss item analysis and elaborate on item difficulty, item validity, item reliability, and the item discrimination index.
3. What ethical measures will you take as a counsellor while conducting a psychological test?

2.10 REFERENCES

- Essentials of Psychological Testing by Susana Urbina (2004). Published by John Wiley & Sons, Inc., Hoboken, New Jersey
- Psychological Testing and Assessment: An Introduction to Tests and Measurement (2018) by Ronal Jay Cohen and Mark E. Swerdlik. Ninth Ed.
- Psychological Testing by Anne Anastasi (1976). Fourth Ed. Published by MacMillan Publishing co., Inc. New York

*****

3

TEST ADMINISTRATION AND REPORTING – I

Unit Structure

3.0 OBJECTIVES

- To learn about the Locus of Control (LOC) Scale by J. B. Rotter.
- To know how to administer, score, interpret and report the LOC Scale
- To know its psychometric properties – item analysis, reliability and validity

3.1 INTRODUCTION: LOCUS OF CONTROL SCALE

In Rotter's own words, locus of control is the degree to which an individual perceives that a reward follows from or is contingent upon their own behaviour or attributes, versus the degree to which they feel that the reward is controlled by the forces outside of him or herself that occur independently of his or her actions (Rotter, 1966). In simple words, locus of control indicates one's belief regarding whether the reward one receives is or is not dependent on his or her own behaviour, that is, whether the reward is or is not a result of their own behaviour.

In this unit, we will study the background and description of the Locus of Control (LoC) scale in brief in this section. In the subsequent sections, we will also learn many other details of the scale, such as its administration, scoring, reporting, and its psychometric properties.

J. B. Rotter proposed social learning theory, which integrated learning theory and personality theory. According to him, the effects of reward or reinforcement on preceding behaviour depend partly on whether the person perceives the reward as dependent on his or her behaviour or independent of it. Thus, attainment and performance differ in situations when situations are perceived as determined by skill versus chance. For example, if someone believes that the reward is a result of his or her own behaviour and skill, he or she will work hard and strive to be successful in his or her field. Similarly, if someone believes that the reward one receives is merely a part of luck or chance, his or her performance will be different than the person who believes the reward is a result of hard work.

In line with this, Rotter stated that persons may also differ in generalized expectancies for internal versus external control of reinforcement. Thus, locus of control is more commonly a known form of the concept of generalized expectancies for control of reinforcement proposed by Rotter. With reference to control of reinforcement, locus of control refers to people's very general, cross-situational beliefs about what determines whether or not they get reinforced in life.

People who have an internal locus of control believe that they can exercise control over events in their life and that outcomes are determined by their own effort and abilities. On the other hand, people who have an external locus of control, believe that their behaviour or decision-making does not have much impact, rather things are controlled by external forces, such as fate, chance, or powerful others.

Reinforcement, reward or gratification, all play an important role in the locus of control. That is, the same event may be regarded as a reward by some people, while others may perceive and react to it differently. According to Rotter, whether the event will be regarded as a reward, will be determined by the degree to which the individuals perceive that the reward follows from, or is dependent on their behaviour or attributes versus the degree to which they feel that the reward is controlled by forces outside of themselves and may occur independently of their own actions. When individuals perceive reinforcement as following their own action, it is typically perceived as the result of luck, chance, or fate, as under the control of others, or as unpredictable because of the great complexity of the forces surrounding them.

Here, the locus of control comes into the picture, which depends on the individuals' perception of whether the event is dependent or not upon their own behaviour or their own relatively permanent characteristics. In other words, if the event is perceived by individuals to be dependent on their own behaviour, it is known to be a belief in internal control. On the other hand, if the event is perceived as not being dependent on individuals' behaviour, it is known to be a belief in external control.

A Brief History of Locus of Control:

Rotter considers the social upheaval of the 1900s, such as the Vietnam War, Watergate, the inner-city riots, and political assassinations as responsible incidents as the ones which might have played some role in the interest in locus of control. The reason for this is that these incidents were themselves very disturbing. At the same time, they were also concerning many people including social scientists, because of the perceived lack of control over their own lives. Hence, Rotter believes that this social upheaval might have led to the interest in the locus of control.

Four Propositions and Locus of Control

Rotter (1990) in his article "Internal versus external control of reinforcement: A case history of a variable" discussed the four propositions that he believed as being responsible for the heuristic value of internal-external control. Rotter also believed that these propositions are particularly relevant to the field of personality theory, personality measurement, and the study of psychology as a whole. These four propositions are listed below with their descriptions in brief:

1) The importance of precise definition:

This first proposition is that the heuristic value of a construct is partially dependent on the precision of its definition.

Rotter (1990) suggests here the need for a good definition, especially of a cognitive or subjective variable, which must be stated in careful and precise language and leads to a common understanding. According to him, it needs to be illustrated with many behavioural examples of its consequences if its presence or absence is not directly observable. The way they are stated should make the operations for its measurement clear and they should be widely accepted as logical and reasonable. Rotter further suggested that even precise definitions should usually and necessarily distinguish between the construct being defined and other constructs used in past and present with which it can be confused. Also these precise definitions should make the connections to other constructs clear, so that previously collected data can be interpreted and built on.

2) The importance of imbedding a construct in a broader theory:

The second proposition is that the heuristic value of a construct is considerably enhanced if it is imbedded in a broader theory of behaviour. Here, Rotter (1990) underlines the importance of recognizing the conept originated both from theoretical and clinical concerns, with social learning theory which organizes our thinking in both cases.

3) Measurement principles should be derived from psychological theory:

The third proposition is that the predictive value of a test is likely to be increased if the principles of measurement are derived from the same theory as the constructs to be measured. Rotter (1990) emphasized the need of having a theory of behaviour, and consequently, of test-taking behaviour, as well as some notions of the theoretical properties of the variable being studied in order to devise a construct valid measure.

4) The dissemination of knowledge:

The fourth proposition is that the research monograph is an ideal form of publication for the dissemination of knowledge. According to Rotter (1990), psychology, in order to advance in understanding human behaviour, needs to emphasize programmatic research (whether theoretical or applied), in which theory and empirical findings are combined. It also needs to build on past research.

Description of the Locus of Control Scale:

The Locus of Control Scale was originally developed by Julian. B. Rotter (1916-2014). It is one of the widely used personality tests, which has been translated into over 40 languages. The Locus of Control Scale, which was referred to as the I-E Scale by Rotter, is a 29-item test, a final version that measures whether an individual has an internal or external locus of control. In other words, it measures the degree to which the individual interprets events as being a result of their own actions or external factors. It is a forced-choice questionnaire and respondents must select a response choice that provides a specific answer to each item based on his or her own belief about the statements in the scale. It can be administered on an individual or even in a group setting.

Each of the 29 items consists of a pair of statements ‘a’ and ‘b’. A careful reading of the items makes it clear that they exclusively deal with the respondents’ belief about the nature of the world. In other words, the test items are concerned with the respondents’ expectations about how reinforcement in terms of reward or punishment is controlled. Consequently, this test is considered to be a measure of generalized expectancy. It may correlate with the value that the respondent places on internal control. However, none of the items is directly addressed to the preferences for internal or external control.

Respondents must select the statement from each of the 29 pairs to which they agree the most and write ‘a’ or ‘b’ accordingly in the blank space provided. The locus of control scale also contains six filler items (1, 8, 14, 19, 24, 27) which maintain the ambiguity of the test purpose. Table 3.1 presents a few items of Rotter’s Locus of Control Scale. The administration and scoring procedure of the test has been explained in Sections 3.2 and 3.3, respectively.

Table 3.1 Items of Locus of Control Scale

Sr. No.	Items	Response
1.	a) Children get into trouble because their parents punish them too much. b) The trouble with most children nowadays is that their parents are too easy with them.	
2.	a) Many of the unhappy things in people’s lives are partly due to a bad luck.	

	b) People's misfortunes result from the mistakes they make.	
5.	a) The idea that teachers are unfair to students is nonsense. b) Most students don't realize the extent to which their grades are influenced by accidental happenings.	
8.	a) Heredity plays the major role in determining one's personality. b) It is one's experiences in life which determine what they are like.	
10.	a) In the case of the well-prepared student, there is rarely if ever such a thing as an unfair test. b) Many times exam questions tend to be so unrelated to course work that studying is really useless.	
19.	a) One should always be willing to admit mistakes. b) It is usually best to cover up one's mistakes.	
22.	a) With enough effort, we can wipe out political corruption. b) It is difficult for people to have much control over the things politicians do in office.	
25.	a) Many times I feel that I have little influence over the things that happen to me. b) It is impossible for me to believe that chance or luck plays an important role in my life.	

{Source: Rotter, J. B. (1966). Generalized expectancies for internal versus external locus of control of reinforcement. Psychological Monographs: General and applied, Whole No. 609, 80(1), 1-28.}

3.1.1 Purpose:

To find out the details of personality aspect in terms of locus of control – that is, the extent of control one perceives to have over the situations in life.

3.2 STANDARDIZATION OF NORMS

The locus of control scale has been widely tested on various population samples, which includes elementary psychology students from Ohio State University, Kansas State University, University of Connecticut National stratified sample of Purdue opinion poll of 10th, 11th, and 12th grades, prisoners from Colorado reformatory, Ohio Federal prisoners aged 18 to 26 years from 8th grade plus reading, Peace corps trainees, Negro students of psychology classes from Florida State University, 18-year-old subjects from Boston area. Most of the work reported by Rotter was completed at the Ohio State University.

3.3 ITEM ANALYSIS, VALIDITY, AND RELIABILITY

Item Analysis and Validity:

This 29-item Locus of Control Scale has been derived from an item analysis of two previous versions of the scale. The earliest version of the scale was developed by Late Shephard Liverant in association with J. B. Rotter and M. Seeman by developing the subscales for different areas, such as achievement, affection, and general social and political attitudes; and control for social desirability. It consisted of 100 forced-choice items and each item compared an external belief with an internal belief. Later, Liverant conducted an item analysis of this 100-item scale and reduced it to a 60-item version based on internal consistency criteria.

Later, item analysis of this 60-item version scale was carried out which indicated that the subscales were not generating separate predictions. Also, achievement items tended to correlate highly with the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1964), while some subscales correlated with other scales at approximately the same level as their internal consistency. As a result, items to measure more specific subareas of internal-external control were discarded. Item correlations of this test with the social desirability scale for different samples were deemed to be very high, ranging from .35 to .40. This 60-item scale was then reduced and purified by S. Liverant, J. B. Rotter, and D. Crowne for which validity data along with internal consistency data from two studies were used. Also, some items were reframed with the changes in their wording to make the items appropriate for non-college adults and upper-level high school students in this final revision.

This scale was later reduced to 23 items by eliminating items either with a high correlation with the Marlowe-Crowne Social Desirability scale, a non-significant relationship with other items, or a correlation approaching zero with both validation criteria. And lastly, a final version of the scale was developed as a 29-item, forced-choice test with 6 filler items intending to make the purpose of the test somewhat more ambiguous.

Internal consistency estimates of the scale are relatively stable. These estimates are only moderately high for a scale of this length (i.e., 29-item scale). However, Rotter (1966) suggested that it should be remembered that the items of the scale are not arranged in a difficulty hierarchy, but they are the samples of attitudes in a wide variety of different situations.

Reliability:

This test is an additive which consists of items that are not comparable. That is why split-half or matched-half reliability (.65 for males, .79 for females, and .73 for total sample) tends to underestimate the internal consistency. Also, this test is a forced-choice scale in which an attempt is made to balance alternatives, so that probabilities of endorsement of either alternative do not include the more extreme splits. Therefore, the Kuder-Richardson reliabilities are somewhat limited in two samples of

Elementary Psychology students from Ohio State University (.70 for males from each of the two samples, .76 and .70 for females, and .73 and .70 for combined sample, that is, male-female together). The Kuder-Richardson reliability was .69 for the combined sample of males and females from the National stratified sample (Franklin, 1963).

Test-retest reliability in two different samples (Elementary psychology students from Ohio State University and Prisoners from Colorado Reformatory) was consistent (.60 and .78 for males) for 1 month period. It was .83 for females and .72 for combined in one of the samples. On the other hand, there were lower test-retest reliabilities (.49 for males, .61 for females, and .55 for combined) for 2 month period, probably due to the types of test administration – first as group administration and second as individual administration. The correlation for this 29-item scale ranged from -.07 to -.35.

3.4 ADMINISTRATION AND SCORING

(Note: Sections 3.4 and 3.5 should be reported in the past tense while writing in the journal as a past activity conducted in the psychology laboratory.)

3.2.1 Tools and Material:

- Locus of Control (LOC) Scale with Answer/Response sheet
- Stationery

3.2.2 Procedure:

The test administrator requires ensure all arrangements for administering the test. Then the test taker/participant is taken to the laboratory and is asked to sit comfortably. Initial general instructions are given to the test taker as follows for rapport building before presenting a Locus of Control Scale, along with a few general questions if required:

"Please be comfortable. I will be presenting scale you with a psychological scale, which requires approximately 5 to 10 minutes to complete. This will help me understand your views in general. I'll be giving you instructions for completing the scale once presented".

A participant then is provided with the answer sheet to record his/ her responses to 29 pairs of statements in the Locus of Control Scale and a pencil. The following standard instructions are as follows as provided by the author J. B. Rotter (1966) in his article "Generalized Expectancies for Internal versus External Control of Reinforcement" to be given to individuals:

"This is a questionnaire to find out the way in which certain important events in our society affect different people. Each item consists of a pair of alternatives lettered a or b. Please select the one statement of each pair (and only one) which you more strongly believe to be more true rather than the one you think you should choose or the one you would like to be true. This is a measure of personal belief, obviously there are no right or wrong answers.

Your answers to the items on this inventory are to be recorded on a separate answer sheet which is loosely inserted in the booklet. REMOVE THIS ANSWER SHEET NOW. Print/Write your name and any other information requested by the examiner on the answer sheet, then finish reading these directions. Do not open the booklet until you are told to do so.

Please answer these items carefully, but do not spend too much time on any one item. Be sure to find an answer for every choice. Find the number of the item on the answer sheet and blacken the space under the number 1 or 2 which you choose as the statement more true.

In some instances, you may discover that you believe both statements or neither one. In such cases, be sure to select the one you more strongly believe to be the case as far as you are concerned. Also, try to respond to each item independently when making your choice; do not be influenced by your previous choices."

After completing the instruction part, the test administrator says, "Have you understood? Shall we begin?" and makes sure that the participant has understood all the instructions, has no difficulty or query regarding it and is ready to start the test. He or she is allowed to start responding to test items.

When the participant finishes the task, the test administrator makes sure that the participant has responded to all test items. If any item has been left without response by any chance, the test administrator requests the test taker to respond to that particular item/s and complete the test.

Once the participant completes the test, the test administrator asks a few questions to the participant regarding the scale. They are as follows:

1) How was the test?

2) Was that easy for you to respond to the test items?

3) What came to your mind while completing the test?

4) What may be the purpose of the test, according to you?

Apart from these post-task questions, other relevant questions may be asked to the participant. Thus, the participant is encouraged to express his/her thoughts, feelings, and views, if any, regarding the task/test. The test taker is then requested to wait outside for some time till the individual test score is obtained.

3.2.3 Scoring:

The responses on the Locus of Control Scale are to be measured with the help of the scoring key. The scoring key indicates the scores as 0 for some responses and 1 for some responses. After completion of the test, out of 29 items responded to by the individuals, all items (except filler items) are scored either as 0 or 1 using the scoring key. The responses for the filler items are not scored at all. Accordingly, some responses are scored as 1 and some are scored as 0. Then, all scores are summed up to obtain the total score. Total scores on the scale range from 0 to 13, excluding the filler items, which are not scored. where lower scores indicate an internal locus of control, while higher scores indicate the external locus of control.

3.5 INTERPRETATION AND REPORTING

Interpretation and reporting the scores and results of the test should always be done properly and in sensitive manner.

3.3.2 Interpretation of the Scores of LOC:

The higher scores reflect the external locus of control, while the lower scores indicate the internal locus of control. The results are reported and communicated to the person accordingly in very sensitive manner for which the test administrator must be having a knowledge of the test, its nature and the interpretation of the test scores (high or low).

3.3.3 Reporting the Results of LOC:

The LOC gives an overall score of the test, which reflects whether the person has internal or external locus of control. The test results should be reported as below:

The participant’s score on the LOC test was ___, which indicates _____ (internal or external) locus of control. That is, it shows that the participant believes that the control of the events occurring in his or her life and around him or is ____ (inside/outside) himself or herself. He or she attributes factors, such as ________________________ as the reason for the resulting events. Also, he or she believes that the reward or punishment that he or she receives are/are not dependent on his or her behaviour.

The same example can be elaborated as below for clear idea about reporting the results:

The participant’s score on the LOC was 4, which indicates that the participant has internal locus of control. That is, it shows that the participant believes that the control of the events occurring in her life and around him or her is inside herself. She attributes the factors, such as efforts, hard work, and other such internal factors as the reason for the resulting events and only these factors are responsible for the reward or punishment that she receives. She does not believe in external factors, like luck, chance, fate. She believes that such external factors have no role in her life. The reward or punishment that she receives are merely dependent on her own behaviour.

Thus, the interpretation and reporting should be done in very careful manner.

3.6 SUMMARY

In this unit, we learned about the Locus of Control Scale which was developed by Julian. B. Rotter. It is one of the widely used personality tests. We learned about its history in brief, four propositions on which it is based, and its description and purpose of this unit. We also had a glance on some test items to understand the nature of the test and its items. We learned about its standardization of norms in brief, that is the various sample populations on which the test was standardized. We learned about the psychometric properties of the test, that is item analysis, validity and reliability. Then finally we moved towards the practical part of the test, that is administration, scoring, interpretation and reporting the test results at the end.

3.7QUESTIONS

1. Describe the concept of Locus of Control with the help of suitable examples.
2. Describe the procedure of administering and scoring LOC.
3. Write short notes on:

(a) Internal and external locus of control

(b) Social learning theory and locus of control

(c) Four propositions and locus of control

(d) Interpretation and reporting LOC results

3.8 REFERENCES

- Rotter, J. B. (1966). Generalized expectancies for internal versus external locus of control of reinforcement. Psychological Monographs: General and applied, Whole No. 609, 80(1), 1-28.
- Mearns, J. (2021). The social learning theory of Julian B. Rotter (1916-2014). Available at http://psych.fullerton.edu/jmearns/rotter.htm
- CES. (). Locus of Control Scale. Retrieved from https://effectiveservices.force.com/s/measure/a007R00000v8Qg2QAE/locus-of-control-scale
- Rotter, J. B. (1990). Internal versus external control of reinforcement: A case history of a variable. American Psychologist, 45(4),489-493.

3.9 APPENDIX

Table 3.1 Items of Locus of Control Scale

Sr. No.	Items	Response
1.	a) Children get into trouble because their parents punish them too much. b) The trouble with most children nowadays is that their parents are too easy with them.	
2.	a) Many of the unhappy things in people's lives are partly due to a bad luck. b) People's misfortunes result from the mistakes they make.	
3.	a) One of the major reasons why we have wars is because people don't take enough interest in politics. b) There will always be wars, no matter how hard people try to prevent them.	
4.	a) In the long run people get the respect they deserve in this world. b) Unfortunately, an individual's worth often passes unrecognized no matter how hard he tries.	
5.	a) The idea that teachers are unfair to students is nonsense. b) Most students don't realize the extent to which their grades are influenced by accidental happenings.	
6.	a) Without the right breaks one cannot be an effective leader. b) Capable people who fail to become leaders have not taken advantage of their opportunities.	
7.	a) No matter how hard you try some people just don't like you. b) People who can't get others to like them don't understand how to get along with others.	
8.	a) Heredity plays the major role in determining one's personality. b) It is one's experiences in life which determine what they are like.	
9.	a) I have often found that what is going to happen will happen. b) Trusting to fate has never trued out as well as for me as making a decision to take a definite course of action.	
10.	a) In the case of the well-prepared student, there is rarely if ever such a thing as an unfair test. b) Many times exam questions tend to be so unrelated to course work that studying is really useless.	
11.	a) Becoming a success is a matter of hard work, luck has little or nothing to do with it. b) Getting a good job depends mainly on being in the right place at the right time.	
12.	a) The average citizens can have an influence in government decisions.	

	b)	This world is run by the few people in power, and there is not much the little guy can do about it.	
13.	a)	When I make plans, I am almost certain that I can make them work.	
	b)	It is not always wise to plan too far ahead because many things turn out to be a matter of good or bad fortune anyhow.	
14.	a)	There are certain people who are just no good.	
	b)	There is some good in everybody.	
15.	a)	In my case, getting what I want has little or nothing to do with luck.	
	b)	Many times we might just as well decide what to do by flipping a coin.	
16.	a)	Who gets to be the boss often depends on who was lucky enough to be in the right place first.	
	b)	Getting people to do the right thing depends upon ability, luck has little or nothing to do with it.	
17.	a)	As far as world affairs are concerned, most of u are the victims of forces we can neither understand, nor control.	
	b)	By taking an active part in political and social affairs the people can control world events.	
18.	a)	Most people don’t realize the extent to which their lives are controlled by accidental happenings.	
	b)	There really is no such thing as “luck”.	
19.	a)	One should always be willing to admit mistakes.	
	b)	It is usually best to cover up one’s mistakes.	
20.	a)	It is hard to know whether or not a person really likes you.	
	b)	How many friends you have depends upon how nice a person you are.	
21.	a)	In the long run the bad things that happen to us are balanced by the good ones.	
	b)	Most misfortunes are the result of lack of ability, ignorance, laziness, or all three.	
22.	a)	With enough effort, we can wipe out political corruption.	
	b)	It is difficult for people to have much control over the things politicians do in office.	
23.	a)	Sometimes I can’t understand how teachers arrive at the grades they give.	
	b)	There is a direct connection between how hard I study and the grades I get.	
24.	a)	A good leader expects people to decide for themselves what they should do.	
	b)	A good leader makes it clear to everybody what their jobs are.	
25.	a)	Many times I feel that I have little influence over the things that happen to me.	
	b)	It is impossible for me to believe that chance or luck plays an	

	important role in my life.	
26.	a) People are lonely because they don't try to be friendly. b) There's not much use in trying too hard to please people, if they like you, they like you.	
27.	a) There is too much emphasis on athletics in high school. b) Team sports are an excellent way to build character.	
28.	a) What happens to me is my own doing. b) Sometimes I feel that I don't have enough control over the direction my life is taking.	
29.	a) Most of the time I can't understand why politicians behave the way they do. b) In the long run the people are responsible for a bad government on a national as well as on a local level.	

{**Source:** Rotter, J. B. (1966). Generalized expectancies for internal versus external locus of control of reinforcement. Psychological Monographs: General and applied, Whole No. 609, 80(1), 1-28.}

*****

4

TEST ADMINISTRATION AND REPORTING – II

Unit Structure

4.0 OBJECTIVES

- To learn about the 16 PF- Fifth Edition Scale by Cattell et al.
- To know how to administer, score, interpret and report the 16 PF- Fifth Edition
- To know its psychometric properties (internal consistency, reliability and validity)

4.1 INTRODUCTION: 16PF FIFTH EDITION

The 16 PF Questionnaire is a widely known instrument measuring personality factors. It was developed by Dr. Raymond B. Cattell over 45 years ago in order to identify the sixteen primary components of personality. Thus, 16PF stands for the sixteen personality factors. It is a comprehensive measure of personality and found to be effective in a variety of settings, where an in-depth assessment of the whole person is required. This test utilizes factor analysis and the 16PF traits are the result of years of factor-analytic research focused on discovering the basic structural elements of personality (Cattell, 1957, 1973). Thus, it is based on the factor analysis of all English-language adjectives which describe human behaviour.

It is widely used internationally and it has been adapted into over 35 languages worldwide since its beginning. These adaptations are beyond simply translations, that is, they are careful cultural adaptations with new norms and reliability and validity research in each new country. The Web-based administration of this test, which took place in 1999, allowed international test-users easy access to administration, scoring, and reports in many different languages by using local norms.

Cattell's Theory of Personality:

Cattell's 16PF Questionnaire was based on his factor-analytic theory (Cattell, 1933, 1946). He identified the five broad dimensions that are a variant of the 'Big Five' Factors (Cattell, 1957, 1970). He proposed three levels of factors through his work as a multi-level, hierarchical structure of personality. These factors are as follows:

1) Primary factors or first-order factors:

Cattell and his colleagues first discovered the primary traits. These factors provide the most basic definition of individual personality differences. They are more specific primary traits, which are more precise in nature and reveal the fine details and shades that make each person unique. They are also more powerful in understanding and predicting the complexity of the actual behaviour of the person (Ashton, 1998; Judge et al., 2002; Mershon & Gorsuch, 1988; Paunonen & Ashton, 2001; Roberts et al., 2005). Cattell and his colleagues then factor-analyzed these primary traits themselves in order to investigate personality structure at a higher level.

2) Second-order factors:

They are global measures that describe personality at a broader and conceptual level. These are the factors that emerged from the factor analysis of the primary traits. They were the original Big Five and were found to define personality at a higher and more theoretical level of personality.

3) Third-order factors:

These factors were the result of the factor analysis of global factors, that is, second-order factors. The third-order factors were at the highest, most abstract level of personality organization (Cattell, 1946, 1957, 1973). These factors are also called the super factors of personality (Cattell & Mead,). Many researchers who attempted to investigate the third-order factor structure of the 16PF by studying different populations and by applying factor analysis, and found similar results with two third-order factors:

- Third-order Factor I, which involves Extraversion and Independence, and
- Third-order Factor II, which involves Self-Control and Tough-mindedness
- The fifth global factor is Anxiety/neuroticism, which loads on both of these two third-order factors.

The results yielded by these researchers are consistent with Cattell's original belief that these third-order factors may not represent personality traits in the usual sense, but they might reflect some broad, abstract level of sociological or biological influences on human temperament (Cattell, 1957, 1973). However, Cattell and Mead () express the need for further investigation in defining and understanding these third-order factors.

The 16PF Questionnaire has a scientific origin, due to which it has a long history of empirical research. It is rooted in a well-established theory of individual differences. This questionnaire has an extensive body of research which dates back over half a century and it provides evidence of its utility in various settings, such as clinical, counselling, industrial-organizational, educational and research (Cattell et al., 1970; Cattell & Schuerger, 2003; Conn & Rieke, 1994; Krug & Johns, 1990; Russell & Karol, 2002). It has been found that the 16PF is among the top five most commonly used normal-range instruments in both research and practice (Butcher & Rouse, 1996; Piotrowski & Zalewski, 1993; Watkins et al., 1995).

The 16PF instrument provides scores on the 16 primary scales, 5 global scales, and 3 response bias scales. All personality scales are bipolar, that is they have clear, meaningful definitions at both ends (test taker and test giver). They are given in the Stens (standardized-ten scores) that range from 1 to 10 with a mean of 5.5 and a standard deviation of 2.0. The latest standardization of the questionnaire, which was published in 2001, includes over 10000 people.

History and Development of the 16PF Questionnaire:

The 16PF Questionnaire was developed from a unique perspective with a scientific quest and it attempted to discover the basic structural elements of personality. Thus, the history of this instrument spans almost the entire history of standardized personality measurement. Cattell's personality research was based on his strong background in physical sciences. He had the goal of discovering the basic elements of personality and investigating universal aspects of personality, for which he wanted to apply scientific methods to the unexplored domain of human personality. It was Cattell's vision for psychology to advance as a science that psychology also needed basic measurement techniques for personality. He believed that the basic dimensions of personality could be discovered and then measured.

Over several decades, Cattell and his colleagues systematically measured the widest possible range of personality dimensions with the belief that all aspects of human personality which are or have been of importance, interest, or utility, have already become recorded in the substance of language (Cattell, 1943). They studied the personality traits in diverse populations using three different methodologies as follows:

1) **L-data:** observation of natural, in-situ life behaviour (E.g., academic grades, number of traffic accidents, or social contacts)

2) **Q-data:** questionnaire from the self-report domain, and

3) **T-data:** objective behaviour measure in standardized, experimental settings (E.g., number of original solutions to problem presented, responses to frustrations).

The 16PF Global Scales and Other Five-Factor Models:

The 16PF has included the broad, second-order dimensions, discussed previously, for over 50 years and they are currently called 'the Big Five'. The 16PF scales and items also played an important role in the development of the other Big Five factor models (Costa and McCrae, 1976, 1985; Norman, 1963; McKenzie et al., 1997; Tupes and Christal, 1961). Since the release of the fourth edition of the 16PF around 1970, all five traits have been clearly identified and scorable from the questionnaire. A range of studies was conducted to compare the five 16PF global factors and the set of NEO Big Five factors. The results of these studies show a striking resemblance between 16PF factors and NEO Big Five factors (Carnivez & Allen, 2005; H.E.P.Cattell, 1996; Conn & Rieke, 1994; Gerbing & Tuley, 1991; Schneewind & Graf, 1998).

The average correlation between the 16PF global factors and their respective NEO five factors is just as high as those between the NEO five factors and the Big Five markers, which the NEO was developed to measure (H.E.P. Cattell, 1996; Goldberg, 1992). However, important differences between the two models also have been found. Also, the particular set of traits of the five global factors has been found to be problematic. The biggest difference between the two approaches is the method of development of the primary-level traits. The first-order primary trait definitions in the 16PF Questionnaire are based on the scientific research carried out for decades, and have been confirmed in a wide range of independent studies. In contrast, the primary-level personality facets of the NEO-PI were decided by consensus among a small group of psychologists (who selected what they felt should appear in each NEO domain).

The extensive research resulted in the 16 unitary traits of the 16PF Questionnaire. They are presented below (Table 4.1):

Table 4.1 16PF Scale Names and Descriptors

DESCRIPTORS OF LOW RANGE	PRIMARY SCALES	DESCRIPTORS OF HIGH RANGE
Reserved, Impersonal, Distant	Warmth (A)	Warm-hearted, Caring, Attentive to Others
Concrete, Lower Mental Capacity	Reasoning (B)	Abstract, Bright, Fast-Learner
Reactive, Affected by Feelings	Emotional Stability (C)	Emotionally Stable, Adaptive, Mature
Deferential, Cooperative, Avoids Conflict	Dominance (E)	Dominant, Forceful, Assertive
Serious, Restrained,	Liveliness (F)	Enthusiastic, Animated,

Careful		Spontaneous
Expedient, Nonconforming	Rule-Consciousness (G)	Rule-Conscious, Dutiful
Shy, Timid, Threat-Sensitive	Social Boldness (H)	Socially Bold, Venturesome, Thick-Skinned
Tough, Objective, Unsentimental	Sensitivity (I)	Sensitive, Aesthetic, Tender-minded
Trusting, Unsuspecting, Accepting	Vigilance (L)	Vigilant, Suspicious, Skeptical, Wary
Practical, Grounded, Down-To-Earth	Abstractedness (M)	Abstracted, Imaginative, Idea-Oriented
Forthright, Genuine, Artless	Privateness (N)	Private, Discrete, Non-Disclosing
Self-Assured, Unworried, Artless	Apprehension (O)	Apprehensive, Self-Doubting, Worried
Traditional, Attached To Familiar	Openness to Change (Q1)	Open to change, Experimenting
Group-Oriented, Affiliative	Self-Reliance (Q2)	Self-Reliant, Solitary, Individualistic
Tolerates Disorder, Unexacting, Flexible	Perfectionism (Q3)	Perfectionistic, Organized, Self-Disciplined
Relaxed, Placid, Patient	Tension (Q4)	Tense, High Energy, Driven
GLOBAL SCALES		
Introverted, Socially Inhibited	Extraversion	Extraverted, Socially Participating
Low Anxiety, Unperturbable	Anxiety Neuroticism	High anxiety, Perturbable
Receptive, Open-Minded, Intuitive	Tough-Mindedness	Tough-Minded, Resolute, Unempathic
Accommodating, Agrreable, Selfless	Independence	Independent, Persuasive, Willful
Unrestrained, Follows Urges	Self-Control	Self-Controlled, Inhibits Urges

{**Source:** Cattell, H.E.P. and Mead, A. D. (). Sixteen Personality Factor Questionnaire (16 PF). In The Sage Handbook of Personality Theory and Assessment (pp. 135-159), The Sage Publication.}

The first publication of the 16PF Questionnaire took place in 1949. Since then there have been four major revisions of this test. The most recent release is the 16PF – Fifth Edition Scale (Cattell et al., 1993).

Uses and Applications:

- The 16PF is used in a wide range of settings, including industrial/organizational, counselling and clinical, basic research, educational, and medical settings, because of its strong scientific background.
- This instrument efficiently provides comprehensive, objective information. This ability of the instrument makes it a powerful tool, particularly for industrial/organization applications, such as employee selection, promotion, development, coaching, or outplacement counselling.
- It is also widely used in career counselling settings.
- Despite being a measure of normal-range personality, this questionnaire can be used in counselling/clinical settings to provide an in-depth, integrated picture of the whole person.
- The 16PF dimensions have proven useful in efficiently developing a comprehensive picture of the whole person (including strengths and weaknesses), facilitating rapport and empathy, helping clients develop greater self-awareness, identifying relevant adjustment issues, choosing appropriate therapeutic strategies, and planning developmental goals (H.B. & H.E.P. Cattell, 1997; Karson et al., 1997).

Description of 16PF – Fifth Edition:

The 16PF – Fifth Edition Scale, as the most recent version, aimed at developing updated, refined item content and collecting a large, new norm sample as its main goal. The item pool of this version included the best items from all five previous forms of the 16PF. Also, new items were written by the test authors and 16PF experts. The items were refined in a four-stage, interactive process using large samples. Thus, this resulting instrument has the following features:

1) Shorter, simpler items with updated language,

2) A more standardized answer format,

3) Reviewed for gender, cultural, and ethnic bias and ADA (Americans with Disabilities Act) compliance,

4) Improved psychometric characteristics,

5) Easier hand scoring, and

6) Standardization with over 10,000 people

7) Recent translations are culturally adapted, with local norms

8) Information on reliability and validity is available in individual manuals.

The 16PF – Fifth Edition Scale contains 185 items with a three-point answer format. These items comprise the 16 primary personality factor scales, along with an Impression Management (IM) index. This index aims at assessing social desirability. Each factor scale contains 10 to 15 items. These factor scales remain denoted by the letters as assigned by Cattell (e.g., "Factor A"). However, they also are designated by more descriptive labels (e.g., "warmth").

The content of its items is non-threatening and items ask about daily behaviour, interests, and opinions. The questionnaire is meant for use with people 16 years and older age and is written in a way providing ease at a fifth-grade reading level. The short ability scale items (Factor B) are grouped at the end of the questionnaire with separate instructions. Like other versions of the 16PF, his particular version provides scores on the 16 primary scales, 5 global scales, and 3 response bias scales.

The following are the parallel versions of the 16PF Questionnaire for younger age ranges, in which 16PF traits are also measured:

- 16PF Adolescent Personality Questionnaire measures the 16PF traits in 12-18 olds (Schuerger, 2001).
- 16PF Select is a shorter (20-minute) version that consists of a subset of somewhat-shortened scales, which was developed for use in employee selection settings (Cattell, R.B. et al., 1999).
- 16PF Express provides a very short, 15-minute measure of all the traits (with four or five items per factor).

Apart from this, PsychEval Personality Questionnaire (PEPQ; Cattell, R. B. et al., 2003), which is a comprehensive instrument that includes both normal and abnormal personality dimensions, also consists of the 16PF traits.

The improvements that have taken place in the 16PF Fifth Edition, are as follows:

- Item content has been revised to reflect modern language usage and to remove ambiguity. Also, it has been reviewed for gender, race, and cultural bias.
- Response choices are consistently organized for all personality items, with the middle response choice, which is always a question mark (?).
- Normative data have been updated, which reflects the 2000 U.S. Census. Combined-gender norms are available as a scoring option in accordance with federal civil rights legislation.

- New administrative indices have been designed to assess response bias. An Impression Management (IM) index is comprised of items that are not found on the 16 primary personality factor scales, and it replaces the “Faking Good” and “Faking Bad” scales of the fourth edition. The current fifth edition also contains indices of Acquiescence (ACQ) and Infrequency (INF). Personality scores are no longer adjusted on the basis of validity indices.

- Psychometric properties also have been improved, which we will discuss in Section 4.6. Also, familiar criterion scores, such as Adjustment and Creativity have been updated, and new ones, such as Empathy and Self-Esteem have been added in this edition.

4.1.1 Purpose:

To find out the details of personality aspects in terms of sixteen personality factors.

4.2 STANDARDIZATION OF NORMS

The standardization of norms for the 16PF Fifth Edition was done on the basis of the U.S. population. The final experimental form of this fifth edition was administered to a large group of people (N = 4,449), when standardization was conducted for the release of the 16PF Fifth Edition, originally. Then a stratified random sampling procedure was used to create the final normative sample of 2500. This sample stratification was based on gender, race, age, and educational variables, with the target number for each variable being derived from 1990 U.S. Census figures (Conn & Rieke, 1994a).

An initial sample of 16PF Fifth Edition protocols (N = 31,244) was taken from IPAT’s Test Services Division in order to create the updated norms released in 2002. These protocols were received between 1999 and 2001. During the process of standardization, all protocols with demographic questions regarding race, Hispanic origin, age, education level, occupation, and location within the U.S. was eliminated. Thus, after this elimination procedure, the initial sample was reduced in size (N = 16,133).

Sample stratification was based on gender, race, age, and educational variables, with the target number for each variable being derived from the 2000 U.S. Census figures. This stratification resulted in a total sample size of 10,261 individuals, and this sample is the basis for the updated norms released in 2002.

The size of the norm sample is 10,261, which consisted of 5,124 males (49.9%) and 5,137 females (50.1%). The ages range from 16 to 82, with a mean age of 32.7 years. Considering education, this sample also ranges from “less than ninth grade” to “having a doctorate” and the majority had at least some college (75.3%). Based on race, the sample included Caucasian, African American, Asian American, American Indian, Multiracial, and Other Races. Across these racial groups, a small percentage (8.6%) identified themselves as being of Hispanic origin. This sample included the residents of the Northeastern states, Southeastern states, South Central States, and Western states.

Apart from this, for this fifth edition of the 16PF, the development of the global factors involved submitting the final primary scales to principle component factor analysis on the basis of the same national sample of 3,498 used in the primary scale development, followed by a Harris-Kaiser oblique rotation and three hand rotations (Cattell, H.E.P., 1994).

4.3 RELIABILITY, INTERNAL CONSISTENCY, AND VALIDITY

Improved psychometric properties of the 16PF Fifth Edition include the consistency of the 16PF results over time as evidenced by test-retest correlations, and the internal consistency or homogeneity, of the test items measured by Cronbach's coefficient alpha.

Reliability:

The stability of the different traits measured by the 16PF over time is evident through test-retest coefficients. When the Pearson Product-Moment Correlations were calculated for two-week test-retest intervals, reliability coefficients for the primary factors ranged from .69 (Factor B – Reasoning) to .86 (Factor Q2 – Self-Reliance), with a mean of .80. Test-retest coefficients for the global factors were higher, ranging from .84 to .91, with a mean of .87.

On the other hand, for the two-month interval, reliability coefficients for the primary factors ranged from .56 (Factor L – Vigilance) to .79 (Factor H – Social Boldness), with a mean of .70. Test-retest coefficients for the global factors ranged from .70 to .82, with a mean of .78.

Internal Consistency:

Internal consistency of the 16PF Fifth Edition can be viewed as reliability estimated from a single test administration. Measurement of the internal reliability of a test shows that all items on a given scale assess the same construct. Based on the norm sample of 10,261 adults, Cronbach alpha coefficients calculated for the 16PF Fifth Edition ranged from .68 (Factor E – Dominance; Factor Q1 – Openness to Change) to .87 (Factor H – Social Boldness), with an average of .76.

Validity:

Construct validity of the 16PF Fifth Edition demonstrates that the test measures 16 distinct personality traits. The results of the factor-analytic methods provide evidence about the construct validity of the fifth edition of the 16PF as a test. The primary factor pattern shows that with a few exceptions, items of the 16PF Fifth Edition from a given primary factor scale load on their particular factor scale, but not on other factor scales. Also, the primary factor scales show a predictable pattern of intercorrelations, because the factors are oblique. The 16PF Fifth Edition is found to be correlated with some other personality tests, such as Personality Research Form (PRF; Jackson), the California Psychological Inventory (CPI; Gough), the NEO PI-R (Costa & McCrae), and the Myers-Briggs Type Indicator (MBTI; Briggs & Myers), based on some personality factors

Criterion validity reflects the ability of the test to predict various criterion scores, such as Self-Esteem and Creative Potential. There are different behavioural criteria predicted from the 16PF Fifth Edition and present correlations of the global and primary factors with the scales of instruments that measure self-esteem, adjustment, social skill, empathy, creative potential, and leadership potential.

4.4 ADMINISTRATION AND SCORING

(**Note:** Sections 4.4 and 4.5 should be reported in the past tense while writing in the journal as a past activity conducted in the psychology laboratory.)

As discussed previously, the 16PF – Fifth Edition is designed to be administered to adults – aged 16 and older. It has an overall readability estimated at the fifth-grade level. This test is untimed and has simple, straightforward instructions. Hence, its administration requires minimal supervision in either individual or group settings. Administration of the questionnaire can take place in the following two formats:

- Paper-and-pencil format: Administration time is about 35-50 minutes.
- Computer/Internet format: Administration time is about 25-35 minutes.

Here, we will focus on the paper-and-pencil format for administration, which is a commonly used format.

4.4.1 Tools and Material:

- 16PF – Fifth Edition Scale with Answer/Response sheet
- Stationery

4.4.2 Procedure of Administration:

The test administrator ensures all arrangements for administering the test. Then the test taker/participant is taken to the laboratory and is asked to sit comfortably. Initial general instructions are given to the test taker as follows for rapport building before presenting a 16PF – Fifth Edition, along with a few general questions if required:

"Please be comfortable. I will be presenting the test that will help me understand some aspects of your personality. There is no time limit for completing this test. However, please try to complete it as quickly and spontaneously as possible. I'll be giving you instructions for completing the scale once presented".

The test administrator then provides the test taker with the test booklet with simple, clear, and instructions printed for the test takers and the corresponding answer sheet. The administrator reads aloud the instructions to test takers and responds to their necessary questions, if any before they start responding to the test.

"Please open the booklet (the test administrator directs the test taker to the page with instructions). I will read aloud the instructions. Please read the instructions with me in your mind." The test administrator reads the instructions aloud starting with basic instructions.

"Do not make any marks in the test booklet, which is reusable. Do not skip any questions and choose the first response that comes to mind rather than spending too much time on any single question. Before starting the test, please write your name and gender in the space provided on the left-hand side of the answer sheet."

After reading aloud instructions and making sure that the test taker has no questions/doubts, the test taker is allowed to start responding to the test. The test administrator checks that the test taker is marking responses appropriately by darkening the response circles completely with a pencil, if the test is going to be computer-scored.

After the test taker completes his responses and submits the answer sheet, the test administrator reviews the answer sheet to ensure that the demographic details have been filled in by the test taker and that all responses are scorable. If required, the test taker is instructed to erase any extraneous marks, complete missing answers, and ensure a single answer to a single item in the response. After ensuring that the administration part has been completed well, the test taker is asked a few post-task questions. For example, "How was the test?", "Did you find any difficulty while responding to the test?", and similar other questions can be asked to know the participant's reactions about taking and completing the test. The participant is also allowed to ask questions, if any. Then, he or she is allowed to leave and escorted to the door.

4.4.3 Scoring the 16PF – Fifth Edition:

The answer sheet of the 16PF Fifth Edition is compatible with both hand- and computer-scoring options. Before scoring the answer sheet with responses to the test, whether by hand or computer, the completeness of the answers is verified and ensured again for the following aspects:

- Demographic details (including sex-norms section) of the test taker as required as indicated on the answer sheet
- Answers to all 185 items of the test
- Sex-specific norms (any one of the "combined sex" or "Sex-specific" circle is marked).

A set of four scoring keys, a norm table, and an individual record form are the required materials for hand-scoring the 16PF. Hand-scoring of the completed answer sheet is done by following these four steps

Once all responses from the answer sheet are completed, the answer sheet is considered to be ready for scoring in the following steps:

Step 1: Score the test:

By using the appropriate scoring keys, the 16 factors and five global factors on the test were scored and a raw score for each factor is obtained. The procedures that are followed are :

- The left edge of the first scoring key over the answer sheet is aligned in such a way that the stars on the right side of the answer sheet appear through the corresponding holes on the right side of the key.
- Marks visible through the holes in the area labelled “Factor A” is counted as 1 or 2 points as indicated by the number adjacent to each hole. The total of the points is obtained and entered in the space for the Factor A raw score (as indicated by an arrow on the scoring key).
- Scoring the remaining four factors corresponding with the first key is continued, following the same procedure as described above.
- The same procedure is followed to score other personality factors and to obtain raw scores for them by using the appropriate corresponding keys. The IM and Factor B raw scores are obtained by using the fourth answer key. Factor B responses are scored as 0 (incorrect) or 1 (correct).
- Appendix C was referred for hand scoring the response style indices of Infrequency (INF) and Acquiescence (ACQ), for which there is no scoring key.

Step 2: Convert raw scores to the Sten scores:

Here, in this step, raw scores are converted into standardized (sten) scores by using the norm table included with the set of hand-scoring keys. Stens are based on a 10-point scale with a mean of 5.5 and a standard deviation of 2. The raw scores are printed in the body of the table, while their corresponding sten scores are located at the top of each column provided in Appendix B of the test manual. Here are the procedures involved in this step, by using which the raw scores of the test are converted into sten scores:

- Whether combined-sex or sex-specific norms are more appropriate for Factor A was determined first.
- The test taker’s raw scores for Factor A corresponding to the norms selected are located: A row (combined-sex norms), Male row, or Female row.

- The raw score for Factor A is found in the column and the corresponding sten score for Factor A is found at the top of the column.
- The norm table is also used in converting the raw scores for the Impression Management (IM) index. But this raw score of IM is converted into a percentile, not a sten.

Step 3: Calculate Global Factor Sten Scores:

In this step, sten scores for the five global factors of personality (Extraversion, Anxiety, Tough-Mindedness, Independence, and Self-Control) are calculated. Global factor sten scores are calculated by following the instructions at the top of Side 1 of the Individual Record Form. The following procedures are followed to calculate global factor sten scores.

- The test taker's primary factor sten scores are transferred from the answer sheet to the left-hand column labelled "Sten" on the Individual Record form.
- Scoring is begun with Factor A, which is the first row. Test taker's Factor A sten score by the decimal in corresponding black boxes are multiplied. The resultant product is entered into the empty box adjacent to the black box.
- The same procedure is repeated to score each global factor. Only one product in some factor rows is calculated and recorded and two in others. Here, some boxes are clear, while others are shaded.
- After calculating and recording the products for all 16 factors, numbers in each pair of vertical columns (clear and shaded) are added separately. While obtaining the total of the decimals, a given constant appearing in the first empty box at the base of the column pair is included. Then, the sum of the decimals from the shaded column is entered in the shaded box at the base of the column pair.
- After the total of all the columns was obtained, each sum in a shaded box is subtracted from the sum in a clear box. Then, the remaining decimal is entered into the empty box. This decimal represents the sten score to the nearest tenth of a sten for the global factor indicated.

Step 4: Profile Sten Scores:

After completing all the previous four steps (1 to 3), a pictorial representation or profile of the test taker's overall personality pattern is achieved by graphing the sten scores for the five global factors. This profile is referred to for the interpretation of test scores, which is quite helpful. The procedures followed are as follows:

- The test taker's primary and global factor sten scores in the sten column are recorded at the left of the profile sheet. The test taker's decimal sten score for each globa factor is rounded to the nearest whole number. Decimal sten scores are determined by completing the global factor scoring worksheet on side 1 of the the Individual Record Form.

- A dot is marked in the appropriate place corresponding to each rounded global factor sten score and to each personality factor sten score.
- The dots are connected using a series of short straight lines.

After completing the scoring by following the procedure explained above, the scores are interpreted and reported as explained in Section 4.5.

4.5 INTERPRETATION AND REPORTING

Guidelines for interpreting the scores on 16PF Fifth Edition

Interpretation of Scores on Primary Factors:

From the interpretation point of view, the 16PF Questionnaire is bipolar in nature (Table 4.1), that is, both high and low scores on the test have meaning. Hence, generally, professionals should not assume that high scores are “good” and low scores are “bad”. The right-side pole or high-score range of the factor is described as the plus (+) pole, while the left-side pole or low-score range is the minus (-) pole.

Let us consider the example of Factor A:

High scorers on Factor A tend to be warm interpersonally. These high scorers on Factor A are described as Warm (A+). On the other hand, low scorers tend to be more reserved interpersonally, and they are described as Reserved (A-). However, in some situations, being reserved might be quite fitting or useful, while being warm might be more suitable in some situations.

Interpretation of Scores on Global Factors:

Like the primary 16 factors, the global factors are also interpreted on both poles and described accordingly. For example, high scorers on the Extraversion factor tend to be extraverted and socially participating. On the other hand, low scorers on this factor tend to be introverted and socially inhibited.

Interpretation of Sten Scores:

The 16PF uses “standardized ten” (i.e., Sten) score scales ranging from 1-10, with a mean score of 5.5 and a standard deviation of 2. Scores that are far away from the mean, are considered more extremes either in the high or the low direction. The more extreme a score is toward a given factor pole, the more likely that the descriptors for the scale’s pole will apply to that particular score and that the trait will be apparent in the test taker’s behaviour.

Historically, 16PF stens fall into the following three types of range:

- Stens 4-7 fall within the average range
- Stens 1-3 fall in the low range
- Stens 8-10 fall in the high range.

The 16PF Fifth Edition continues with the same ranges, in which a sten score of 4 is described as "low-average" and a sten score of 7 is considered as "high-average". Similar categorizations are used in the profile sheet and interpretive reports in this edition.

The recommended interpretive strategy for the 16PF profile involves evaluating three types of indices in the following sequence indicated:

1) Response style indices: They are evaluated first as a check for atypical test-response styles.
2) Global factor scales: They are examined next, because they provide a broad picture of the person.
3) Primary factor scales: They are evaluated to obtain details of the personality picture.

Thus, both the global dimensions and the primary scales are scored.

Once the scores of the test taker on all 16 primary and 5 global factors are interpreted by following the above guidelines, the results of the test are reported as below (Since the reporting is on the past activities of conducting, scoring and interpreting test, the results should be reported in the past tense as mentioned before right in the beginning of the section and as the sample provided below). Keep in mind that you are supposed to enter the scores and interpretation of your participant in the blank space provided while reporting. While reporting the scores and their interpretation for each factor, Table 4.1 in this unit or Figures 3 and 4 and other important information from the test manual can be referred to.

The 16PF Fifth Edition was conducted on the participant. The following results were obtained after entering all sten scores of the 16 personality factors and 5 global factors in the 16PF test profile. The test manual gives an in-depth interpretation procedure of the obtained scores. However, here we will focus on interpreting and reporting the sten scores at the primary level, not going into much detail.

• **Reporting the results on 16 Primary factors:**

Factor A: The sten score was ___, being low/average/high, which indicates the participant is ____ (Describe what the score implies in the blank space).

Similarly, report the scores of the participant on other remaining 15 primary factors (i.e. **Factor B, Factor C, Factor E, Factor F, Factor G,**

Factor H, Factor I, Factor L, Factor M, Factor N, Factor O, Factor Q1, Factor Q2, Factor Q3, and Factor Q4).

Let us understand this reporting part with an elaborated example considering Factor A again:

The sten score for Factor A was 8, which was high. This indicates that the participant is towards the plus pole (A+), in the high score range. This means that the participant is warm, outgoing, and attentive to others. Thus, in general, the person has more interest in people. He (or she) tends to prefer occupations dealing with people.

• Reporting the results on 5 Global factors:

Factor EX: The sten score was ___, being low/average/high, which indicates the participant is

____

(Describe what the score implies in the blank space).

Similarly, report the scores of the participant on the other remaining 4 global factors (i.e. **Factor AX, Factor TM, Factor IN, and Factor SC**).

Let us understand this reporting part with an elaborated example considering Factor EX again:

The sten score for Factor EX (extraversion) was 7, which is average, but still on the higher side. This also indicates that the participant is towards the plus pole (E+), towards the high score range. This score is consistent with the participant's score on Factor A, which shows that the person is extraverted and fond of people.

Note: While interpreting and reporting the scores on this particular test, always keep in mind that whether the score is high or low, it does not mean that it necessarily reflects the good or bad quality of the person. In simple words, high scores are not necessarily associated with good quality. Similarly, a low or average score is not necessarily associated with bad quality.

All scores have a range of quality which is associated with that particular factor and is normal, whether primary or general factors or global factors. So the interpretation of the scores should be very comprehensive while reporting and even communicating the results of the test with the participant/client. Also, they should be reported and communicated to the test taker/participant in a very sensitive manner.

• Reporting the results regarding Impression Management (IM) factor:

The corresponding percentile for the IM raw score of 5 is 10. This shows the lower social desirability reflected in the participant's responses. It means the low score on IM reflects low social desirability on the participant's side.

4.6 SUMMARY

In this unit, we learned several details of the 16PF Fifth Edition Questionnaire. Right in the beginning, we discussed some important background of this test which included Dr. Raymond Cattell's theory of personality based on which he developed this test, the history and development of the 16PF Questionnaire, comparison between the 16PF global factors and other five-factor models, uses and applications of the 16PF Questionnaire.

We also studied the description of the 16PF Fifth Edition scale in terms of its number of items, translations in multiple languages, its important features along with the improvements that took place in the most recent fifth versions of the scale. We discussed the stanadardization of the test norms, and the psychometric properties of the test (i.e., reliability, validity, internal consistency). Then finally, we focused on its administration and scoring, interpretation and reporting along with the precautions that should be taken while conducting and after conducting the test on the test taker.

4.7 QUESTIONS

1. Explain Cattell's theory of personality.
2. Describe the recent version of the 16PF Questionnaire.
3. Explain the administration and scoring procedure of the 16PF Fifth Edition in brief.
4. Write short notes on:

a) Uses and applications of the 16PF Questionnaire

b) History and Development of the 16PF Questionnaire

c) 16PF global scales and other five-factor models

d) Guidelines on interpreting the scores on the 16PF Fifth Edition Questionnaire

4.8 REFERENCES

- Russell, M. & Karol, D. (2002). 16PF – Fifth Edition with Updated Norms: Administration Manual. Illinois: Institute for Personality and Ability Testing, Inc.
- Cattell, H.E.P. and Mead, A. D. (). Sixteen Personality Factor Questionnaire (16 PF). In The Sage Handbook of Personality Theory and Assessment (pp. 135-159), The Sage Publication. Retrieved from https://people.wku.edu/richard.miller/520%2016PF%20Cattell%20and%20Mead.pdf

******