Table of Contents

Lesson 01

Psychological Assessment and Tests

Psychological Tests: Why Do We Need Them?

"Psychology is the scientific study of behavior and mental processes …. Human or Animal" (Feldman)

There are three important terms used in this definition…..scientific study, behavior, and mental processes. Behavior and mental processes constitute the content or subject matter of psychology, whereas scientific study refers to the methodology used by psychologists. Scientific method ensures that the results or conclusions of an investigation are objective and consistent. For this purpose psychologists use carefully designed tools of data collection. Psychological tests are one of those tools.

In some of our lectures in the foundation course, we had discussed that the main and important goals of psychology, or in other words of understanding human behavior and mental processes, are;

1. To understand the nature and mechanisms of behavior and mental processes
2. To develop an understanding of the relationship between behavior and mental processes
3. To apply this understanding to real life situations and, on the basis of this understanding, predict for the future
4. To employ the scientific approach for developing this understanding

We also studied that the main goals of psychology can be categorized as:

- Observation,
- Description,
- Understanding,
- Explanation,
- Prediction, and
- Control of human behavior and mental processes

Psychological tests help and assist psychologists in attaining all of these goals.

After doing a degree course in psychology one may join a variety of work settings, the most common being:

- Education/teaching
- Research
- Hospitals/clinics
- Recruiting/screening agencies
- Specialized professional settings e.g. armed forces, social welfare etc.

In all of the above mentioned professional settings, some form of testing and assessment is used; for measuring achievement, for data collection, for assessment of personality, intellect, or pathology, for selecting the most suitable candidates for a position, or short listing students for admissions on the basis of aptitude.

Course Description: Psychological Testing and Measurement:

The course will encompass basic concepts of psychological measurement. The main focus of the course will be on introducing essential terminology, theories, concepts, types of psychological tests, measurement procedures, socio-cultural variables affecting measurement, and modern trends.

Practical application of knowledge, besides developing a knowhow of theoretical constructs, will be encouraged.

Course Objectives:

As suggested by HEC, the objectives of this course will be:

- To introduce the students to the basic theoretical psychometric concepts and use of psychological tests.
- An understanding of the basic principles of psychological measurement and the techniques of test construction, administration, and validation. Test rationale, construction, characteristics and the use of evaluation is emphasized.
- To understand statistical concepts, including scales of measurement, used in psychological measurement.
- To understand reliability, validity, and

- To understand social and cultural factors related to the assessment and evaluation of individuals, groups, and specific populations.
- An understanding of the techniques of test construction, administration, and validation. Test rationale, construction, characteristics and the use of evaluation are emphasized.
- To understand social and cultural factors related to the assessment and evaluation of individuals, groups, and specific populations.

Difference between Testing and Assessment

What is a Test?

"A test is a measurement device or technique used to quantify behavior or aid in the understanding and prediction of behavior" (Kaplan, & Saccuzzo, 2001)

Assessment includes more than just tests. A typical assessment plan may include a test or a battery of tests, interview, behavioral observation, and case history data. In many cases even more sources of information regarding a person's behavior and mental processes are also used e.g. portfolios containing samples of a person's (mostly a student or a job candidate) skill or work such as photographs, drawings, stories, or essays.

Remember!!!

- A test is just one tool. In order to make more precise and accurate predictions one should supplement tests with other sources of data as well, e.g., observation, case history etc.
- Psychological tests do not, and they cannot, always present a 100% accurate picture of behavior and/or mental processes. There can always be some chance of error; and we should be able to gauge the amount of error.
- Tests present a picture of only those characteristics or variables that have been covered in the content of the test. They do not capture behavior in entirety.
- The precision of test results or conclusion depends, to a great extent, on the subjects' state of mind and attitude toward the test as well as the testing process.

Types of Tests:

There are a large number of varieties of tests available for a wide range of purposes. Tests can be categorized on the basis of:

- The purpose or the type of behavior/characteristics to be measured: personality, aptitude, intelligence, achievement etc.
- The administration procedure: individual versus group tests
- Speed versus ability tests
- Aptitude tests, achievement tests, or intelligence tests
- Ability versus personality tests
- Structured/objective tests versus projective tests
- Original versus translated and adapted tests
- Translated tests

Essential Characteristics of Psychological Tests:

A good psychological test should have these qualities:

- **Validity:** A test should measure what it is intended to measure.
- **Reliability:** A test should give consistent results. It should give same or similar results every time it is administered to the same subjects in same conditions.
- Norm development and standardization

Ethics in Psychological Testing:

Just like in psychological research and psychotherapy, we have to keep in mind the ethical standards while using psychological tests for assessing people's behavior, personality, or other characteristics. Confidentiality, respect for the client/subject's privacy, and use of test results only according to their will and ethicality.

The testing processes, interpretation, or reporting should in no way harm the subject.

Some Sources of Information on Tests:

- **See APA (American Psychological Association) Divisions by Number and Name:** Division 5, Evaluation, Measurement, and Statistics; http//www.apa.org/about/division/div5.html
- **See APA Divisions by Topic: Division** 5, Evaluation, Measurement, and Statistics, and division 14, Society for Industrial and Organizational Psychology http//www.apa.org/about/division/div14.html
- Test manuals and catalogues
- Internet sites

Lesson 02

Historical Background of Psychological Testing

Like most modern disciplines, major developments in psychological testing took place in the west, mostly in the U.S. However, if we try to trace the roots of psychological testing we will end up in the orient rather than the west. Researchers and historians agree that China was perhaps the first country to develop and use tests in the formal sense of mental measurement.

Man has always been interested in knowing and understanding other human beings. The 'how' and 'why' of human psyche and behavior have always fascinated man. Mental measurement, formal or informal, has been one of the tools used for this purpose. And mental measurement is what psychological tests do. As students of psychology we know that the Greek philosophers Plato and Aristotle proposed their ideas about individual differences some centuries before Christ (300-400 BC). Other Greek philosophers talked about temperaments and humors. However we should also know that the Chinese had developed a system of mental measurement even in 2200 years BC.

Even more than 4000 years ago the **Chinese** had developed a civil service testing program (DuBois, 1970, 1972). Under the Chinese emperors, oral examinations were arranged every third year in order to determine work evaluations and promotion decisions.

Later rulers are known to have used test batteries i.e., using a number of measures. This practice was quite common in the **Han Dynasty** (206 B.CE. to 220 C.E). These early tests pertained a variety of topics e.g. revenue, civil law, military affairs, agriculture, and geography.

The civil service evaluation system and the test became broader based in the Chan dynasty, beginning in 1115 B.C.E. It covered the evaluation of proficiency in such divergent areas as archery, horsemanship, music, writing, arithmetic, civil law, agriculture, revenue, military affairs, geography, and skill in the rites of public and social life.

By the **Ming Dynasty** (1368-1644 C.E.) very well developed tests were being used. The evaluation process involved a national multistage testing program. During this period, tests were held at local as well as regional testing centers. These centers had special testing booths. The people being tested would go for more extensive essay examinations to provincial capitals if they were successful in the examination at local level. This second testing was followed by another round. The candidates who had the highest test scores in the second round went on to the nation's capital for a final round. Only those who passed this third set of tests were considered eligible for public office.

The civil service examination system prevailed till 1905.

The Chinese pattern was followed and copied by other nations, and it was adopted soon by the Western world. British missionaries and diplomats wrote reports about this system. In **1832** the **English East India Company** copied the Chinese model. They started using this system as a method of selecting employees for overseas duty.

When the success of this system under the British rule became well known, the German and the French governments also adopted it.

The U.S. government also introduced a similar mechanism. In **1883,** the American Civil Service Commission was established. This commission developed and administered competitive examinations for certain government jobs. This was a time when the testing movement became significant and grew rapidly in the Western world (J.S. Wiggins, 1973).

By the **sixteenth century, greater awareness of European society** had become more advanced and capitalistic. Individuality of people was being recognized more than before. However, the major developments were observed in the **Renaissance.** Rebirth of individualism took place in this time.

By the **early nineteenth century** most human knowledge was gathered through human observation. Physical phenomena were observed and recorded in such a way that the quality and accuracy of info depended on the perceptual abilities of the observers/ trained researchers/ data collectors. In order to make their investigations and findings more objective and precise, the physical scientists were concentrating on the development of instruments that were more precise and errorless, and that could be used by the observers. But the most significant shift took place after **Charles Darwin** wrote On the Origins of Species in 1859. By this time psychology was also taking the form of a scientific discipline. The awareness and study of individual differences is one of the most popular areas of research and writing. Darwin had proposed that humans had descended from the ape. This, he believed, was a result of chance variation.

He wrote that chance variation in species would be selected or rejected by nature on the basis of adaptability and survival value. He also proposed concept of 'survival of the fittest', that only those species would survive who are capable of adapting to the natural conditions and who were strong enough to bear the atrocities of these conditions.

Darwin is considered to be the one who generated the interest of scientists in the study of individual differences. He wrote: "The many slight differences which appear in the offspring from the same parents…… may be called individual differences…….. These individual differences are of the highest importance…….[for they] afford materials for natural selection to act on" (Darwin, 1859 p. 125)

Round about the same time, **Gustav Fechner, Wilhelm Wundt, Hermann Ebbinghaus, and other German** experimental psychologists had been conducting studies and experimenting. They had shown that psychological phenomena could be expressed in quantitative, rational terms. Psychologists in the **United States** had also been studying and reporting on individual differences and developments that paved way to the emergence of psychological tests. As a result of increased attention given to written examinations in the U.S. school system, the American experts had started developing **standardized measures of scholastic achievement**.
Developments on similar grounds were taking place in Psychology and psychiatry in France. **French psychiatrists and psychologists** were studying and writing on mental disorders. This research influenced the development of clinical assessment techniques and tests.
The most significant names that initiated and contributed to the study of individual differences and test development in the 19th century included Sir **Francis Galton, James Mc Keen Cattell, and Alfred Binet.**

Galton was a cousin of Charles Darwin. He was born in the family of geniuses and he himself was a genius having an IQ of more than 200. He was a geographer, meteorologist, tropical explorer, "founder of differential psychology", inventor of fingerprint identification, pioneer of statistical correlation and regression, convinced of hereditarianism, eugenics, proto-genetics and a best-selling author.
He gave the concept of "hereditary genius". According to Francis Galton ("Hereditary Genius, 1869) "gifted individuals" tended to come from families, which had other, gifted individuals. He went on to analyze biographical dictionaries and encyclopedias, and became convinced that talent in science, the professions, and the arts, ran in families. His was the first systematic attempt to measure intelligence by investigating the role of heredity and its impact on intellectual abilities. He further attempted to measure human trait quantitatively in order to determine the distribution of heredity in it. For this he used **"word association test", and "mental imagery".**

Galton argued that it would be "quite practicable to produce a highly gifted race of men by judicious marriages during several consecutive generations". Eugenics, he said, was the study of the agencies under social control that may improve or repair the racial qualities of future generations, either physically or mentally.
For Galton "What Nature does blindly, slowly, and ruthlessly, man may do providently, quickly, and kindly". He also said that "Intelligence must be bred, not trained".
Such arguments appealed many and some people took this approach to extremes; this way of thinking had drastic social consequences and was used to support apartheid policies, sterilization programs, and other acts of withholding basic human rights from minority groups.
Galton was interested in the hereditary basis of intelligence and in techniques for measuring abilities. A particular concern of Galton was the inheritance of genius, but he also constructed a number of sensorimotor tests and devised several methods for investigating individual differences in abilities and temperament. Using these simple tests, he collected measurements of over 9000 people ranging in age from 5 to 80 years. Among the many methodological contributions made by Galton was the technique of "co-relations", which has continued to be a popular method for analyzing test scores.
James McKeen Cattell is an American psychologist who gave more importance to the mental processes. He was the first ever to use the term "mental test" for devices used to measure intelligence. He developed tasks that were aimed to measure reaction time, word association test, keenness of vision and weight discrimination. These tests were proved to be a failure as they were not comprehensive and complex enough to measure intelligence. James Cattell joined Galton's in his methods and tests. He tried relating scores on these mental tests of reaction time and sensory discrimination to school marks. It remained for Frenchman, Alfred Binet to construct the first mental test that proved to be an effective predictor of scholastic achievement.

Alfred Binet: The first formal measure of intelligence was developed by French psychologist Alfred Binet and **Theodore Simon**, in 1905 in France. The test or the scale was developed in order to assist the education ministry and department in identifying "dull" students in the Paris school system, so that they could be provided remedial aid.
The main idea was that intelligence can be measured in terms of performance of a child. Using the same concept Binet developed the first intelligence test in 1905. The test could identify more intelligent children within a particular age group. It could differentiate intelligent children from the less intelligent ones.
The test was devised for locating the 'dullest' students in the Paris school system so that remedial assistance could be provided to them before they were denied instruction.
The testing procedure adopted by Binet and Simon:

- Initially Binet developed a number of tasks.
- Then he took groups of students who were categorized or labeled as 'dull' or 'bright' by their teachers.
- The tasks were presented to them. The tasks that could be completed by the 'bright' students were retained; the rest were discarded.
- The idea was to retain tasks that could be completed by the bright students, as these were considered to be indicative of the child's intelligence.
- With further work, dull or bright children could be identified with reference to their age.
- The scale could, thus, identify bright or dull students within particular age groups.

The original Binet- Simon scale was revised a number of times. The American psychologist, **Lewis Terman** gave the first Stanford revision of the scale in 1916. This revision comparison American standards is from age 3 to adulthood. Further revisions were made in 1937 and 1960. Stanford- Binet is one of the most widely used tests even today.

Binet and Simon constructed and individually administered test consisting of 30 problems arranged in order of ascending difficult. The problems on this first workable intelligence test, which was published in 1905, emphasized the ability to judge, understand and reason. A revision of the test containing a large number of subtests grouped by age levels from 3 to 13 years was published in 1908. In scoring the 1908 revision of the Binet-Simon Intelligence Scale, the concept of mental age was introduced as a way of quantifying an examinee's overall performance on the test.

Among other pioneers in psychological testing and assessment was C**harles Spearman** in test theory, **Edward L. Thorndike** in achievement testing, **Lewis Terman** in intelligence testing, **Robert Woodworth and Hermann Rorschach i**n personality testing, and **Edward Strong** in interest measurement. The work of **Arthur Otis** on group-administered tests of intelligence led directly to the construction of the Army Examinations Alpha and Beta by a committee of psychologist during World War I.

Lesson 03

Historical Background of Psychological Testing

We know that the Chinese were the first ones to have a formal system of assessment and evaluation of the civil servants. However other societies also had been using some form or the other of examination and evaluation. The Greeks for example used to use tests in the educational process. It was an important part of the educational system.

When Europe came out of the **dark ages** and became enlightened, they started developing universities. The examination system became an important part of their educational system as well. European universities that were established in the beginning of the middle ages, developed their formal examination system. Degrees and honors were awards awarded on the basis of examination.

Psychological tests, in the form as we see them today, began in the 19th century. As discussed earlier, the most significant names associated with test development in that period were, those of Sir **Francis Galton, James Mc Keen Cattell, and Alfred Binet.**

We remember that Alfred Binet and his associate Theodore Simon developed the first proper test to measure intelligence. He had proposed that children who could not benefit from normal schooling, and did not respond to education the way they should have, needed to be identified, If they were found to be educable, they could be sent to special classes. As a result of Binet and his associates the Ministerial Commission for the Study of Retarded Children was established. These developments ultimately led to the development of the first intelligence test. But much before Binet, many other professionals and thinkers had been working on the diagnosis and treatment of mental illness that developed an increased realization of the need for diagnostic tools.

Different psychologists, psychiatrists, and medical practitioners were dealing with people with psychological problems in different parts of the world. A lot was being done in France. In the earlier centuries, people suffering from such problems were not treated in a humane manner; in fact they were treated inhumanly. In the 19th century there was a greater realization of the need to detect the mentally ill, differentiate them from the 'normal', to assess the extent of the problem, and to treat accordingly. A need was felt to differentiate between the mentally retarded (now called 'special') and the 'insane' (as they used to call the psychotics). Mentally retarded were the people who were born with intellectual deficit. Those who were called insane were the ones who had extreme emotional problems. Some of the French scholars and practitioners made significant contributions to the cause of rehabilitation and treatment of such persons. Contributions of some of these practitioners led to the development of psychological tests in later years.

Esquirol: He was a French physician, published his famous two volume book in 1838. This work proved to be a milestone in the understanding and treatment of mental retardation. He gave a comprehensive account of such conditions and symptoms of the mentally ill that were called 'mental retardation' later on. He said that mental retardation can be seen along a continuum ranging from 'normality' to 'low grade idiocy'. According to Esquirol, mental retardation was found in various degrees. His work contained more than 100 pages on this topic. He felt that language use by the person being diagnosed for mental retardation was the most reliable criterion for assessing intellectual level. He had proposed this notion after having tried other procedures for the same purpose. This idea was found to have weight and most intelligence tests that we come across today have verbal ability as a major part of their content.

Seguin: Another French physician who made similar contributions was Seguin. He was based in France while he made his initial contributions, but then he emigrated to the U.S in 1848. His contribution to the training of the mentally retarded persons provided a basis for the inclusion of the performance based items or performance part of intelligence tests. He was the one to establish the first school for the mentally retarded in 1837. He developed the 'physiological method of training' and tested it for many years. He did not agree with the idea that mental retardation was incurable. He had developed a number of muscle training and sense training techniques. His techniques were adopted by institutions for the mental retardation. One of the techniques developed by Seguin was the 'Seguin Form Board'. It contained a number of variously shaped slots same shaped corresponding blocks. The subject had to insert the blocks into the corresponding slots or holes. This type of items were adopted by psychologists and were made a part of the performance based tests of intelligence or performance part of intelligence tests that included both verbal and performance parts.

As mentioned earlier, Sir Francis Galton was the main and most prominent figure in the testing movement in the 19th century. He included a number of traits or abilities in his tests for measuring intellectual ability. He was primarily interested in the inheritance of genius which he investigated in depth. He also developed some sensorimotor tests. Using these tests, he collected data from over 9000 people within an age range of 5 to 80 years. He established his anthropometric lab at the International exposition of 1884. The lab was later shifted to South Kensington Museum after the exposition closed. It remained there for another six years. Galton was already interested in and studying the hereditary nature of human characteristics. He would study traits/ characteristics in siblings, twins, cousins, and other relatives. He looked into characteristics in related and unrelated people. The tests that he used involved physical traits. He believed that: "The only information that reaches us concerning outward events appears to pass through the avenue of our senses; and the more perceptive the senses are of difference, the larger is the field upon which our judgment and intelligence can act" (Galton, 1883).

The tests that he included the sort of:

- Galton whistle for determining the highest audible pitch
- Galton bar of visual discrimination of length
- Graduated series of weights for measuring kinesthetic discrimination

He introduced other test formats and tests also e.g. rating scales, questionnaire method, the use of free association technique

However, the first person to use the term 'mental tests', in psychology, was James Mc Keen Cattell. He used this term in 1890 in an article. Cattell had written his doctoral dissertation on reaction time in Leipzig, Germany, under the direction of Wilhelm Wundt. In later years while he was lecturing in Cambridge, he got in touch with Sir Francis Galton. This interaction with a popular scholar of his time, whose ideas on individual differences appealed him, strengthened Cattell's interest in the measurement of individual differences. In the article where he had introduced the term '**mental tests',** he had written about a series of tests that were used with college students to measure their intellectual ability. These tests were administered every year. **Cattell and Galton held the** common understanding that intellectual function could be measured by measuring sensory discrimination and reaction time. The tests that he had mentioned involved measurement of speed of movement, muscular strength, sensitivity to pain, keenness of hearing and vision, memory, weight discrimination, reaction time etc.

Some other psychologists had also developed tests that measured similar faculties. However, these tests were not found to be good predictors of intellectual functioning. Their results neither corresponded with teachers' ratings nor academic grades, or with other similar tests. Hence these tests met little popularity and acceptance.

During the same time period, European psychologists like Emil Kraepline and Herman Ebbinghaus also developed similar tests. More detailed discussion of these and other tests will be made in later sections.
The experimental psychologists of the 19th century also contributed to the development of psychological tests, though indirectly. The experimental psychologists were trying to develop objective measures for studying behavior and other phenomena of interest. The earlier experiments in the psychological laboratories were investigating different aspects of thinking and behavior. Leipzig lab of Wundt was a pioneer in this regard. The main focus of these labs was not on individual differences as such. They sought to explore uniform patterns of behavior. However the experiments did generate information regarding individual differences. It was observed that different people yielded different patterns of behavior when tested or studied under same conditions. These experiments also generated the realization that objectivity, success, and accuracy of psychological experiments required the use of carefully designed measuring instruments and the use of controlled conditions, i.e., making sure that all other relevant variables were kept under control while the effect of independent variable was being examined. As a result of these developments, formation of standardized procedures began. The same principles and procedures were adopted in test development and test administration.

20th Century:
Binet's scale, as already said, was the first formal test of intelligence. The first version appearing in 1905 consisted of 30 problems or tests. These were arranged in order of ascending difficulty. Empirical procedures were followed for determining the difficulty level. The scale was given to 50 normal children, 3 to 11 years of age.

Some mentally retarded children and adults were also included. The items covered sensory and perceptual tests along with verbal content. The later carried more importance than the former. Judgment, comprehension, and reasoning were specially emphasized.

In the 1908 scale, many unsatisfactory tests were removed and new ones added. This time the scale containing many subtests was given to 300 normal children aged 3 to 13 years, and on the basis of this try out the test was grouped into age levels. The tests that were passed by most children at a certain scale level were placed in the scale level meant for that age group. For example the tests that were passed by 80-90% of normal 3 year old children were kept in the three year level.

Now the scale could determine the 'mental level' of a child which indicated the age of normal children who could successfully do the tests. Binet preferred to stick to the term 'mental level' whereas many other psychologists used 'mental age' instead. Another revision was made in 1911. Many translations and adaptations were also made. But the most significant was the one by L. M. Terman and associates at Stanford University. This revision is called Stanford- Binet. The American psychologist Terman gave the first Stanford revision of the scale in 1916. Further revisions were made in 1937 and 1960. Stanford- Binet is one of the most widely used tests even today.

The idea of intelligence quotient was first used in this version. Intelligence quotient or IQ was described as a ratio between mental age and chronological age. Mental age of a person can be different from his or her chronological age i.e., it can be above or below that. It could reflect whether or not a child was performing at a level at which his age mates were. BUT it gave rise to a problem. How could we compare people belonging to different age groups? Will a 22 year old with a mental age of 25 be equally intelligent as an 7- year old having a mental age of 10? In order to remove problems like this statistical concepts and procedures were employed that will be discussed in the section on intelligence testing.

Group Testing:

A new concept in psychological testing emerged prominently in the early 20th century. This was about group testing. All previously existing tests were supposed to be individually administered. Their nature was such that the subjects had to be attended to one by one. In some tests or items the response time was also to be measured. In that sense these tests were quite time consuming. In some cases the oral responses were also to be recorded, or performance materials were to be individually administered. The tests required not only one to one administration but highly skilled and trained examiners as well. These tests therefore could not be used for group administration. A need was felt for tests that could be administered to groups of people together so that readings were quick and available for large numbers of people.

This became significantly important in the time of World War in 1917 when the U.S became a party involved in the war. American Psychological Association (APA) appointed a committee to examine how psychology could be of help. The government and the army needed the psychological services in many ways. First of all they needed to classify the recruits according to their intellectual level. There were about 1.5 million recruits whose screening and short listing required testing in large batches. On the basis of intellectual ability testing they could be selected, retained, or discharged. This could also help in decisions regarding allocation of specific duties, type of training and assignments.

Army psychologist were gathering tests and test items from all sourced. Arthur. S. Otis had prepared an unpublished group intelligence test in which he had used an objective test pattern and included multiple choice questions. He gave that test to the army. Looking into all the available tests and items, including Otis' test, the army psychologists developed two tests called Army Alpha and Army Beta. These were the first group intelligence tests. After this there was no limit to test development. Tests for all purposes an varying nature were developed, and today it is impossible to count the number of psychological tests available for use.

Psychological tests were translated and adapted in other countries in their native language. They developed their own indigenous tests also.

The following table describes some of the significant milestones in the history of psychological testing:

Significant Milestones in the History of Psychological Testing:

2200 B.C	Chinese emperors set up civil-service testing program in China.
A.D 1219	First formal oral examinations in law held at University of Bologna.
1575	J.Huarte publishes book, *Examinen de Ingenios*, concerned with individual differences in mental abilities.
1636	Oral examinations for degree certification used at Oxford University.

1795	Astronomer Maskelyne of Greenwich Observatory fires assistant Kinnebrook when their observations of the transit time of Venus disagree.
1845	Printed examinations first used by Boston School Committee under guidance of the educator Horace Mann.
1864	George Fisher, an English schoolmaster, constructs a series of scales consisting of sample questions and answers as guides for evaluating students' answers to essay test questions.
1865	Establishment of the New York State Regents' Examination.
1869	Scientific study of individual differences begins with publication of Francis Galton's *Classification of Men According to Their Natural Gifts.*
1879	Founding of first psychological laboratory in the world by Wilhelm Wundt at University of Leipzig in Germany.
1884	Francis Galton opens Anthropometric Laboratory fro international Health Exhibition in London.
1887	Gustav Fechner formulates first psychological law.
1888	J.M Cattell opens testing laboratory at the University of Pennsylvania.
1893	Joseph Jastrow displays sensorimotor tests at Columbian Exhibition in Chicago.
1896	Emil Kraeplin proposes new classification of mental disorders. Hermann Ebbinghaus develops first completion test.
1897	J.M Rice publishes research findings on spelling abilities of U.S school children.
1900	College Entrance Examination Board founded.
1904	Charles Spearman describes two-factor theory of mental abilities. First major textbook on education measurement, E.L Thorndike's *introduction to the Theory of Mental and Social Measurement, published.*
1905	First Binet-Simon intelligence Scale published. Carl G.Jung uses word-association test for analysis of mental complexes.
1908	Revision of Binet-Simon Intelligence scale published.
1908-1909	Obejective arithmetic tests published by J.C Stone and S.A Courtis.
1908-1914	E.L Thorndike develops standardized tests of arithmetic, handwriting, language, and spelling, including *Scale for handwriting of Children* (1910).
1914	Arthur Otis develops first group test of intelligence, based on Terman's Standford Revision of the Binet-Simon Scales.
1916	Standford-Binet Intelligence Scale published by Lewis Terman.
1917	Army Alpha and Army Beta, first group intelligence tests, constructed and administered to U.S Army recruits.
1919	Louis Thustone's Psychological Examination for College Freshmen published.
1920	National Intelligence Scale published. Hermann Rorschach;s Inkblot Test first published.
1923	Standford Achievement Test first published. Pintner- Cunningham Primary Mental Test first published.
1924	Truman Kelly's *Statistical Method* published.
1925	Arthur Otis's *Statistical Method in Educational Measurment* published.
1926	Scholastic Aptitude Test first administered.
1927	First edition of Strong Vocational Interest Blank for Men published. Kuhlmann-Anderson Intelligence Tests first published.
1935	Development of the first IBM test-scoring machine.
1936	First volume of *Psychometrika* published.
1937	Revision of Standford-Binet Intelligence Scale published.
1938	Henry Murray publishes Explorations in Personality. O.K. Buros publishes first Mental Measurements Yearbook.
1939	Wechsler-Bellevue Intelligence Scale published.
1942	Minnesota Multiphasic Personality Inventory published.
1947	Educational Testing Service founded.
1949	Wechsler Intelligence Scale for Children published.
1960	Form L-M of Standford-Binet Intelligence Scale published.

1969	Arthur Jensen's paper on racial inheritance of IQ published in *Harvard Educational Review.*
1970	Increasing use of computer in designing, administering, scoring, analyzing, and interpreting tests.
1971	Federal court decision requiring tests used for personnel selection purposes to be job relevant. (*Griggs v. Duke power).*
1974	Wechsler Intelligence Scale for Children-Revised published.
1975	Growth of behavioral assessment techniques.
1980	Development of item response theory.
1981	Wechsler Adult Intelligence Scale-Revised published.
1985	Standards for *Educational and Psychological Testing* published.
1987	California Psychological Inventory-Revised published. DSM-III-R published.
1989	MMPI-II published.
1990	Wechsler Intelligence Scale for Children-III published.
1992	Eleventh edition of *The Mental Measurement Yearbook* published.

Lesson 04

Types of Tests and Their Significance

Cohen and Swerdlik (1999) have enlisted a number of assumptions that are basic to psychological testing and assessment:

1. **Psychological traits and states exist:**
   Traits refer to characteristics or psychological features that differentiate one person from another. According to Guilford (1959, p. 6) a trait is "any distinguishable relatively enduring way in which one individual varies from another".
2. **Psychological traits and states can be quantified and measured:**
   Quantities of course refer to numbers and quantification means that when we use psychological test we gather and present the findings in terms of numbers. Quantification makes the whole process objective. Also, quantities are comparable. If quantitative data regarding a particular trait are available then one can compare the scores of various people and assess as to who is higher and who is lower; who is more intelligent and who is less intelligent; who has depressive tendencies and who does not.
   The second concept in this assumption is that of measurement; it can be defined as "the act of assigning numbers or symbols to characteristics of subjects (people, events, whatever) according to rules" (Cohen & Swerdlik, 1999, p. 19).
   For the sake of quantitative measurement, we use scales. A scale, according to the same author, is "a set of numbers (or other symbols) whose properties model empirical properties of the objects or traits to which numbers are assigned.
3. **Various approaches to measuring aspects of the same thing can be useful:**
   The same characteristic, property, trait, aptitude, ability, interest, state, construct, or anything that one is interested in, can be measured in many ways. Different tools, different types of items, different scales can be measuring the same phenomenon in many ways. One can see that a very wide variety of tests and measures are available to us. There is no 'one and only' single tool for measuring any one aspect of behavior or thinking process.
4. **Assessment can provide answers to some of life's most momentous questions:** one fall out of this assumption is that test/assessment/tools/contents and procedures will be designed, produced and refined all the time.
5. **Assessment can pin point phenomena that require further attention or study:** This implies that tests can be used for diagnostic purposes as well. Diagnosis may be done for therapeutic purpose, for forensic reason, for placement and task allocation, for educational counseling or any other purpose.
6. **Various source of data enrich and are part of the assessment process:** Assessment involves more than one source of information for a complete picture.
7. **Various sources of error are part of the assessment process.**
   Many variables, even under controlled conditions, may intervene in the testing process as well as test results.
8. **Tests and other measurement techniques have strengths and weaknesses.**
9. **Test-related behaviors predict non-test related behaviors:**
   It is not necessary that the tasks that a person performs during a test are indicators or measures of the person's performance on the same tasks. The tasks may actually be indicators of something totally different, for example in HTP (house, tree, person) it is not a person's drawing ability or talent that are being tested. Rather the contents are assumed to present cues to a picture of the subject's personality.
10. **Present day behavior sampling predicts future:**
    The test results do not stand true for a person's behavior on the day of test administration alone. It is assumed that today's results stand true for future too.
11. **Testing and assessment can be conducted in a fair and unbiased manner.**
12. **Testing and assessment benefit society.**
    Looking at assumption 1 – 5 one can see that these imply that we have a variety of tests available to us, and that for every question regarding a person's behavior or mental processes we have some test that can help us find the answer. And if a test is not available in some situation, there is always a possibility of development of some new test or tool. That is why test construction, revision, and adaptation remain an ongoing process.

Types of Tests:

Although there is no limit to constructs, traits, attributes, or objectives for which test are available, one can categorize psychological tests on the following basis:

1. Individual versus group tests
2. Ability, achievement, or aptitude tests
3. Intelligence versus personality tests
4. Speed test versus ability tests
5. Structured personality tests versus projective tests
6. Verbal versus non-verbal/performance tests.
7. Commercial copyrighted tests versus available to all tests

1- Individual Versus Group Tests:

Some tests are meant to be administered to only one person at a time, whereas some can be taken by a number of persons together on one occasion. Individual tests require one examiner and one subject. In case of group tests one examiner works with many subjects together. For example, WAIS and WISC, Stanford-Binet Scales, and Kaufman Scales are tests that are to be administered individually. Many paper and pencil tests are also available for group administration. Raven's Progressive Matrices (RPM) are available for individual as well as group administration. Three forms of RPM are available. These forms differ in difficulty level. Otis self-administering tests of mental ability are group tests.

Group tests are quick, easy to administer, and easy to score. These are usually based on multiple choice items. The examiner's role is not as important as it is in case of individual tests. Even tape recorded instructions can be used for administration. Computer administration is quite common. However such tests are not suitable for situations where subject-examiner rapport is important.

2- Ability, Achievement, or Aptitude Tests:

Ability tests measure intellectual ability or cognitive behavior. Intelligence tests fall in this category. Such tests yield a value or score for IQ of the examinee. A test of ability covers a sample of what the person knows at the time of being tested. It usually may cover more than one ability; and is an indicator of the level of development attained in those abilities.

Achievement tests on the other hand are meant and designed for measuring the effect of educational programs or trainings. Programs of training and instructions are conceived and tailored with specific objectives to be achieved. Achievement tests are given at the end of the program, or after certain stages or sub-sections of the same are covered, to measure if the objectives have been achieved. SAT is an example of such tests.

Aptitude tests measure cumulative influence of multiplicity of experiences in daily living (Anastasi, and Urbina, 1997). In short aptitude is based upon learning under uncontrolled, general conditions in life. It is not rooted in any program of instructions as such. An aptitude test is used for predicting future performance.

Aptitude tests measure the potential for learning a specific skill. If a child has learnt mathematics and he is given a test that measures how many mathematics based problems can he solve, then it will be a test of mathematical achievement. On the other hand if a test measures as to how many problems, or how well, can he solve if provided the requisite training and education, then it will be a test of mathematical aptitude. Differential aptitude tests (DAT) are the most widely used aptitude test batteries.

3. Intelligence Verses Personality Tests:

Although there is no need to make this distinction because the very titles suggest what the tests measure, we still need to clarify the concept. Many people confuse the two types as one. We should be clear that personality tests do not yield information regarding I.Q and I.Q tests cannot be used for assessing personality traits. Varieties of both types of tests are available. Both types of tests are available in large numbers.

4. Speed Tests versus Ability Tests:

In case of ability tests, the level or amount of ability is measured, for example I.Q in intelligence tests. Time for completion of test is not of utmost importance. In some tests there is no time limit as such e.g. Raven's Progressive Matrices.

On the other hand, speed of performance matters in many other tests. Individual differences depend upon this variable e.g. Clerical Speed and Accuracy Test in DAT.

A similar variety is that of power tests. Such tests do have a time limit, but that is long enough to allow every man to complete all items.

5. Structured Personality Tests versus Projective Tests:

Personality tests are found in a number of varieties. These tests measure typical behavior e.g. dispositions and traits.

There are two broad categories of measures of personality; structured personality tests (objective tests of personality) and projective tests.

In case of structured tests, fixed response options are provided for each item and the examinees choose or mark the one that describes them, or that represents them the best. The item response options may be of alternate response style e.g., true/false, or may have three, four or more options e.g. MCQs.

Such tests are usually "self-report" type.

Example:

I love animals	Yes ☐	No ☐
Animal are one's best friends	Yes ☐	No ☐

The interpretation and scoring of structured tests is quite simple. Anybody can score these tools. All that is required is an answer key. That is why these are called objective tests. EPPS or Edwards Personal Preference Schedule is an example of this variety of personality tests.

On the contrary, the projective tests are not easy to score or interpret. Their administration too requires care. A vague or ambiguous test stimulus is presented to the subjects who describe or explain whatever they see or find in the stimulus e.g. Rorschach's Inkblots. The subject's descriptions are then interpreted. In some cases the subject has to narrate or write a story about whatever they perceive as happening in a picture, e.g. TAT or Thematic Apperception Test. The narration is supposed to reflect subject's personality.

The subject may also be required to draw something e.g. HTP or House, Tree, Person. The drawings are analyzed to identify the subject's traits or predispositions. In some other tests the subject has to complete an incomplete sentence e.g. RISB or Rotter's Incomplete Sentence Blank, or has to give prompt answer to stimulus words presented by the examinee one by one e.g., WAT or Word Association Test.

6. Verbal versus Nonverbal/ Performance Tests:

Most tests are either fully verbal in nature, or have verbal ability or the use of language as a major component. This is truer of IQ tests. In case of structured personality tests, they all depend upon the use of language. But some tests do not involve the use or measurement of verbal ability e.g. RPM or Raven's Progressive Matrices.

7. Commercial-Copyrighted Tests Versus "Available To All"/Online Tests:

Most standardized IQ tests are copyrighted. They cannot and should not be reproduced, photocopied, or used where prohibited e.g. WAIS or WISC. Such tests have to be purchased from the author or the agency which has the rights to sell. Disclosing items of these tests to general public would be unethical and would make their use meaningless if the examinees are already familiar with the contents.

However some tests are 'available to all'. Contents of these tests are available online as well as in many textbooks. But these tests are primarily meant for research purpose and not for diagnostic or screening purpose. Tests or scales measuring personality traits and other aspects are more commonly available online as compared to IQ tests. For example the authors of Multidimensional Health Locus of Control Scale (Wallston and Walston) and Self Efficacy Scale (Schwarzer and Jerusalem) have placed their scales on websites and anybody can borrow and use them. These can be downloaded, copied, and printed. The use of these scales has yielded very valuable empirical evidence in many spheres.

Major Contexts of Current Test Use:

Although psychological tests are designed for, and are used in numerous life situations, we can see that there are three most significant areas where testing becomes an essential component of whatever process is being undertaken. The three major context of test use in the modern world are:

i) Educational testing
ii) Occupational testing
iii) Clinical and Counseling Psychology (Anastasi, and Utnins, 1997).

i) Educational Testing:

Testing is used in the educational systems and set ups at all levels and for various purposes. However school is the setting where maximum use of tests takes place; the type of tests used in this context will depend on who uses the test; the school counselor, the school psychologist or the teacher. Achievement, intelligence, special and multiple aptitude, and personality tests are used in the educational system.
Tests may be used in the following forms in the educational settings:

a. General Achievement batteries

These batteries generate profiles of scores on individual subsets or in major academic areas. These can be used by those in the primary grades up to adult level. Varieties of these tests are available and for various levels. Batteries comprising combinations of tests are also used. Examples include Stanford Achievement Test Series with the Otis-Lennon School Ability Test; The IOWA Test Series and Tests of Achievement and Proficiency with the Cognitive Abilities Test; California Achievement Test and Comprehensive Test of Basic Skills with the Test of Cognitive Skills.

b. Tests of Minimum Competency in Basic Skills

These tests measure mastery of basic skills in children as well as adults. An example of these tests is the TABE battery or Tests of Adult Basic Education. This battery of tests covers five graduated levels of difficulty across five different content areas. These areas include reading, language and applied mathematics.
The results are to be found in two forms; competency-based information, and as norm referenced scores.

c. Teacher-made class room tests

Teacher made tests can be objective as well as subjective; great care is required in designing test and writing test items.

d. Tests for the college level

The most widely known test program of this type is the Scholastic Assessment Tests (SAT) Program of the College Board. The tests include the Reasoning Test (SAT I) and the Subject Tests (SAT II).

e. Graduate School Admission

Tests are also available for use at the time of admission to graduate and professional schools. The test that most people are familiar with is GRE or Graduate Record Examinations. The test is used in most countries of the world. It is primarily meant for admission to American universities/colleges, but some local universities also use it for screening purpose. Also, it is used for candidates' selection for honors and scholarships. The test has a general tests as well as subject tests in a variety of disciplines.

ii) Occupational Testing:

Psychological tests are used for a variety of objectives to be attained in occupational or workplace settings. These tests are available for screening and short listing at the time of induction, assessment of performance, for job analysis, prediction of job performance, and similar reasons.

- Academic Intelligence Tests: e.g. Wonderlic Personnel Test
- Aptitude tests: e.g. GATB or General Aptitude Test Battery; ASVAB or Armed Services Vocational Aptitude Battery.
- Personality tests: e.g. Five Factor Model Personality Inventories are used. MMPI is also used when there is a need to trace or identify psychopathology; CPI and HPI e.g, Hogan Personality Inventory are also used.

iii) Tests used in Clinical and Counseling Psychology:

In these settings psychological tests are used for diagnosis, induction in treatment groups or hospitals, for general assessment, and for gauging the rate of recovery. All intelligence and personality tests may be used. For example HTP can depict psychopathology.
Tests for neuropsychological assessment are also available. Two more commonly used such tests are The Bender Visual Motor Gestalt Test (Bender Gestalt Test/BGT) and Benton Visual Retention Test (BVRT). Some tests are used for diagnosing specific learning disabilities e.g. Kaufman Test of Educational Achievement (K-TEA).

Lesson 05

The Testing Process: Test Administration and Test Taking

Test Administration Process:

The main purpose of testing is to generalize the results obtained from the sample to those in non-test situations. Any influences that are specific to the test situation constitute error variance and reduce test validity. It is therefore important to recognize any test-related influences that may limit or weaken the generalizability of test results. The testing process itself, examiner related variables, and the subject/examinee related variables all are potential confounding variables that need to be looked into, watched, and thoroughly controlled.

1. Advance Preparation Of Examiners:

Besides basic academic qualification and practice in psychological testing some other testing situation related variables need to be kept in mind. Advance preparation is one of such variables. Advance preparation for testing is required for the uniformity of the testing procedure.

The preparation for testing session takes many forms.

- In most of individual testing, verbal instructions are essential to memorize. Even in group tests, some previous familiarity with the statements to be read prevents misreading and hesitation and permits a more natural, informal manner during test administration.
- The preparation of test materials is another important preliminary step. In individual testing such preparation involves organizing the necessary material so that when actually using the test, mishandling can be avoided.
- Materials to be used during the test should be in easy reach.
- The testing procedure should be known and clear before test administration.
- Training for test administration is also essential.

2. Testing Conditions:

- The environment for testing should also be appropriate.
- Noise-free room, proper seating arrangement, and adequate light should be provided to test-takers.
- It is important to recognize the conditions which might affect the test scores e.g. noise, privacy, traffic in the testing room etc. There is also evidence that the type of answer sheet employed may affect test scores (F.O. Bell, Hoff, & Hoyt, 1964)
- Many subtle testing conditions affect the performance on ability as well as on performance tests. Whether examiner is a stranger or someone familiar to the test takers might make a significant difference in test takers' scores (Sacks, 1952; Tsudzuki, Hata, & Kuze, 1957).
- In another study, the general manner and behavior of the examiner, as illustrated by smiling, nodding, and making such comments as "good", or "fine" were shown to have a decided effect on test results (Wickes, 1956).

3. Introducing the Test: Rapport and Test-Taker Orientation:

In test administration, rapport refers to the examiner's efforts to arouse the test taker's interest in the test, elicit their cooperation, and encourage them to respond in a manner appropriate to the objectives of the test e.g. in ability tests what is required is careful concentration on the given tasks and making one's best efforts to perform well.

The training of the examiner involves techniques for establishing rapport which vary with the nature of the test as well as with the age and personal characteristics of the persons tested. This may be taken to suggest that the examiner should have a flexible personality, ready to change and adopt different communication styles while catering to clients/ subjects with different backgrounds.

The examiner has to introduce the task or test to the subject, make him feel comfortable, try to help him overcome possible anxiety, and encourage him to concentrate and complete it.

Special motivational problems may be encountered in testing emotionally disturbed persons, prisoners, or juvenile delinquents. Especially when examined in an institutional setting, such persons are likely to manifest a number of unfavorable attitudes such as suspicion, insecurity, fear, or cynical indifference. Special conditions in their past experiences are also likely to influence their test performance adversely. As a result of earlier failures and frustrations in school, for example, they may have developed feelings of hostility and inferiority toward

academic tasks, which the test resembles. The experienced examiner may make special efforts to establish rapport under these conditions.

Examiner and Situational Variables:
Comprehensive surveys of the effects of examiner and situational variables on test scores have been published periodically (Lutey & Copeland, 1982; Masling, 1960; S.B. Sarason, 1954; Sattler, 1970, 1988, Sattler & Theye, 1967). The situational variables have been found to have an effect on test taking behavior and test performance. These variables are more likely to operate with unstructured and ambiguous stimuli, as well as with difficult and novel tasks, than with clearly defined and well-learned functions. In general, children are more susceptible to examiner and situational influences than are adults.
The test results may be influenced by the examiner's behavior immediately preceding and during test administration. For example controlled investigations have yielded significant differences in intelligence test performance as a result of a "warm" versus a "cold" interpersonal relationship between examiner and examinees.
Another way in which the examiner may inadvertently affect the test taker's responses is through the examiners' own expectations. The test taker's activities immediately preceding the test may also affect their performance, especially when such activities produce emotional disturbance and fatigue. In an investigation with third- and fourth grade school children, there was some evidence to suggest that **IQ** on the Draw-a-Man Test was influenced by the children's preceding class room activity (McCarthy, 1944).
Several studies have been concerned with effects of feedback regarding test scores on individual's subsequent test performance. In a particularly well- designed investigation with seventh-grade students, Bridgeman (1974) found that "success" feedback was followed by significantly higher performance on a similar test than was "failure" feedback in students who had actually performed equally well to begin with.

Characteristics of a Good Examiner:
- Understanding of the nature of test to be used
- Well trained and experienced
- Knows the instructions
- Clear speech
- Empathetic
- Sharp senses, particularly hearing and vision
- Quick in understanding and responding
- Professional honesty and integrity

Test Taker's Perspective: Examinee Variables

Test Anxiety:
Test anxiety is the response of test takers that was noticed by examiners in early studies. Test anxiety is a reaction of test taker in a testing situation. It was observed that this response was stimulated by the negative effect on test performance. Many practices/procedures have been introduced to build up the rapport and to reduce examinee anxiety. These practices reduce the strangeness of test takers to testing situation and help them to deal with their test anxiety. Examiner's own smooth and well organized way of handling the testing process also help to reduce the test anxiety.

Individual Differences in Test Anxiety:
It has been seen through research that there are individual differences in the responses of test anxiety of school and college students. For examining these differences in students, different questionnaires were developed. For example the children questionnaire includes these types of items.
1- Do you worry a lot before taking a test?
2- While you are taking a test, do you usually think you are not doing well?
The initial findings of researches indicated that intelligence and achievement tests had negative correlation with test anxiety in school and college students. It is obvious that these findings do not reveal the **causal relationship.** It may happen that after the poor performance on tests, resulting in failure, student develops test anxiety. Research evidence also suggests the opposite side of this idea. It has been seen that students' performance on the test is affected because of their anxiety. In a study, low anxious and high anxious children

with equal intelligence scores were tested. After many repeated learning tasks it was found **that low anxious children significantly improved** as compared to high anxious children.

Anxious and Relaxed States in Test Performance:

Researchers have also compared test performance under two different states: "anxious" and "relaxed" states. A study by Mandler and Sarason (1952) revealed that ego involving instructions e.g. telling test takers that everyone is expected to finish his/her test in a given time has some positive effect on the test performance of low anxious students. But these instructions had negative affect on high anxious students. Similar findings have been revealed in other studies involving the interaction between testing situation and individual characteristics:

- There is no doubt that a chronically high anxiety level exerts a detrimental effect on school learning and intellectual development.
- The competition among college bound high school students for getting admission in colleges and universities is also a contributory factor in test anxiety (take the example of bright students who fail in the medical college entry test). French (1962) compared the performance of high school students on a test given as part of the regular administration of the SAT with the performance on a parallel form of the test administered at a different time under "relaxed" conditions. The result of the study indicated that performance was no poorer during the standard administration than during the relaxed administration.
- Research findings suggest that students scoring high on a test anxiety scale obtain lower GPAs. Also, those scoring low on test anxiety had better study habits in comparison to those scoring high on test anxiety (Culler and Holahan, 1980).

Research on Nature, Measurement, and Treatment Of Test Anxiety:

- Looking at the nature of anxiety, it is believed to include two important components: emotionality and worry. Emotionality includes feelings and physiological reaction like increased heart rate and tension. Worry component, on the other hand, is understood to include the negative self-oriented thought about failure and its consequences. This is the cognitive component. These cognitions disrupt test performance by drawing attention away from the task oriented behavior.
- These components of anxiety can be measured with the help of several test inventories like Test Anxiety Inventory developed by Spielberger and his coworkers.
- Treatment of the test anxiety includes the behavior therapy procedures with the combination of the cognitive therapies.

Comprehensive Investigation of Test-Taker Views:

A book edited by Baruch Nevo and R. S. Jager provides a detailed description of examinee's reaction to testing in educational, clinical, and counseling settings. These are based upon the works of a number of researchers.
The fifteen chapters of book provide extensive research on various topics by the author's own research work as well as the researches of many other authors. The said book provides the solution of social and practical problems of testing.

Effects of Training on Test Performance:

The effect of earlier practice, exposure to same or similar items, and guidance has also been a topic of interest for psychologists involved in testing. There is no dearth of research evidence in this area.

Coaching:

Several studies were conducted by British Psychologists, with special reference to the effects of practice and coaching on the tests formerly used in assigning 11 year old children to different types of secondary schools (Yates et al., 1953-1954). It was seen that:
Individual with deficient educational background are more likely to benefit from special coaching than are those who have has superior educational opportunities and are already prepared to do well on the tests.

- How much will someone benefit from coaching and subsequently improve depends upon:
    1. Ability of the test taker
    2. Earlier educational experiences
    3. Nature of the test, and
    4. The amount and type of coaching

5. The more the test material and the coaching materials have in common the greater will be the improvement.

- Many coaching studies interpreted yield ambiguous and un-interpretable results because of serious methodological shortcomings. The main shortcoming is failure to employ non-coached control group.

Test Sophistication:

Research evidence has shown that test sophistication or test- taking practice has a positive effect on test performance.

1. It has been seen that in studies where alternate forms of the same test were used, there was a tendency for the second score to be higher. Similar findings were reported when alternate forms were administered after varying intervals ranging from one day to three days (Donlon, 1984; Droege, 1966; Peel, 1951, 1952).
2. Short orientation and practice sessions can also be affective in equalizing test sophistication. These familiar trainings reduce the effect of prior differences in test taking experience as such. This approach is illustrated by the College Board publication entitled taking the SAT I: Reasoning test, a booklet. This booklet offer suggestions for effective test-taking behavior. Another example of this is GRE that also provides test familiarization materials.
3. The test familiarization is not limited to the printed media but includes transparencies, slides, films, videocassettes, and computer software.

Instructions in Broad Cognitive Skills:

Some researchers have been exploring the opposite approach to the improvement of test performance. They emphasized to provide education rather than coaching. Some of these researchers have been working with the educable mentally retarded children and adolescents.

Many of the training procedures employed in these programs are designed to develop effective problem-solving behavior such as careful analysis of problems, consideration of all alternatives, and relevant details.

However these programs are still in an exploratory stage and more research is needed to establish the improvement in these programs.

Lesson 06

Test Norms: Interpreting Test Results

Norms are used for interpreting the test scores. The term norm means a standard that we use for comparative as well as normative functions. Test norms are those scores on a measure that are used as standards against which the scores of any test taker are compared. Raw scores of psychological tests have no meaning unless they are interpreted with additional supporting information or data. For example, a person's score of 50 on an intelligence test does not tell us whether he or she has an average, below or above I.Q level unless we know what the average score, or the norm is. From this score it is not possible for a researcher to judge his/her standing in a distribution of scores. This score of 50 may be equal to a score of 70 on another intelligence test because the difficulty level of the items in both intelligence tests may be different. The difficulty level of the items in a test determines the meaning of the scores of this test. The scores on any test become meaningful in the presence of norms that have been developed for that test.

Norms can be defined as:

> *"The test performance data of a particular group of test takers that are designed for use as a reference for evaluating or interpreting individual test scores".*

In a more simple way it can be said that norms provide standards to which the results of the test takers on different measurements can be compared.

Norms are established by administering a specific test to a sample representative of the population of interest, and by obtaining the distribution of score of the same sample. For example, a researcher develops a scale for measuring the stress level of university students. The test is administered to "normative sample" of university students. After the statistical analysis, it is found that average score of these students is 25. This average score then serves as a norm for university students. Another university student when takes this test and obtains a score of 30, then it can be concluded that this particular student has an above average score on the stress measure

.

Norms are the test performance of the standardization sample. **Standardization sample** is a group of people whose performance on a specific test is taken as a standard or norm for comparison. All the other individuals' performance on this specific test is compared with the scores of this standardized sample. A person's performance on a test is interpreted with reference to the distribution of scores in the sample used as a representative of the population. This is done to discover where he/she stands in the distribution preferably called the relative standing of a person.

Relative standing refers to position of a person where he or she lies in distribution of scores in relation to the rest of other persons in a particular population. The raw scores obtained by people need to be given a meaningful and comparable form. These scores are converted into relative measures, or derived scores. These measures are obtained for two reasons:

1- To learn about an individual's standing in a distribution of scores with the reference to the previously established norms. This relative standing allows evaluation of the test taker's performance in comparison to other persons who have already taken the test.
2- The second purpose is to make a direct comparison of individual scores with the standard comparable measures.

For fulfilling these two purposes, raw scores are converted into relative measures in several different ways. These measures yield test performance related information and are expressed in two forms:

- Developmental level, and
- Relative position within a specific group

Development of norms takes place after a series of steps followed in a special sequence that will be discussed later. In order to understand establishment of norms, the processes involved, the meaning of raw scores, and other related concepts and processes one needs to understand some basic statistical concepts. Basic statistical concepts are used in the development and utilization of norms. The knowledge of these statistical concepts is necessary for everyone for a better understanding of research literature in testing as well as any other area of psychology.

Statistical Concepts Used In Psychological Testing:

The aim of the statistical method is to organize and summarize the quantitative data. The raw scores are arranged for the sake of better understanding.

The raw scores first of all tabulated in order to give them some meaningful shape.

The first statistical concept used here is that of a frequency distribution.

1-Frequency distribution is the method that arranges large data in an organized way. In the frequency distribution, data is grouped into class intervals and tallying each scores in the appropriate level. After entering the data these tallies are counted for the purpose of frequency i.e., total number of cases in the data. By adding all the frequencies in a frequency distribution table, total number of cases that is equal to N is obtained. For example, the raw scores of 100 individuals on achievement test ranges from 50 to 104. For summarizing the data we make class intervals of these raw scores of five points. These scores presents in the table given below:

Frequency Distribution Scores Of 100 Individuals On Achievement Test:

Class interval	Frequency
50-54	2
55-59	5
60-64	7
65-69	10
70-74	14
75-79	13
80-84	20
85-89	11
90-94	4
95-99	6
100-104	8
N=100	

The scores of frequency distribution can also be presented in the form of different graphs. The two types of graphs that are used for this purpose

1. Histogram
2. Frequency polygon

Histogram:

It is a graph that shows the distribution of measured scores in the form of class intervals. In histogram, horizontal axis (or x-axis) presents the class interval scores and vertical axis (or y-axis) presents the number of cases or frequencies falling within each class interval. The height of the column presents the number of frequency and the width of column covered with the length of intervals.

Frequency Polygon:

Scores of frequency distribution also presents in the form of frequency polygon. In frequency polygon, the number of people indicated by taking a mid-point in each class interval. These points or dots then connected by a straight line. Like histogram, the scores in this graph are plotted on horizontal axis and frequencies are plotted on vertical axis.

2-Normal Bell-Shaped Curve:

Statistical data can also be seen in the form of normal distribution curve. This curve facilitates with all basic statistical analysis. The curved area of graph covers the average of some value or characteristics like I.Q, height, weight or some personality traits and the two sides of the graph indicate the extreme cases of any value or characteristics. The average of some value or characteristics means that a large number of populations expressed or have this thing. As the number of people increases there are more chances of distribution scores to resemble with the original normal curve.

The raw data scores can also interpret in some other ways. The one way to describe these raw scores in meaningful scores is measures of central tendency.

3-Measures of Central Tendency:

Measures of central included these concepts of statistics.

1. Mean
2. Median
3. Mode
4. Variability and Standard Deviation

Mean:

Mean or average is calculated by adding all scores and then dividing this sum by frequencies (number of cases or people). The formula of the mean is $M=\sum X/N$. For example, children obtained the following marks in a English test 7, 8, 4, 6, 5, 9, 3. We calculate the mean or average this score by using the above stated formula. The sum of these scores is 42 which are then divided by N that is 7. The mean is 6.

Median:

A second measure in central tendency is Median. It is the middle most value or score in a group of data. The median is the point that divides the distribution into half above and half below scores. Median can be simply calculated by using this formula $n+1/2$. The median of the above scores is $7+1/2 = 4$; hence the median will be the 4^{th} value that is 6.

Mode:

It is the third measure in the central tendency. Mode is the highest frequency value in scores. In a normal distribution cure it represents by the highest point. In a group of scores like 7,8,7,7,9,4,7, mode value is 7 because it occurs most often in the data.

Variability and Standard Deviation:

Test scores can be further divided by measures of variability which is the dispersion of scores around the mean.

Variance: Variance is known as the "average squared deviation" around the mean. The formula of variance is:

$$\sum(X-\bar{X})^2/n$$

At first, the mean of the scores is subtracted from each individual score for finding the deviation. The values of the deviation then squared and lastly the mean of these squared deviation values is calculated. The one important thing is that sum of the deviation will always equal to zero because the positive and negative (e.g. +20, -20) values around the mean always cancel each other.

Standard deviation: Another more adequate measure of variability is Standard Deviation which is denoted by S.D or **σ.** It is calculated by taking the square root of the variance. The more individual differences reveal the larger S.D while small individual differences indicate the low values of S.D. In a normal or approximately normal distribution curve, interpretations of S.D are very clear.

Lesson 07

Types of Norms

As discussed previously, raw test scores need to be turned into a more meaningful form. Therefore they are converted into relative measures or derived scores. Not only do these derived scores tell us about the relative standing of any one who takes this test, they also provide comparable measures. Using these scores a person's performance on various tests can be compared.
Derived scores are expressed either in terms of developmental level attained or relative position of a person (or his score) in a certain group.

Developmental Norms:
Test scores can be expressed in terms of developmental norms. Developmental norms can be defined as the typical patterns or characteristics, and age specific tasks or skills of development at **any age** or stage of development.
Developmental norms are established keeping in view development and maturation. The underlying assumption is that people, children and adults, are capable of performing at specific levels at different stages of life. When most people can perform certain tasks at a certain age level then it is considered as the norm for that age level. This is also considered the mental age of persons at that age level. Subsequently if a 13 year old girl can perform the tasks accurately and completes a test that most 13 year olds can do, then her mental age will be stated as 13 (MA = 13). If an adult can perform only the tasks that a six year old can do, then his MA will be 6. Considering his physical age, he is mentally deficient, backward, or special. The MA of a 9 year old who can perform tasks meant for a 16 year old will be interpreted in the same manner, but as deviating in the positive direction. His mental age will be higher than his biological age.
We can also say that a fifth grade child has 7^{th} grade ability when tested in a specific mathematic ability because he could solve most of the problems that a seventh grader could do. Developmental norms may also be based upon highly qualitative descriptions of behavior in specific functions such as sensorimotor activities or concept formation, expressed in qualitative terms.
This is one way of comparing one's performance with the norms. But this approach is not as easy to apply as it may seem to be. At times people take tests that measure different abilities. Even the subtests of the same test may be measuring a variety of skills or abilities. In such cases it is not necessary that everybody attains the same MA in all tests or subtests. This makes comparison in terms of developmental norms difficult.

Although developmental norms are not an unpopular form of norms, test scores based on developmental norms are not psychometrically sound. However they are used commonly for descriptive purposes, especially in clinical and research settings.

Mental Age:
The term mental age is widely used after the development of Binet-Simon scales. Binet himself used a more neutral term "mental level" in his own scale. In Binet's scale items were grouped in year levels e.g items that were passed by the majority of 8 years old children in the standardization sample were included in the 8- year level. Similarly items that were passed by a large number of 10 year old children were placed in 10- year level and so on. Stanford- Binet and other similar scales are age scales.
While the scale was being used very frequently, the problem of 'scatter' of scores was observed. It meant that many subjects did not show uniform performance on all subtests of the scale/test. Some individuals would failed the tests that were below their age level while at the same time they had passed the tests that were above their age levels. For overcoming this problem the concept of **"basal age"** was introduced.
Basal age refers to the highest year at which a person passes all items. For all the tests that were passed at higher year levels, the subject was given partial credits in months. These were then added to basal age .Basal age and additional months of credit were added together to yield the child's mental age.
Mental age norms are also used with the tests that are not formatted or designed according to the year levels. In these tests, mean raw scores of children of specific age groups in the standardization sample are the norms of the tests for the corresponding age groups. The mental age of a child is determined by comparing her raw score with the age norm available. For example if the raw score of a 12 year old girl is equal to the 12 year norm then she would be said to have mental age of 12 years.

One major shortcoming of using mental age as indicator of intellectual ability is that mental age does not mean the same thing at different stages of life. MA of 4 of a 5 year old is not the same as the MA 24 of a 25 year old. As age progresses the unit of MA tends to shrink. According to Anastasi and Urbina (2007), a child who has mental age of 3 at the age of 4, would be 3 years retarded at the age of 12. Mental growth of one year from 3 to 4 years of age is equivalent to 3 years of mental growth from 9 to 12 years. Therefore positive or negative deviation from norm at different age groups does not mean the same. Deviation at a very young age means a lot as compared to that at older age.

Grade Equivalents:

Grade equivalents represent the scores on educational achievement tests attained by children in a certain grade. These norms are obtained by calculating the mean raw scores of children in the standardization sample representing each grade. If 6th grade children in the standardization sample obtained a mean score of 35 in arithmetic test then this raw score has a grade equivalent of 6. Hence a student obtaining 35 on the same test will be said to have a grade equivalent of 6.

Most schools have an academic year spanning over ten months. A whole year is represented by the corresponding grade. However the measurement may be made after some months have passed after the grade started. In such cases the successive months can be expressed in decimal points. For example, the grade equivalent of 7.0 refers to the average performance of a 7th grader at the beginning of session. 7.5 present the average at the middle of the session and so on.

Grade norms have several limitations. According to Anastasi and Urbina (2007), grade units are unequal and these inequalities occur irregularly in different subject matter areas. They are only applicable for the common subjects taught throughout the grade levels covered by the tests; but not for different subjects taught for only one to two years in high schools or colleges. Even when the same courses are covered in the tests, it cannot be ensured that they received identical importance, attention, and learning in all grades.

Yet another complication, it may happen that a child progresses in one subject more rapidly than another subject during the same grade.

Grade norms tend to be incorrectly regarded as the performance level of students. Because of grade norms it is possible that a teacher of 6th grade assumes that all the students in class will obtain scores equal to or near to 6th grade norm in achievement tests. However individual differences in any grade can be so large that scores on achievement test will vary over several grades.

Ordinal Scales:

Another approach to developmental norms develops from the research in the area of child psychology. Psychologists exploring development in infants and young children made interesting observations. They gave descriptions of the behavioral functions of infants and children that were typical of successive ages. These behaviors included functions like sensory discrimination, linguistic communication and concept formation. These empirical observations proved to be valuable in the understanding of human development.

An example of this research is the work of Gesell and his associates. Their main emphasis was on the sequential patterning of early behavior development. The Gesell Developmental Schedules were developed to see the approximate developmental level in months that a child has achieved. The attainment of this development is shown in each of four major areas of behavior; motor, adaptive, language, and personal-social. Eight key ages from 4 to 36 weeks are used as standards of developmental level. The developmental level of a child is determined by comparing his behavior with the behavior typical of each level.

Gesell and his associates, who focused on the sequential patterning of early behavior development, claimed that the children's development involved: a) orderly progression of behavior changes and b) uniformities of developmental sequence. For example, a chronological sequence can be observed in visual fixation and in hand and finger movements when they are reacting to a small object placed in front of them. The way palm, thumb, or fingers etc. are moving and the manner in which they are used varies from one stage of development to another.

This approach had the underlying notion that developmental stages follow a certain sequential order. Hence the scales used for measuring these are ordinal scales that yield information regarding the stage where a child stands. The use of these also involves the understanding that successful performance of a child at one level implies success at all lower levels of age.

Jean Piaget did extensive work in child and developmental psychology in the 1960s. He gave his theory of cognitive development. He talked about stages of cognitive development as falling in a sequence, and said that age levels for these stages were arbitrary. His theory covered ages from infancy till mid-teens. Rather than broad

abilities, he was interested in studying specific concepts. He introduced many specific concepts e.g. Object permanence, conservation, and perspective etc.

- In object permanence the child is aware of the object existence, when they are out of sight.
- Conservation is the recognition that an attribute remains constant over changes in perceptual appearance e.g. the quantity of a liquid will remain constant whenever it is poured in a different shaped container.
- Perspective means the knowledge that objects appear differently when at a distance and seen in perspective.
- In order to assess cognitive development, Piagetian tasks are used. These are designed in such a manner that they reveal the dominant aspect of each developmental stage.

In short, ordinal scales gauge the uniform progression of development through successive stages by measuring attainment of specific functions.

Within-Group Norms:

Within-group norms evaluate a person's performance with the most nearly comparable standardization group like a child's raw score is compared with the children of his age or grade. These norms are so popular that now all test scores provide within-group norms in some types of form. Within-group scores employ many statistical procedures because of their clearly defined quantitative meaning.

Percentiles:

"A percentile indicates the individual's relative position in the standardization sample". The percentage of a person in standardization sample expressed in terms of percentiles scores. For example, if 50 % people obtained 25 score in an analytical reasoning test then this score corresponds to 50th percentile.

- Percentiles can also describe ranks in a group of 100 people i.e a person who is at the top in the group given the rank of one; likewise a person who is at the bottom in a group will be given a poorer rank.
- 50th percentile refers to the median as percentile; if a score of 50 was at the 50th percentile then a score above 50 represents the above average score while a score below 50 indicates the below average scores. The 25th and 75th percentiles are known as the first and third quartile points (Q1 and Q3) because they cut off the lowest and highest quarters of the distribution.
- The difference between percentage and percentile is that percentage is a raw score while percentiles are derived scores.
- 50th percentile refers to the median as percentile; if a score of 50 was at the 50th percentile then a score above 50 represents the above average score while a score below 50 indicates the below average scores. The 25th and 75th percentiles are known as the first and third quartile points (Q1 and Q3) because they cut off the lowest and highest quarters of the distribution.
- The difference between percentage and percentile is that percentage is a raw score while percentiles are derived scores.

Standard Scores:

Standard scores are the scores that express the individual's distance from the mean in terms of the standard deviation of the distribution.

Standard scores can be calculated by the linear and non-linear transformation of the raw scores. Linearly derived scores are also known as "z-scores". In z-scores the mean of the normative sample is subtracted from the raw score and then divided by the standard deviation of this sample.

Computation of Standard Scores:

$$z = \frac{X - M}{SD}$$

If X = 100, M = 80 and SD= 10

By putting the given values in formula

$$z = \frac{100 - 80}{10}$$

$$z = 2$$

Any raw score that is equal to mean will end up in to a z-score of zero. A negative derived score indicates that a person's score is below average; positive scores indicate that it is an above average score.

Most modern tests use standard scores and interpretation of their scores is made with reference to standard scores.

Normal Standard Scores:

These are standard scores expressed in terms of a distribution that has been transformed to fit a normal curve. Normal standard scores are obtained by finding the percentage of a person in standardization sample. Then this percentage is located in the normal curve frequency table, and the standard score is obtained.

Normal standard score can also be put in any convenient form. If the normalized standard score is multiplied by 10 and added or subtracted from the 50, it is converted into a T score. It was first proposed by McCall (1922). In this scale, an individual score of 50 corresponds to mean and score of 60 to 1 SD above the mean and so on.

Normalized standard scores should be applied when the sample is large and representative and when this is confirmed that deviation from normal results is due to the some drawback in the test rather than from the characteristics of the sample.

Another variation of such transformation of scores is on the Stanine scale. United States Air force developed this scale during the Second World War. Stanine is based upon the words 'standard nine' and the fact that scores run from one to nine. A single digit system is employed with a mean of 5 and standard deviation of approximately 2.

The Deviation IQ:

The term IQ (Intelligence Quotient) was introduced in early intelligence tests. It is simply obtained by dividing the MA by chronological age, and multiplied by 100:

$IQ = MA/ CA \times 100$

If the child's MA equals to CA then the child's IQ will be exactly 100. IQ of 100 represents the average or mean performance. IQ below 100 indicated below average scores that are moving toward retardation and above 100 presents the acceleration.

However it is proved that Ratio IQ has some major technical problems. The problem with the IQ level is that it is not comparable with different age levels unless the SD of IQ distribution remains constant with the age. For example if a child can read at the age of 3 which is his chronological age and an average child starts reading at the age of 6 which is mental age than his/her IQ will be scored 200. For this reason ratio IQ is replaced by the so called deviation IQ.

Deviation IQ is a standard score with a mean of 100 and an SD that approximates the SD of the Stanford-Binet IQ distribution. It compares people of the same age and assumes that IQ of individuals is normally distributed.

Relativity of Norms:

Interest Comparisons:

An IQ of a person should always be described by the name of the test on which it was obtained. For example, the IQ of one person is 110 and another person is 90. It cannot be accepted without further detailed information. The relative standing of the IQ of both persons can change with the exchange of particular tests.

An individual's relative standing in different functions may be misrepresented by the lack of comparability of the test norms. For example, an individual has been given a verbal comprehension test and a spatial aptitude test to determine his/her relative standing in the two fields. If the verbal ability test was standardized on a random sample of high school students, while the spatial test was standardized on a selected group of students attending elective courses, the examiner might erroneously conclude that the individual is much more able on verbal ability than spatial ability, when the reverse may actually be the case.

In longitudinal comparisons, individual's scores on a specific test obtained over time. For example, if the child scores are 110, 115, and 120 at the fourth, fifth, and sixth grades, it can be said that these differences in the scores may be due to the different tests. There are three reasons for these variations among the scores of the same individual performance on different tests.

1. Intelligence tests can differ in content with the same label. Like one test may include only verbal content, other includes numerical content and so on.
2. The scales' units may not be comparable e.g. IQ on one test may have SD of 12 while IQ on another test has SD of 18.
3. The compositions of the standardization samples used in establishing norms for different tests may vary. The same individual will appear to have performed better when compared with a less able group than when compared with a more able group.

Scales of Measurement:

In order for us to describe test scores in a quantitative form, we have to design tests in such a manner that they yield results in a numeric form. They either should be originally obtained in the form of numbers or should allow conversion in that form. Psychological measurement involves rules according to which objects are assigned numbers and "quality" is expressed in numeric form. For example, in a personality test, an item asks "do you like to be in the company of young age mates most of the time? The answer is allowed in terms of degrees e.g. "always", "often" "could not say", "rarely" and "never". The subject has to choose one option that best describes her. Now the response is going to be in a qualitative form. Comparison with others is not possible in this form. Therefore a certain number is assigned to these options. Ranging from one to five, option "never" is allocated one, and option "always" is assigned 5. Now all the responses of the subject can be quantified, and these quantities can be subjected to st5atistical treatment.

In short tests where every question has a right answer, like in ability tests, the total numbers of corrected responses is counted and that yields the test score.

Lesson 08

Test Norms and Related Concepts

Standardization:

Norms are established for the sake of standardization of any test. Standardization is the process whereby a test is administered to a representative sample of population whom the test is meant for, for the sake of establishing norms. A standardized test is the one that has normative data, as well as clearly specified administration and scoring procedures.

The Normative Sample:

When we are using test scores for any purposes, or interpreting them and making judgment about the test taker, we should keep in mind, that norms that are being referred to be representative of a particular population from which the standardization or normative sample was selected. The mean scores of that sample are assumed to represent the parent population. Therefore if the sample comprised women alone, than their mean performance score should be used as a norm for women's raw scores alone. And if women from any specific cultural/regional background alone were used for norm development, then one should be very careful in interpreting the scores of women belonging to completely different cultures. It is therefore advisable that the standardization sample should include maximum characteristics of the population. Also the sample should be of a large enough size to ensure stable values. However the standardization sample can be as small as one person (Cohen, & Swerdlik, 19) depending upon the population of interest. Nevertheless the significance of the size of sample cannot be ignored. As the size of sample increases, the chance of making error reduces. This is because when the sample is small and does not include all characteristics of the population of interest, then many possible sources of error may become intervening variables. Therefore the size of the sample should be good enough to generate a meaningful distribution of raw scores. Also it should not be ignored that over inclusion or under inclusion may take place. Therefore great care is required in sample selection.

In order to obtain a true representative sample, careful sampling procedures need to be followed. Most populations comprise sub groups or strata. The sample therefore should include members from all strata. Proportionate stratified random sampling is the best approach for selecting a representative sample. In this type of sampling members from each subgroup/stratum are included in the sample in the same proportion in which they are found in population. The characteristics of the sample will indicate the type of population to which the results can be generalized. Ideally speaking, we should decide and define our population in first place. Then as a second step we can select the representative sample. But practically speaking this can be somewhat difficult. Identifying all possible characteristics from the population and then selecting a sample with all those qualities can be difficult. The second, more practical, option can be that we take a purposive sample that we believe contains all characteristics that we are interested in catering for. Norms may be developed for that sample, and the population can be defined accordingly.

For example if the standardization sample included children aged 12 to 16 years, with six years of schooling, then the norms will be meant for a population of children within the same age range and similar educational background. There is no test that provides norms for all sorts of population altogether. "No test provides norms for the human species"! (Anastasi, & Swerdlik). We have a common tendency to use tests developed for, and standardized in, western countries with our local population. We should be cautious while interpreting the results of such tests and making judgments about personality or ability of local tests takers because the available norms were not established for either population of their origin, or even for a very similar population. Another point that needs to be considered while interpreting scores with reference to the norms is that whether any such specific conditions prevailed at the time of norm development that could have affected the performance or scores of the members of normative sample. These could be any special societal conditions or any special selective variables (Anastasi, 1985).

National Norms:

If a test is standardized using a nationally representative sample of the population, than it would be called a "national sample". The sample containing all characteristics of interest is chosen from different geographical region/location, communities, socioeconomic status, institutions etc. For example, if we were to establish norms for a test meant for measuring achievement of university students in Pakistan, then we will have to select a normative sample that represents university students from all region of the country.

National Anchor Norms:

We have a variety of tests that measure the same ability or human trait. People who are tested on the same ability through different tests, many obtain different scores on all of these tests. The psychologist, or professional who is going to interpret the scores will need information regarding the equivalence of these scores i.e., how do we compare and interpret a score of 25 on test ABC, and a score of 34 on test XYZ of verbal ability?

National anchor norms provide solution to this problem. These norms provide equivalency table for scores on the various tests of the same ability. Equivalency of scores on various tests is calculated using the 'equipercentile method'. Test scores are compared and their equivalence is determined with reference to their corresponding percentile scores.

A score on two tests will be considered to be equal only when they have equal percentiles in the group being studied. Therefore a score of 34 on test ABC carries 85th percentile, and so does the same score in test XYZ then they are equivalent. But if a score of 35 on ABC had 85th percentile, and a score of 29 on XYZ had the same percentile, then 35 on ABC will be equivalent to 29 on XYZ.

National anchor norms are very helpful is assessing the equivalence of scores, but they should not be used as a single and fully dependable source of judgment. The difficulty level, detailed contents, and the sample from which scores have been obtained is very important. It is a prerequisite that every member of the sample should have taken all the tests whose equivalence is being determined. Tests should be interchangeable in the true sense before they are described as equated or fully equivalent.

Specific Norms:

As discussed earlier, national anchor norms do provide information regarding equivalence of test scores, but relying completely on these may be problematic. An alternate solution to the problem of non-equivalence of tests and their comparability is to use specific norms. This solution requires that tests should be standardized on more narrowly defined population. It should be chosen in a manner that it suits to the specific purposes of each test.

Rather than using broadly defined samples from broadly defined populations, tests can be standardized on narrowly defined samples. Normative samples may be selected on the basis of purposive sampling, including precisely that type of subjects that fit into the purpose for which the test is making measurements. Such samples are chosen on the basic of specific purpose of a test or subtests.

In case of using specific samples, the chances of controlling nonequivalence are reduced. However when the norms for such tests are reported, a clear report of the limits of the normative sample has also to be made. Additionally the use of such tests should be avoided with samples chosen from populations that are beyond the limits of a specific normative sample.

Highly specific norms are considered to be useful for most testing purposes. Even when representative norms from broadly defined populations are available, availability of separately reported 'sub group' norms is considered very helpful (Anastasi & Swerdlik). If a large group, population of interest, includes distinguishable subgroups then it is better to have overall groups' norms as well as specific norms. For example, medical profession comprises a large community of doctors. Within this broadly defined population, it is believed that doctors working in different wards or areas of specialization undergo different types and levels of stress and public dealing. It is believed that the experiences of doctors working with dying patients, burn victims, and newborns in an obstetrics nursery are entirely different. Therefore whereas we can have one single inventory to measure occupational and personality variables in doctors as one community, we may also develop separate subscales or other measures to assess variables of interest in doctors working under different levels of stress and in different working conditions.

At times norms may be even more narrowly defined than specific norms. There are occasions when institutions or organization prefer to develop their own norms. These norms, developed by test takers themselves are called **local norms**. For example a university may decide to develop norms on its students; norms may be accumulated for students entering the first year and then these may be used to predict achievement in following years. An organization may establish norms for selectees or new recruits and on the basis of it their future performance may be predicted. For this purpose data regarding performance and progress will also be gathered.

Fixed Reference Group:

Although conventionally developed norms, whichever type, give a good reference for interpretation and comparability of test scores, there are other approaches to interpretation as well. At times non normative scales are used. In one such type a fixed reference group is used. This is called the 'fixed reference group scoring system'. In this system the distribution of scores obtained on the test from one group of people who took the

test is used as the basis for the calculation of test for future administration of the test. The group from which the scores were obtained is called the 'fixed reference group'.
This system does not provide normative evaluation of performance, but ensures comparability and continuity of scores.
The College Board Scholastic Aptitude Test or SAT is an example of this system. The test was later on renamed as Scholastic Assessment Test (SAT). The first administration of SAT took place in 1926. At that time its norms were based on the mean and standard deviation of people who took the test. Till 1941 SAT scores were expressed on a normative scale in terms of mean and standard deviation of test takers at every administration.
With the passage of time more and more colleges became members of the College Board.
The variety of colleges also expanded. It was felt that there was a need to bring changes into the normative scale because of two reasons:

a) The element of scale continuity needed to be maintained. Failing this the test taker's scores would depend on the characteristics of the group tested during a particular year.
b) It was observed that there was variation in students' scores in tests taken at different times during the year. Students performed less well at certain times of the year than those who took SAT at other times. It was concluded that this was a function of the time of year when test was administered.

It was speculated that different factors operated at different times when the test was administered. The system was therefore changed in 1941. In the same year, approximately 11,000 candidates had taken the test. The distribution of scores of this sample was taken as a standard, and all SAT scores were expressed in terms of mean and standard deviation of these candidates. This standard was used for future conversion of raw scores. For subsequent forms of the test, these 11,000 candidates constituted the fixed reference group. A score of 500 corresponded to the mean of this group; 600 meant one SD above mean, and 400 was one SD below.
In each form of the SAT, a short anchor test (set of common items) was included in order to allow translation of raw scores on any form of the SAT into these fixed reference scores. A chain of items extending back to the 1941 form was developed. This happened as each new form was linked to one or two earlier forms which in turn were linked with other forms, thus ending into a chain. These non-normative scores could then be interpreted through comparison with any appropriate distribution of scores e.g., a particular college, a type of college, region etc.
In 1995, a new fixed reference group began to be used. This one comprises those more than a million (2 million, according to Cohen) who took the SAT in 1990. After April 1, 1995, the scores of SAT takers are reported on the "recentered" scale derived from the 1990 reference group. In order to assist test users in converting individual and aggregate scores from the former scale and vice versa, interpretive aids and materials have been developed.

Item Response Theory:
Item response theory can be understood in terms of the 'latent trait models'. Beginning from the 1970's, psychologists have been increasingly interested in a class of mathematically sophisticated procedures for scaling the difficulty of test items.
The availability of high speed computers made such procedures possible. The general title of 'latent trait models' was used for these approaches. The basic measure used by these is the probability that a test taker with a talent trait (or a specific ability) succeeds on an item of specified difficulty. There is no implication regarding the existence of the trait as such. The latent traits are mathematically derived statistical constructs. These are derived from empirically observed relations among test responses. The total score that a test taker obtains on the test is a rough, initial estimate of their latent trait.
The term latent trait model was later replaced by Item Response Theory (IRT) because 'latent trait' created a false impression of a specific trait.
The purpose of IRT models is to establish a "sample-free" scale of measurement that is uniform, is applicable to individuals and groups having widely varying ability levels, and to test contents that vary widely in terms of difficulty levels. Rather than using the mean and standard deviation of a specific reference group to define the origin and the unit size of the scale, IRT models set origin and unit size in terms of data representing a wide range of ability and item difficulty. This may be obtained from many samples rather than a single sample.

Domain Referenced Test Interpretation

Nature and Uses:

So far we have been discussing the concept, use and significance of norms for the interpretation of psychological test results. But we should remember that norms are not the only way of assessing, measuring, and interpreting individuals' abilities or traits.

The tests that provide normative data for the sake of score interpretation are norm-referenced tests. A norm-referenced test is "a test that evaluates each individual relative to a normative group" (Kaplan, & Saccuzzo, 2001). A norm referenced test is administered to a representative sample of the population of interest, raw scores are gathered and analyzed, and in the light of the analysis norms are established for future test takers. In such type of tests, the norms or standards with which every individual's performance is compared are the scores obtained by other persons. If one has attained scores equivalent to the normative samples' average scores then one's performance is considered average. If he fails to do so then his score is below average, and in case of a score higher than the normative average he is considered above average.

There is another type of tests that employs a certain criterion or standards for describing a persons' performance. A person has to perform within a domain in order to be considered proficient in a skill, behavior, or ability. These are called the domain referenced tests.

The terms criterion-referenced and domain-referenced are used interchangeably. The latter is used more commonly. A criterion referenced test is "a test that describes the specific types of skills, tasks, knowledge of an individual relative to a well-defined mastery criterion. "The content of criterion-referenced test is limited to certain well-defined objectives". (Kaplan, & Saccuzzo, 2001)

Glaser (1963) was first one to use the term 'criterion-referenced testing'. Alternative terms like 'domain-referenced' and 'content-referenced' testing were also proposed and used by writers. The interpretive frame of reference in a domain-referenced test employs a specific 'content' domain rather than a specified population of 'persons'. The test results are reported in terms of what the person/ test taker knows, how proficient he is, to what extent he has command over a certain content domain. For example, according to Anastasi (2007), test takers' performance may be reported in terms specific kinds of arithmetic operation which they have mastered, the estimated size of their vocabulary, or difficulty level of reading matter that they can comprehend (from comic books to literary classics). It can also be expressed in terms of the chances of a person's achieving a designated performance level on an external criterion (educational and occupational).

Who Decides and Determines the Domain or Criterion?

The domain is primarily derived from the values or standards of an individual or organization. There are so many real life situations and professional scenarios where evaluation of a person's ability, proficiency, or performance with reference to a norm is meaningless. Rather what is required is command over the domain. This is what is of value. For example, in training of surgeons how better can one perform a surgery than others or how much above or below average his skills are is not what is of importance. What is required is that one should have acquired the skill of surgery, to a specified extent.

Similarly in the evaluation of a pilot, what is most important to be assessed is whether the pilot can fly a plane and how well can he do that. Whether he or she is equal, above, or below others is not of prime importance.

Norm-referenced tests assess and describe how well test takers have performed in relation to other people. Domain-referenced tests on the other hand tell us about what test takers can do. In other words we can say that domain-referenced tests focus on the potential, and norm-referenced tests focus on the performance of a test taker in comparison to others.

> *Can you think of a situation where you would prefer a norm referenced test and a situation where you would need a domain- referenced test ……what about learning a table of 5 by a 4 year old, and being able to repair an electrical connection?*

In some contexts, domain-referenced tests are also called "mastery tests".

These are the situations where the test is used to assess the mastery or achievement of certain skills or contents. The focus of attention is content rather than a specific population. Such tests, particularly in the educational context, became very popular in the 1970's. Where domain referenced tests are used, the performance of test takers may be reported in terms of mastery or command over skills e.g. specific kind of arithmetic operation mastered by them; the estimated size of their vocabulary; difficulty level of reading matter that they can

comprehend; or the chances of achieving a designated performance level on an external criterion that may be educational or occupational.

> *Can u think of situations where you would be interested in mastery of a person over a skill rather than where he stands in comparison to others?*
> *What would you like to see in a cricketer? Whether he can hit the ball very well or what is his position in comparison to other cricketers? And if you were to travel by air, what would interest you; is your pilot a good pilot or his relative position among pilots?*

Domain referenced tests are commonly used in education; such tests have been used in educational innovations. According to Anastasi and Urbina (2007), their major applications have been made in computer-assisted, computer-managed, and other individualized, self-placed instructional systems. Testing is closely integrated in these systems, with instruction being introduced before, during and after completion of each instructional unit to check on pre-requisite skills, diagnose possible learning difficulties, and prescribe subsequent instructional procedures. Broad surveys of educational accomplishment have also employed domain referenced tests e.g. National Assessment of Educational Progress in the U.S.

Another area of application of these tests is when mastery of small number of clearly defined job skills is to be assessed for the sake of testing job proficiency e.g. in military occupational specialties. A similar application is when domain referenced assessment is made for evaluating the attainment of minimum requirements e.g., qualifying for drivers' or pilots' license.

In addition to these areas of application, it is believed that domain referenced tests can be helpful in improving the traditional teacher-made tests if the test developers are familiar with the concept of, and philosophy behind, domain referenced tests.

Content Meaning:

The most significant feature of domain referenced tests is that test performance is interpreted in terms of content meaning. As previously said, the goal is not to find out about relative standing of a person, but to learn what he/she knows, and what he/she can do.

In developing and designing this type of tests, the content has to be carefully chosen, treated, and presented. While constructing such a test, a clearly defined domain of knowledge or skills to be assessed is the primary requirement. The content to be tested will be selected from a content domain. This domain should be an important one, and must be generally accepted as important. In order to develop items for making assessment of mastery over the content, the content has to be subdivided into smaller units. These small units are defined in performance terms i.e. what performance or behavior will indicate that a certain element or section of content has been learnt or mastered. When this approach is employed in educational settings and content is subdivided into smaller units, then these units correspond to behaviorally define instructional objectives. The mastery over each unit is described in terms of instructional objectives. For example, "divides numbers carrying zeros by numbers carrying one zero by canceling one zero at the end", "can convert grams into ounces", "can convert centigrade into Fahrenheit".

Instructional objectives not only specify the learning outcomes, they also affect the way a course is taught, and the way assessments will be made.

After the instructional objectives have been finally shaped, the difficult task of item development for each objective follows. It may take quite long to prepare items for sampling individual objectives, because each item has to be a good representative of the domain to be assessed. Careful formulation of objectives and clear statement of concepts and methodologies is also very important in this regard. The test developer's own expertise, experiences, and judgment, all matter a lot to the various steps of item construction.

Domain-referenced tests can be most useful for testing basic skills like reading and arithmetic at elementary levels. According to Anastasi and Urbina (2007) an ordinal hierarchy is usually adopted for arranging instructional objectives for testing basic skills. For the acquisition of higher level skills, it is necessary to acquire elementary skills. Therefore they will be arranged in that order. Another point that one needs to keep in mind here is that the nature of objectives will depend on the nature of content or subject being considered. It is not advisable to formulate highly specific objectives for advanced levels of knowledge in less highly structured subjects. At elementary levels the content as well as the sequence or learning will be mostly flexible.

Mastery Testing:

Mastery testing is an important characteristic of domain-referenced tests. The procedure of testing mastery provides us an all-or-none score. This means that the test score will yield information about the presence or absence of mastery. It can be about presence or absence or about the attainment of the pre-established level of mastery. A generally expected level is complete mastery that may be up to 80 % or 85 %.

Another way of reporting mastery is to use a three way-distinction. Under this system the report is in terms of mastery, non-mastery, and an intermediate, doubtful, or "review" interval.

In mastery tests, individual differences are not a matter of concern. A number of educators believe that individual differences become meaningless in these tests. They argue that if suitable instructional methods are used and enough time is given, then almost everyone will be able to exhibit complete mastery of the domain and achieve the instructional objectives. The area where individual differences can be noted in traditional educational tests is the time that subjects take in learning the content according to the objectives. Therefore, it is said that the effect of individual differences can be reduced to minimum after appropriate training.

Mastery testing is used in a number of individualized instructions. Published domain-referenced tests of basic skills for elementary school also employ this approach.

Two issues need to be considered in the construction of such tests; how many items should be used, and what proportion of items has to be correct before a reliable assessment is made. Initially these issues were tackled by the test developers using their own judgment. But now a number of procedures and statistical techniques are available for resolving these issues.

Develop a test to assess if your friend has learnt certain content.

Relation to Norm-Referenced Testing:

Although mastery testing has its advantages, its usefulness is more in case of basic skills at elementary levels. In case of areas where complete mastery is not the focus of interest, mastery testing is not the choice for evaluation of the subject.

As we go to higher levels, and talk of more advanced and less structured subjects, mastery testing does not prove to be the best and the most suitable approach. In areas like understanding, critical thinking, appreciation, and originality there is no way of assessing complete attainment, or mastery. In fact there is no end, extent, limit, or direction of learning or progress. In such cases norm-referenced tests are useful. There are in fact no cutoff points/ scores to show complete absence of these faculties or skills.

We also have some published tests available that allow norm-referenced as well as domain-referenced applications e.g., Stanford diagnostic tests in reading and in mathematics. Both, appropriate norms and a system for qualitative analysis of child's attainment of detailed instructional objectives are available in these tests.

Some authors write that even when we are using domain-referenced tests, we cannot say that they have nothing to do with norms. The concept of norms in any case is operative in the domain referenced tests. There is an underlying realization that a continuum of abilities does exists.

Minimum Qualification and Cutoff scores:

As we said earlier, mastery tests may adopt an all-or-none approach or make a three-way distinction. But there are situations where clear cut cutoff score points or scores need to be specified. There are numerous situations where minimum qualifications have to be specified and implemented e.g., when someone is to be granted a driving or flying license; when workers are being selected at a war zone where sharp learning and vision are required; when workers are being selected for a nuclear plant; when students graduate from one college and are to be chosen for a medical school.

Different tests use their own cutoff scores. It is recommended that the following points should be kept in mind using cutoff scores for decision making.

a) One should not use the scores of a single test as cutoff. A band of scores (from more than one test) should be used.
b) Multiple sources of information should be used for decision making regarding test takers. Relevant performance on tests other than the one in question, whether from past or present, should be used.
c) If a panel of judges set the cut off points, then those judges should be experts in test construction as well as the areas of task performance.

d) Whenever possible, cutoff scores should be established and verified with the support of empirical information.
e) Test scores should be obtained from groups that are clearly different from each other on the relevant criterion behavior (Anastasi & Urbina, 2007).

An empirical method for setting cutoff scores can be in the form of expectancy tables.

Expectancy Tables:

An expectancy table contains probability of different criterion outcomes for persons who obtain each test score. These tables are based upon statistical information regarding the relationship of tests/variables as yielded past administrations.

An expectancy table is something like this:

Relationship between scores on test XYZ and course grades:

Score on XYZ	*Number of students*	*Percentage in each grade*			
		A	**B**	**C**	**D**
50-60	12	-	1	8	4
60-70	15	3	6	6	-
70-80	23	12	9	2	-

Lesson 10

Test Construction

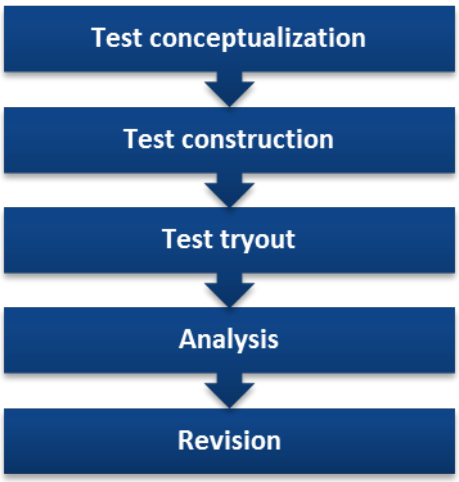
Test Development:

A good test is created by implying established principles of test construction. The process of test development occurs in five steps

1. Test conceptualization
2. Test construction
3. Test try-out
4. Item analysis
5. Test revision

Once the idea for a test is conceived (test conceptualization), items for the test are drafted (test construction). This first draft of the test is then tried out on a group of sample test takers (test tryout). Once the data from the tryout are in, test-takers' performance on the test as a whole and on each of the test's items will be analyzed. Statistical procedures collectively referred to as item analysis, will be employed to assist in making judgments about which items are good as they are, which items may need to be revised, and which items should be discarded. The analysis of test items may include analyses of item reliability, item validity, item discrimination, and -depending upon the type of test it is- item difficulty level. On the basis of item analysis and related considerations, a revision or second draft of the test will be created. This revised version of the test is then tried out on a new sample of test-takers, the results will be analyzed, the test further revised *if* necessary- and so it goes.

Test Development Process:



1. Test Conceptualization

Test development is a result of test developers' idea of developing a tool to measure a particular construct. The stimulus for developing a test can be anything. For example,

- Literature on an already developed test might create the need for further work on the psychometric soundness of the test, and the would-be test developer thinks that he/she can do better.
- The emergence to prominence of some social phenomenon or pattern of behavior might serve as the stimulus for development of a new test.

Apart from the stimulus for developing a new test, a number of questions immediately confront the prospective test developer. Some of these questions include

- What is the test designed to measure?
- What is the purpose of developing the test?
- Is there any need for this test?
- What will be the sample of the test?
- What should be the test content?
- What should be the procedure for test administration?
- What should the ideal format of the test be?
- Should more than one forms of test be developed?

- What special training will be required of test users for its administration and interpretation?
- What type of responses will be required by test takers
- Who will get benefit from its administration
- Is there any potential for harm as the result of an administration of the test?
- How will meaning be attributed to scores on test?

The last question points out the issue of norm versus criterion referenced tests.
There are different approaches to test development depending upon whether they are criterion referenced or norm referenced tests. A good item on a norm=referenced achievement test is an item for which high scorers on the test respond correctly; low scorers on the test tend to get that very same item incorrectly. Whereas, development of a criterion oriented test or technique entails pilot work with at least two groups of test takers' one group known to have mastered the knowledge or skill being measured and another group known to have not mastered such knowledge or skill. The items that best discriminates between these two groups would be considered "good" items.

Pilot Work:
In the context of test development, pilot study/research refers to preliminary research surrounding the creation of a prototype of the test. Test items may be pilot studied to evaluate whether they should be included in the final form of the instrument. In pilot work, the test developer typically attempts how to best measure the targeted construct. The process may involve the creation, revision, and deletion of many test items. Once pilot work has been completed, the process of test construction begins. However, in future the need for pilot research is always a possibility because of the test's requirement for updates and revisions.

2. Test Construction:
Scaling may be defined as the process of setting rules for assigning numbers in measurement. In other words scaling is the process in which values are assigned to different amounts of attributes being measured.

Types of scales:
The scales can be categorized along a continuum of level of measurement and referred to as nominal, ordinal, interval, or ratio scales. But scales can be categorized in other ways.

- **Age scale:** if the test takers' performance on a test as function of age is of critical interest**,** then the test might be referred to as an age scale.
- **Grade scale:** if the test takers' on a test as function of grade is of critical interest**,** then the test might be referred to as grade scale.
- **Stanine scale:** if all raw scores on the test are to be transformed into scores that can range from 1 to 9, then the test might be referred to as a stanine scale.

Scaling Methods:
The Likert scale is used to scale attitudes. Likert scales are relatively easy to construct. Each item presents test taker with five alternative responses, usually on agree/disagree. Or approve/disapprove type of continuum. Likert (1932) after different experiments concluded, that assigning weights of 1 through 5 generally works best.
Another scaling method is the **method of paired comparisons**. Test takers are presented with pairs of stimuli which are asked to compare. They then must select one of the stimuli more appealing than the other, and so on. For each pair of options the test taker would receive a higher score if they selected the option that was considered more justifiable by the majority of a group pf judges.
Another way of deriving ordinal information through scaling system entails **sorting tasks.** In these approaches, printed cards, drawings, photographs, objects, or other such stimuli are typically presented to test takers for evaluation. One method of sorting, comparative scaling entails judgments of a stimulus in comparison with every other stimulus on the scale.

Categorical scaling is another scaling system that relies on sorting. Stimuli are placed into one of two or more alternative categories that differ quantitatively with respect to some continuum.
All the foregoing methods yield ordinal data. The method of equal=appearing intervals, first described by Thurstone (1929) is one scaling method used to obtain data that are presumed to be interval.

Writing Items:

The process of test construction also involves the ideas related to item writing. The three important considerations in this regard are:

- What range of content should the items cover?
- Which of many different types of item formats should be employed?
- How many items should be written?

When a standardized test is developed which is based on multiple-choice response format, it is usually advisable that the number of items for the first draft of a standardized test contain approximately twice the number of items that the final version of the test will contain. Sampling provides a basis for content validity of the final version of the test. Because half of the items are eliminated from the test, the test developer should keep in mind that the final version of the test should sample the domain.

The test developer may write a large number of items from personal experience. Help from experts can also be taken for item writing. In addition to experts, information for item writing can also be obtained from the sample to be studied. Literature searches may also be a valuable source of inquiry for item writing.

Considerations related to variables such as the purpose of the test and the number of examinees to be tested at one time, enter into decisions regarding the format of the test. A good response format for the test is one in which the participant have many choices to answer. This is called selected response format.

Item Formats:

There are two types of format; selected response format, and constructed response format.

Selected response format:

This type of format presents the examinee with a choice of answers and requires selection of one alternative. The types of selected-response format are multiple choice, matching, and true/false items

Constructed response format:

It is the response format that requires the examinee to provide or create the correct answer than just selecting it. Three types of constructed-response items are the completion item, the short answer, and the essay.

- A completion item requires the examinee to provide a word or phrase that completes sentences.
- A good short answer item is written clearly enough that the test taker can indeed respond briefly, with a short answer.
- An essay is a type of response format in which the examinees are asked to describe in detail a single topic which is asked from them.

Scoring Items:

There are many scoring models but the most common is the *cumulative* model. The concept underlying this model is that the higher the score on the test, the higher the ability or the trait, being measured, is. For each test taker's response to targeted items made in a particular way, the test taker earns cumulative credit with regard to a particular construct.

The second model is a *class* model in which test takers' responses earn credit toward placement in a particular class or category with other test takers whose pattern of score is presumably similar in some way.

A third scoring model is *ipsative* scoring. A typical objective in ipsative scoring is the comparison of a test taker's score on one scale within a test, with another scale within that same test.

Once all of the groundwork for a test has been laid and a draft of the test is ready for administration, the next step is, logically enough, test tryout.

3. Test Tryout:

Having created a pool of items from which the final version of the test will be developed, the test developer next tries out the test. The test is tried out on the sample for which it is constructed. It is also important to consider that on how many subjects the test should be tried out. It is usually considered that no fewer than five subjects and preferably as many as ten subjects, for every one item on the test. The more the subjects in try out, the better it is. Test tryout should be executed in the same conditions that are same as possible to the conditions under which standardized test will be administered. Test instructions, and everything from the time limits allotted for completing the test, to the atmosphere at the test site should be similar as possible.

What is a Good Item?

Characteristics of a good test are considered to be characteristics of a good item.

- A good test is one that is reliable and valid; similarly a good test item should be valid and reliable.
- Further, a good test items helps to discriminate test takers; a good test item is one that high scorers on the test as a whole get right.
- An item that high scorers on the test as a whole do not get right is probably not a good item.
- A good test item can also be described as one that low scorers on the test as a whole get wrong; an item that low scorers on the test as a whole get right may not be a good item.

After the first draft has been administered to a representative group of examinees, it remains for the test developer to analyze test scores and responses to individual items. At this stage the test undergo different types of statistical analyses which are collectively called as "item analysis".

4. Item Analysis:

Statistical procedures are collectively known as item analysis. For item analysis different statistical procedures are employed in order to select the best items from a pool of tryout items. Among the tools that test developer employs to analyze and select items is an index of an item's difficulty, an item-validity index, an item-reliability index, and an index of item discrimination.

Qualitative Items Analysis: Though statistical procedures are employed for item analysis, there are some non-quantitative methods that employ verbal rather than mathematical techniques. Through use of simple questionnaires or individual or group discussions with test takers, any test user can obtain valuable information on how the test could be improved.

5. Test Revision:

A great amount of information is gathered at the time of item-analysis stage. On the basis of that information, some items from the original pool will be eliminated and others will be re-written. One approach would be to characterize each item according to its strengths and weaknesses. Test developer may sometimes find it necessary to balance the strengths and weaknesses across items. For example, if many otherwise good items tend to be somewhat easy, the test developer may purposefully include some more difficult items.

Having balanced all the concerns, the test developer come out of revision stage with a test of improved quality. The next step is to administer the revised test under standardized conditions. On the basis of item analysis of the data derived from this administration of the second draft of the test, the test developer may consider the test in its finished form.

Lesson 11

Item Writing

The process of test construction also involves careful planning, considering a number of factors to be kept in mind before actually writing test items. Three important considerations in this regard are:

- What range of content should the items cover?
- Which of many different types of item formats should be employed?
- How many items should be written?

Nature and Range of Content:
The test developer will have to decide about the nature and extent of content to be included in the test. This will primarily be decided with reference to the objectives laid down for measurement. Each item should be measuring some aspect of the content area.

Type of Items:
Different types of formats are available to test developers. The type of format will also depend upon the type of test and construct or trait being measured. For example a projective item may be suitable for a personality test but not for an achievement test. Details of item formats are given in the following sections.

Number of Items:
The test developer also has to consider the number of items to be included in the test. It will affect the length of the test. Also, whether the test is going to be administered individually or in group may also be affected by this decision. Another variable that may affect decision about the number of items is whether the test is going to measure a single trait/ ability/ domain or multiple traits/abilities/domains. Fewer items will be required if a single domain is to be measured. In case of multiple traits/ abilities a larger number of items will be required so that there are enough items available to measure each one of them. We have very lengthy inventories like MMPI containing a few hundred items, and short scale such as self-efficacy scale containing 3 to 10 items.
When a standardized test is developed which is based on multiple-choice response format, it is usually advisable that the number of items for the first draft of a standardized test contain approximately twice the number of items that the final version of the test will contain.
The test developer may write a large number of items from personal experience. Help from experts can also be taken for item writing. Literature searches may also be a valuable source of inquiry for item writing.
Considerations related to variables such as the purpose of the test and the number of examinees to be tested at one time, enter into decisions regarding the format of the test.

Item Formats:
Kaplan and Saccuzzo (2001) have given a very good description of test item formats. The test developer can choose from the following formats described by them:

a. The dichotomous format
b. The polytomous format
c. The Likert format
d. The category format
e. Checklists and Q –sorts

Test formats are classified in many other ways as well; recall type and recognition type; constructed response type and identification type.
A very common distinction is made between objective type and essay type formats.
Another way is to divide test formats as selected response format, and constructed response format.

Selected Response Format:
This type of format presents the examinee with a choice of answers and requires selection of one alternative e.g. on achievement test, the test taker is required to select the correct option. The types of selected-response format are multiple choice, matching, and true/false items. Test formats described by Kaplan and Saccuzzo (2001) fall into this category. We will discuss these types in the following section.

a. The Dichotomous Format:

This is the alternate response format in which the test taker is provided with two response options to choose from. It can be used in a number of ways. In a common way, one of the two options is right and the other one wrong. For choosing the right option a certain mark or score, usually one, is given e.g. ten right options marked will carry a score of ten. Another way is to use the 'yes' and 'no' options, as in many personality inventories. Here the subject tells if a statement describes him or not. An even more popular use is the true/false format. In this format the test takers are presented with a number of statements and they have to tell whether these are true or false. In case of teacher made tests, certain statements from the textbooks or other materials are used in these items. Items using these statements are 'true' whereas text materials are altered to develop items that are 'false'.

Such items have the following advantages:

- They are easy to develop. The test developer does not have to make effort to think of a number of options that seem to be correct.
- Development of such tests does not take too much time.
- These items are easy to score.

There are some disadvantages as well:

- A test taker may be able to accurately answer at least 50% answers even when he does not know the right answers. The probability of answering an item correctly by chance is ½. If the teacher had used equal number of true and false items in the test and if the test taker marks all items as true (or marks all as false) then he will be marking half of the items correctly.
- Such items may encourage rote memorization of course content in case of teacher made tests. When students know that the items will be based on the text of the course content then they try to memorize the content rather than learning the concepts.

Nevertheless these items are used quite often. In order to control the disadvantages of these tests a large number of items covering a wider range of content should be used.

> ***Activity:*** *Make two test items with alternate response options. One should be true/false type and the other one yes/no type.*

b. The Polytomous (Polychotomous) Format:

In this type of items multiple response options, rather than alternate response options are given. Such an item includes these elements:

- A stem
- A correct alternative option
- Several incorrect alternatives or options which are called "distracters" or "foils".

Example:

Stem → A psychological test, an interview, and a case study are:
Correct Alt. → a. psychological assessment tools
Distractors → b. standardized behavioral samples
c. reliable assessment instruments
d. theory-linked instruments

These response options include one right answer and the others are distractors. The question or the item statement is called the stem. The stem should clearly state the question or the problem. The distractors should be framed in such a manner that each one of them should appear to be the right answer.

In multiple choice items the problem of answering half of the items without knowing the right answers is controlled to a great extent. However there is still a possibility that someone may be able to answer a certain percentage of questions even without knowing a word of it. If the items in a test have four options each then there is a possibility that one may be able to answer 25% of items correctly simply by making a right guess by chance. If three options are used then one may manage to mark 33% of items correctly by chance.

This problem can be overcome by increasing the options, but practically speaking it is very difficult to make a number of options which all seem to be the right answer. Testing experts believe that it is good to use three or four options for each item.

At times a correction for guessing is used employing the following formula:

Corrected score= R- W/n-1

Where, R= the number of right responses
W= the number of wrong responses
n= the number of choices for each item

For example a test taker scores 40 out 100 in a test that used a four option format. The score of 40 will be corrected like this:

Corrected score= 40- 60/4-1
= 40- 60/3
= 40-20= 20

This example shows that one should understand that guessing in a test may have more serious consequences than one might foresee. If the scoring procedure involves correction for guessing then guessing is to be completely avoided. If one has little or no knowledge of the content then the likelihood of making wrong guesses will also be very high. Looking at the formula, it can be seen that the greater is the number of wrong answers the greater will be the number subtracted from the right answers, which means the smaller will be the corrected score.

The test developer should include all response options in the same proportions. Some teachers have a tendency to include some particular option as their favorite and they would use that option frequently as the right option e.g. 'C' or 'D' more often than 'A' or 'B'. If students realize this trend then they will be more likely to choose the teacher's favorite options and mark many answers correctly even without knowing the answer.

> ***Activity:*** *Try to develop a multiple choice item with five options. Can you make five such options that look like correct answers of the stem? If you find it hard to make five, then try making three options.*

Matching Item:

A matching item is the format different from multiple choice format. The test taker is presented with two columns of responses and the test taker has to determine which item from the first column matches the item from the second column.

Both columns may have equal number of items, hence making it easy for the examinee to match the items to which he does not know how to respond. E.g. if a respondent is unsure about one of the options give, he may deduce the right answer by matching all the other options first. Thus a perfect score would be obtained. Providing more options than are needed is designed to minimize such a possibility.

c. The Likert Format:

A very popularly used tool for measuring personality and attitudes is the Likert's scale. Likert's scale format provides test takers an opportunity to endorse degree of their agreement to a statement. These statements pertain to attitude or personality. Likert used it as part of his method of attitude scale construction. Likert scales provide five response options ranging from 'strongly disagree' to 'strongly agree', with 'neutral' in between.

For example: "I like to make friends who are older to me". Choose from the following options:

Strongly disagree _____, disagree_____, neutral _____, agree____, strongly agree_____

At times people have a tendency to mark 'neutral' rather than giving any clear cut answers. Therefore six options, rather than five may also be used:

Strongly disagree _____, moderately disagree_____, mildly disagree_____, mildly agree____, moderately agree_____, strongly agree_____

The responses are summed to determine a person's score. In case of negatively worded items are reverse scored and added to the total.

d. The Category Format:

This format is similar to the Likert's scale, but has more options than five. A 10-point scale is more commonly used. However the number of response categories may be more or less than that.

This scale is used to rate something e.g. performance of a team. It is felt that people may not be very accurate in rating a player or a team's performance using this scale. It may happen that when a person watches a player in comparison to a superb player then he might rate the player to be rated at a lower level, whereas he may rate him very high when comparing with a poor performer. Some authors have recommended that in order to control this

factor the persons who will be rating may be shown videos of performances that could be rated '10' and those that 'deserve a '1'(Kaplan & Ernst, 1983).

e. Checklists and Q–sorts:

Adjective checklists contain a list of adjectives. The test taker checks the adjectives that are true about him. Checklists may be used for indicating one's own characteristics or those of other people. These are mostly used in personality assessment.

In Q – sort the test taker is provided with a number of statements and has to sort them in nine piles depending on to what extent they stand true of her. This can be used for rating other people as well.

> ***Activity:*** *Imagine you have to assess your classmates' ratings of a certain teacher using Q- sort format. What five statements will you write to be used by the subjects? …… and if you had to rate a TV serial what characteristics will you consider?*

Constructed Response Format:

It is the response format that requires the examinee to provide or create the correct answer than just selecting it. Three types of constructed-response items are the completion item, the short answer, and the essay.

- A completion item requires the examinee to provide a word or phrase that completes sentences.
- A good completion item should be clearly worded so that the correct answer is specified..
- A good short answer item is written clearly enough that the test taker can indeed respond briefly, with a short answer. There is no hard and fast rule that point out how short an answer should be.
- An essay is a type of response format in which the examinees are asked to describe in detail a single topic which is asked from them. The skills measured by essay type items are different from other type of item formats e.g. an essay requires recall, organization, planning, and writing ability, the other types of items require only recognition.

Lesson 12

Item Writing: Guidelines For Item Writing

Although every test developer has specific aims and objectives in mind while designing and developing a test there are a few points that all test developers have to keep in mind. DeVellis (1991) has given a number of guidelines for writing items. Some of those are as follows:

- **Clear definitions and item specificity:**
  Whatever is to be measured should be clearly defined. Items should be as specific as possible.
- **Developing an item pool:**
  Test developer should prepare a large number of items first and then choose items from this item pool. These items should all be representing the content area to be covered. Initially, the test developer may develop 3-4 items for each item to be included in the final version.
- **Appropriate level of reading difficulty:**
  The level of test takers should be kept in mind while preparing test items. Test's reading difficulty level should be appropriate for their level. For example ' I have an inherent predisposition to procrastinate when my life gets stuck at crucial junctures and I have to take serious decisions to tackle and cope' includes many words that are not included in a layman's everyday vocabulary.
- **Reasonable length of items:**
  Items should not be too long. Exceptionally long items are hardly found to be good.
- **Avoid 'double- barreled' items:**
  Double- barreled items are the ones that include two or more ideas at the same time. Practically speaking, these may be converted into two or more items depending on the number of ideas being conveyed. "I often help the poor because I believe in serving humanity and I am a follower of my leader who himself is known for social service."
  This item conveys a number of ideas; "I often help the poor because I believe in serving humanity and I am a follower of my leader who himself is known for social service."
- **Mixing of negatively and positively worded items:**
  At times the test takers have a tendency to agree with most items. This is called 'acquiescence response set'. In order to overcome this problem positively and negatively worded items should be mixed and used alternately. If all items are negative e.g. 'I hate liars', 'Most people betray', 'I feel depressed most of the time' then there is a chance that the test taker will develop a response set and will mark every statement as true, or may be false. Alternately, positively directioned items may be added to the sequence in order to avoid the possibility of a response set.
  There are some other points that also need to be considered in test planning and development, for example:
  Avoid double negatives: e.g. 'I do not disagree to the fact that people should not be stopped from playing cricket', or 'Don't you think that unsupervised internet use by children is not a good thing'. It is difficult for the test taker to understand what is being asked. These could have been rephrased as: 'I agree to the fact that people should not be stopped from playing cricket', and 'Do you think that unsupervised internet use by children is not a good thing' or even better 'Do you think that internet use by children should be supervised?'
  Besides this, the cultural background of would be test takers should also be kept in mind. Cultural, racial, and gender bias should be completely avoided.

Initial Test Plan and Design:

- From whatever we have discussed so far, we understand that development of any test involves the following steps:
- Determining and formulating the observations of the test
- Deciding the domain and content to be covered
- Deciding about the format of the test and the test items
- Developing/ writing the items.
- Developing much more items than required
- Trying out the initial version

- Analyzing the results of the try out
- Reshaping and refining the first version
- Another try out if needed
- Norm development, if required (required in case of standardized tests).

Apart from these steps that are common to most tests, we have considerations that are specific to certain types of tests.

Development of Educational Achievement Tests:

Other than the basic requirements of tests, what is most important in educational tests is the formulation and use of educational objectives. As discussed earlier, test objectives are stated in behavioral terms. Educational objectives define behaviors that will indicate whether the content in question has been learnt or not. There are a number of formal systems of educational objectives available to help a test developer. These systems are known as the taxonomies of educational objectives.

The most popular system is the one developed by Benjamin. S. Bloom and colleagues entitled Taxonomy of Educational Objectives: The Cognitive Domain (Bloom & Krathwohl, 1956). The major categories of instructional objectives in this taxonomy include the following:

- Knowledge
- Comprehension
- Application
- Analysis
- Synthesis
- Evaluation

Educational Testing Service's (1965) taxonomy includes the following categories:

- Remembering
- Understanding
- Thinking

Gerlach and Sullivan's (1967) system includes these categories:

- Identifying
- Naming
- Describing
- Constructing
- Ordering
- Demonstrating

Another system developed by Ebel (1979) covers a wider variety of objective:

- Understanding of terminology (or vocabulary)
- Understanding of fact and principle (or generalization)
- Ability to explain or illustrate (understanding of relationships)
- Ability to calculate (numerical problems)
- Ability to predict (what is likely to happen underspecified conditions)
- Ability to recommend appropriate action (or some specific practical problem situations)
- Ability to make an evaluative judgment

When test developers are planning a test they often make a table of specification to specify content areas and the corresponding objectives. This is a two way table that has the behavioral objectives on the vertical axis (row headings) and content or topics written on the horizontal axis (column headings).The table shows the type of bjectives corresponding to specific content areas and the total number of items in each box. The table looks something like this:

Behavioral Objectives	Content/ topic		
	Topic I	**Topic II**	**Topic III**
Application			
Comprehension			
Vocabulary			
Knowledge of facts			
Total	**Number of items**	**Number of items**	**Number of items**

Development of Personality Tests:
In these tests primarily the theoretical approach that the test developer belongs to matters. The items are based on the constructs that are to be measured.

Development of Intelligence Tests:
In planning and developing these tests a number of variable s have to be considered; what aspects of intelligence are to be measured? What criteria will be used? e.g. age, grade, or other, the test format considered suitable, the target population etc.

Development of Screening Tests:
Generally aptitude tests are used for screening purpose. The objective of screening tests is to identify candidates who are most suitable for the job, and those who fulfill the requirements of the position. Therefore, first of all a job analysis or task analysis is done. The various components of the job are identified and listed. The test items are based on these tasks and job related situations or 'critical incidents'.

Test Scoring:
Different tests are scored differently. In case of objective tests the scoring is done with the help of a scoring key. Scoring can be done manually as well as by using computers with attached scanners.
The scoring of essay type items is difficult. It may be done using 'holistic' or 'global' scoring approach in which the scoring is done for the whole response. The other option is to do analytic scoring in which different components or parts of a response are scored separately. The latter is the better approach. If possible, some other person may also be involved for rescoring in order to make scoring more objective.
The standardized, and many other formal, tests have their scoring procedures specified in the test manual.

Scoring Items:
There are many scoring models but the most common is the **cumulative model**. The concept underlying this model is that the higher the score on the test, the higher the ability or the trait, being measured, is. For each test taker's response to targeted items made in a particular way, the test taker earns cumulative credit with regard to a particular construct.
The second model is a **class model** in which test takers' responses earn credit toward placement in a particular class or category with other test takers whose pattern of score is presumably similar in some way.
A third scoring model is **ipsative scoring**. A typical objective in ipsative scoring is the comparison of a test taker's score on one scale within a test, with another scale within that same test.
Once all of the groundwork for a test has been laid and a draft of the test is ready for administration, the next step is, logically enough, test tryout.

Some Cautions:
- Avoid jargon. Use vocabulary that most people can understand
- The test should not be too long
- It should not be too easy that everyone can do it. Nor should it be that difficult that no one can do it.
- Cultural biases and stereotypical ideas should be avoided
- If any cultural differences are suspected to be present that may affect test results then nonverbal and culture free tests should be used.

But remember that even culture free tests may not be fully culture fair. People who have previous experience and exposure to such items may outperform those from such cultures where people are not familiar with tasks involving drawings and images.

Lesson 13

Reliability

Whatever attribute, characteristic, trait, object, or phenomenon one aims to measure, one wants the measurement to be reliable. In case of measurement in physical sciences reliability of measures is not a big issue. The measure or instrument will either be reliable or not. An instrument that is not completely reliable will not be used for measurement. On the other hand, in psychology we may be using measures whose reliability may be affected by a number of variables. Psychologists deal with phenomena that may not remain stable and consistent all the time. Therefore, the tools or instruments used for measuring these phenomena may not give us the same results every time we use them.

There may always be a chance of some degree of error in the measurement or the findings in any investigation. We, or the tools we use, may end up in underestimation or over estimation of a given phenomenon. Hence it is very important for us to estimate or calculate the chance and amount of error that may be involved in our assessment. This issue becomes even more significant when our assessment or measurement is to be used for serious decisions about someone's future, education, profession, diagnosis of a condition, or major life decisions. Therefore psychologists involved in test development do two things try to make tests as reliable as possible. Also, they report the reliability of their measure.

There are three basic qualities of a good psychological test namely, reliability, validity, and standardization. A test that is not reliable is not a trustworthy test. A test that is not valid will not measure what it is supposed to measure; and a test that has not been standardized will not give us results that we can confidently generalize to various other groups. We have discussed the concept of norms and standardization in the previous sessions and now we will look into reliability, its types, and applications.

By definition reliability means " the consistency of the scores obtained by the same persons when they are reexamined with the same test on different occasions, or with different sets of equivalent items, or under other variable examining conditions" (Anastasi & Urbina,2007). In order to understand the concept reliability we have to understand two other concepts; correlation and error in measurement. According to Kaplan & Saccuzzo (2001) reliability is "the extent to which a score or measure is free from measurement error. Theoretically, reliability is the ratio of true score variance to observed score variance."

The classical test score theory implies that everyone can obtain a true score on any test or measure if the measure is free of error. But we know that there is perhaps no measure that can be considered as totally error free. There is always some chance of error. This further implies that the scores that we obtain on different measures are not the true scores of the test taker. These are the observed scores plus error. In other words the observed score is not the true score of a test taker, implying that the observed score may not be a 100% true representative of a person's traits or abilities. Here we need to realize that the term 'error' is not used to indicate that something has gone wrong or we have made some 'mistake'. Error, in this context, refers to the amount and extent of variance that may be expected in results.

The observed score is therefore the sum of true score and error i.e.

$$X= T+E$$

Here X refers to the observed score, T is the true score, and E is the error that can be expected in the measurement.

Sources of Error in Test Scores:

Psychologists and educationists try to estimate the amount and degree of error that may be expected in their measurement. That is the main reason why test developers and test administrators emphasize on uniform testing procedures besides controlling other possible sources of error. The following are a few of the variables that may be possibly causing error in measurement:

a. Test Related Factors:

- Difficulty level (too difficult or too easy)
- Length of the test (too long causing fatigue or boredom)
- Domain or content (not suitable for all test takers…..suitable for some but not all)
- Items may not be representing the domain or content
- Time limit (may cause stress or a handicap)

b. Test Administration Process:

- Poor and not uniform testing conditions (physical setting and environment)
- Poor, faulty, improperly worded, improperly delivered, not uniform instructions
- Test administrator's personality (different administrators in different situations having different personality styles)
- Rapport (poor, too much, or too little)

c. Examinee Related Variables:

- Prior learning and experience
- Individual differences (within group differences; Personality styles, stress tolerance level, IQ level, emotionality, motivation, knowledge)
- Difference from the normative sample (no group is identical to the normative sample; every group is different from every other group)
- Within- person differences (the same persons may change over time (life, academic, and professional experiences; physical conditions, health, motivational level, emotional state etc.)

The test developers, psychologists, or other professionals involved in test construction try their best to make sure that these variables are controlled or kept constant in test administration. But bringing in complete consistency in the testing process is not possible, and there is no test that is hundred percent reliable. However, the test developers do report the coefficient of reliability of their measures as well as the characteristics of the normative sample from which the coefficient was obtained. This provides a guideline to the test users about the type of people or groups to whom the measure may be administered.

Correlation and Reliability:

We understand that reliability is about the consistency of scores of the same persons on same measures in different conditions; or it is about the consistency observed in the scores of different persons or different groups having similar characteristics on the same measure. Calculation of reliability involves the concept of correlation. Coefficient of correlation is the value yielded by the calculation of correlation. Coefficient of correlation is denoted by letter 'r'. This coefficient indicates the relationship between two variables. In the context of testing, it expresses the correspondence between two scores. As students of psychology we know that the coefficient of correlation tells us two things about a relationship; the magnitude of relationship and the direction of relationship. Magnitude means the size of correlation whereas direction indicates whether the correlation is positive or negative.

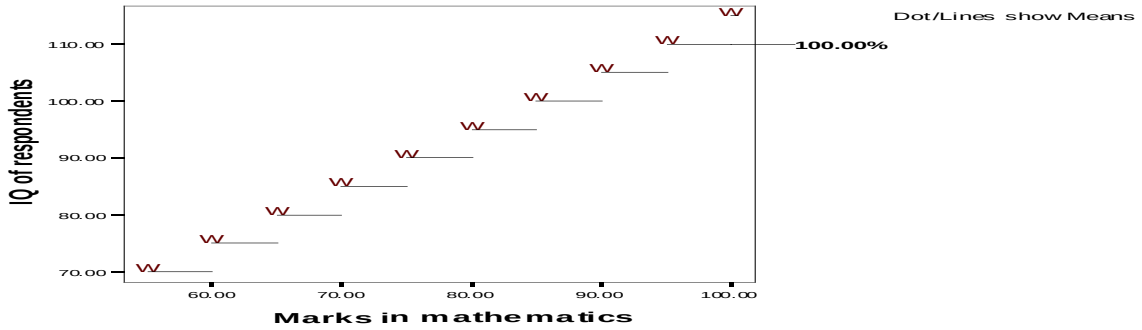
The size of a correlation ranges between -1 and +1, with a zero value in between. The value of a coefficient of correlation is always one or less than one. A coefficient of correlation of +1 means a perfect positive correlation. A coefficient of correlation equal to -1 indicates a perfect negative correlation, whereas a zero value indicates no correlation. The closer is a value of correlation to one, the stronger is the correlation e.g. r= 0.9. The closer a coefficient of correlation is to zero, the weaker will be the relationship. If scores on two sets of scores increase and decrease together then it is a positive correlation. On the contrary, if when scores on set-I increase the scores on set-II decrease then it is a negative correlation.

The concept of correlation can best be understood by looking at the concept of a scatter plot or graph. If we plot values of two sets of scores in a graph, then the appearance of the graph or the scatter of the scores will indicate the relationship between the two sets of scores. If the person who scored the lowest in set-I is also the one to score the lowest on set-II, and the one having the highest score on first set is also the highest scorer on set-II; and if every other person has the same position on both sets then it is a perfect positive correlation. On the other hand if the situation is the other way round i.e., the lowest scorer on set-I is the top scorer on set-II, and the top scorer on set-I is the lowest scorer on set-II, and all other persons' scores follow the same pattern, in the same order on both sets, then it shows a negative correlation.

For example look at the following sets of scores obtained from a group of 10 students regarding their IQ, marks in annual exam of English and math, and the classes that they had missed in a year.

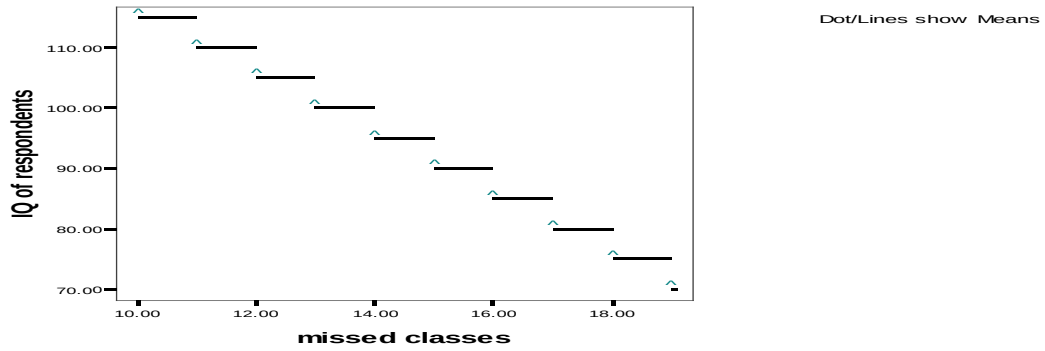
IQ	Marks in Math	Marks in English	Missed classes
70.00	55.00	20.00	19.00
75.00	60.00	13.00	18.00
80.00	65.00	45.00	17.00
85.00	70.00	6.00	16.00
90.00	75.00	67.00	15.00
95.00	80.00	78.00	14.00
100.00	85.00	20.00	13.00
105.00	90.00	12.00	12.00
110.00	95.00	.00	11.00
115.00	100.00	32.00	10.00

When these scores are plotted in a graph, we see that a perfect positive correlation exists between the IQ and the marks in math. Everyone has obtained exactly similar scores on the two measures. A graph moving from the lower right corner to the upper left corner expresses a positive correlation.



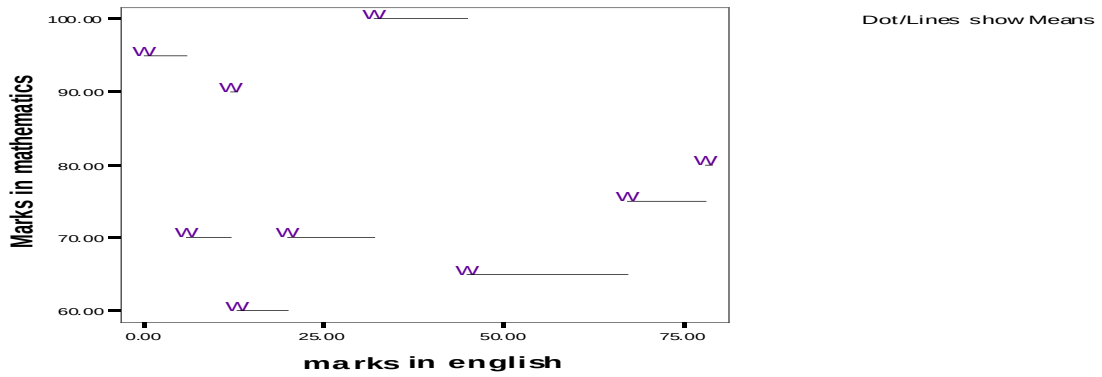
Positive correlation: IQ and marks in math

A perfect negative correlation can be seen in case of IQ and classes missed in an academic year. Every one's scores on IQ are moving in the opposite direction of the number of classes that they had missed. A graph moving from the upper right corner to the lower left corner expresses a negative correlation.



Negative correlation: IQ and classes missed in a semester

The following graph shows a zero correlation between marks in English and marks in math. No significant trend of scores can be seen from the graph.



.

Zero correlation: Marks in math and marks in English

Although more than one way of computing the coefficient of correlation is available the most commonly used procedure is the Pearson Product moment method. The size of correlation coefficient generally acceptable for the purpose of reliability is around .80 or .90. In order to see if our obtained coefficient of correlation is significant we can use tables of significance of correlation found at the end of most statistics and research math textbooks. In case the data have come from a large number of people a coefficient less than .80 may also be acceptable. However a coefficient less than .60 is usually not acceptable.

Lesson 14

Types of Reliability

Reliability of a measure can be measured in a number of ways. The selection of method of computing a coefficient of reliability depends on what reliability is understood to mean is it the stability of scales over time, is it consistency between items, or something else.

According to the definition given by Anastasi & Urbina (2007) reliability refers to "the consistency of the scores obtained by the same persons when they are reexamined with the same test on different occasions or with different sets of equivalent items, or under other variable examining conditions".

Scrutinizing this definition one can see that reliability is about consistency of scores that may be obtained in of any of these forms:

- Consistency of the scores obtained by the same persons when they are reexamined with the same test on different occasions, or
- Consistency of the scores obtained by the same persons when they are reexamined with different sets of equivalent items, or
- When the two previous strategies are used under other variable examining conditions

In this section we will discuss some commonly used types of reliability.

a) Test- Retest Reliability:

As the very name suggests, test- retest reliability deals with two performances of the same test by the same persons on two different occasions. If reliability refers to the consistency and stability of scores over time then it will be measured using this method. The test- retest coefficient is also known as the 'coefficient of stability'. The test takers' scores on first administration of the test are correlated with their scores obtained on the second administration of the same test. a major advantage of this method is that there is maximum control over the test taker and test item variables. The subjects are the same and the test items are also the same.

This type of reliability does take into account the errors of measure that may affect the findings. The possible sources of error primarily include three types of variable; those pertaining to the test taker, those related with the testing conditions, and those stemming from the very nature of the test. Some of the sources of possible sources are as follows:

- It may happen that the testing conditions do not remain the same as they were on the first occasion. Physical and environmental changes may have taken place the second time e.g. noise, too high or too low temperature.
- Many changes may take place in the test takers. This is perhaps the major drawback of this method which is observed in most situations. Bodily, emotional, motivational, social, financial and other changes may interfere with the administration and test taking on the second occasion.
- Fatigue effect or practice effect may account for error variance. The test taker may find it boring to repeat the test taking process; alternately, he/she may be facilitated on the second occasion because of familiarization with the content.
- The test may be of such type that repeated measurement would affect its nature. The test taker may have learnt the way to attempt and solve problems for example. Some may remember the test items. This factor may seriously affect performance on items measuring analytical ability, logical reasoning, spatial reasoning, and even mathematical or verbal ability.
- Length of the time period between the two administrations may also matter. It has been seen that this interval affects the magnitude of a test- retest reliability coefficient. If the time gap between the two administrations is short the magnitude of the coefficient will be larger than that computed after a longer duration. That is why test developer should report test- retest reliability with reference to the interval in between two administrations of the same tests. It is always better to do the second administration after a short time gap i.e., a week or a few weeks only.

Test- retest reliability is denoted as $r_{tt.}$

b) Alternate-Form Reliability:

In order to overcome the problems caused by using the same version of the test on two occasions a second approach is often adopted. This involves computing the alternate- form reliability. In this approach the test developer develops two alternate or parallel forms of the same test. Ideally speaking the two forms should be

independently developed and completely *parallel* or *equivalent forms* of the same measure. They should match each other in all respects including:

- Same specifications
- Same instructions
- Same time limit
- Same content
- Same number of items
- Same item format
- Difficulty level

A number of other test variables may also have to be considered in constructing such forms e.g. group or individual administration, vocabulary, terminology, examples used etc. In short the test developer has to make sure that the alternate forms are replaceable by each other's.

When alternate forms are used in immediate succession, without a time gap, then the reliability coefficient obtained is the 'coefficient of equivalence' or 'parallel form coefficient. This tells us about the correlation between the scores on two alternate forms. This is reliability of the measure across two forms. But we are also interested in reliability across occasions i.e., coefficient of stability. In order to compute stability over time the two forms are administered with a reasonable gap in between. The following pattern of administration may be adopted in order to obtain a 'coefficient of stability and equivalence':

	Form A	**Form B**
Occasion I	**Group I** (50% of total)	**Group II** (50% of total)
Time interval		
Occasion II	**Group II** (50% of total)	**Group I** (50% of total)

The sample or the group to be tested is divided into two groups and the groups are administered the two forms alternately. Group-I takes Form A first and group-II takes form B first. The process is repeated with alternate forms on the second administration after a time interval. This procedure makes it possible to calculate a correlation that yields 'coefficient of stability and equivalence'. This procedure considers two sources of error; use of two different sets of items, and time interval between two administrations.

Although a good approach to computing reliability and used more commonly than test- retest approach, alternate forms reliability also has some shortcomings. Firstly it is not easy to prepare to forms of the same test that are equivalent in all respects. Secondly, the general practice effect after taking one form of the test may affect performance the on the second form. Nevertheless, this approach is considered to be a good estimate of reliability.

c) Split- Half Reliability:

Alternate form reliability is a popular type of reliability but with some obvious limitation; Construction of two completely parallel forms is not easy and requires a lot of time and effort; even when to alternate forms are available, the practice effect and prior experience may affect performance on the second occasion. The time gap between two administrations is another intervening variable. To overcome these problems, in fact to remove these problems, split- half reliability is computed.

In this approach, the test is assumed to be divided into two equal halves. When it is administered two scores are obtained for each test taker. The test is arbitrarily divided into two halves. Scores on one half are correlated with scores on the other, using Pearson's r. The obtained value is the 'coefficient of internal consistency'. In this approach error caused by the temporal variation or different sets of items is controlled.

The test can be divided in many ways:

Divide test into equal number of items. The first 50% items are considered as one half and the remaining 50% as the other half. Scores are calculated separately for the two halves. There is one limitation of dividing the test this way. There is likelihood that the two halves may be having different difficulty levels. The initial items may be easier than the items in the second half. Also, items in the two halves formed in this way may pertain to different content areas

- A better option can be to consider odd numbered items in one half, and the even numbered items in the other half. This can ensure that the items in the two halves have similar difficulty level and are closely similar in terms of content area. In fact the test should be constructed keeping in mind that every pair of items will be split in two items to be considered in separate test halves.
- If the test contains any such items that will appear only once e.g., specific problems to be solved, specific concepts, diagrams etc., then it is not possible to have the same or similar items in both halves. This may affect the reliability coefficient.

The reliability coefficient is computed using the regular procedure for calculating correlation. However, as one can see the calculated coefficient is for half of the test because half of the items were treated as set X and half as set Y. The length of the test affects the reliability of the test. The half test relationship is adjusted using Spearman – Brown formula. This formula makes it possible to estimate the internal consistency reliability of the test from the reliability of the two halves of the test. The formula is used for the computation of reliability while taking care of the fact that a test has been shortened or lengthened. The formula also considers the number of times the test has been shortened or lengthened. Spearman- Brown formula is given below:

$$r_{nn=}\frac{nr_{tt}}{1+(n-1)r_{tt}}$$

rnn= estimated coefficient
rtt= obtained coefficient
n= number of times the test has been lengthened or shortened e.g. two times, three times etc.

In split- half reliability the test is reduced to half, therefore the reliability of the full test is calculated using the Spearman- Brown formula that involves doubling the length of the test, as follows:

$$r_{tt=}\frac{2r_{hh}}{1+2r_{hh}}$$

r_{hh} = Reliability of the half tests
See Anastasi & Urbina (2007) for further reference.

d) Kuder- Richardson Reliability:
Kuder and Richardson (1937) introduced a procedure that measures reliability through measuring inter item consistency. It is similar to split- half method in the sense that measurement of reliability involves a single administration of a single test. However the two approaches are different in the way consistency is measured. Rather than dividing the test in two halves, the Kuder- Richardson method is based on an examination of performance on each item (Anastasi & Urbina,2007). It in fact is the mean of all split- half coefficients that can be obtained from a test.

$$r_{tt=}\left(\frac{n}{n-1}\right)\frac{SD_t^2-\sum pq}{SD_t^2}$$

r_{tt}= Coefficient of reliability of whole test
n= the number of items in the test
SD_t^2= Standard deviation of total scores on the test is SD_t, squared (=variance)
p= proportion of persons who pass each item
q= proportion of persons who do not pass each item
pq= product of (p) and (q) for each item
Σpq= The sum of products of (p) and (q) for all item

e) Coefficient Alpha:
Another approach quite similar to the Kuder- Richardson technique is to calculate 'coefficient alpha'. Kuder-Richardson formula can be used only for tests in which items are scored as either zero or one. It is not applicable to tests where answers to items are assigned two or more scoring weights e.g. personality inventories where a number of response options with attached weights are available (never=0, occasionally=1, often=2 and so on). Coefficient alpha is a general formula that caters for such tests.
The following formula is used for calculating this coefficient:

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

r_{tt}= Coefficient of reliability of whole test
SD_t^2= Standard deviation of total scores on the test is SD_t squared (=variance)
SD_i^2= Variances for item scores
$\sum^{(SD_i^2)}$= sum of the variances of item scores

Inter-scorer Reliability / Inter-rater Reliability:

At times the test may have to be evaluated by more than just one examiner. In case of objective tests a test taker will obtain the same marks no matter who, and how many people, evaluate it. On the other hand when essay type or other open- ended items are to be evaluated, a need for multiple scorers may arise. When more than one persons are evaluating inter-scorer reliability is required. A number of procedures are available for this type of reliability. Some common procedures include the following:

- Inter-rater or inter-scorer reliability: Two examiners evaluate and mark a test of a group of test takers. The resulting two sets of scores are used for computing correlation between the two sets.
- Intra-class coefficient/ coefficient of concordance: Two or more examiners score the performance of a number of test takers and the coefficient is computed from these scores.

Lesson 15

Reliability in Specific Conditions and Allied Issues

Looking back at whatever we have discussed so far, you have learnt the concept and application of reliability as well as various ways of computing reliability. We know that reliability is about the consistency of the scores. The coefficient of reliability indicates the extent to which we can expect to obtain the same scores from the same group of people on different occasions; or on two sets of very similar items; or under other similar conditions.
We have seen that there are a number of methods available for calculating reliability, including:

a) Test- Retest reliability
b) Alternate- Form Reliability
c) Split- Half Reliability
d) Kuder- Richardson Reliability
e) Coefficient Alpha

Method (a) of reliability tells us about the consistency of the scores obtained by the same persons when they are reexamined with the same test on more than one occasion. Method (b) is about the consistency of the scores obtained by the same persons when they are reexamined with different sets of equivalent items. When the two previous strategies are used under other variable examining conditions then we use methods (c) to (e).
No matter in what situation the tests are used, the basic methodology for computing reliability will be chosen from these methods. However the manner in which these methods are used may vary in certain specific cases.

Some Factors Influencing Reliability of a Test:

Reliability of Oral Tests:
The issues of inter scorer reliability is a significant matter in case of oral tests. In such tests, the examinee orally answers the questions/items and the raters/scorers/ or examiners make the assessment. If we are using written tests particularly the objective ones, inter scorer reliability is not a matter of concern. An examinee will obtain the same score no matter who marks the test because answers to all the questions are provided and every answer will either be right or wrong. Similarly, scoring of oral test can be problematic, considering inter-rater or inter scorer reliability. The reliability of oral test is usually lower than the reliability of equivalent written tests. However, a number of measures can be adopted in order to improve their reliability. Some experts have suggested that the test takers should be told to give their responses only after they have thought about the question for a while (Meredith, 1978).The use of electronic recordings of the examinee's responses has also been suggested. These recordings can be played back later on and can be used for reevaluation by the examiners or raters.
Some other recommendations include:

- Oral tests should be carefully designed
- Before the actual administration of the test starts, some model questions should be constructed
- More than one rater should be used

There is evidence available to suggest that when these factors where considered the obtained reliability coefficients were in the .60s and .70s (Carter, 1962; Levine, & McGuire, 1970; Hitchman, 1966).

Reliability of Speed Tests:
Speed tests are one of those tests whose reliability deserves extra concern and care. We know that some tests are by nature speed tests and some power tests. As the very name suggests speed tests focus on the speed with which a person completes a test. There is a specified time limit for taking the test and it is up to the speed of the test takers whether they can complete the test in the allocated time or not. The test items are not difficult, in fact quite easy. What makes such tests difficult is the time allowed for the test.
A good description of such tests has been given by Anastasi & Urbina (2007, p.116):
"A pure speed test is one in which individual differences depend entirely on speed of performance. Such a test is constructed from items of uniformly low difficulty, all of which are well within the ability level of the persons for whom the test is designed. The time limit is made so short that no one can finish all the items. Under these conditions, each person's score reflects only the speed with which he or she worked."
The power tests are somewhat different in nature from these tests. In a power test the test takers are given enough time for completing the test. However the difficulty level of the items is such that no one can attempt all the items correctly. In Anastasi & Urbina's (200, p. 116) words:

"A power test, on the other hand, has a time limit long enough to permit everyone to attempt all items. The difficulty of items is steeply graded, and the test includes some items too difficult for anyone to solve, so that no one can get perfect score."

One can realize from the description of these tests that they have been designed in such a manner that they prevent people from attaining a perfect score. This is quite understandable because tests aim to differentiate between those who know and those who do not; those who can correctly solve a problem and those who can not. If all tests are doable for everyone then what are we trying to test?

Rather than being purely speed or purely power tests, most tests contain both features. The degree and extent in which these features are present varies from test to test. The test manual should provide this information alongside information regarding the reliability of a test. Whether a test is speed test or not affects the reliability and the procedure for computing reliability of the test. Some procedures of computing coefficient of reliability are not suitable for speed tests.

Split half procedure using odd- even format is not suitable for speed tests. These tests, we know, contain uniformly easy items. So whatever number of items does a person attempt correctly may, hypothetically speaking, mean that the person has correctly done equal number of odd and even numbered items. This is because items have a uniform level of difficulty. This is an extreme example. However one may expect to get similar results in actual situations. Going back to the hypothetical situation, we can see that if we will calculate the reliability coefficient using split- half odd-even method, then we may end up with an inflated coefficient, may be r= +1. The reason is that anyone with a low score on odd list will have the same score on even list as well. The test may not be actually that highly reliable, but the nature of the test was such that an inflated value was obtained. Therefore this method should not be used with speed tests. It is always better to use reliability procedures where repeated administrations tale place. The following approaches may be used for calculating the reliability of speed tests, while keeping in view the nature of the test and applicability of the procedure:

a) Test- retest reliability
b) Equivalent- half reliability
c) Variations of Split- half method; the two halves should be administered separately as if two separate tests were used, with half of the total time for each segment. For example if the complete test contains 50 items and time allowed for the test is 60 minutes then the following pattern may be used:

Odd numbered items	Even numbered items
Half- I Items= 25 Time= 30 minutes	Half- II Items= 25 Time= 30 minutes

Relationship between the Sample Tested and Reliability:

The characteristics of the sample from which reliability coefficient has been obtained may affect reliability of a measure. Basically two such characteristics are important with reference to reliability:

- Variability
- Ability level

Variability: Variability of score, or in other words individual differences, make a difference to reliability coefficient. We know that the measurement of reliability involves the use of correlation. The coefficient of reliability is based on the coefficient of correlation. Correlation can be calculated when scores vary, increase or decrease, together as well as independently. If the scores in one group are identical or very similar, the resulting value of correlation may turn out to be zero. Therefore, the sample from which we calculate reliability should have a wide range of scores, or a wider variety of test takers. The test manual should provide detailed information about the sample used for computation of reliability.

Ability level: Whereas lack on variability may cause problems, too much variation between groups may also be a source of error in calculation of reliability. For example if the sample consists of subgroups that are very different in terms of their ability level (particularly in tests where difficulty level and prior learning or skills matter) then it may affect the distribution of scores from which correlation is to be calculated. To control this factor the subgroups within the larger group should be chosen with great care.

Standard Error of Measurement (SEM):

SEM is defined as "an index of the amount of error in a test or measure. The standard error of measurement is a standard deviation of a set of observations for the same test" (Kaplan & Saccuzzo, 2001). SEM is one way of expressing reliability. The following formula is used for calculating SEM. If we have the reliability coefficient of the test then we can calculate SEM using a simple formula.

In this formula SD refers to the standard deviation of the test score; r tt is the reliability coefficient.

$$SEM = SD_t\sqrt{1 - r_{tt}}$$

The SEM value indicates the standard deviation of the normal distribution of the test scores, which is hypothesized to be attained if it were taken by a test taker for infinite number of times. It is assumed that the mean of this hypothetical distribution is the true score of the person. The calculated SEM will be the standard deviation.

For example, if a test had a standard deviation of 12 and the reliability coefficient is .90, then SEM will be= 12√ 1- .9= 3.79 .The SEM is the standard deviation of the hypothetical distribution. If XYZ attains an average IQ 0f 90 in the hypothetical situation it will be her true score that may range between 90- 3.79 and 90+ 3.79.we remember that in normal distribution about 68% people fall between -1 and +1 standard deviations from the men. Therefore applying the same concept we can say that 68% people may be expected to score within this range of scores.

Acceptable Values of Reliability:

The closer the reliability coefficient is to one, the better is it considered. However, for most research purposes coefficients within the range of .70- .80 are considered good enough (Kaplan & Saccuzzo, 2001).

Enhancing Reliability:

Increasing the number of items in a test may improve the reliability of a test. Also, the test items should be carefully phrased and it should be made sure that all items measure the same content which is to be measured by the test in question.

Lesson 16

Validity

Validity is an essential ingredient of a test. The use of a test that is not valid will not only be a waste of time and energy, but it may also result in erroneous judgments and decisions. Validity, just like reliability, is a characteristic and quality that we seek for in most life situations where we have to acquire something, make it a part of our life, and base decision making on that. For example if we need to buy an air conditioner, we would like to make sure that it cools the room, providing the service that it is meant to provide (validity); we would also like to ensure that every time we will switch it on it will turn on and start working (reliability). If the machine fails in cooling, and/or does not turn on when we switched it on, then we do not need it; if we already have acquired it then we would try to get rid of it.

Take an extreme example of human relationships. Think about making a friend. What are the characteristics that you want to see in your friend, and you may not make someone your friend if he/she does not possess them? Validity and reliability besides may be some others. We want our friends to be our 'friends' in the true sense, warm, sympathetic, understanding, caring etc i.e., what a friend ought to be (validity). On the other hand we like that our friend is there to support us, help us, and be with us every time we needed him/her (reliability).

Similarly psychological tests need these features as an essential integral part.

Definition:

- "Traditionally, the validity of a test has been defined as the extent to which a test measures what it was designed to measure" (Aiken, 1994, p.95).
- "The extent to which a test measures the quality it purports to measure. Types of validity evidence include content validity, criterion validity, and construct validity evidence" (Kaplan & Saccuzzo, 2001, p.640).

From these definitions one can realize that validity of a test is about the nature of a test with reference to the content of the test as well as the content or domain in which it is rooted.

It is about what the test measures and how well it measures. Although the title or name of a test gives us clue of what the test measures, we may not have accurate idea of what it actually measures until and unless we have appropriate information in this regard. This information may be reported in the test manual or may be calculated by us.

Whether or not a test measures a trait that it claims to measure can be determined only through an examination of the objective sources of information and empirical operations utilized in establishing its validity. A test's validity must be established with reference to the particular use for which the test is being considered.

The validity of a test should not be taken casually, and cannot be reported in general terms either. Validity is not, and should not be measured and reported as "highly" valid or "low" validity. It is reported in specific terms. Procedures for determining test validity are concerned with the relationships between the test content and the domain that it represents; between the test content and the objectives that lead to test construction; or test performance and performance on other independent measures that are known to measure the same phenomenon/trait/ability/domain etc.

Psychological tests are developed primarily for two reasons:

- Achievement testing: In order to see if a specific content area has been learnt and/ or mastery is acquired, and
- Predictive function: predicting future performance on the basis of the present test's performance.

In both cases the ultimate objective cannot be achieved if the test is not valid.

An achievement test is evaluated by comparing its content with the content domain it is designed to assess. These tests may be used as end-of-course examination in school and by licensing tests for driving a car or qualifying with a skill for a specified occupation. These focus on content domain.

The performance tests used to assess and predict future behavior are designed on certain criterion. To measure the validity of such tests the correlation coefficient between test scores and a direct and independent measure of that criterion is used.

Another area with reference to validity is about testing the construct in a particular test. Constructs are broad categories, derived from the common features shared by directly observable behavioral variables.

Considering these uses and objectives of tests, one can see that there are three broad categorizations of tests possible. Subsequently three types of validity are measured:

Content Validity Evidence:
"The evidence that the content of a test represents the conceptual domain it is designed to cover" (Kaplan & Saccuzzo, 2001, p.635).

Construct Validity Evidence:
"A process used to establish the meaning of a test through a series of studies. To evaluate evidence for construct validity, a researcher simultaneously defines some construct and develops the instrumentation to measure it. In the studies, observed correlations between the test and other measures provide evidence for the meaning of the test" (Kaplan & Saccuzzo, 2001, p.635).

Criterion Validity Evidence:
"The evidence that a test score corresponds to an accurate measure of interest. The measure of interest is called the criterion" (Kaplan & Saccuzzo, 2001, p.635).

Content-Description Procedures:
Nature:

- A systematic examination of the test content is made in order to determine whether it covers a representative sample of the behavior domain to be measured.
- Such validation procedures measure how well the individual has mastered a specific skill or course of study.

"Content validity is built into a test from the outset through the choice of appropriate items. For educational tests, the preparation of is preceded by a thorough and systematic examination of relevant course syllabi and textbooks, as well as consultation with subject matter experts. On the basis of the information thus gathered, test specifications are drawn up for the item writers" (Anastasi & Urbina, 2007, p, 129)

- Certain considerations are needed for content validation measure:
  - The test shod cover all major aspects and items in correct proportion (the overloading and under-representation of elements should be avoided).
  - The domain under consideration must be fully described in advance rather than after test preparation.
  - Content should be broadly covering the major objectives, application and interpretation and also factual knowledge.

Specific Procedures:
The choice of appropriate items is important in content validity. A number of careful decisions are taken at the time of test development e.g.:

- For educational tests, items are prepared form relevant course syllabi and textbooks and in some cases consultation with subject experts is used.
- *Test specifications* are given to item writers.
  - It includes instructional objectives, relative importance of topics/processes.
  - It should clearly indicate the number of items for each topic.
  - It may also provide sample material.
- The process of content validation should include description of all procedures in the manual; that ensure that test is appropriate and representative.
  - For example in subject-specialists has participated, the names, qualification and number of individuals should be stated.
- The test total score and achievement score of an individual can be checked for grade progress.

Application:

- Content validity is basic to the validity of educational and occupational achievement tests.
- This validity measure is applicable to occupational tests designed for employee selection and classification.
- For aptitude and personality test content validation in inappropriate.

Content Validity versus Face Validity:

We should not confuse content validity with face validity. Many students of psychology take them to be one and the same, whereas actually they are different. Face validity is about the test appearing to be valid. It is not validity in the technical sense; nor is it technically measured. Face validity is defined as

"The extent to which items on a test appear to be meaningful and relevant, actually not evidence for validity because face validity is not a basis for inference" (Aiken, 1994, p.636).

Face validity may be used where prediction r inference are not involved e.g. questionnaire or scale for a survey.

Criterion Prediction Procedures:

Criterion validity evidence is "The evidence that a test score corresponds to an accurate measure of interest. The measure of interest is called the criterion" (Kaplan & Saccuzzo, 2001, p.635).

Lesson 17

Criterion Validity

Validity is no doubt the most essential feature of a test. If one were in a situation where one had to choose only one primary ingredient of a psychological test, one will have to opt for validity. A test can serve our purpose without being reliable and without having established norms, but a test will be useless without being valid. A test can be reliable without being valid, but a test cannot be valid without being reliable. This is because of technical reasons, but this also supports what has been said in previous lines. A valid test will be reliable too, but it is not necessary that a reliable test will always be valid. Therefore if a test is highly reliable but does not measure what it is supposed to measure then it is a useless test, at least for the use for which it was planned.

We have discussed that validity is about the extent to which a test measures what it is meant to measure. The validity of a test can be assessed in more than one way:

a. Assessment of validity by analyzing the content of the test. If the test content represents the domain which is to be tested, and is according to the specifications and/or the instructional or behavioral objectives laid down at the time of test planning, then the test is considered valid.
b. Finding out the relationship between the test scores and scores on a relevant criterion. The idea is that the test and the criterion serve the same purpose, and if there is a positive correlation between the two then the test in question is valid.
c. Assessment of the validity of the test through an examination of the particular constructs measured by the test.

The first approach (a) refers to the content validity of the test; b refers to criterion related validity; and the third one (c) describes construct validity.

In the previous section we discussed content validity in detail and saw how it is assessed.

In the present section we will look into the criterion related and construct based validity.

Criterion-related Validity:

Whenever we plan and design a test we have a certain standard in mind that we want to meet. We relate score on our test with those achieved on the criterion or standard.

One approach to assessment of validity is through comparing the test results with those on a criterion.

Why do we need a criterion to estimate the validity of our measure? To understand this concept let's look at two simple everyday examples. Consider a situation where you are looking for a foot ruler to measure the length of a large sheet of paper. Someone brings a stick to you and tells you that the stick is foot long and you can use it instead of a ruler. You doubt this and want to ensure that this stick is really foot long. Now you look for another object that you are sure is foot long, though cannot be used for measuring another object. You see that the tiles in your kitchen are 6x6 inches square. You measure the stick against the tiles and you see that the stick is equal to two tiles in length i.e., 12 inch long. What you have done here is, you measured your tool with reference to a criterion and the result of this exercise showed that your tool was valid. You can use it instead of a ruler.

Take another example. Your friend claims that he is 6.5 feet tall. You want to test if he is right but you do not have an inches tape or another tool to measure. Then an idea strikes your mind that the doorway to your room is 6.5 feet high. You ask your friend to stand by the door. If his height is the same as the doors, then you can say that your friend made a right claim. The door was a criterion that you used as an indicator of height.

Criterion- related validity refers to a procedure where scores o a test being used are correlated with scores on a criterion. Criterion validity evidence can be defined as "The evidence that a test score corresponds to an accurate measure of interest. The measure of interest is called the criterion" (Kaplan & Saccuzzo, 2001, p.635). In other words, for example, if a teacher develops a test of mathematical ability, how is she going to determine its validity? How will she judge that her test is measuring mathematical ability and nothing else? Most probably she will try to examine the relationship between scores on her test with another authentic test of the same ability, or may be a test of similar skills. By doing so, she will be assessing the criterion- related validity of her test.

The criteria used for this purpose may be in the form of scores on psychological tests, mental and behavioral measurement, classifications, grades, teacher's or supervisor's ratings etc.

According to Aiken (1994, p. 96) "Traditionally, however, criterion- related validity has been restricted to validation procedures in which the test scores of a group of examinees are compared with ratings, classifications,

or other behavioral or mental measurements. Examples of criteria against which tests are validated are school marks, supervisor's ratings, numbers of dollar amounts of sales."

Criterion Prediction Procedures:

Two procedures may be adopted for this purpose:

- Predictive validity evidence
- Concurrent validity evidence

Predictive Validity Evidence: "The evidence that a test forecasts score on the criterion at some future time"(Kaplan & Saccuzzo, 2001, p.638). Predictive validity pertains to prediction over a time interval.

Predictive validation yields information that is most relevant to tests that are used for predicting future performance of an individual. There are situations where tests are used to classify, select, and even reject individuals. For example tests are used for selecting students for an academic program, for personnel selection and job placement, or for choosing staff members for a specific training or higher education. Similarly there are times where rather than being used for selection; tests are used for screening out individuals. The tests are used to predict if the candidate is likely to develop an undesirable behavior, characteristic, mental or emotional problem etc. in future.

Concurrent Validity Evidence:

In many situations estimating validity on predictive validation basis is not too practical. It involves a certain time interval and requires a suitable problem selection sample. In such situations it is preferable to use the concurrent validation approach. In this approach the test is administered to such a sample whose data on the criterion are already available.

Concurrent valid evidence is defined as "evidence for criterion validity in which the test and the criterion are administered at the same point in time" (Kaplan & Saccuzzo, 2001, p.635). When the validity of a test is measured with reference to a criterion, and both are administered at nearly the same time, then the resulting validity will be concurrent validity.

Concurrent validity approach is used in situations where a test is administered to people in various categories. This procedure helps determine if there are significant differences between the average scores of different categories of individuals. The primary idea is whether this test can differentiate between people with specific characteristics. If the test can do so, then it can be used as an efficient tool for categorizing people according certain criteria or classifications. One of the best examples of such a tool can be diagnostic tests e.g. Minnesota Multiphasic Personality Inventory (MMPI).

The criterion data are always available at the time of testing.

Anastasi & Urbina (2007) have described the difference between predictive validation and concurrent validation in a very clear cut manner:

"The logical distinction between predictive and concurrent validation is based not on time but on the objectives of testing. Concurrent validation is relevant to tests employed for diagnosis of existing status rather than prediction of future outcomes. The difference can be illustrated by asking 'Does Smith qualify as a satisfactory pilot?' or 'does smith have the prerequisite s to become a satisfactory pilot?' The first question calls for concurrent validation; the second, for predictive validation."

Criterion Measures:

There is no limit to the type of criterion that may be used for the purpose of test validation. However some are used more frequently and popularly than others. Tests may involve a number of criteria. Some of the criteria are mentioned here:

a. Academic achievement ; used for intellectual tests and similar tests
b. Matriculation marks at the time of admission to higher classes; used for selection and screening.
c. For those who have completed, or discontinued education, years of education can be good criteria.
d. Performance in specialized training; used for selection, admission, achievement.
e. Instructor's ratings
f. Grades in a semester

Characteristics of a Criterion:

According to Cohen (1999) criterion should be:

- ***Reliable;*** the scores on the criterion as well as those on the test should be reliable. The reliability coefficients of the two sets of scores affect the coefficient of validity. This can be understood from the following pattern of relationship:

$$r_{xy} \leq \sqrt{(r_{xx})(r_{yy})}$$

R_{xy} stands for the coefficient of validity; test reliability and criterion reliability are represented by $r_{xx \text{ and }} r_{yy.}$ It can be seen that the two coefficients of reliability contribute to the validity coefficient. Weak/little/or no reliability will automatically affect validity.

- ***Valid;*** it should measure what it is supposed to measure.
- ***Relevant;*** it should be relevant to the purpose for which it is being used.
- ***Uncontaminated;*** we should try to use uncontaminated criterion. There are situations when the criterion itself has been, partially or fully, based on predictor measures. This is called criterion contamination. For example you use a test as a criterion for the diagnosis of patients examined in a psychiatric clinic. A probe into the diagnostic procedure indicates that the diagnosis itself involved the use of the same test. In this case the criterion will be contaminated and will not be a suitable criterion.

Lesson 18

Construct Validity

Some years back I developed a test for the assessment of science concept acquisition, called SCAT. The purpose of developing this test was to develop a tool that could measure the extent to which children in the final year of their school have learnt the application of science concepts that they were taught in the past 4-5 years. The reliability and validity of the test were also measured. In order to see if the test was valid or not, three assumptions were formulated:

- The SCAT scores will have a high-positive correlation with scores on a test of similar nature and content.
- The SCAT scores will have a high-positive correlation with scores on a test of reasoning ability.
- The SCAT scores will have a moderate-positive correlation with marks on school science achievement test.
- The SCAT scores will have either no or a negative correlation with scores on a test of rote memorization of meaningless material like non-sense syllables.

In order to test these four assumptions, four different tests were used for these assumptions respectively: Dallas Times Herald Test (scale test), Basic Reasoning skills Test, General Science Test used in school's annual examination, and list of non-sense syllables.
A sample of school children were administered the main test, SCAT, and the other tests. Correlation between SCAT scores and the four other sets of score was computed. The resulting information was used as the validity index of SCAT.
You are familiar with two other types of validity, content and criterion- related validity. In the computation of those two forms of validity other tests were not used the way they have been used in the case of SCAT. The procedure described here pertains to a third approach to calculating validity of a test. This is called construct validity.

Construct Validity Evidence:
Construct validity is measured keeping in view the particular construct that the test is supposed to measure. If the test is found to measure the construct in question then it is a valid test, and it does not measure the trait then the test is not accepted as valid.
Construct validity evidence refers to "A process used to establish the meaning of a test through a series of studies. To evaluate evidence for construct validity, a researcher simultaneously defines some construct and develops the instrumentation to measure it. In the studies, observed correlations between the test and other measures provide evidence for the meaning of the test" (Kaplan & Saccuzzo, 2001, p.635).

Another description of construct validity has been given by Cohen; "construct validity is a judgment 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 constructed to describe or explain behavior". "Constructs are unobservable, presupposed (underlying) traits that a test developer may invoke to describe test behavior or criterion performance" (Cohen, 1999).

Anastasi & Urbina (2007) have provided an all-encompassing description of construct validity: "The construct validity of a test is the extent to which the test may be said to measure a theoretical construct or trait. Examples of such constructs are scholastic aptitude, mechanical comprehension, and verbal fluency, speed of walking, neuroticism, and anxiety. Each construct is developed to explain and organize observed response consistencies. It derives from established interrelationships among behavioral measures. Construct validation requires the gradual accumulation of information from a variety of sources."
The significant elements of this description are as follows:

- A theoretical construct or trait.
- Each construct is developed to explain and organize observed response consistencies.
- It derives from established interrelationships among behavioral measures.
- Requires the gradual accumulation of information from a variety of sources.

These elements give us an idea about the steps involved in construct validation. But before looking into the procedure of construct validation, let us see what underlying assumption underlie this approach.

Assumptions Underlying Construct Validity:

a. Test scores will be highly correlated with tests measuring same construct/similar construct. This is convergent validity.
b. Test scores will have weak/low (or at times may be negative) correlation with tests meant to measure construct which are different from the one measured by the main test. This is discriminant validity.

These assumptions end up in the calculation of two type of construct validity, convergent and discriminant validity.

Convergent Validity Evidence: This aspect of validity shows us the extent to which a test measures the same construct/trait/ attribute as do other measures that have been designed and developed to measure the same construct/trait/ attribute

Discriminant Validity Evidence: Discriminant validity provides us information that the test in question does not measure what other tests measure, or it measures something different from what other available tests measure.

Steps in Construct Validity Process:

- The test in question has to be there
- Clearly defining the construct to be examined
- Assumptions about the construct
- Assumptions about the similar constructs and the different constructs
- Identification of measures that will be used to test the constructs in question
- Test administration, may be in various steps
- Calculation of correlations between the test scores
- Analysis of correlations in order to reach final conclusions regarding construct validity

Multitrait- Multimethod Approach To Construct Validity:

One of the most commonly used models of construct validity is the convergent and discrimination model, the multitrait- multimethod approach proposed by Campbell, and Fiske (1959). They proposed that while estimating validity of a test we should not only look into the measures that our test is related with, but also the ones that it is not related with.

According to this model the following relationship analysis can provide the required information regarding the convergent and discriminant validity:

a. Correlation between the same construct, using the same method
b. Correlation between different constructs using the same method
c. Correlation between different methods used for same construct
d. Correlation between different constructs using different methods

The example of the validation of SCAT that we discussed earlier, employed this approach

Computation of Construct Validity:

Primarily correlations are calculated for all sets of scores. Where a larger number of tests are involved, factor analysis is applied to the obtained statistics.

When the multitrait-multimethod approach is employed, a multitrait-multimethod matrix is developed. This is based on a systematic experimental design (Campbell & Fiske, 1959). It is employed when two or more traits tested by two or more methods.

Lesson 19

Decision Theory

How Much Valid Is "Valid"?

We use test that are valid. A test will be avoided or not used at all if:

- The information about validity is missing
- If validity is too little
- If contradictory information about validity is available

The Size of Validity:

The magnitude of validity is very important, but perhaps less important than is the size of reliability coefficient in case of reliability of a test. What is even more important is the level of its significance of validity. According to Kaplan & Saccuzzo (2001, p.138), "there are no hard- and- fast rules about how large a validity coefficient must be to be meaningful. In practice, one rarely sees a validity coefficient larger than .60, and validity coefficients in the range of .30 to .40 are commonly considered high".

Validity of a test becomes more valuable as the likelihood of having attained it by chance becomes less. The validity coefficient as we know is a value of correlation and in the words of Anastasi & Urbina (2007, p.156), "The obtained correlation, of course, should be high enough to be *statistically significant* at some acceptable level, such as the .01 or .05 levels" and " In other words, before drawing any conclusions about the validity of a test, we should be reasonably certain that the obtained validity coefficient could not have arisen through chance fluctuations of sampling from a population correlation of zero."

The acceptable size of validity also depends on the purpose of test. In some situations tests with a small validity coefficient are also acceptable. Stringent standards for validity of a test may not be a serious issue in situations where test results are used for admissions to academic programs. A certain score is acceptable for admission and below that score the candidates will not be admitted. In such situations test's predictive power or its failure to accurately predict future performance of students may not have as drastic consequences as in case of some other selection situations. In these situations, what may happen at the most is that some candidates may be chosen for a program of instruction for which they were not suitable; or some candidates could not be selected for a course in which they could have shown excellent performance.

On the other hand there are situations where one would not like to have even the tiniest doubt in the validity of a selection/screening instrument. For example if you were to choose a fighter pilot, anesthetist, or a cardiac surgeon from a bunch of candidates, you would not like to compromise on the validity of the instrument.

Similarly think of a situation involving forensic testimony. What if the lie detector is not valid? What if a professional killer is wrongly diagnosed as mentally ill and acquitted? And even worse than that, what if a schizophrenic person is given capital punishment because the instrument failed to diagnose his disorder? In such situations we would be looking for tests that have a high validity coefficient and that have generally been known to be good predictors.

Standard Error of Estimate (SE_{est}):

The concept of validity and its magnitude entail another concept i.e., the amount of error involved. This is the amount of error that may be involved in decisions based on a test whose validity is not perfect It is described as "The error of estimate shows the margin of error to be expected in the individual's predicted criterion score, as a result of the imperfect validity of the test"(Anastasi & Urbina,2007, p.157).

The formula for $\mathbf{SE_{est}}$ is as follows:

r_{xy}^2 = validity coefficient squared

SDy = the standard deviation of the criterion score

Decision Theory and Decision Analysis:

Talking about the validity, the predictive strength of a test, and the chance of error one needs another important piece of information about the test other than the validity coefficient. It is about the amount of information the test provides when used, as compared to when it was not used. In other words is it worthwhile to spend time, money, and effort on using a test for selection/ screening purposes? Are we better off using this test and relying on it for selection procedures as compared to when we selected people without using this test? Is our success rate better after employing this test than we did without it? When we are looking at this type of information we are talking about two things: Base Rates, and Hit Rates.

Base Rates:

"In decision analysis, the proportion of people expected to succeed on a criterion if they are chosen at random" (Kaplan & Saccuzzo, 2001, p.635). There are occasions when accurate choices can be made without using a test. 'Would be' good workers, or professionals can be chosen on the basis of other information available e.g. previous grades or reports. So why and when should we decide to use ate for the same purpose?; Obviously when we have information that our selection procedure, in terms of accuracy of prediction, will become even better when the test will be used. This information comes from the hit rate of a test.

Hit Rates:

Hit rate refers to "in test decision analysis, the proportion of cases in which a test accurately predicts success or failure" (Kaplan & Saccuzzo, 2001, p. 636). It is usually in terms of percentage of people whose success or failure was accurately predicted by the test.

In some situations dichotomous decisions are made on the basis of tests. Dichotomous decisions are decisions in which we have to opt for one of two given decision options. Such situations may be of a wide variety, but become more significant with reference to employment set ups where individuals are either selected or not selected/rejected.

The decision to use a test as 'deciding factor' depends obviously on how good a test is in identifying the right persons for the job. The other side of identifying the right person for the job is identifying the persons who should not be selected. In such situations we have to decide a test score above which will be the selection level and below which people will be rejected. This score is called the cut- off score. Test taker's scores above this will be kept in the plus category and the values lower than this will fall into the minus category. Whether the employer is going to select everyone in the plus category or not, is another decision that depends on a number of factors and of course the employers themselves.

How good a test is in selecting the right persons entails two types of information:

1. The relationship between the individual's present performance, on test used for selection or screening, and the criterion e.g. future success in the profession or position for which the person was chosen (related with hit rate).
2. The individual's scores on the criterion when the test in question was not used for selection. Some other criterion of selection was used (related with base rate).

Different situations arise when we are using tests for dichotomous decisions. When we give a test, primarily two outcomes are possible:

Selection on the basis of a test cut- off score	
1 **Selected**	**2** **Not selected**

But what are we predicting? Is everybody who is selected going to be successful in the job? Will everyone who is rejected, fail if given a chance to work? The answer is may be yes and may be no. Hypothetically speaking, the following outcomes may take place in such a situation:

	Decision on the basis of a test cut- off score	
Future performance	**Selected**	**Rejected**
Success	**1** Right decision Valid acceptance	**2** Wrong decision False rejections
Failure	**3** Wrong decision False acceptance	**4** Right decision Valid rejections

The amount or percentage of right and wrong decisions matters. How well can a test predict is about its hit rate. At the same time the base rate is also very important.

Taylor Russell Tables:

When we have to choose and use a test we want a test that offers us more than simply being better than selection based on mere chance. Taylor and Russell (1939) devised a procedure for evaluating the validity of tests. The procedure takes into account the amount of information the test can yield beyond the available base rate. We know that right decisions can be made even without using a test. If that were not true then all selections that did not involve the use of tests would have ended up in failures alone. What we require is better decision making than the one without formal assessment procedures. Taylor and Russell developed tables that help in evaluating test validity. According to Anastasi & Urbina (2007, p.130) these tables "permit a determination of the net gain in the selection accuracy attributable to the use of a test". These tables present the probability that a person selected on the basis of the test score will actually be successful. Each table is with reference to a certain base rate; there are different tables for different base rates.

The use of these tables requires three pieces of information:

1. Validity coefficient of the test
2. Selection ratio: the proportion or percentage of applicants who must be accepted.
3. Base rate: the proportion of successful applicants selected without the use of a test i.e., what percentage of applicants or candidates would turn out to be successful even when their selection was not made on the basis of the test.

Besides these three, another piece of information is also important. There should be a clear description or definition of 'success' i.e., what would indicate that the applicant has succeeded. Success may be understood and interpreted differently by different people. Therefore it should be clearly laid out as to what outcomes will show that the selected applicant proved to be successful.

Example of a Taylor- Russell table:

***Proportion of "Successes" Expected Through the Use of Test of Given Validity and Given Selection Ratio, for Base Rate .60**

Selection Ratio											
Validity	**.05**	**.10**	**.20**	**.30**	**.40**	**.50**	**.60**	**.70**	**.80**	**.90**	**.95**
.00	.60	.60	.60	.60	.60	.60	.60	.60	.60	.60	.60
.05	.64	.63	.63	.62	.62	.62	.61	.61	.61	.60	.60
.10	..	..	..	..	..	..	..	..	..	..	..
.15	All	blank	boxes	contain	values	..	..	..	..	..	..
.20	..	..	..	..	..	..	..	..	..	..	..
.25	..	..	..	..	..	..	..	..	..	..	..
.30	.82	.79	.76	.73	.71	.69	.68	.66	.64	.62	.61
..	..	..	..	..	..	..	..	..	..	..	..
..	..	..	..	..	..	..	..	..	..	..	..
..	..	..	..	..	..	..	..	..	..	..	..
..	..	..	..	..	..	..	..	..	..	..	..
..	..	..	..	..	..	..	..	..	..	..	..
.85	1.00	1.00	.99	.97	.95	.91	.86	.80	.73	.66	.63
.90	1.00	1.00	1.00	.99	.97	.94	.88	.82	.74	.67	.63
.95	1.00	1.00	1.00	1.00	.99	.97	.92	.84	.75	.67	.63
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.86	.75	.67	.63

* For complete set of tables see H. C Taylor and Russell (1939)

Lesson 20

Threats to Validity and Related Issues

Factors Affecting Test Validity:

There can a number of variables that can be a source of error in the validity of a test. One should be aware of these variables and their possible effect. The effect of these variables should be controlled in planning, designing, developing, and administering the test. Anastasi & Urbina (2007) have described some such variables.

a) Nature Of The Group:

Tests are not administered to identical or very similar groups every time. The group on which the test was administered may not be the same as the other groups with whom the test is used. When we are using a test on the basis of its validity, we need information regarding the details of the sample on which it was validated. It may happen that a test with high validity coefficient obtained from one group does not turn out to be an efficient predictor with another sample. For example a test of mathematical ability is validated on a sample comprising people of mixed abilities and background. At a later stage it is administered to a group of accountants who are experts in math; on the other hand the test is given to a group of fine artists who have never learnt math after junior school. In both cases one may not expect to get results similar to those of the original sample. Test manuals should clearly specify the all details of the sample from whom validity was obtained and also the type of population to which the test can be generalized.

b) Heterogeneity Of The Sample:

Validity coefficient is a value of correlation, and we know that correlation is calculated from two sets of scores, the predictor and the criterion variable. Statistically speaking the correlation between the two sets of scores will be higher if the scores come from a wider range i.e., the correlation will be higher in heterogeneous groups as compared to homogeneous groups.

c) Preselection:

Preselection may cause a problem when criterion related validity of a test to be used in job selection situations is being measured. The test may be validated on a sample of newly selected employees as their performance on the criterion measure will also be available. The very fact that they were selected for the job may affect the validity coefficient because they are a bunch of people who have already succeeded as they were more capable than others, and similar to each other in ability. Thus the range of score is likely to be limited and the validity coefficient may be lower.

d) The Form Of Relationship Between Test And Criterion:

One basic assumption in computation of Pearson correlation coefficient is that the relationship between the test and the criterion is linear and uniform across the whole range of scores. If the relationship is not linear and the scores are clustered at certain points (as can be seen in a scatter plot) then it is going to affect the validity coefficient.

Evaluation of Validity Coefficient:

The joint committee of the American Psychological Association, American Educational Research Association, and the National Council on Measurement in Education (1990), in booklet titled Standards of Educational and Psychological Testing, has described a number of issues to be considered while validity coefficients are being interpreted. The issues are as follows:

a) Look for changes in the cause of relationships.
The cause of relationship between the test and the criterion will remain the same in future. However, situations may change and so may this relationship. This needs to be kept in mind in future use and interpretation.

b) What does the criterion mean?
The criterion should be carefully chosen. Any test with unknown or dubious validity should not be chosen as criterion. The criterion should relate specifically to the use of the test.

Review the subject population in the validity study:
One should know on what type of population was the test validated. It may happen that the test is going to be used with groups who are not represented in the population from which the validity coefficient was obtained.

c) Be sure the sample size was adequate
Tests should not be validated on very small samples. The sample should be reasonably sized.

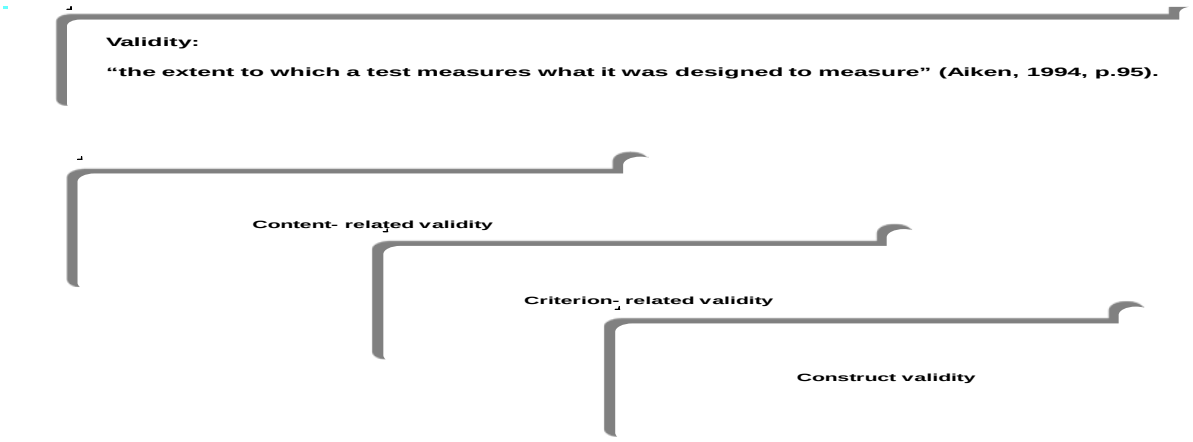
d) Never confuse the criterion with the predictor
At times people, institutions, or organizations using tests may confuse criterion with predictors. They would use the criterion first (as if it were the test/predictor) and the predictor later. For example universities give admissions to students who have not passed the specified graduate level ability test, but require that the students will have to clear the test before the result is declared. As you can see this test was actually supposed to be the predictor on whose basis the students were to be chosen, but now it is following the criterion i.e., performance, that has already taken place.

e) Check for restricted range on both predictor and criterion
The range of scores should not be restricted. We know that a wider range of scores would yield better correlation/ validity.

f) Review evidence for validity generalization
The test may be used with different groups or people every time. The test takers may be different from the ones from whom validity was derived. The test user must have the information about whether the test may be used with a different population or not. And if the new sample is different from the original one then do we have validity evidence from a relevant sample? If not, then the user will have to reconsider the choice of the test.

g) Consider differential prediction
At times the use of criterion- related validity is not the right choice for determining the validity of a test. For example if a good criterion is not available then using a weak or wrong criterion would not be advisable. In such situations a better option would be to use construct- related evidence for validity.

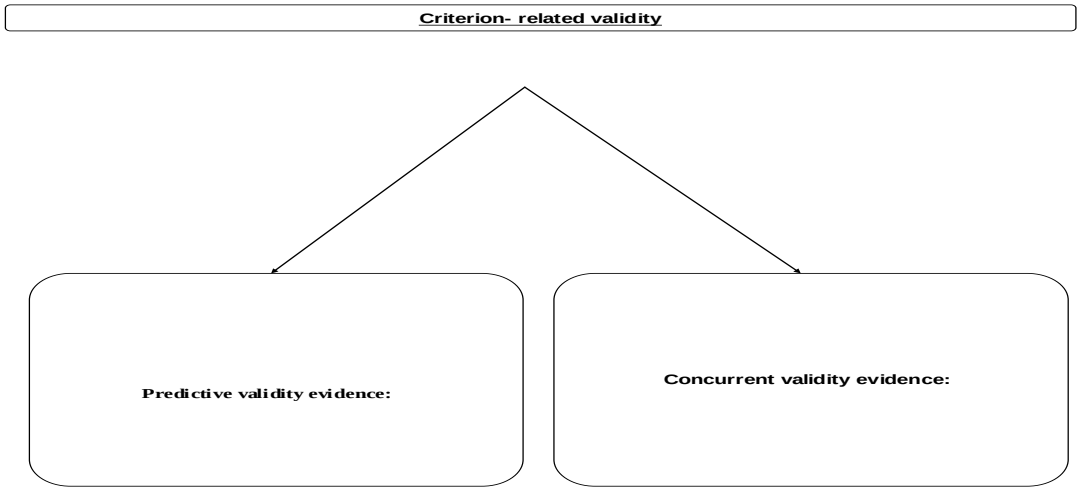
Review of Validity and Its Types:



Content Validity Evidence:"The evidence that the content of a test represents the conceptual domain it is designed to cover" (Kaplan & Saccuzzo, 2001, p.635).

Criterion- Related Validity refers to a procedure where scores o a test being used are correlated with scores on a criterion. Criterion validity evidence can be defined as "The evidence that a test score corresponds to an accurate measure of interest. The measure of interest is called the criterion" (Kaplan & Saccuzzo, 2001, p.635).

Construct Validity Evidence: to "A process used to establish the meaning of a test through a series of studies. To evaluate evidence for construct validity, a researcher simultaneously defines some construct and develops the instrumentation to measure it. In the studies, observed correlations between the test and other measures provide evidence for the meaning of the test" (Kaplan & Saccuzzo, 2001, p.635)



Predictive Validity Evidence: "The evidence that a test forecasts score on the criterion at some future time" (Kaplan & Saccuzzo, 2001, p.638). Predictive validity pertains to prediction over a time interval.

Concurrent Valid Evidence: "evidence for criterion validity in which the test and the criterion are administered at the same point in time" (Kaplan & Saccuzzo, 2001, p.635).

Lesson 21

Item Analysis

The task of the test developer is not over after the test items have been written according to the specifications and the domain area to be covered; nor does trying out the test on a representative sample mean that nothing else is left that needs to be looked after. In fact a major and most significant part of the test development process follows the try out or the pilot study. It is the analysis of the test as a whole as well as each one of the individual items. The test can be analyzed both qualitatively and quantitatively.

Considering the qualitative aspect, the test developer takes all measures to select and include the right content, use the most suitable format, write the items carefully, and arrange the items in carefully designed sequence according to order of difficulty. In other words the test developer tries to make sure that the test has maximum content validity. However after test administration for the try out the test takers' opinion may also be gathered regarding issues like time allocated for test, vocabulary used, or other similar aspects.

The quantitative analysis involves more complex procedures.

Imagine that you are a test developer who has developed a 60- item test of mathematical ability for children aged 14-16. The test is meant for identifying students who are good in mathematical ability so that they are admitted to prestigious state colleges. You took real pain in writing items for the test. Prior to this you went through the mathematics text books that children are taught at school till 16 years of age. You determined the behavioral objectives with reference to the ability in question. Table of specifications was designed and developed, and items from all sections were written. You tried your best to write items that were perfect in all respects, for which you consulted a number of experts as well. Item stems and distracters were carefully chosen and phrased. Great care was exercised in arranging items according to their difficulty level, and the number of items was determined keeping in view the time allocated for completing the task. Now can we say that the test has taken a final form and is ready for use? Perhaps not, because we still do not know a number of things about the test. We need to know if there are items that no one will be able to attempt correctly, if so then do we need those items? We also do not know if there are items that everyone will be able to do accurately, if so then why should we retain that item? Such items do not differentiate between the ones who know and the ones who do not. They are not sensitive to individual differences. Similarly we would like to know whether the items are arranged exactly according to their difficulty level or not. Once we have answers to these questions then we will be able to give a final shape and form to the test. Additionally, you may expect to drop a few items from the original 60 items and retain the ones that serve the purpose for which the test was developed, better than the rest. You are looking for the best items.

In order to find answers to these and other similar questions we conduct try- outs or pilot studies of tests. The quantitative analysis of the results of the try out helps us revise and refine the initial form of a test. This process is called item analysis.

Item Analysis is "A set of methods used to evaluate test items. The most common techniques involve assessment of item difficulty and item discriminability". (Kaplan & Saccuzzo, 2001, p. 637)

Item Difficulty:

"A form of item analysis used to assess how difficult items are. The most common index of difficulty is the percentage of test takers who respond with the correct choice" (Kaplan & Saccuzzo, 2001, p.637). The item difficulty index can be calculated in the form of proportions as well.

The purpose of measuring item difficulty is:

- To examine the difficulty level of test items; whether the difficulty level set up at the time of test construction was right or not.
- To see if there are any items which are correctly attempted by everyone. If there are any such items that are 'too easy' then they need to be removed, replaced, or revised.
- Similarly, if there are any items which are not correctly attempted by anyone then such 'too difficult' items also need to be removed, replaced, or revised.

Item-Difficulty Index:

As mentioned earlier, item difficulty index is either in form of percentages or proportions of the total number of test takers who attempted an item correctly. Item difficulty is calculated separately for every item. It is denoted by a lowercase italicized '*p*'. A number attached to this '*p*' indicates the item number whose difficulty level is

described. For example p_1 indicates item difficulty of item number 1, p_2 is the difficulty level of item number 2 and so on. On occasions you may come across the term *'facility index' rather than difficulty index that refers to the percentage of responses to correct choices. Both terms refer to the same procedure.*

Theoretically speaking, the value of item difficulty index may range from zero to 1. Item difficulty index of .60 means that the item was correctly attempted by 60% test takers (60/100=.60) and .50 means 50% (50/100=.50). It can be any value between zero and 1.00 depending upon the accurate responses to individual items. A zero item difficulty index would mean that nobody was able to attempt the item correctly. Therefore the item in question is a bad or poor item. On the other hand an item with difficulty index of 1.00 is also a bad item because it was correctly attempted by 100 % test takers. Good items are the ones which are correctly attempted by some of the test takers, while others failed in doing so. A large item difficulty index indicates that the item is easy; the smaller an index is the more difficult is the item. The common item difficulty index range of test items is between .3 and .8, but the test developer may exercise her discretion.

Average Index of Difficulty Of A Test:

Once the item difficulty index has been calculated for all items in a test, you can calculate the average difficulty index of the whole test as well. Simply add up the indices for individual items and divide the summation with the total number of items. "For maximum discrimination among the abilities of the test takers, the optimal average difficulty is approximately .5, with individual items on the test ranging in difficulty from about .3 to .8" (Cohen & Swerdlik, 1999).

Taking Care of Guessing:

In case of some tests the test taker can give the right answer simply by guessing. This happens mostly in case of items where multiple response options are provided. In such tests the desired or acceptable proportion of accurate answers is set higher than that in free response tests. The optimal average difficulty for MCQ type tests is calculated by taking the mid-point between chance success proportion and 1.00. Chance success proportion is the likelihood of giving a correct response simply by guessing; in MCQs with 4 options it is .25; with 5 options it is .20, and with three options it will be .33. The mid-point can be calculated by adding the chance success proportion and 1.00 and dividing the sum by 2. For example in a test with 4 response options in each item, the chance success proportion is .20. The optimal item difficulty will be:

.20+ 1.00/ = 1.20/2= .6.

For a test with 5 options the average proportion correct should be .69 (Lord, 1959).

Item Difficulty Index In Different Types Of Tests:

Item difficulty index may not be based on the success rate/passing proportion/ number of accurate responses every time. In case of achievement tests or intellectual ability tests the index is based on the proportion or percentage of people who gave the correct answer. However, in case of tests where there are no right or wrong answers, item- endorsement index may be used. Here the analysis can help identify items which nobody replied to, which received the same response from every one etc.

Item Discrimination:

A test is supposed to discriminate between those who know and those who do not know; those who score high and those who score low; those who have acquired a skill and those who have not. A test will not be a good test if the people who are supposed to know the correct answer fail and those who are not supposed to know succeed. A good test differentiates between the high and low scorers. If some items are correctly answered by high scorers and some by low scorers then something is wrong with the test.

Item Discrimination Index:

Every test has its discrimination power. To see if the test discriminates between high and low achievers a certain percentage of the high and low achievers are taken. The discrepancy between their attempts to correct responses is calculated in terms of percentages. Item discrimination index is denoted by a lowercase italicized '*d*'.

"The item- discrimination index is a measure of the difference between the proportion of high scorers answering an item correctly and the proportion of low scorers answering the item correctly; the higher the value of d, the greater the number of high scorers answering the item correctly" (Cohen & Swerdlik, 1999).

The item discrimination index is calculated by considering the number of people in high scorers (U) and the number of people in low scorers (L) who correctly answered an item.

Item number	U	L	U-L	n	U-L/n= d
1	10	10	0	10	0
2	10	0	10	10	1
3	0	10	-10	10	-1
4	9	2	7	10	.7

The value of 'd' may range from -1 to +1. However +1 would be an ideal situation where all test takers in the upper scoring group gave correct answers and none from the lower scorers did so; , whereas no test developer would like to get d= -1 which means that all high scorers failed and all low scorers passed this item. A value of d= 0 indicates that the item does not differentiate between the two groups; they had identical proportion of success. The larger the value of 'd' of an item the more discrimination it is making between the two groups.

A pertinent question arises here. What percentage or proportion of the high and low scorers should be considered for this analysis? A shortcut procedure described by Aiken (1994) provides answer to this question: "A shortcut procedure is to divide the examinees into three groups according to their scores on the test as a whole: an upper group consisting of the 27 percent of examinees making the highest test scores, a lower group of the 27 percent making the lowest test scores, and the remaining 46 percent in the middle group. When the number of examinees is small, upper and lower 50 percent groups on total test scores may be used."

Lesson 22

Item Analysis

Item Analysis of Item Distracters:

When item discrimination index is calculated in item analysis, we gather information about how far the individual items differentiate between the high scorers and the low scorers. It is assumed that more high scorers will be giving correct answers to questions and the correct- response pattern of the low scorers will be the other way round. High scorers are supposed to be the ones who 'know' and the low scorers are supposed to 'not know' the right answers. If items are discriminating between the two groups then they will be retained in the test, but if they do not then they needed to either be discarded or improved. But this is not all that one would like to know and determine about the test. The response options or the distracters used in each item also need to be item analyzed. We know that a good multiple choice item is the one in which every alternative or distracter appears to be the right or plausible answer to examinees who do not know the exact right answer. The high scorers, on the other hand, will obviously know the keyed option. It is therefore important that all the alternatives provided are good distracters. This can decrease the likelihood of correctly responding without actually knowing the right answer; if some distracter is too weak an alternative that most people will realize that this cannot be the right answer then they may not choose it as their answer; on the contrary if an option is strikingly better than the other options and stands out to be the keyed option then many examinees would mark it even when they did not know the correct answer. Therefore the analysis and discrimination power of item alternatives needs to be done.

The 'd' value or the item discrimination index does provide information that whether or not the item discriminates between the high scorers and low scorers. A similar analysis is done to see how each response option, other than the keyed one, is chosen by the examinees. The number of times each distracter is chosen as the right answer by members in the high scoring and low scoring groups is counted. This is the response valence. To determine the valence of each option for an item, percentage of responses to each keyed response as well as each distracter are calculated. As said earlier, the high scorers will mostly be selecting the keyed answer and more of the low scorers will be going for distracters. If a distracter is selected by a large number of high scorers and by very few low scorers then there is something wrong with the distracter. It needs to be removed, improved, or replaced. If all distracters attract the same number of test takers then they are very good response options.

A good item, when the high scorers and low scorers include 25 members each:

	Option (a) ✓	*Option (b)*	*Option (c)*	*Option (d)*
High scorers	18	3	3	*3*
Low scorers	6	6	7	6

A poor distracter! Here the distracter (c) has attracted a significant number of the high scorers, more than the low scorers who selected this option:

	Option (a) ✓	*Option (b)*	*Option (c)*	*Option (d)*
High scorers	13	2	10	*2*
Low scorers	6	6	7	6

Example of a very weak, in fact bad, item: Here all of the distracters have been selected by more high scorers than the low scorers. The low scorers have also not selected all options equally:

	Option (a) ✓	*Option (b)*	*Option (c)*	*Option (d)*
High scorers	4	5	8	*8*
Low scorers	9	3	6	7

The weak items and distracters may be improved upon, whereas it would be better to remove the bad items.

Lesson 23

Item Analysis

So far we have discussed the major concepts related with item analysis. However, there are some other concepts too that you should be familiar with, though you might be using them at a later, higher, level of your studies or whenever you will work on test development.

In this section you will be introduced to the following concepts:

1. Item response theory
2. Item- characteristic curves
3. Cross validation
4. Qualitative analysis of tests

Item Response Theory:

Item response theory is an approach that takes into consideration the probability of answering, right or wrong, each individual item in a test. The information regarding each item is plotted graphically. This approach is known as the item response theory or IRT. It is also known as Item- Characteristic Curve Theory and Latent Trait Theory. The graph containing information about the items is called the item- characteristic curve. Decisions regarding the items of a test can be based upon this information.

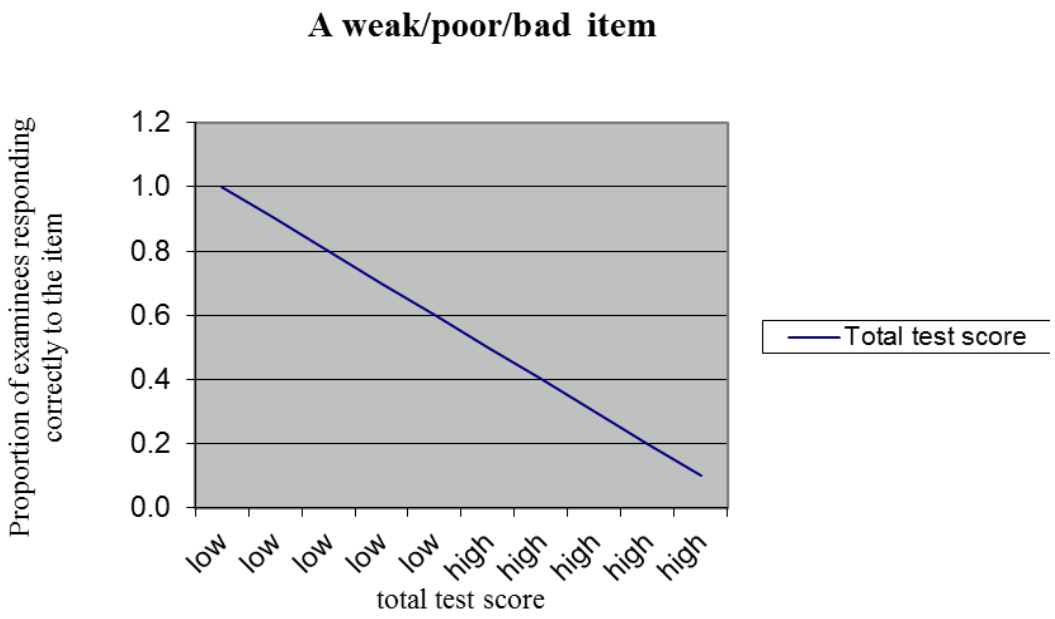
Item- Characteristic Curves:

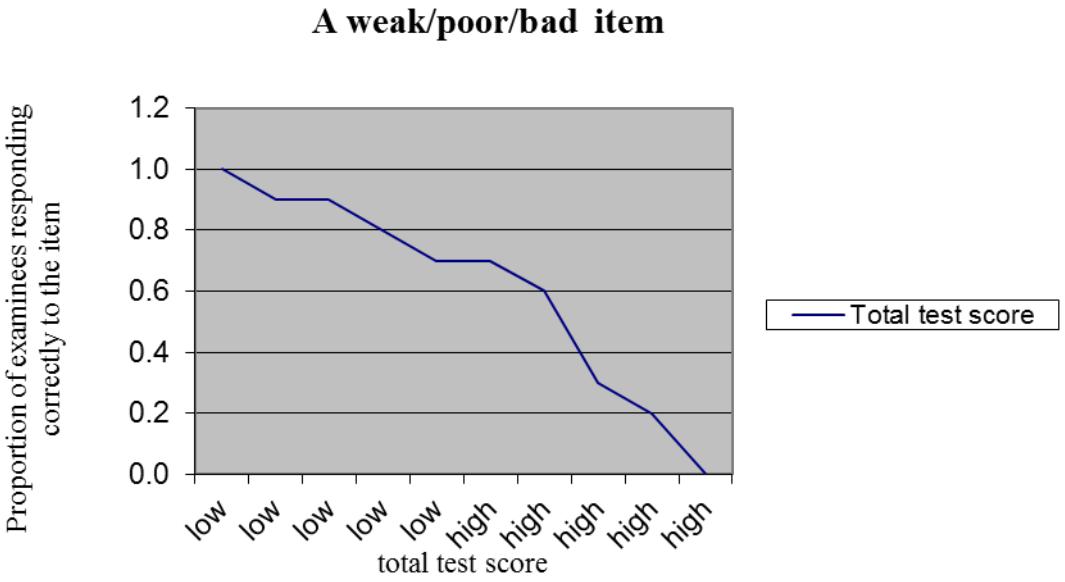
Item difficulty and item discrimination can be presented graphically also. Item- characteristic curves are the graphs that represent these characteristics of a test. The horizontal axis represents the ability being tested whereas the vertical axis contains the probability correct responses or the proportion of examiners responding correctly to the item. In the words of Kaplan and Saccuzzo (2001), it is "a graph prepared as part of the process of item analysis. One graph is prepared for each item and shows the total test score on the X axis and the proportion of test takers passing the item on Y axis" (p. 637). The shape or slope of the graph or curve indicates whether the item is a good one or not, how far does it discriminate high scorers from low scorers. A steep slope indicates that the test discriminates between the two groups. Scores of a highly discriminating test will yield a very steep slope.

A good item has a positive slope. As can be seen from the following graph, the proportion of high scorers responding correctly is higher than the proportion of low scorers. More of the low scorers are not responding correctly. It can be said that the item that yielded such a curve is a good item.

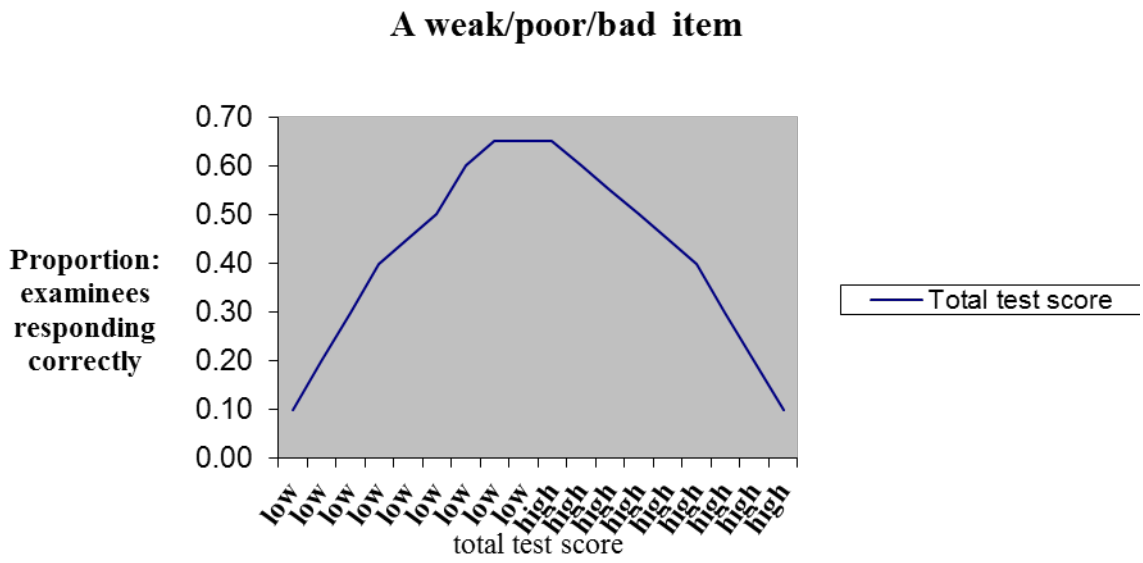
A negative slope like the following one, or a similar one in the same direction, means that the item is not a good one. The item does discriminate between the two groups, high scorers and low scorers, but in the opposite way. We do not expect and accept this type of discrimination. This graph indicates that more of the low scorers did the item correctly than the high scorers. This means that the probability of answering the item correctly is higher

for the low scorers rather than the high scorers. This, therefore, is a bad or poor item that needs to be removed or replaced.





Another type of items can be the one in which the majority of neither the top scorers nor the low scorers do the item correctly. It is the middle, moderate, scorers who attain the maximum proportion of correct responses. This type of item is also a bad item.



Cross Validation:

The validity of a test, we know, may be determined from a sample that was used for item selection. However, in order to have a better estimate of the validity of the test, the entire test needs to be validated on different samples as well. This process is called cross validation. "The term cross validation refers to a revalidation of a test on a sample of test takers other than the ones on whom test performance was originally found to be a valid predictor of some criterion" (Cohen and Swerdlik,1999, p.246). The regression equation is used to predict performance in a sample of test takers who are different from the ones on whom the test was validated.

If validity of a test is computed from the original sample used for sample selection, then there are chances that the validity index will be higher than the one expected to be obtained from a new or different sample of test takers. It is expected so because of the possible chance variations. It is expected that validity will shrink in the process of cross validation. "The amount of decrease in the strength of the relationship from the original sample to the sample with which the equation is used is known as shrinkage" (Kaplan & Saccuzzo, 2001).

The factors that may affect the amount of shrinkage:

1. The size of the original item pool
2. Proportion of test items retained
3. Sample size

High validity coefficient can be expected if the original item pool was large while the proportion of retained items is small. The size of cross validation sample also affects shrinkage. Greater validity shrinkage may be expected if smaller samples are used.

Qualitative Analysis:

Qualitative analysis of a test may also be conducted along with quantitative analysis. After test administration is over, the test takers may be asked questions about various aspects of the test. These questions can be asked and answered orally or in writing. Different formats can be adopted for this purpose e.g. interview, discussions etc. The respondents' responses can be of great help in improving the individual test items, test format, and the entire test itself. Cohen and Swerdlik (1999) have pinpointed some areas that may be explored

- Cultural sensitivity
- Face validity
- Test administrator
- Test environment
- Test fairness
- Test language
- Test length
- Test taker's guessing
- Test taker's integrity

- Test taker's mental/physical state upon entry
- Test taker's mental/physical sate during the test
- Test taker's overall impressions
- Test taker's preferences
- Test taker's preparation

Review of Item Analysis:
Methods used for assessing and evaluating characteristics of test items and the test itself. Primarily two characteristics are measured, item difficulty and discriminability.

Item-Difficulty Index:
As mentioned earlier, item difficulty index is either in form of percentages or proportions of the total number of test takers who attempted an item correctly. Item difficulty is calculated separately for every item. It is denoted by a lowercase italicized '*p*'. A number attached as subscript to this '*p*' indicates the item number whose difficulty level is described. For example p_1 indicates item difficulty of item number 1, p_2 is the difficulty level of item number 2 and so on. On occasions you may come across the term 'facility index' rather than difficulty index that refers to the percentage of responses to correct choices. Both terms refer to the same procedure.

Item Discrimination:
A test is supposed to discriminate between those who know and those who do not know; those who score high and those who score low; those who have acquired a skill and those who have not. A test will not be a good test if the people who are supposed to know the correct answer fail and those who are not supposed to know succeed. A good test differentiates between the high and low scorers. If some items are correctly answered by high scorers and some by low scorers then something is wrong with the test.

Item Discrimination Index:
Every test has its discrimination power. To see if the test discriminates between high and low achievers a certain percentage of the high and low achievers is taken. The discrepancy between their attempts to correct responses is calculated in terms of percentages. Item discrimination index is denoted by a lowercase italicize

Lesson 24

Assessment of Intellectual and Cognitive Abilities

Thinking about intelligence and intelligence testing, a number of questions come to one's mind:

- *What is intelligence?*
- *Why do we try to understand intelligence?*
- *Why do we measure intelligence?*
- *Why cannot we use the same measure for everyone?*
- *Can intelligence tests be taken as reliable measures of intellectual ability?*
- *Are these tests valid?*
- *Are intelligence tests the only way to measure intellectual ability of a person?*

Moving on to a relatively personal side, one may ask:

- *Am I an intelligent person?*
- *Is intelligence about a skill, specific skills, or a general ability that I possess?*
- *How do I judge using my observation if someone is intelligent or not?*
- *Is intelligence an inborn ability or is it a learned phenomenon?*

None of these questions can be answered in one word, one sentence or one answer.
Before we start our discussion on intelligence testing, we need to understand what intelligence is. This is important because we can understand the significance, logic, and process of intelligence testing if we have understood what intelligence means to both the test developer and the user. Different authors and researchers have proposed a variety of descriptions of intelligence.

Intelligence:

As discussed in earlier courses also, we know that "intelligence is the capacity to understand the world, think rationally, and use resources effectively when faced with challenges"(Feldman, 2002, p.261).
Intelligence refers to the ability to adapt, to reason, to solve problems, and think in an abstract manner; it also includes learning and experiencing new things and understanding from the past experiences.
Intelligence or the intellectual ability of a person is based upon a constant and ongoing interaction between environmental factors and inherited potentials in order to have better understanding of how to 'use' and 'apply' the potentials in a meaningful manner.
Modern psychology considers both environment and heredity and their interaction to be influential.

Theories of intelligence:

One of the earliest contributions to the measurement of intelligence in the 19th century was made by Sir Francis Galton, the cousin of Charles Darwin. Born in the family of geniuses he himself was a genius having a very high IQ. Besides being a geographer, meteorologist, and tropical explorer, he was the "founder of differential psychology".
According to Francis Galton ("Hereditary Genius, 1869) "gifted individuals" tended to come from families, which had other, gifted individuals.
His was the first systematic attempt to measure intelligence by investigating the role of heredity and its impact on intellectual abilities. He attempted to measure human trait quantitatively in order to determine the distribution of heredity in it. He also talked about the relationship between intelligence and the size of one's head, but the idea received no empirical support.
Cattell, an American psychologist, gave more importance to the mental processes. He used the term "mental test" for the first time for devices used to measure intelligence. He developed tasks that were aimed to measure reaction time, word association test, keenness of vision and weight discrimination.
Another name without which the history of intelligence testing will never be complete is that of **Alfred Binet** who developed the first proper formal intelligence test in 1905.his test will be discussed in detail in the following sessions. All these people talked more about intelligence tests and little about what constituted intelligence itself. Today we see that there are a number of approaches to explaining and understanding the concept of intelligence.

Spearman's g-factor theory:

Psychologists have always been interested in studying if intelligence is a single, specific, ability or a general ability that contains multiple components and is reflected in all aspects of one's life. One of the earliest theories of intelligence proposed the idea that intelligence was a single, general factor. British psychologist, Charles Spearman who gave his theory in the early 1900s, observed that people who scored high on one mental test also tend to score on the other as well. The same applies to the low scorers. Using "factor analysis" a statistical technique on the basis of which he proposed two factors that can account for the individual differences; g- factor and s- factor. The "g" factor referred to "general intelligence" and the "s" factor meant "specific intelligence". Spearman and other psychologists with similar approach believed that there was a single factor, a general factor, for mental ability, the "g" factor.

This factor, Spearman proposed, can account for the general ability that is common in all people: as observed from the mental tests, whereas the's' factor can account for the specific abilities that are different in different people. Spearman's theory seemed to be impressive to many people but not all. There were psychologists who believed otherwise and proposed alternate explanations of intelligence.

Thorndike's Social Intelligence:

Thorndike was critical of Spearman's g- factor approach. He argued that instead of only one 'g' factor, there are a number of factors that make and influence intelligence; factors that cannot generally be found out but that are expressed in human actions.

According to him intelligence covers three main divisions:

1. Social intelligence; that enables one to understand and manage relationships
2. Abstract intelligence; that enables one to understand and manage ideas such a algebra, mathematics, or abstract concepts
3. Concrete intelligence; that enables us to manage concrete and mechanical concepts and ideas e.g. accounting, economics, architecture, banking etc.

Thurstone's approach: Primary Mental Abilities:

Some psychologists such as American psychologist Louis L. Thurstone (1938), argued that intelligence is not a general factor, but it is composed of small independent factors or elements. Thurstone called these factors "primary mental abilities".Thurstone and his wife prepared a set of 56 tests for the identification and verification of these abilities. These were administered to 240 college students. The results were analyzed through factor analysis. The analysis yielded seven primary mental abilities:

- Verbal comprehension: An ability to understand and define words
- Word fluency: An ability or speed of thinking of verbal material such as rhyming, or naming words in a given category
- Spatial visualization: Ability to recognize and manipulate objects or things in three dimensions such as drafting and blue print reading
- Perceptual speed: An ability to quickly perceive and detect the visual details and differentiate between the similarities and differences between designs
- Reasoning/ inductive reasoning: A logical ability of deriving general ideas from specific information
- Numbers/ Arithmetic ability: Capability of doing work easily on numbers such as doing simple arithmetic tasks fast and rapidly
- Memory: An ability or capacity of remembering and retaining the material such as words, letters and ability to recall and associate different words.

Crystallized and Fluid Intelligence: R.B Cattell:

After in depth research and observation some psychologists have proposed the idea that there are two types of intelligence, crystalized intelligence and fluid intelligence.

1. Crystallized intelligence: The capability of using information that has been learnt through experience. Education and culture affect this type of intelligence. Whatever knowledge, skills, techniques, and arts one learns over years of one's life are accumulated. All of these are applied in problem solving situations. This type of intelligence keeps on increasing with age, or the learning experiences of a person.
2. Fluid intelligence: Largely influenced by biological factors, it is the capability of information- processing, solving problems which depends more on the neurological development of a person such as reasoning and

memory, which decline with age. If we are learning a multiplication table, grouping objects according to some classification system, or solving an equation then we will be using fluid intelligence.

It "reflects information- processing capabilities, reasoning, and memory" (Feldman, 2002, p. 269).

Guilford's theory of the Structure of Intellect (SOI):
It is a model of intelligence according to which intelligence is the result of the interaction of operations, contents and products.
He believed that intelligence is a much more complex phenomenon than one thinks of it; it is difficult to define it as a 'g' factor or in terms of 'primary mental abilities. Guilford talks of 150 such abilities/ factors (Guilford, 1985).

The different components of intelligence are:
Operations: it is the potential of different ways of thinking including:
- Evaluation
- Convergent thinkingDivergent thinking
- Memory retention
- Memory recording
- Cognition

Contents: A potential of what we think about something. Contents include:
- Visual
- Symbolic
- Semantic
- Behavioral

Products: The results obtained by applying certain operations to certain contents, or the ability of thinking in a certain manner about a certain thing. Products include:
- Units
- ClassesRelations
- SystemsTransformation
- Implication

Multiple Intelligences: Howard Gardner's Approach:
Howard Gardner (1985) maintained that intelligence does not consist of a single factor. Intelligence, he proposed, consists of eight independent intelligences. These eight are possessed by all individuals though in varying degree. The said eight kinds of intelligences include:
1. Linguistics
1. Logical- mathematical
2. Spatial intelligence
3. Musical intelligence
4. Bodily- kinesthetic
5. Interpersonal intelligenceIntrapersonal intelligence
7. Naturalistic intelligence

Sternberg's Triarchic Theory:
According to Robert Sternberg's triarchic or three- dimensional theory given in the 1980s, intelligence consists of three main components:
- Analytic intelligence
- Creative intelligence
- Practical intelligence

No other psychologist has talked about practical intelligence and its significance the way Sternberg has. Practical intelligence is related to overall success in living (Sternberg, 2000). According to Sternberg, the traditional tests measure academic success whereas career success requires practical intelligence. Practical intelligence is a learnt

phenomenon. It results from the observation of others' behavior, unlike academic success that results from knowledge of a particular information base obtained from reading and listening.

Emotional Intelligence:

A modern approach is to judge the ability of people from their emotional intelligence (Goleman, 1995) . Emotional intelligence is about the ability to go along with others according to Daniel Goleman (1995). Accurate assessment, evaluation, expression, and regulation of emotions have certain underlying skills and these skills stem from emotional intelligence.

Emotional intelligence involves the realization and regulation of personal emotions as well as empathy and understanding of others' emotions. Social skills, self-awareness, and empathy are important aspects of one's emotional sphere, and all three require emotional intelligence.

Piaget's View of Intelligence:

- Intellectual development can be defined in terms of qualitative changes in thinking which are clearly apparent in children of particular age.
- His theory is more concerned with the universal patterns of intellectual development and functioning. He maintained a comprehensive theory that emphasized on 'how' children acquire knowledge and use it to solve logical problems
- He was more interested in how children exhibit intelligence in different stages of life as he proposed the four stages of cognitive development, which he termed as universal and invariant (occurring in the same sequence). The stages are: sensorimotor, preoperational, concrete operational and formal operational.

Lesson 25

Measurement of Intelligence

A variety of tests are available to us for the assessment of intelligence. The tests may be chosen according to the specific purpose for which the test is to be used. The tests may be used independently or in combination with other tests i.e., as part of a battery of tests. Similarly, we have the option to choose from individual tests or group tests. The prevalent, modern, approaches to measure intelligence are based upon the contribution of Alfred Binet.

In this section we will discuss some intelligence tests that are commonly used. First of all we will look at the historical evolution of Binet Scale in order to get a feel of how tests are developed and how their evolution may continue over decades.

The Binet Scale:

French psychologist Alfred Binet and Theodore Simon, in 1905 in France, developed the first formal measure of intelligence. The purpose of their scale was to assist the education ministry and department in identifying "dull" students in the Paris school system, so that they could be provided remedial assistance or training. They felt that the children's performance could be used as an indicator of their intelligence. In other words, they believed that intelligence can be measured in terms of performance of a child.

The idea was that children can perform different types of tasks at each age level. With growing age the tasks that a child can perform become more difficult and complex in nature. If a child can perform the tasks that children of his age can perform then he is an intelligent child; if he cannot perform the way his age mates do, then he possesses below average intelligence; the one who can perform tasks that children older than him can perform then he has above average intelligence. However, when Binet developed his scale, the main focus was to identify children who could not perform according to their age level. For the next many decades, 1905 onwards, a number of revisions were made in the original scale. The first intelligence test, Binet and Simon's scale could identify more intelligent children within a particular age group and could differentiate intelligent children from the less intelligent ones.

The Development of Binet and Simon Scale:

First of all a number of tasks were developed. Then groups of students who were categorized or labeled as 'dull' or 'bright' by their teachers were selected. The tasks were presented to them. The tasks that could be completed by the 'bright' students were retained; the rest were discarded. The idea was to retain tasks that could be completed by the bright students, as these were considered to be indicative of the child's intelligence. The realization was there that not all tasks could be performed by all children, even the bright ones; tasks were age related. Children of any age group could perform tasks specific tasks. Children of a lower age group could not do tasks meant for a higher level. If a child could perform tasks meant for a higher age group, she was considered to be of above average intelligence. On the contrary, if a child could not perform tasks meant for her age then she was considered as a dull child. Using the same approach dull or bright children could be identified with reference to their age.

Binet's scale became popular very soon and work was being done in various parts of the world on its translation and adaptation. The U. S was a country that took lead in this regard. A training school in New Jersey was using it in 1908 (Goddard, 1910). A modified version was published in 1912 (Kuhlman, 1912). This version included a widened age range that went down to three months of age.

The major developments took place at Stanford University. The major milestones in the history of this scale are as follows:

The 1905 Scale:

The original, 1905, scale comprised 30 items arranged according to increasing order of difficulty. The norms were obtained from a sample of only 50 children who were reported to be 'normal' considering their average school performance. Though normative and validity related information for this scale was not sufficient, this scale was a major milestone in the history of psychological measurement.

The 1908 Scale:

The 1908 scale was similar to the 1905 scale in that it was an age scale that retained the principle of age differentiation. The concept of mental age was introduced in this revision. The performance of a test taker was

compared with the average performance of other persons of the same chronological age, as discussed earlier. The standardization sample for this revision included 203 individuals.

The 1916 Revision: The Stanford- Binet Intelligence Scale:
The American psychologist, Lewis Terman gave the first Stanford revision of the scale in 1916, known as Stanford- Binet Intelligence Scale. The revision was standardized on American sample and was meant for age 3 years to 14 years plus average and superior adults. The standardization sample was increased in size, though containing only white- native Californian children. This fact was a reason for criticism against the scale.
This scale had three significant characteristics; (a) it used the concept of IQ and was the first American test to do so, (b) It included the concept of 'alternate item' i.e., an item that could be used in place of a regular item on occasions when the original item could not be used properly due to any reason, and (c) it provided detailed and organized instructions for administration and scoring.

The 1937 Revision:
The next revision took place in 1937. It was the result of a project that began in 1926. Terman and Merrill, Terman's colleague at Stanford, worked on this revision. The 1937 revision contained new tasks meant for preschool level and for adult level. It comprised two equivalent forms, 'L' and 'M' (the initials of the first names of the authors). The test was appreciated for its validity and reliability. The standardization sample was much larger in size than the previous samples. It consisted of 3184 individuals, selected from 11 U.S. states. However, it was criticized for lack of representativeness because the subjects were all whites and more from urban than rural areas. Nevertheless it was more improved than the previous versions.

The Concept of Mental Age:

Children taking the Binet- Simon test were assigned a score that corresponded to the age group they belonged to. This score indicated their "mental age". Mental age referred to the average age o children who secured the same score. Mental age can be understood as the typical intelligence level found for people at a given chronological age. •Mental age of a person can be different from his or her chronological age i.e., it can be above or below that. It could reflect whether or not a child was performing at a level at which his age mates were

The Concept of Intelligence Quotient or IQ:

As a result of problems with depending merely on mental age, a solution was devised in terms of intelligent quotient, a concept whereby the chronological age of the person is also given due consideration. It is an indicator of intelligence that takes into consideration a person's mental as well as chronological age. The formula for IQ is:

$$\text{IQ score} = \text{MA} / \text{CA} \times 100$$

This formula is basically a ratio of a person's mental and chronological age. It is multiplied with 100 for eliminating decimal points. Using this formula means that if the mental and chronological age of a person is the same, then he or she will have an IQ of 100. If one is below his chronological age then the IQ will fall below 100 and vice versa.

The 1960 Revision:
This revision was followed by another one in 1960. The 1960 edition was being worked upon when Terman died in 1956. One major change that took place in this edition was that instead of two forms, it comprised a single form, 'L-M'. No new items were added and the test included the best items from the two previous forms. This form of the test used the concept of deviation IQ rather than ratio IQ. The previous test manuals contained ratio IQ tables, whereas the 1960 edition's manual included deviation IQ tables. Use of deviation IQ meant that the performance of a test taker was compared with the performance of other people of the same age level in the standardization sample. The mean score is taken to be 100 with a standard deviation of 16. The score of a test taker is converted into a standard score using these values. Using the ratio IQ one could assess a person's intelligence but not his standing in comparison to other test takers. But the use of deviation IQ made it possible to estimate a test taker's relative position with reference to the standardization sample. However, the criticism against unrepresentativeness of the test continued.
The 1960 revision did not involve re-standardization or a new normative sample.

The 1972 Revision:

In 1972 an improved normative sample was taken which comprised 2100 subjects. The sample included nonwhites as well. About 100 subjects for each Stanford- Binet age level were included. Still the sample was criticized for not taking enough non-whites.

The 1986 Version: (The Stanford Binet: 4th Edition):

This version of the Binet Scale overcame the problems for which it was criticized. A standardization sample of 5000 subjects was used. The subjects belonged to 47 states of the U.S. and the District of Columbia. Geographic region, community size, ethnic group, age, and gender were considered for stratification of the sample.

The following content areas are covered in the latest scale:

1. Verbal reasoning
2. abstract/ visual reasoning
3. quantitative reasoning
4. short- term memory

The subtests of the scale are as follows:

1. Verbal reasoning
    i. Vocabulary
    ii. Comprehension
    iii. Absurdities
    iv. Verbal relations
2. Abstract/ visual reasoning
    i. Pattern analysis
    ii. Copying
    iii. Matrices
    iv. Paper folding and cutting
3. Quantitative reasoning
    i. Quantitative subtest
    ii. Number series
    iii. Equation building
4. Short- term memory
    i. Bead memory
    ii. Memory of sentences
    iii. Memory of digits
    iv. Memory of objects

Administration of Stanford-Binet Test:

Individual-oral administration is used. The examiner begins from a mental level at which he finds out the subject to be. Items from succeeding levels are asked. The test ends when they reach a level where no items are successfully attempted. The administrator establishes the 'basal age' and a 'ceiling' for each test. Basal age refers to "the lowest level or point where two consecutive items of approximately equal difficulty are passed" and ceiling is "the point at which at least three out of four items are missed" (Kaplan & Saccuzzo, 2001, p. 271). Scores for all 15 tests are attained, and are converted into standard age scores, with mean of 50 and SD of 8. Four area- content scores result from the grouping of individual tests into content areas. The mean in this case is 100 and SD is 16. a composite score is also calculated.

Some Sample Items from Early Versions Of Simon-Binet Scale:

Three years: Shows nose, eyes and mouth. Repeats two digits, describes objects in a picture, gives family name and repeats a sentence of six syllables.

Four years: Gives own sex, names key, knife, and penny, repeats three digits, compares the length of two lines.

Five years: Compares two weights, copies a square, repeats a sentence of ten syllables and counts four pennies.

Six years: Distinguishes between morning and afternoon, defines objects in terms of their use, copies a shape, counts 13 pennies and compares faces from the aesthetic point of view.

Seven years: Identifies right hand and left ear and describes a picture, follows precise directions and names four colors.

Eight years: Compares two remembered objects, counts from 20 to 0, indicates omissions in pictures, gives day and date and repeats five digits.At the highest level of **Fifteen years:** Repeats seven digits, gives three rhymes, repeats a sentence of 26 syllables, interprets a picture and solves a problem from several facts

The Wechsler Scales:

The Wechsler scales are perhaps the most commonly used intelligence tests. These were developed by psychologist David Wechsler. Three Wechsler intelligence tests that are available to us at present are:

i. **Wechsler Adult Intelligence Scale: third edition or WAIS-III:** this scale is meant for ages 16 years to 89 years
ii. **Wechsler Intelligence Scale for Children: third edition or WISC-III:** this scale is meant for children aged 6 to 16
iii. **Wechsler Preschool and Primary Scale of Intelligence-Revised or WPPSI-R:** this scale is for children aged 3 years to 7 years three months.

The first scale was developed by David Wechsler was published in 1939, known as the W-B I or Wechsler - Bellevue. Wechsler was employed by Bellevue Hospital in Manhattan. The test was not an age scale like the Binet scale. Rather it was a point scale in which credit for every correct response was given. In 1942 another form, an equivalent one, was also developed. Known as W-B II, this form is rarely talked about. The original scale had some problems, particularly those related with standardization. The standardization sample used for this scale comprised 1081 whites as subjects. Most of them belonged to New York. But Wechsler soon removed the initial flaws, revised W-B I, and developed the WAIS (Wechsler Adult intelligence Scale) in 1955. The WAIS was revised again and WAIS-R was introduced in 1981. The latest version is the WAIS- III that was developed in 1997.

The scale consists of two categories of scales, verbal and performance. The details of the two types of subtests are as follows:

Verbal Scale:

i. Vocabulary
ii. Similarities
iii. Arithmetic
iv. Digit span
v. Information
vi. Comprehension
vii. Letter- numbering sequencing

Performance Scale:

i. Picture completion
ii. Digit symbol- coding
iii. Block design
iv. Matrix reasoning
v. Picture arrangement
vi. Symbol search
vii. Object assembly

Administration of WAIS and WISC or the latest prevalent forms is time consuming because it requires individual administration.

Psychometric Properties of WAIS-III:

A standardization sample comprising 2450 subjects was used. The subjects were all adults. Thirteen age groups were taken, starting from 16-17 and going up to 85-89. The sample's stratification was done on the basis of gender, race, education, and geographic region. This information was obtained from the 1995 census of the U.S.

The Meaning of IQ Test Scores:

The commonly followed standards of interpreting IQ scores are as follows:

IQ score	*Rating*
< 70	Retarded
85	Borderline
100	Average
Above 115	Superior
Above 140	Gifted

Lesson 26

Intelligence Tests

The Kaufman Scales:
The husband and wife duo, Alan and Nadeem Kaufman, have contributed to intelligence testing by developing the following tests:

i. K-ABC: Kaufman Assessment Battery for Children (Kaufman & Kaufman, 1983)
ii. K-BIT: Kaufman Brief Intelligence Test (Kaufman & Kaufman, 1990)
iii. KAIT: Kaufman Adolescent and Adult Intelligence Scale (Kaufman & Kaufman , 1993)

K-ABC: Kaufman Assessment Battery for Children:
The following global scales are included in this test battery:

i. Sequential processing. It includes subtests; Hand movement, number recall, word order.
ii. Simultaneous processing. It includes subtests; Magic window, face recognition, Gestalt closure, triangles, matrix analogies, spatial memory and photo series.
iii. Mental processing Composite (combining i and ii)
iv. Achievement. It includes subtests; expressive vocabulary, faces and places, arithmetic, riddles, reading/decoding, reading/ understanding
v. Nonverbal. It includes subtests; face recognition, hand movements, triangles, matrix analogies, spatial memory and photo series.

The scales further consist of subtests, 16 in number other than the nonverbal scale.
As can been seen from these four areas, the battery is focusing on information processing. The information processing approach looks into the way information is processed. A national sample of 2000 American children was used for standardization of this battery. The ages of the children ranged between 2.5 to 12.5 years. The sample considered a number of characteristics of the subjects; age, gender, geographic region, parental education, community size, educational placement, and race. In addition, gifted and talented, mentally retarded, and learning disabled were also included on the basis of their proportion in general public.

K-BIT: Kaufman Brief Intelligence Test:
The K- BIT, meant to be used with ages 4 to 90 years, is a quick screening instrument. It is an individual test for assessment of intellectual functioning that involves individual administration. The K-BIT yields three scores namely, verbal, non-verbal, and composite. The verbal subtest includes 45 Expressive Vocabulary items and 37 Definitions. The non-verbal subtest contains 48 matrices. This test used nearly 20% of the KAIT standardization sample.

KAIT: Kaufman Adolescent and Adult Intelligence Scale:
The KAIT was developed for measuring intelligence of subjects aged 11 to 85 plus. It consists of a Crystallized scale and a fluid scale. The former measures concepts leaned from schooling and acculturation. The latter is about the ability to solve new problems. Three subtests are used in each scale in the Core Battery. An additional plus point is the possibility of using an Expanded Battery for subject who are suspected to have neurological damage. In such case, any of four specified subtests are added to the original battery. For the test takers who are cognitively impaired to the extent that they cannot complete the whole battery then a brief mental status test is also included. This test assesses attention and orientation.

Differential Ability Scales:
The Differential Ability Scales or DAS were developed by C. D. Elliot (1990) in gReat Britain. It is actually a revised form of the British Ability Scales (Elliot, Murray, &Pearson, 1979).
There are three major components of the DAS that contain 20 subtests in all.

The Core Subtests:

i. Block Building
ii. Verbal Comprehension
iii. Picture Similarities
iv. Naming Vocabulary
v. Early number Concepts

vi. Copying
vii. Pattern Construction
viii. Recall of Designs
ix. Word Definitions
x. Matrices
xi. Similarities
xii. Sequential and Quantitative reasoning

Diagnostic Subtests:
i. Matching Letter-Like forms
ii. Recall of digits
iii. Recall of objects
iv. Recognition of Pictures
v. Speed of Information Processing

Achievement Test:
i. Basic number skills
ii. Spelling
iii. Word reading

The four core subtests are used with preschoolers, preschoolers aged 2 years and 6 months to 3 years and 5 months. Six core subtests are meant for preschoolers of age 3 years and 6 months to 5 years and 11 months. Six subtests are for school level, ages 6 years to 17 years and 11 months.
The standardization sample included 3475 persons, aged 2 years and 6 months to 17 years and 11 months, representing the target population of all non-institutionalized English- proficient individuals in the U.S. The subjects' age, gender, race/ethnic origin, parental education, and geographic region were taken into consideration in sample selection for the sake of stratification.

Cultural Biases And Intelligence Tests:
Tests used to assess people's intelligence have been frequently criticized for being biased against particular groups of people. Culture-fair IQ tests are developed and used for overcoming this problem. These tests do not discriminate against any minority or cultural group, e.g. Raven's Progressive Matrices

Some Significant Questions Pertaining To the Use of IQ Tests:
Is the test a validity test?
Is it a reliable test?
Was it standardized?
Is it being used with people similar to those included in the standardization sample?
Is it being used with people different from those included in the standardization sample?
Are the differences drastic and serious?
Have the consequences been anticipated and weighted against the expected benefits?
Can the cultural background of the test taker affect the results?
Can the test takers ethnic origin affect the test results?
Can the administration be problematic due to personal, environmental, physical or other reasons?
All these questions need to be considered, answered, and tackled in case problems are seen or foreseen.

Alternative Formulations:
These include:
- Moral intelligence
- Social intelligence
- Emotional intelligence

Moral Intelligence:
Given by Coles (1997) and Hass (1998)
- It is the ability to differentiate between right and wrong

- More comprehensively, it is the capacity of making right decisions that are not only beneficial for one self but to others as well

Social Intelligence:

Given by Hough, 2001; Riggio, Murphy, & Pirozzolo, 2002)

- Manifested as SQ
- Ability to understand and deal with people; salesmen, politicians, teachers, clinicians, and religious leaders exhibit this type of intelligence
- It is also the ability to understand and deal with in own self by identifying one's thoughts, feeling, attitudes and behaviors

Emotional Intelligence (EI):

- It is the type of social intelligence which is the ability to cope with one's own and other's emotions, to differentiate between them and use information for guiding one's thoughts and actions.
- Indicated by the EQ of a person.

It includes these aspects:

1. Self-awareness
2. Managing emotions
3. Empathy
4. Handling relationships

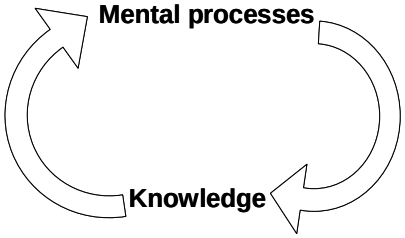
Lesson 27

Piagetian Approach: Measurement of Cognitive Development

Different from the traditional psychological measurement, we see an approach that measures cognitive development but does not employ tests the way we do in routine. This approach is the Piagetian approach, the approach introduced by Swiss psychologist Jean Piaget. Piaget presented his theory of cognitive development and introduced his methodology for studying and understanding the same. Cognitive development is the process whereby the development of children understands of the world as a function of age and experience takes place.

In order to understand the Piagetian approach and methodology, let us refresh our knowledge about cognitive development. Cognition is the process of knowing as well as what is known. It includes "knowledge" which is innate/ inborn and present in the form of brain structures and functions. We 'remember' the physical environment in which we were brought up and develop perceptual constructs or knowledge accordingly (seeing, hearing, sounds etc).

Cognition refers to 'mental processes' that people use to gather/ acquire knowledge, and also the knowledge that has been gathered/ acquired subsequently used in mental processes. Cognition and knowledge, therefore, can be said to have a circular relationship.



Cognitive development involves:

- Language,
- Mental imagery,
- Thinking,
- Reasoning,
- Problem solving and
- Memory development
-

Jean Piaget's Theory of Cognitive Development:

Piaget (1896-1980) was a Swiss psychologist, who became interested in epistemology i.e., knowledge and knowing as a result of his study of philosophy and logic. This interest in observation and epistemology laid foundation of his theory of cognitive development.

Piaget was influenced by Henri Bergson's Creative Evolution, unlike most of the other psychologists who were impressed by Darwin's theory of evolution. Bergson believed in divine agency instead of chance as the force behind evolution: life possesses an inherent creative impulse. After having secured a position in Alfred Binet's laboratory in Paris he got a chance to observe children's performance, their right and wrong answers. Piaget's work and observation generated an interest in children's mental processes. The real shift took place when he started observing his own children from birth onwards. He kept records of their behavior and used them to trace the origins of children's thoughts to their behavior as babies; later on he became interested in the thought of adolescents as well

These experiences resulted in two significant consequences:

1. Piaget's theory of cognitive development
2. Piagetian method of study

Piagetian Method of Investigation:

Piaget's method is known as the clinical approach which is a form of a structured observation. Piaget used to present problems/tasks to children of different ages, asked them to explain their answers. Their explanations were further probed through carefully phrased question.

Piaget's Theory of Cognitive Development:

Cognitive Development takes place in four stages in a set sequence. The sequence of stages is invariant. The age range of each stage is also described but the age is not invariant. Age specification is arbitrary and different children may perform the same task at different age levels. The organization of behavior is qualitatively different

in different Stages. Children throughout the world pass through a series of four stages of cognitive development in a fixed order.

Piaget's Stages of Cognitive Development:
1. Sensorimotor stage
2. Preoperational stage
3. Concrete operational stage
4. Formal operational stage

Sensorimotor Stage: Infancy: Birth-2 years
The child's thought is egocentric and confined to action schemes. Development is very rapid in this stage but thought processes are limited to the immediate world of the child. Development of object permanence and development of motor skills takes place. The child has little or no capacity for symbolic representation.

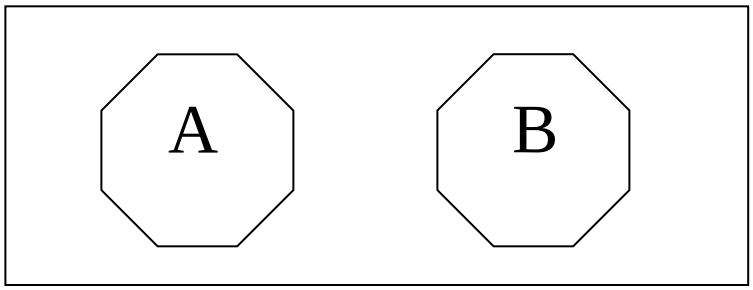
Preoperational Stage: Preschool: 2-7 years
Development of representational thought takes place. The child's thinking is intuitive not logical. A significant aspect of development at this stage is the development of language and symbolic thinking. Thinking remains egocentric.

Concrete Operational Stage: Childhood: 7-11 years
At this stage the child's thinking becomes systematic and logical, but only with regard to concrete objects. Development of conservation and mastery of concept of reversibility takes place.

Formal Operational Stage: Adolescence and adulthood: 11 years onward
Abstract and logical thought develops at this stage. The person can deal with the abstract and the absent.

Some Piagetian Tasks:
These tasks measure the acquisition of various concepts. The acquisition of concepts is progressive. Children of different age levels or children belonging to different stages of cognitive development show different levels of acquisition,

The 'A-B' Search Task:
This task is meant for sensorimotor stage children. It is about the acquisition of the concept of object permanence. In this task two hiding places, A and B, are used. The places are in front of the child. The places can be something like two place mats or napkins on a table under which an object may be hidden. An object is hid under either of two and the child has to look for it. The children's responses vary according to the stage of development at which they are. Even within the same stage children of different age levels give different responses. Their responses may be something like this:

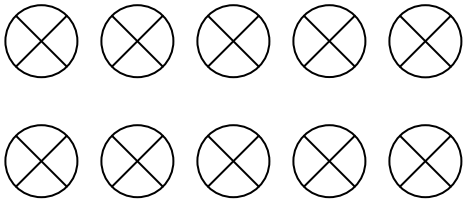


4- 8 month olds: these children will not search for the object even when it is hidden under A in front of them.
8-12 month olds: they will try to search for the object and will find it under A. when the object is shifted from under A to under B, the child will still look for it under "A" even when it was hidden in front of the child. This shows egocentric thinking. It is also known as 'A- not B' error.
12-18 months: The child can accurately search for the object.
18-24 months: The child can not only search for the object but can also experiment with it and other similar objects.

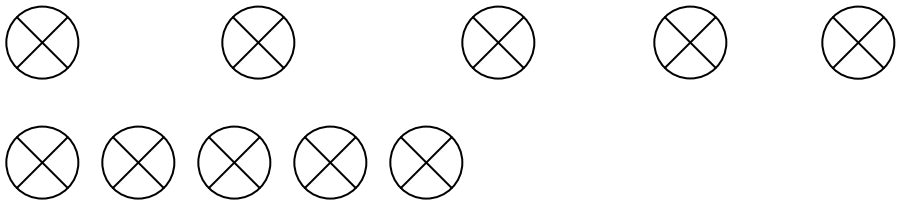
Conservation Tasks:
Conservation is a concept according to which the some properties of an object/ mass matter remain unchanged or invariant while some others have been changed. For example the weight of an object will remain the same when its shape has been changed; the number of objects remains unchanged while their arrangement is changed. Children learn the conservation of mass and number earlier (around 5-6 years of age) than conservation of weight (around 8-9 years of age).

Conservation of Mass:
Play dough, plasticine, or clay can be used for this task. Take two same sized balls of the dough and ask the child if the two have the same amount of dough/clay in them. Let the child feel it and then answer. When he says yes they are the same amount then flatten one ball like a pan cake or chapatti and ask if they still have the same amount of the pliable material in them. Children at different cognitive levels will respond differently. If the child says the two had different amounts of mass, and then ask why does he think so.

Conservation of Number:
Take ten coins and arrange them in two parallel rows. Ask the child if there are buttons in same number in each row.

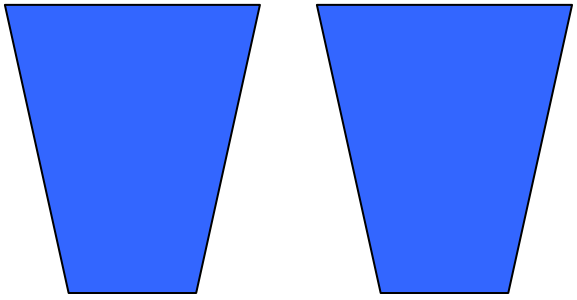


When the child says yes, then rearrange the buttons and spread buttons in one row distantly so that the row appears to be longer than the other one. Now ask the child if the two rows contained the same number of buttons. Children belonging to different levels will respond differently. Those who have not acquired the concept of conservation of number will say that one row was longer than the other one.

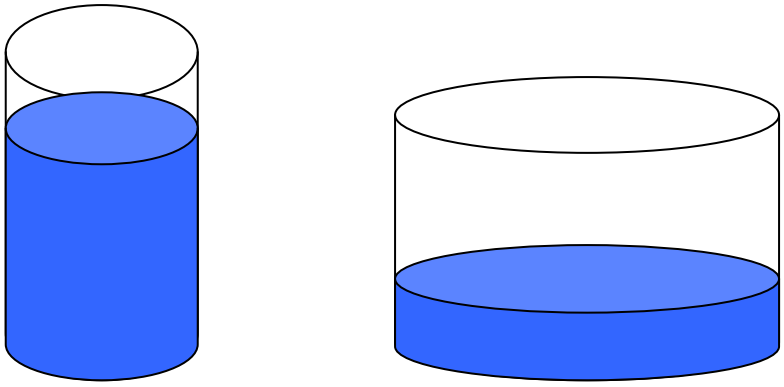


Conservation of Weight:
Once again two, same weight, play dough or clay balls may be used. Ask the child if the two balls had the same weight. Once the child agrees then change the shape of one of the two balls and convert it into an oblong. Now ask the child if they were of the same weight. Children belonging to different levels will respond differently. Those who have not acquired the concept of conservation of weight will say that the ball and the oblong were of different weight, whereas those who have acquired the concept of conservation of weight will say that the two objects were of the same weight.

Conservation of Volume:
Put equal amount of water in two same sized glasses or jars.

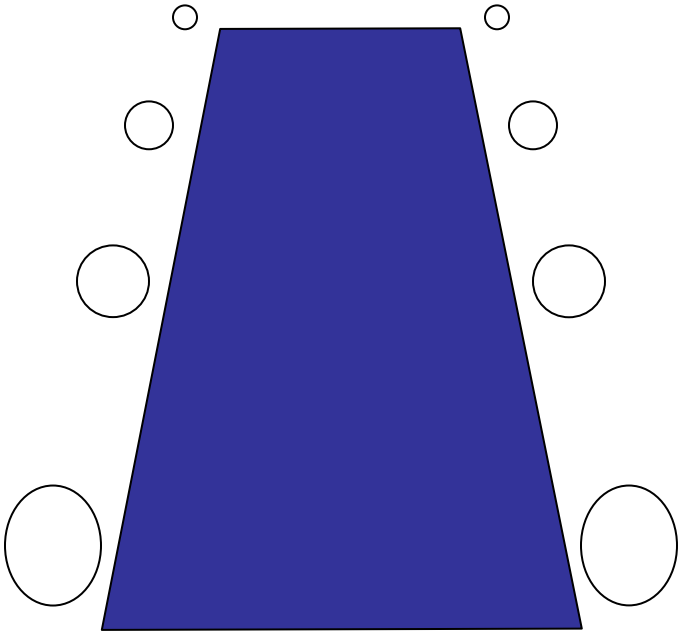


Ask the child if the two contained the same amount of water. When the child agrees and says yes, then take two other, differently sized, containers and pour water in them, one will have a lower level of water and the other one will have higher level. Ask the child if the two containers had the same amount of water in them. Once again children at different levels of cognitive development, with reference to conservation of volume, will give different answers.



Perspective:

The child is asked to imagine if she is standing at the beginning of a long road, and there are trees on both sides of the road. She is asked to draw and tell how the road and the trees would look from her position.



Significant Influences on Cognition:

Socio- Cultural Factor:

- Given and debated in the early 1900s socio-cultural approach has now regained interest among cognitive scientists
- It states that cognitive ability does not start with the anatomy/ biology of the individual or only with the environment: the culture and society into which the individual is born provide the most important resources/ clues for human cognitive development.
- They provide the context into which the individual begins his experience of the world.
- Social groups help in person's cognitive development by placing value/ importance on learning certain skills, thereby providing all important motivation that the person needs and requires in order to learn and exhibit those skills or behaviors. This results in cognitive development
- One perspective about cognitive ability suggest that there is some sort of innate potential existing within an individual

- Another suggests that there is potential within the socio- cultural context for development of the individual. The individual is born into a society of potential intellect. Knowledge will develop largely based on the evolution of intellect within the society and culture.

Motivation, Cognition and Learning:

- It is believed that cognitive ability alone cannot account for achievement; motivation is also important in acquiring/ attaining cognitive skills and abilities.
- People learn information that corresponds to, and is in accordance with, their view of the world. They learn skills that are meaningful to them. e.g. children who are born in a poor family may not give any attention or importance to the formal education and as adults, they may pass on similar beliefs and attitudes to their off springs.
- Motivation determines whether or not one is capable of learning. Whether one learns well or not, depends on one's own view and that affects the ability to learn. The motivational condition largely depends on the way the culture responds to achievements and failures. There are culturally developed attitudes about the probability of learning successfully after one has initially failed to learn. These attitudes can greatly affect future learning.

Lesson 28

Individual Tests of Ability for Specific Purposes

A numbers of tests have been designed for individuals with special needs like learning disability and/or memory problems.

Learning Disabilities:
One of the areas that cause most concern for psychologists and educationists is learning disabilities. In all mainstream schools one may expect to find children who face problems in routine education because of certain learning disabilities. A child's average achievement scores (marks) at school may be lower than the expected score for his age level due to the same disability. If a child of average cognitive development or IQ cannot perform or achieve what other average children of same intelligence may do, then the child may be suspected to be a learning disabled child. For this purpose such tests will be employed that can detect learning disability. The difference between IQ and achievement amounting to 1.5 to 2 standard deviations is considered to be indicative of learning disability.

Illinois Test of Psycholinguistic Abilities (ITPA):
The test is based on modern concepts of information processing. ITPA is based on the theory that inability to respond correctly to stimuli does not result from defective output alone. The input has a role to play as well. Input refers to the information-processing system. The input comes from an external stimulus. Our response to it or information processing takes place in three stages:
Stage 1: incoming information is received through senses
Stage 2: analysis or processing of information is done
Stage 3: the response takes place

The Illinois test provides the independent measure of all these three stages. The three subtests measure individual's ability to receive visual, auditory or tactile inputs. Three further subtests provide independent measure of processing in these sensory modalities. There are other measures of motor and verbal output as well.
The test is designed for children aged 2-10 years. ITPA is widely used among educators, psychologists, learning disability specialists and researchers. But the psychometric properties of this test are widely criticized. The test provides no validity and reliability data. The test norms have been obtained from middle class population and it contains culturally loaded content. Therefore it may not be appropriate for lower-class or minority groups.
The ITPA subtests include the following:

- Auditory Reception
- Visual Reception
- Auditory Association
- Visual Association
- Verbal Expression
- Manual Expression
- Grammatic Closure
- Visual Closure
- Auditory Sequential Memory
- Visual Sequential Memory
- Auditory Closure
- Sound Blending

Woodcock-Johnson Psycho-Educational Battery – Revised:
The Woodcock-Johnson Psycho-Educational Battery - Revised is a commonly used measure of children's achievement measuring various aspects of scholastic ability. The test measures cognitive abilities, aptitudes, achievement and interests. Learning problems can be identified by comparing subjects' cognitive ability score with their achievement.
If a different of 1.5 to 2 SD is found between the cognitive ability and achievement of a child, then it will be considered a major discrepancy that is taken to be indicative of learning disability.
The tests of cognitive ability include:

- Picture vocabulary
- Spatial relations
- Memory for sentences
- Concept formation
- Analogies, and
- A variety of mathematical problems

The achievement tests include:

- Letter and word recognition
- Reading comprehension
- Proofing
- Calculation
- Science and
- Social science and humanities

The tests of interest level cover math, language and physical and social interests. The scores can be described in terms of percentiles, which can further be converted into standard scores, with a mean of 100 and SD equal to 15. If a child is on the 50th percentile, or mean, in cognitive score, with an achievement score that is 2 SD below the cognitive score, then it can be taken as indicating learning disability. With normative data of more than 4700 including members of different race, gender, urban and rural status, these tests have good psychometric properties.

Visiographic Tests:
Visiographic tests require a subject to copy various designs. These test are useful form many kinds of brain damage.

Benton Visual Retention Test (BVRT):
This test is used for measuring brain damage and psychological deficit. The BVRT assumes that brain damage easily impairs visual memory ability. The test is designed for individuals aged 8 and older. The subjects have to reproduce geometric designs presented briefly before them and then removed. The subject loses points for mistakes and omission. As the number of errors increases subject approaches the organic (brain-damage) range.

Bender Visual Motor Gestalt Test (BVMGT):
The Bender Visual Motor Gestalt Test (BVMGT) is one of the most popular individual tests. The test has nine geometric figures that subject has to copy. The test is scored on the bases of errors. Norms are available for children aged 5- 8 years. The one or two errors by the age of 9 years are considered normal. But if individuals over 9 make more errors this may be indication of some deficit. The individuals with more errors can be said to have mental age less than 9 years (low intelligence), brain damage or emotional problems. Though the test has number of scoring systems the reliability of the test is questioned.

Memory-for-Designs Test (MFD):
The MFD is a short time administered (10 minutes only) simple drawing test. The test measures perceptual motor coordination of individuals from 8 ½ to 60 years of age. The subjects are shown simple designs for a short duration and are then asked to reproduce them. The drawings are given scores from 0 to 3 depending on how close or similar they were to the original designs.
The scoring of the 15 drawings can indicate brain injury and brain disease with the help of provided reference tables according to age and intelligence. The test has good psychometric properties with additional needs for validity.

Torrance Tests of Creative Thinking (TTTCT):
Creativity can be defined as "the ability to be original, to combine known facts in new ways, or to find new relationships between known facts" (Kaplan & Saccuzzo, 2001, p.333). The Torrance Tests of Creative Thinking (TTTCT) measure different aspects of creativity including fluency, originality and flexibility.

Fluency: Fluency is about the ability to generate a variety of solutions to problem. People would score high on fluency if their solutions are distinct. The more distinct are the solutions, the higher is the score. The individual's fluency is measured by his/her ability to provide as many solutions as on can find.

Originality: Originality has to do with the uniqueness, novel ness, and unusual nature of solutions. One can be said to be a creative person if one can come up with novel ideas or solutions. The unusual and unique solutions, which are different from the usual, conventional, and expected ones, add to originality score of an individual.

Flexibility: flexibility is about the ability to shift from one stand point or strategy to another for finding solutions of problems. People can be said to have a flexible approach if they do not mine shifting positions in problem solving. If one strategy does not seem to be working, such people would try other strategies. Flexibility is measured by gauging a person's ability to switch to different approaches of problem solution. The test is useful for applied practitioners but needs more research for enhancing its psychometric properties.

Wide Range Achievement Test- 3 (WRAT-3):

The intelligence tests measure what an individual may achieve whereas what an individual has actually achieved is measured through achievement tests. The IQ tests are about the potential while achievement tests are about the use of potential. The scores on achievement may be indicative of people's intelligence test but not necessarily always. Factors such as interest, motivation, training, prior knowledge or previous exposure may affect the way one has used one's potential i.e., intelligence. Therefore both IQ and achievement tests are used for assessing one's ability, depending on the situation and purpose. Usually achievement tests are used in groups, but some individual achievement tests are also available. The Wide Range Achievement Test- 3 (WRAT-3) is most widely used achievement test that measures the grade-level functioning in reading, spelling and arithmetic. The test can be used for children aged 5 and older. The WRAT-3 is widely criticized for its grade-level reading ability.

Lesson 29

Group Testing

Group versus Individual Tests:

So far you have learnt about varieties of ability or intelligence tests that are primarily individual tests. These tests are administered individually and are chosen, used, and interpreted according the needs of individual subjects. We also mentioned some tests in the previous sections that can be used in group administration as well. In the present section we will be exclusively discussing group tests. Such tests are designed and developed for group administration. No doubt, the individual tests are their own advantages which cannot be denied. For example there is a certain degree of rapport between the examiner and the test taker. Test administration is comparatively flexible in the sense that instructions and questions can be repeated and at time rephrased according to the educational/cultural/ anxiety level of the examinee. In addition, if testing is being done for diagnostic purposes then individual testing is the best and perhaps the only approach used. However, there are situations where we prefer to use group testing rather than individual testing.

Which type of test administration shall we use, individual or group, will depend on the purpose of the test. There will be occasions when the disadvantages will not affect much the findings and the advantages will be considered to be more important. On the other hand there will be cases where individual testing will be the preferred approach. Therefore, as is quite understandable, that the mode of test administration will be decided on the basis of purpose for which the test will be used.

Characteristics of Group Tests:

Group tests are the tests that have been designed in such a manner that a large number of test takers can do them simultaneously. The instructions address the entire group not individuals, are delivered only once at the beginning of the test and are not repeated for anyone. These tests are usually timed tests and the examiner does not do or say anything that may interrupt the testing process. Group tests are generally paper- pencil tests, although computerized administration is also becoming popular in developed countries.

Group tests mostly have a multiple choice format. Usually blank circles, or spaces,
are provided in front of every response option. The test taker blackens the circle, or places a tick mark in blank space, or encircles the option number (a, b, c, d etc.). The answer sheets are marked with the help of a key. Answer keys can be of different types but one common type is the one which is designed like a stencil. A sheet like the test's sheet is used in which holes or slots are made on the right option. The stencil is placed on top of the answer sheet. If the space under the slots has been darkened then the item is counted as right. The answer sheets can be marked by computers also, which is the commonly used approach nowadays. Group tests are also administered with the help of computers. In computerized testing, a mouse or joy stick is used to mark the answers. But this approach can be adopted only when sufficient number of computers is available.

Advantages of Group Tests:

- Such tests save time of administration
- A large number of test takers can be examined simultaneously
- Group tests are a good source of quick data collection for research projects
- If quick decisions are to be made, such as screening or school admissions, then group tests are very useful
- Test administration is easy from the point of view of the examiner, especially because there is little pressure on the examiner for taking notes of individual expressions, explanations, or clarifications on part of the examinee.
- There is little impact of the personality of the examiner on the performance of the examinees, which can be the other way round in case of individual testing.
- The role of the examiner is the minimum. If desired, tape recorded instructions may also be used.

Disadvantages of Group Tests:

In spite of a number of advantages, the group administered tests entail some disadvantages as well:

- Group tests are very impersonal. The personal touch and rapport that is an ingredient of individual tests is not found in group tests.
- Group administration is not flexible like individual testing. In individual administration, the examiner may repeat instructions, or may rephrase them according to the needs of the test taker.

- In group testing all subjects attempt all items, whereas in individual testing considering basal rule and ceiling rule the testing process may be tailored according to the ability level of the test taker.
- In group testing the test takers' motivation level cannot be judged, maintained, or enhanced.
- The examiner is deprived of the verbal and non-verbal cues that can reflect the examinees' anxiety, confusion, or discomfort etc. Nor can the examiner do anything to put individual examinees at ease.
- Most group tests require the subjects to have reading skills and the ability or practice in paper- pencil manipulation to record their responses. There may be some test takers in the group who do not have these abilities and who may remain undetected because of the nature of the testing process. Therefore their scores will not be accurately representing their true potential.

Group Tests and Batteries:
Following is a description of some of the commonly used and popular group administered intelligence tests. These are just a few of the large number of available intelligence tests. This will also give you a flavor of the various purposes for which these tests may be used.

Kuhlmann- Anderson Test- 8th Edition (KAT):
KAT is a group intelligence test. It can be used with kindergarten to 12th grade children. It is primarily a non-verbal test and it is so for all grade levels. The test has eight levels that cover these school grade levels. A number of tests are found in each level. KAT is popularly used as a very good test of mental ability. The test has strong psychometric properties,
Its norms obtained from a sample of more than 10000, with high reliability and validity coefficients.

Henmon-Nelson Test (H-NT):
H-NT is also considered to be a good test of mental ability or intelligence. The test has two sets of norms, grade-wise and age- wise. The 90 item test can be completed in around 30 minutes and can be used for all grade levels. Rather than considering multiple intelligences, this test provides a single score relating to Spearman's g factor. H-NT also has high reliability in the 90s, and validity coefficients of 50s to 90s.

Cognitive Abilities Test (COGAT):
The COGAT yields three scores; verbal, nonverbal, and quantitative. The test has been designed very carefully and attempts have been made to remove sources of difficulty such as cultural bias. It was developed specially for poor readers, those who were poorly educated, and the ones who had English as a second language. The statistical analysis of items was done to identify if there were any items that predicted differentially for white and minority students. Such items were removed from the test. the purpose was to eliminate bias against any group from test content. There are three subtests which can be completed in 32-34 minutes each. However the test manual recommends that this testing be completed in 2-3 days. Good psychometric properties have been reported for the test. The most positive point about this test is that it is a tool better than other tests for the assessment of culturally diverse, minority, and economically disadvantaged children (Kaplan & Saccuzzo, 2001).
Some research evidence suggests that verbal underachievement can be measured well with this tool (Langdon, Rosenblatt, & Mellanby, 1998). It has also been reported as a sensitive discriminator for giftedness (Harry, Adkins, & Sherwood, 1984) and a tool that can make good predictions about future performance (Henry & Bardo, 1990).
The reliability coefficients reported for COGAT are very high, in the 90s.

Specific Purposes Tests

The Scholastic Assessment Test (SAT-I)
The Scholastic Assessment Test or SAT-I, was previously known as Scholastic Aptitude Test or SAT. First used in 1926, the test is the most commonly used college entrance test in the U.S. SAT-I has two parts that contain the reasoning tests which comprise further subtests.
Verbal Reasoning:
This part consists of 78 questions in all, to be completed in 75 minutes. The distribution of items is as follows:

- Sentence Completion; 19 questions
- Critical Reading; 40 questions
- Analogies; 19 questions

Mathematical Reasoning:
This art contains 60 questions. The questions are distributed in the following subtests:

- Regular Mathematics; 35 multiple choice questions
- Student- Produced Responses; 10 questions
- Quantitative Comparisons; 15 questions

The test norms were obtained from a large representative sample. SAT-II is also available that comprises Subject Tests including:

- A direct writing test
- Tests in Asian languages
- English-as-a- second Language Proficiency Test

Graduate and Professional School Entrance Tests:
The graduate school entrance tests are widely used for admission in graduate school and professional degree programs like medicine, art, and law etc.

Graduate Record Examination Aptitude Test (GRE):
Graduate Record Examination Aptitude Test is the most commonly used graduate-school entrance test. The general scholastic ability is measured by GRE along with grade point average and letter of recommendation as process of general selection in school. The three sections of GRE include verbal (GRE-V), quantitative (GRE-Q) and analytic (GRE-A). The verbal section includes measurement of reasoning, antonyms, analogies and paragraph comprehension. GRE-Q purports to measure reasoning, algebra and geometry. In addition to that GRE also measures the general achievement in at least 20 majors like psychology, history and chemistry.
Though the psychometric properties of GRE are not very impressive it is used as a relatively strong instrument. The studies on relationship of Grade Point Average (GPA) and GRE have shown the correlation from .22 to .33. The GRE, in addition with GPA, has been proved to be a good predictor of graduate success.

Miller Analogies Test:
The Miller Analogies Test (MAT) is the second major, widely used, scholastic aptitude test. The MAT is a 50 minute verbal test that measures student's ability to find logical relationships for 100 different analogy problems. The MAT offers special norms for various fields. The research has indicated that MAT has an age bias as its scores over predicted the GPAs for 25-34 years group and under predicted for the age 35-44 years.
The psychometric properties are adequate for MAT but it does not predict research ability, creativity, and other factors important in graduate-school.

Nonverbal group ability tests:
Nonverbal tests are used for evaluation of individuals without the use of language. The individuals are usually asked to perform some tasks like drawing, solving maze or identify problem figure from set of figures presented before them.

Raven Progressive Matrices:

Raven Progressive Matrices (RPM) is a non-verbal multiple choice measures of the general intelligence. In each test item, the subject is asked to identify the missing element that completes a pattern. The test can be administered to groups or individuals of 5 years old to older adults. The Raven's test contains 60 matrices with a missing part presented in graded difficulty. The subject has to select appropriate pattern from a group of eight options.

The research has shown that test measures general intelligence along with the capacity to think clearly and make sense of complex data. In spite of criticism over psychometric properties RPM has widespread use for children, language-handicapped and the culturally deprived. The updated manual of Raven Progressive Matrices provides us the comparison of performance of children from major cities of the world. The RPM has minimized the effects of language and culture.

Goodenough-Harris Drawing Test:

The Goodenough-Harris Drawing Test is the quickest, easiest and less expensive nonverbal test for measuring intelligence. The subject is asked to draw a whole human figure. The test is scored for each item included in drawing. The subject can get out of 70 possible points. The G-HDT scoring follows the age differentiation principle; older children tend to get more points because of greater accuracy. The test has good psychometric properties. The scores on the G-HDT can be related to Wechsler IQ scores. The test can be more appropriately used in combination with other tests of intelligence.

IPAT Culture Fair Intelligence Test:

The purpose of nonverbal and performance tests is to remove cultural influences in intelligence and learning. The IPAT Culture Fair Intelligence Test is a paper pencil test for three levels; age 4-8 and mentally disabled adults, age 8-12 and randomly selected adults and high-school age and above-average adults.

Group Tests for Specific Purposes:

Along with large body of group tests for intelligence, academic aptitudes, personnel and occupation selection, a number of tests are available for specific populations. For example Black Intelligence Test of Cultural Homogeneity (BITCH) is used as culture-fair intelligence test for African Americans.

Tests For Use In Industry: Wonderlic Personnel Test:

This test helps in making decisions concerning employment, placement and promotion. The Wonderlic Personnel Test (WPT) is based on population Otis Self-Administering Tests of Mental Ability. The WPT is a quick (12-minute) test of mental ability in adults. The test lacks in validity documentation. The test is widely used for employee-related decisions in industry.

Tests for Assessing Occupation Aptitude:

The General Aptitude Test Battery (GATB) is widely used ability test among a variety of available group tests; to measure the aptitude for various occupations. The U.S Employment Service developed GATB for employment decisions in government agencies. The test provides scores for motor coordination, perception and clerical perception along with verbal numerical and spatial aptitudes. The test is criticized for its normative data.

The Differential Aptitude Test, The Bennett Mechanical Comprehension Test and Revised Minnesota Paper Form Board Test are used for mechanical ability and clerical competence.

Armed Services Vocational Aptitude Battery:

Armed Services Vocational Aptitude Battery (ASVAB) was designed for postsecondary school student and students of 11-12 grades originally for use in Defense Department. The test scores can be used for both in educational and military settings. The ten subtests of ASVAB include: general science, arithmetic reasoning, word knowledge, paragraph comprehension, numeral operations, coding speed, auto and shop information, mathematics knowledge, mechanical comprehension and electronics information. These ten subtests are grouped into composites: academic composites (academic ability, verbal and math); four occupation composites (mechanical and crafts, business and clerical, electronic and electrical and health and social); and overall general ability. The ASVAB has very good psychometric properties and recently the military has started using this test adaptively through new computerized format than traditional paper-based test.

Tests for Choosing Careers:

There are a number of psychological tests available for helping individuals in choosing the right career depending on their interests and aptitude. The first step of career selection is evaluation of interest. Carnegie Interest Inventory is the first interest inventory introduced in 1921 that provided measure for 15 different interests. Nowadays more than 80 interest inventories are available for interest measurement.

The Strong Vocational Interest Blank (SVIB):

The most widely used interest test is the Strong Vocational Interest Blank was developed in 1927. By the use of criterion-group approach strong developed SVIB; the subject's interest was matched with criterion group of people who were happy in their selected careers. The 399 items of SVIB are related to 54 occupations for men and 32 occupations for women presented separately. Items in the SVIB were weighted according to how frequently an interest occurred in a particular occupation group.

The test provides strong psychometric characteristics with most interesting findings that patterns of interest remain relatively stable over time. The studies have also shown that interests' patterns usually established by age 17 years. Though the SVIB has been used widely it is criticized for gender bias having used different scales for men and women and lack of theoretical information associated with SVIB.

The Kuder Occupational Interest Survey:

The second most popular interest test is The Kuder Occupational Interest Survey (KOIS).

The test taker has to select most preferred and least preferred activity among 100 triads of alternative activities. The similarity between test taker's interest and those who are employed in various occupations is assessed in KOIS.

The separate norms are available for men and women in addition with separate scales for college majors. So, KOIS helps students for choosing majors. A series of new scales is added to KOIS for nontraditional occupations. In spite of lack of research data the KOIS is useful for guidance decision for high-school and college students.

Lesson 31

Tests for Special Populations

Scales for Infants:

Brazelton Neonatal Assessment Scale (BNAS)

Population: infants

1. *Age group:* 3 days- 4 weeks
2. *Aim:* to measure a newborn's competence
3. *Total scores:* 47 (20 elicited responses and 27 behavioral items)
4. *Areas of functioning covered:* social, behavioral, and neurological
5. *Examples of factors assessed:* reflexes, motor maturity, ability to habituate to sensory stimuli, startle reactions, cuddliness, responses to stress, hand mouth coordination

The BNAS is considered as a good assessment and research tool. However, its test retest reliability is not satisfactory. Also if prediction of future intelligence is required then this scale cannot be of help.

Gesell developmental Schedules (GDS):

1. *Population:* infants and children
2. *Age group:* 2.5- 6 years
3. *Aim:* To measure infants' and children's intelligence. The purpose is to measure the subject's developmental status
4. *Final score:* Developmental quotient or DQ that can be used to calculate IQ. The formula is the same as the one for IQ. The value of MA is replaced by DQ.
5. *Areas:* gross motor, fine motor, adaptive, language, personal-social
6. *Other names:* Gesell Maturity Scale, the Gesell Norms of Development, and The Yale Tests of Child

Development: Originally developed in 1925, this tool has been used quite popularly, however it entails certain shortcomings. For example it cannot be used as a predictor of future intelligence. Also it is criticized for poor psychometric properties, inadequate sample, and weak standardization.

Bayley Scales of Infant Development: 2nd Edition (BSID-II):

1. *Population:* infants
2. *Age group:* 2- 30 months
3. *Aim:* To measure infants' cognitive and motor functions
4. *Final score:* Two main scores, mental and motor
5. *Areas:* gross motor, fine motor, adaptive, language, personal-social

The Bayley scale is valued and appreciated for adequate and good standardization. Although this scale can also not predict future intelligence, it is considered as a useful tool that can predict well for children who are mentally retarded.

Cattell Infant Intelligence Scale (CIIS):

Population: infants
Group: 2- 30 months
Aim: To measure infants' intelligence
Areas: gross motor, fine motor, adaptive, language, personal-social

The CIIS is an age scale that uses the concept of mental age and IQ and is designed after the Binet Scale. It is referred to as a downward extension of Binet's Scale.

Assessment of the Mentally Retarded:

Vineland Adaptive Behavior Scales (VABS):

These are the latest version of the Vineland Social Maturity scale developed by Edger Doll in the 1930s. Doll developed it after observing differences among the mentally retarded patients. Doll developed a standardized record form for the assessment of developmental level. The level was determined by considering the following:

- Looking after their practical needs, and

- Taking responsibility in daily living

The VABS is available in three versions. The following domains and subdomains are covered in this tool:

1. **Communication**
   - Receptive
   - Expressive
   - Written
2. **Daily Living Skills**
   - Personal
   - Domestic
   - Community
3. **Socialization**
   - Interpersonal relationships
   - Play and leisure time
   - Coping skills
4. **Motor Skills**
   - Gross
   - Fine
5. **Adaptive Behavior Composite**
6. **Maladaptive behavior**

Issues of Intelligence Testing:

No matter how we understand intelligence and what approach to its measurement do we adopt, we should keep these issues in mind:

- Is intelligence innate or acquired?
- Is intelligence a stable phenomenon?
- Can the gender, race, culture, or region of the subjects affect their scores?

Standardized intelligence testing has been called one of psychology's greatest successes. But intelligence testing has also been met with number of issues including race, gender, class and culture; of minimizing the importance of creativity, character and practical know-how; and of propagating the idea that people are born with an unchangeable intellectual potential that determines their success in life. Some issues of intelligence testing include following:

Nature versus Nurture:

The number of researches has shown the importance of heredity and environment issues in intelligence. The twin studies have shown that identical twins that have been reared apart have shown similar intelligence test scores. Children born to poverty stricken parents but adopted and brought up with better-educated and middle class families tend to have higher intelligence test scores. The natural mothers with higher IQs tend to have children with high intelligence regardless of the place and families they have brought up.

The proponents of nurture view emphasize the role of prenatal and postnatal environment, socioeconomic status, educational opportunities and parental modeling with respect to intelligence testing. However the interactionist position propagates that intelligence is result of the interaction between heredity and environment.

T**he Measurement Process:**

The process of intelligence testing itself involve number of issues like instrument used, standardization sample, test administrator and accuracy of test scoring. The test taking attitude prior coaching of examinee or test administrator's training are the factors that may affect the scores of intelligence testing.

Personality:

Researches have shown that several personality and intelligence tests overlap each other. Wechsler (1958) believed that tests of intelligence measure traits of temperament and personality such as drive, energy level, impulsiveness, persistence and goal awareness.

Longitudinal studies have shown positive various personality characteristics and intelligence measurement. Higher intelligence scores are found to be associated with high need of achievement, competitive striving, self-confidence and emotional stability. On the other hand low intelligence measurement is expected among individuals with passivity, dependence and maladjustment.

Gender:

The extensive research has been done to find out the gender differences in intelligence. However the findings have revealed that the differences found in gender are result of psychosocial and physiological factors.

Family Environment:

The family environment includes both aspects of nature and nurture. The twin studies have shown the significance importance of heredity and family environment. The issues like parental use of language, parental stress for achievement, access to resources and exposure to the world and parental influences over discipline and policies in home environment are also matter of interest for researchers.

The studies on maternal age and social class have shown that aged mothers tend to have children with higher IQ.

Lesson 32

Personality Testing

Before we discuss the various approaches to the assessment of personality, we need to understand what personality is. Personality has been defined in many ways:

- The sum total of characteristics on the basis of which people can be differentiated from each other.
- The stability in a person's behavior across different situations.
- Characteristic ways in which people behave.
- Characteristics that are relatively enduring, and that make us behave in a consistent and predictable way.

Methods of Assessment of Personality:

- Interview
- Observation and behavioral assessment
- Psychological tests

1. Interview:

Interview is direct, face to face encounter and interaction between the psychologist and the person being assessed. Verbal as well as non-verbal information is available to the psychologist. Interviews are usually used to supplement information gathered through other sources. Skill of the interviewer is very important since the worth and utility of the interview depends on how well he can draw relevant information from the interviewee.

2. Behavioral Assessment:

It refers to direct observation of behavior, for investigating, understanding, and describing personality characteristics. Skill and expertise of the observer are the most significant ingredients of the observation process.

3. Psychological Tests:

Psychological tests are standard measures devised in order to objectively assess personality and behavior. Like any other type of psychological tests personality tests also have to be valid and reliable. Availability of norms is an additional characteristic.

Psychological tests are generally of two types:

1. Objective tests/ personality inventories/ self- report measures
2. Projective tests

Objective Tests/ Personality Inventories/ Self- Report Measures:

Measures wherein the subjects are asked questions about a sample of their behavior

For example MMPI (Minnesota Multiphasic Personality Inventory) is the most frequently used personality test. It was initially developed to identify people having specific sorts of psychological difficulties. But it can predict a variety of other behaviors too. It can identify problems and tendencies like Depression, Hysteria, Paranoia, and Schizophrenia.

Projective Tests/ Techniques:

Tests in which the subject is first shown an ambiguous stimulus and then he has to describe it or tell a story about it is known as projective tests.

The most famous and frequently used projective tests are:

- Rorschach test, and
- TAT or Thematic Apperception Test

How Is The Content Of A Personality Test Decided?

What the test will contain and how it will measure personality will be affected by the theoretical orientation of the test developer. Similarly the choice of a test to assess personality also depends on how one defines personality

1. Psychodynamic Approach:

This approach focuses upon the unconscious determinants of personality i.e., psychologists belonging to this approach believe that unconscious forces determine our personality. Unconscious is the part of personality which we are not aware of. Unconscious contains instinctual drives: Infantile wishes, Desires, Demands, and Needs.

Therefore the test based on this orientation will try to unfold and explore the unconscious.

2. Trait Approaches:

These are the approaches that propose that there are certain traits that form the basis of an individual's personality. These approaches seek to identify the basic traits necessary to describe and understand personality. Traits are enduring dimensions of personality characteristics that differentiate a person from others. Trait theories do not imply the absence or presence of different traits in different people i.e., either/or situation.

These assume that some people are relatively high on some traits whereas, some are low on the same traits.

- *Trait theories based upon factor analysis:*

  Factor analysis: a statistical method whereby relationships between a large numbers of variables are summarized into fewer patterns. These patterns are more general in nature.

  For example a researcher prepares a list of traits that people may like in an ideal man.

  The extensive list is then administered to a large number of people, who are asked to choose traits that may describe an ideal man. Through the factor analysis, the responses are statistically combined and the traits associated with one another in the same set (or person) are computed. Thus the most fundamental patterns are identified. These patterns are called factors.

- *Raymond Cattell's Sixteen Personality Factors:*

  After using factor analysis Cattell proposed that two types of characteristics form our personality; Surface traits, and source traits

- *Eysenck's Dimensions of Personality:*

  According to Eysenck, personality can be understood and described in terms of just two major dimensions; Introversion-extroversion, and neuroticism-stability

  On the first dimension, people can be rated ranging from introverts to extroverts: the rest of the traits fall in between. The second dimension is independent of the first one, and ranges from being neurotic to being stable. Introverts are quiet, passive, and careful people. Extroverts are outgoing, sociable, and active people. Neurotics are moody, touchy, and anxious people. Stable are calm, care-free, and even-tempered people.

  Eysenck evaluated a number of people along these dimensions. Using the information thus obtained, he could accurately predict people's behavior in a variety of situations.

3. Social Cognitive Approach to Personality:

This approach emphasizes upon the role of people's cognitions in determining their personalities. Cognitions include: people's thoughts, feelings, expectations, and values. These approaches consider the "inner" variables to be important in determining one's personality. These approaches emphasize the reciprocity between individuals and their environment. There exists a web of reciprocity, consisting of the interaction of environment and people's behavior. Our environment affects our behavior, and our behavior in turn influences our environment and causes modifications in the environment. The modified environment in turn, affects our behavior.

Lesson 33

Objective / Structured Tests of Personality

The objective measures of personality are also known as the structured measures.
"Structured measures of personality are characterized by structure and lack of ambiguity. A clear and definite stimulus is provided, and the requirements of the subject are evident and specific" (Kaplan & Saccuzzo, 2001, p. 406).

Advantages of Structured Tests:

- They are easily administered.
- The subject can endorse own responses using paper and pencil.
- They are easy to score.
- They can be group administered.
- Their scoring is uniform for all and the scorers' likes, dislikes, or theoretical orientations do not interfere with the results.
- They are time- economical.

Some Popular Personality Inventories:

Woodworth Personal Data Sheet:

Developed in:	**Developed during the first World War. Its final form was published after the war (Woodworth, 1920).**
Developed for:	Identifying recruits who were likely to break down in combat. The recruits who reported many symptoms were called for interview. They were the ones who were most likely to be rejected.
Forms:	Single
Format:	It was like a paper-pencil psychiatric interview. The items were chosen from psychiatrists' questions asked in screening interviews and lists of known symptoms of emotional disorders.
Items:	116 questions with a 'yes' 'no' format.
Score:	A single score was obtained from the data sheet as it was designed as a global measure of functioning.
Individual/ group:	The Woodworth was used for mass screening. Many new tests followed its foot-steps.

Mooney Problem Checklist:

Developed in:	**1950**
Developed for:	Identifying problems experienced/ faced by the subject.
Items:	Checklist of problems. Problems included in the checklist are chosen out of problems reported in statements of around 4000 high- school students as well as well in clinical case history data.
Score:	The checked items indicate the problems experienced by the respondent.
Weak points:	For the sake of interpretation, one has to rely on the reported problems. There is no way to check if the reports are true or not. All that counts is face validity of responses.

Minnesota Multiphasic Personality Inventory (MMPI):
MMPI has been developed by S. R. Hatahaway and J. C. Mckinley first published in the 1940s (Hatahaway & Mckinley, 1940, 1942, 1943, 1951) and was mental to be used with people 14 years of age and above. It was first published by University of Minnesota press in 1943.

Developed in:	1940s
Developed for:	It was initially developed to identify people having specific sorts of psychological difficulties or to detect major psychiatric or psychological disorders. It can predict a variety of other behaviors too. It can identify problems and tendencies like Depression, Hysteria, Paranoia, and Schizophrenia. It can be said that the main goal is to differentiate abnormal persons from normal.
Forms:	MMPI, MMPI-2 and MMPI -A
Items:	It is a self-report measure. It contains statements that have a true/false format. The subject has to tell if the different statements are 'true' about them or 'false'. There are a total of 566 items in MMPI and 567 in MMPI- 2.
Difference between MMPI and MMPI-2:	MMPI had a total of 566 items of which 16 were repeated items. In MMPI-2 a number of items were dropped; the 16 repeated items of MMPI, 77 items from 399 to 550, and 13 items from the clinical scales. 460 items were retained from the original test. 127 new items were added in MMPI-2; two critical items, were added to identify severe pathology, 81 items added to new content scales, and 24 items added for experimental purpose. These 24 were the unscored items. MMPI- A is the version developed for adolescents. It was observed over years that the response/ score pattern of adolescents on earlier versions reflected some problems. This group tended to score higher than adults on the clinical scales. Therefore a need was felt for a scale specifically for adolescents. MMPI- A contains 478 items, 88 items less than the earlier versions. It can be used in school settings, educational counseling, and psychiatric set ups. The format and scales are the same as the other versions, and it follows a true/ false format.
Similarities between MMPI and MMPI-2:	The use and interpretation is the same for both tests.
MMPI scales:	It contains three validity scales and ten clinical scales.
Score:	The scores on the MMPI scales are used to plot a profile of the subject which shows tendencies/pathologies on which a subject is high or low.
Validity scales:	• Lie scale, 15 itemed, to detect naïve attempt to 'fake good'. • K scale, 30 itemed, to identify defensiveness. • F scale, 64 itemed, to detect attempt to 'fake bad'.
Clinical scales:	• Hypochondriasis, 33 itemed, for physical complaints • Depression, 60 itemed, for detecting depression • Hysteria , 60 itemed, for detecting immaturity • Psychopathic deviate , 50 itemed, for authority conflict • Masculinity- femininity , 60 itemed, for masculine or feminine interests • Paranoia , 40 itemed, for identifying suspicion and hostility • Psychasthenia , 48 itemed, for detecting anxiety • Schizophrenia , 78 itemed, for detecting alienation and withdrawal • Hypomania, 46 itemed, indicates high energy and elated mood level • Social introversion, 70 itemed, yields score for shyness and introversion.
Individual/ group:	It has been used in both ways.
Prerequisite:	The subject should have IQ within the normal range. Also, MMPI requires reading ability of grade 6 whereas MMPI- 2 requires that the subject has the reading ability of grade 8 level.
Strong points:	The inventory includes validity scales/ lie scale that help identify faking. It can be identified if people have been wrongly reporting pathological inclinations/symptoms or intentionally avoiding pathological content. The standardization of all versions was done on large samples.
Weak points:	It is a lengthy test and requires a long duration of time for administration.

California Psychological Inventory-Revised Edition (CPI):

CPI (Gough, 1987) is one of the popularly used personality inventories. It follows the pattern of MMPI and many of its items are the same as the ones in MMPI.

Developed for:	**Personality assessment in normally adjusted individuals.**
Scales:	CPI has 20 scales. There are four classes of scales. Each of its 18 scales is grouped into one of four classes.
What the classes of scales measure :	Class I scales: Poise, self-assurance, and interpersonal effectiveness. • High score: resourceful, active, competitive, outgoing, spontaneous, self-confident, at ease in interpersonal situations. Class II: Socialization, maturity, and responsibility. • High score: Honest, dependable, conscientious, calm, practical, cooperative. Also, alertness to social issues. Class III: Achievement potential and intellectual efficiency. • High score: Efficient, organized, capable, forceful, knowledgeable, mature, and sincere. Class IV: Interest modes. • High score: Socially well adapted, responsive to others' needs.
Items:	462 items
Individual/ group:	Can be used in both settings.
Advantages:	CPI can be used with normal subjects.

Cattell's Sixteen Personality Factor Questionnaire (16 PF):

At least five editions of 16PF are available. This test/questionnaire covers the following primary source traits:

A. Cool-warm
B. Concrete thinking –Abstract thinking
C. Affected by feelings-emotionally stable
D. submissive- Dominant
E. Sober- Enthusiastic
F. Expedient- conscientious
G. Shy- bold
H. Tough minded- tender minded
I. Trusting- Suspicious
J. Practical- Imaginative
K. Forthright- Shrewd
L. Self-assured- Apprehensive
   Q_1: Conservative- Experimenting
   Q_2: Group oriented- Self sufficient
   Q_3: undisciplined self-conflict-following self-image
   Q_4: Relaxed- Tense

Shortcomings of Structured Tests:

- These tests are of less help if in depth information about the subject's personality is required.
- Their format is fixed and cannot be molded according to the respondent's needs. For example it is difficult to find out if the subject is facing problems in understanding the wording of test items, and if so no alterations can be made in the wording or format. This is more so if the test is being group administered.
- The accurate endorsement of the responses depends on the skill and understanding of the subject as well as the skill of the psychologist/ examiner in test administration.
- Response bias: at times people may have a tendency to mark all items in the same pattern e.g., all true or all false responses.

Lesson 34

Projective Personality Tests

The projective tests are the tests in which the subject is first shown an ambiguous stimulus and then he has to describe it or tell a story about it. Two most famous and frequently used projective tests are:

I. Rorschach test, and
II. TAT or Thematic Apperception Test

Rorschach Ink Blot Test:

- The test consists of inkblot presses. These have no definite shape.
- The shapes are symmetrical, and are presented to the subject on separate cards.
- Some cards are black and white and some colored.

Procedure of Rorschach administration:

The subject is shown the stimulus card and then asked as to what the figures represent to them. The responses are recorded.

Using a complex set of clinical judgments, the subjects are classified into different personality types. The skill and the clinical judgment of the psychologist or the examiner are very important.

Thematic Apperception Test (TAT)	
Developed in:	The TAT was developed by Christiana D. Morgan and Henry Murray in 1935 during their working at Harvard Psychological Clinic.
Population:	Originally the TAT was developed for patients in psychoanalysis to obtain raw data.
Developed for:	The main objective of TAT is to evaluate a person's patterns of thought, attitudes, observational capacity, and emotional responses to ambiguous test materials.
Scoring:	The interpretive system of TAT identifies the story with individual/person described in story, needs and demands of environment created by storyteller. Among variety of scoring interoperations the scoring system is based on Murray's personality theory.
Items:	The test contains 30 black-and-white picture cards with variety of situations including human figures and situations. Some cards are suggested to use with adult males or females and some are used with children.
Forms:	In clinical practice TAT is used depending on the client's need and situation. The practitioner may use the 20 cards as prescribed number for presentation or depending on client's story-telling capacity; 1-2 or 30 cards may be used.
Strong Points:	The test has great intuitive appeal. It helps to identify emotions and motivations of the story-teller projected by unambiguous stimuli. The test can be used with ample liberty of administration and scoring for practitioners.
Weak Points:	The psychometric properties of the TAT are debated like other projective techniques. It has general lack of standardization in administration, scoring and interpretations procedures.

Other Picture-Story Tests:

Children's Apperception Test (CAT)	
Developed in:	The TAT was developed by Leopold Bellak in 1949.
Population:	The CAT was developed for children ages 3-10 years.
Developed for:	The main objective of CAT is to measure the personality traits, attitudes, and psychodynamic processes evident in children.
Scoring:	Scoring of the Children's Apperception Test is not based on objective scales; it must be performed by a trained test administrator or scorer. The scorer's interpretation should take into account: the story's primary theme; the story's hero or heroine; the needs or drives of the hero or heroine; the environment in which the story takes place; the child's perception of the figures in the picture; the main conflicts in the story; the anxieties and defenses expressed in the story; the function of the child's superego; and the integration of the child's ego.

Items:	The test contains black-and-white picture cards and use animal figures instead of humans.
Forms:	In addition with original CAT an alternative version of CAT called CAT-H is also published.
The Picture Story Test	
Developed in:	The Picture Story Test was developed by Symonds in 1949.
Population:	The test was developed to use with adolescents.
Developed for:	The main objective of Picture Story Test is to elicit stories related to specific situations like coming home late, leaving home and planning for the future.
Items:	The test contains 20 picture cards.
The Education Apperception Test and the School Apperception Method	
Developed in:	The Education Apperception Test was developed by Thompson and Sones in 1973 and the School Apperception Method was designed by Solomon and Starr in 1968.
Developed for:	These two picture instruments were designed to tap children's attitude toward school and learning.
The Michigan Picture Test	
Developed in:	The Michigan Picture Test was developed in 1953.
Developed for:	It is used to elicit various responses, ranging from conflicts with authority figures to feelings of personal inadequacy.
Population:	The test was developed for use with children between the ages of 8 and 14 years.
Items:	The test contains 16 pictures.
Make A Picture Story Method	
Developed in:	Make A Picture Story Method was developed in 1952.
Developed for:	The test helps to indicate the thinking, feelings of test-taker's projections.
Items:	The test contains 67 cut-up figures of people and animals that may be presented on any of 22 pictorial backgrounds. The number of figures with blank faces and different background settings like living room, street, nursery, stage, bridge etc. are available to use.

Tests Using Pictures As Projective Stimuli:

The Hand Test	
Developed in:	The Hand Test was developed in 1983 by Wagner.
Items:	The test contains nine cards with pictures of hands on them and tenth blank card.
Scoring:	The test taker is asked what the hands on each card might be doing. When presented with the blank card, the test taker is instructed to imagine a pair of hands on the card and then describe what they might be doing. The responses are interpreted according to 24 categories such as affection, dependence and aggression.
The Rosenzweig Picture-Frustration Study	
Developed in:	The Rosenzweig Picture-Frustration Study was originally developed in 1947 by Rosenzweig.
Developed for:	The test is based on assumption that the test taker will identify with the person being frustrated.
Forms:	The test is available in forms for children, adolescents and adults.
Scoring:	The task of test is to fill in the response of cartoon figure being frustrated. The responses are scored in terms of the type of reaction elicited and the direction of the aggression expressed.

Words as Projective Stimuli:

The first attempt of using words as projective measure was made by Galton in 1879. Afterwards Cattell and Bryant in 1889, Kraepelin in 1896 and Jung in 1910 use the words as tests. The task of these tests is to interpret the responses of words.	
Word Association Tests	
Developed in:	The test was developed by Rapaport, Gill and Schafer in 1946.
Developed for:	The Word Association Test tries to evaluate the responses with respect to variables like popularity, reaction time, content and test-retest responses.
Scoring:	The length of response time is recorded each time. The examinee is asked to clarify the relationship exist between the original word and response word.
Items:	The test consists of 60 words that are presented before the examinee in two parts firstly the examinee is asked to respond quickly with first word came in mind. In second part he is

	again presented with words and asked to reproduce original response.
The Kent-Rosanoff Free Association Test	
Developed in:	The test was developed in 1910.
Developed for:	The test attempts at standardizing the response of individuals to specific words. The purpose of the test is to identify the individuality of response that may be influence by psychopathology and many other variables.
Items:	The test consists of 100 stimulus words.

Sentence Completion Tests:

Sentence completion tests are some other tests that use verbal material as projective stimuli. The number of standardized tests is available for use.

Rotter Incomplete Sentence Blank (RISB)	
Developed in:	RISB is standardized test developed in 1950.
Population:	The test was developed for use with populations from grade 9 through adulthood.
Scoring:	The manual of the RISB suggests that responses on the test be interpreted according to several categories: family attitudes, social and sexual attitudes, general attitudes, and character traits. Each response is evaluated on 7-point scale ranges from "need for therapy" to "extremely good adjustment".
Items:	The test consists of 40 incomplete sentences.
Strong Points:	The test may be used for obtaining diverse information relating to an individual's interests, educational aspirations, future goals, fears, conflicts, needs and so forth.
Weak Points:	The sentence completion test is most vulnerable of all the projective methods to faking on the part of the examinee intent on making a good or bad impression.

Production Figure Drawings

The use of drawings in clinical and research settings has extended beyond the area of personality assessment. It attempts to use artistic productions as a source of information agbout intelligence, neurological intactness, visual-motor coordination, cognitive development and learning disabilities. The figure drawings are appealing source of diagnostic data.

Draw A Person test (DAP)	
Developed in:	DAP is developed on the working of Karen Machover (1949).
Scoring:	The drawings of person made by examinee are evaluated for various characteristics including, placement, size of the figure, pencil pressure used, symmetry, line quality shading the presence of erasures, facial expressions, posture clothing and overall appearance.
Items:	The test needs simple pencil and 8 ½ by 11 inch paper and person is asked to draw a person.
The House-Tree-Person test (HTP)	
Developed in:	The HTP was developed and popularized by Buck in 1948.
Developed for:	The drawings of house tree and person are use as reflective source of psychological functioning.
Items:	The pencil and white paper is required for test and person is instructed to draw a picture of house, a tree and a person.

Advantages of Projective Tests:

- In depth investigation
- Flexible nature
- Subject's liberty to respond in whatever way
- Psychologist has access to nonverbal cues
- Used for psychodynamic examination

Disadvantages of Projective Tests:

- The psychologist has to be highly skillful
- These tests may be very time consuming

Lesson 35

Personality: Measurement of Interests and Attitudes

The First World War brought about many changes in the way psychologists were functioning. Many new avenues of action and research were open. New tests were developed, the scope of psychology became wider, and the application of this discipline became more practically oriented. One of the areas thus explored and worked upon was the study and assessment of interests. As a result many new tests were developed in the following years.

Strong Vocational Interest Blank (SVIB):

I. E. K. Strong, Jr., and colleagues studied the activities of persons belonging to different professions. They observed that different professionals had different likes and dislikes for different activities. Their interests followed different patterns. It was observed that hobbies of people working in same profession were also similar. They tended to indulge in similar past time activities. Using the criterion- group approach, Strong developed a test to see if the interests of a person/ test taker matched with the interests and values of people who were happy in their chosen careers. The criterion group included persons belonging to various professions. The test is called Strong vocational interest blank or SVIB. The 399 items of SVIB are related to 54 occupations for men and 32 occupations for women presented separately. Items in the SVIB were weighted according to the frequency of occurrence of an interest in a particular occupational group as compared to how frequently it occurred in the general population..

The test provides strong psychometric characteristics. The normative samples used for the development of this test were well sized. Around 300 people were used in each criterion group. The raw scores obtained on SVIB were converted in standard scores, with a mean of 50 and SD of 10. Research has shown that interests' patterns usually established by age 17 years. It was also observed that patterns of interest remain relatively stable over time. In a study of Stanford University students taking this test first in the 1930s and then later on also, it was found that the interests remained relatively unchanged even after 22 years. Though the SVIB has been used widely it is criticized for gender bias having used different scales for men and women and lack of theoretical information associated with SVIB. It was often criticized for not having a theoretical basis to explain why people belonging to different professions were likely to have similar interests.

Holland (1975) had presented his theory of vocational choice. According to that theory people's personality is expressed in their interests. Furthermore, considering people's interests we can classify them in into one or more of the following six categories of personality factors:

a. Realistic
b. Investigative
c. Artistic
d. Social
e. Enterprising
f. Conventional

These factors were gender- bias free and could be used with both men as well as women. There was a similarity between Holland's factors and the patterns of interests yielded by research with SVIB. This appealed Campbell who incorporated this theory in his version of SVIB.

The Strong- Campbell interest Inventory (SCII): The SVIB had certain features for which it was criticized. A new version was developed by D. P. Campbell and was named as The Strong- Campbell interest Inventory or SCII.	
Developed in:	1974
Population:	Both men and women. It is a gender free version.
Special feature:	Unlike the SVIB, the SCII does not have separate forms for men and women. The gender bias in the SVIB was removed and the items from the male and female forms were merged into one form. The scales in this form are free of any gender bias.
Developed for:	The measurement of interest.
Items:	There are 325 items. The response options include 'like', 'dislike', or 'indifferent'. The present version of SCII contains seven parts including: Occupations: 131 items

	1. School subjects: 36 items 2. Activities: 51 items 3. Amusements: 39 items 4. Types of people: 24 items 5. Preference between two activities: 30 items 6. Your characteristics: 14 items
Forms:	Now it has one form for everyone, male or female.
Scoring:	Several scores are obtained for each profile. 1. General themes based on Holland's (1999) six personality types 2. Scoring for the administrative indexes 3. A Person's basic interests. 4. Occupational scales
Strong feature:	The test follows Holland's theory of vocational choice and therefore has a theoretical basis. Also it is not gender biased.

The Kuder Occupational Interest Survey (KOIS):

As mentioned earlier, in lecture 30, another most popular interest test is The Kuder Occupational Interest Survey (KOIS).

The test taker has to select most preferred and least preferred activity among 100 triads of alternative activities. A triad is a set of three alternates. The similarity between test taker's interest and those who are employed in various occupations are assessed in KOIS.

This instrument has separate norms for men and women. Separate scales for college majors are also available. The KOIS helps in two major ways. Firstly, it suggests as to which occupational group may be best suited to a person's interests, and secondly it can assist students in choosing their majors. A series of new scales is added to KOIS for nontraditional occupations. In spite of lack of research data the KOIS is useful for guidance decision for high-school and college students.

The Jackson Vocational Interest Survey (JVIS):	
Developed in:	The JVIS was developed by D. N. Jackson. The version revised in 1995 is in use commonly.
Population:	The test is used for the career education and counseling of high-school and college students.
Developed for:	It can be used for career planning of adults and for those seeking mid-life career changes.
Items:	It contains 289 statements about job-related activities. The JVIS can be completed in around 45 minutes. The items follow a forced choice format and the respondent has to select one interest that he/she prefers over the other.
Forms and scoring:	Available in both hand-scored and machine-scored forms.
Strong Points:	The test has strong psychometric properties. It carefully avoided gender bias.

The Minnesota Vocational Interest Inventory (MVII):	
Population:	The MVII is designed for men who are not oriented toward college. Skilled and semiskilled trades are emphasized.
Developed for:	The MVII has been used extensively by the military and by guidance programs for individuals not going to college.
Items:	The MVII has nine basic interest areas, and 21 specific occupational scales. The basic interest areas include areas like mechanical interests, electronics, and food service. The occupational scales cover occupations like plumber, carpenter and truck driver.

The Caree Assessment Inventory (CAI):	
Developed in:	The test was developed by Charles B. Johansson in 1976.
Population:	The test is developed for American citizens with less than four years of postsecondary education. This segment of the population comprises around 80 % of the U.S citizens (Kaplan & Saccuzzo, 2001). The test requires reading ability of grade-6 reading level.
Developed	Measuring interests.

for:	
Items:	Test taker is evaluated on Holland's six occupational theme scales. Basic interests in 22 areas are also assessed. 89 occupation scales are used.
Strong Points:	The test has good validity and reliability. The test developer also tried to make the CAI culturally fair and eliminate gender bias.

The Self -Directed Search (SDS):	
Developed in:	J.L Holland developed the Self-Directed Search.
Developed for:	The test attempts to simulate the counseling process by allowing respondents to list occupational aspirations, indicate occupational preference in six areas, and rate abilities and skills in these areas.
Items:	The test has 228 items. There are six scales that have 11 items each that describe activities. Competencies are assessed by 66 items. Another six scales with 14 items each evaluate occupations. The respondents list their occupational aspirations, and indicate occupational preferences in six areas. The also rate abilities and skills in these areas.
Scoring:	The SDS is self-administered, self-scored and self-interpreted vocational interest inventory. The test takers score the inventory and calculate six summary scores. These scores and further codes reflect areas of highest interest.

Eliminating Gender Bias in Interest Measurement:

The advocates of women's rights justifiably pointed out discrimination against women in early interest inventories. The Associate for Evaluation in Guidance appointed the Commission on Sex Bias in Measurement, which concluded that interest inventories contributed to the policy of guiding young men and women into gender-typed careers.

The SVIB has separate form for women but careers for women tended to be lower in status and to command lower salaries. Because career choices for many women are complex, interest inventories alone may be inadequate and more comprehensive approaches are needed.

Sentence completion tests are some other tests that use verbal material as projective stimuli. The number of standardized tests is available for use.

Rotter Incomplete Sentence Blank (RISB)	
Developed in:	RISB is standardized test developed in 1950.
Population:	The test was developed for use with populations from grade 9 through adulthood.
Scoring:	The manual of the RISB suggests that responses on the test be interpreted according to several categories: family attitudes, social and sexual attitudes, general attitudes, and character traits. Each response is evaluated on 7-point scale ranges from "need for therapy" to "extremely good adjustment".
Items:	The test consists of 40 incomplete sentences.
Strong Points:	The test may be used for obtaining diverse information relating to an individual's interests, educational aspirations, future goals, fears, conflicts, needs and so forth.
Weak Points:	The sentence completion test is most vulnerable of all the projective methods to faking on the part of the examinee intent on making a good or bad impression.

Production Figure Drawings:

The use of drawings in clinical and research settings has extended beyond the area of personality assessment. It attempts to use artistic productions as a source of information about intelligence, neurological intactness, visual-motor coordination, cognitive development and learning disabilities. The figure drawings are appealing source of diagnostics data.

Draw A Person test (DAP)	
Developed in:	DAP is developed on the working of Karen Machover (1949).

Scoring:	The drawings of person made by examinee are evaluated for various characteristics including, placement, size of the figure, pencil pressure used, symmetry, line quality shading the presence of erasures, facial expressions, posture clothing and overall appearance.
Items:	The test needs simple pencil and 8 ½ by 11 inch paper and person is asked to draw a person.
The House-Tree-Person test (HTP)	
Developed in:	The HTP was developed and popularized by Buck in 1948.
Developed for:	The drawings of house tree and person are use as reflective source of psychological functioning.
Items:	The pencil and white paper is required for test and person is instructed to draw a picture of house, a tree and a person.

Advantages of Projective Tests:

- In depth investigation
- Flexible nature
- Subjects liberty to respond in whatever way
- Used for psychodynamic examination

Disadvantages of Projective Tests:

- The psychologist has to be highly skillful
- These tests may be very time consuming

Measurement of Attitudes, Opinions, Locus of Control, Multidimensional Health Self-efficacy

Measurement of Attitudes:
Besides measuring different personality characteristics, personality dynamics, and interests etc. psychologists assess and investigate many other aspects of personality too. Attitudes and opinions are two such aspects. An attitude can be defined as "a tendency to react favorably or unfavorably toward a designated class of stimuli, such as a national or ethnic group, a custom, or an institution" (Anastasi & Urbina (2007, p. 418-419). Another related aspect is 'opinions'. Attitudes and opinions are closely related since one's opinions are determined and directed by one's attitudes. The measures commonly used for the measurement of attitudes are called attitude scales. There are different types of scales available that we can use. However in most attitude or opinion surveys one needs to develop attitude scales or other instruments according to the theme of the research. One can develop scales following the formats available. Some of the scales whose format is commonly followed are as follows:

Thurstone (1931) Scale/ Equal Appearing Intervals:
In this scale the scale development is most important. A large number of attitude- related statements are developed. The statements can be positive and negative toward the object of attitude. A panel of judges rates each statement from one to eleven. One means highly negative on the subject and eleven indicates highly positive. The ratings of all judges are processed and mean rating for each statement is gauged. When respondents are given scores according to the mean values obtained from the judges. They score the scale value of each item agreed with.

Likert(1932) Scale :
In this type of scale a number of statements are developed regarding the object of attitude. The statements are both favorable and unfavorable. They are rated according to the following scale:

5	*4*	*3*	*2*	**1**
Strongly agree	Agree	Undecided	Disagree	Strongly disagree

The value of a chosen response is the score of the person on that item. The total or overall score is calculated from the score on individual items.
Some other popular scales include the Guttman (1950) scale, the Semantic Differential Scale (Osgood et al., 1957), and the Social Distance Scale (Bogardus, 1925).

Locus of Control:
Locus of control refers to a person's perceptions or beliefs about the location of responsibility for his or her life; circumstances, happenings, events, conditions. The perception of who is in-charge of one's life, who decides one's fate, and who is responsible for whatever the person is experiencing, is determined by the person's locus of control (LOC). People's perceptions of success and failure, of health and illness, ability or inability, all reflect their locus of control. In other words, the concept of LOC refers to perceived control; the perception of how much a person feels in control of life
(Lefcourt, 1982). The earliest formal investigations of the concept were reported by Julian Rotter (1966, 1975). Rotter showed that people have different views about things that happen to them. People have their own beliefs or generalized expectations about where the control of life, or events, resides.
Rotter's original formulation of LOC had a dualistic approach. People's beliefs regarding who is responsible for events, and who influences life, were classified along a bipolar dimension. Rotter showed that people have different views about the source of control, and the things happening to them; these beliefs, and therefore people holding them, were seen as falling into two categories namely, internal and external, that could be measured with the I-E scale. Later on Levenson and others added to this construct as well as its measure. However, perhaps the most quoted contribution in this regard is in the assessment of *Multidimensional Health Locus of Control.*

Measurement of Multidimensional Health Locus of Control (MHLC):

Health locus of control can be measured by using specifically designed instruments. Its quantification gives an edge to this construct over many other constructs pertaining to health beliefs. The Multidimensional Health Locus of Control (MHLC) scales were developed by B. S. Wallston and K. A.Wallston, and their colleagues (Wallston, Wallston, Kaplan, & Maides, 1976; Wallston, Wallston, & De Vellis ,1978). These are the most popularly used measures for quantifying HLC.

Krischt and his colleagues (Dabbs & Krischt, 1971; Krischt, 1972) were the ones who produced the first published version of an LOC measure that was specifically meant for use in the domain of health and illness. Due to some inherent flaws in this measure however, it could not gain popularity and a need was felt for more precise measures.

B. S.Wallston, and K. A. Wallston have so far made highly significant contributions to the measurement of HLC.

Multidimensional Health Locus of Control (MHLC) Scales:

The MHLC follows the 6-point Likert response pattern, and includes three scales. In conceiving these scales, Levenson's (1973, 1981) multidimensional approach was followed, in which the dimension of externality was split into two components. Hence the scales for powerful others health locus of control (PHLC), and chance health locus of control (CHLC).Instead of treating EHLC as a single dimension that covers the influence of powerful others and that of chance as one and the same dimension, the new versions treated PHLC and CHLC as two separate components. PHLC, that measures the powerful others health locus of control, pertains to a person's beliefs about the influence of other people on her health. The people who are believed to have the power to determine one's health may include the family, friends, doctors, hospital staff and others.

The extent to which a person believes in the influence of chance, luck, or fate is measured by the CHLC Scale. CHLC measures the beliefs about health or illness being related to chance, fate or luck, instead of being related to one's own responsibility.

The three MHLC scales include:

1. Internal Health Locus of Control (LHLC)
2. Powerful Others Health Locus of Control (PHLC)
3. Chance Health Locus of Control (CHLC)

A six-point rating scale is used with the MHLC items, where one can make a choice from response options ranging from 'strongly disagree' to 'strongly agree'. The scale contains three subscales and eighteen statements in all. Each subscale carries six items/ statements. The items pertaining to the three subscales have been mixed up. The scoring procedure is very simple. The values of the marked responses to each item in a subscale are added up. The sum indicates the person's score on that subscale. The subscale on which the person obtains the highest score indicates the type of health locus of control that the person has. One may choose any one from forms A or B. The IHLC Scale assesses the internal health locus of control that is the extent to which a person believes that her health or illness is determined by internal factors. K. A.Wallston, and B. S. Wallston (1982) have asserted that the dimensions measured by the scales are more or less statistically independent. Therefore a low IHLC score does not necessarily indicate that the person believes in the influence of external factors, powerful others, or chance. A low IHLC score may be understood to mean that the person's belief in the influence of internal factors is low (or may be non-existent in some cases).

The Measurement of Self Efficacy:

Self-efficacy, as the very name suggests, is the perception of one's own ability to produce some desired outcomes.

Researchers have used the construct of self-efficacy for assessing the impact of people's perceptions of personal control and capability on their behavior in a variety of situations. A divergent range of self-efficacy measures is available to researchers interested in investigating the relationship between thought and action. Although a considerable majority of studies in this regard have investigated the influence of perceived self-efficacy on people's health-related behaviors, measures like collective teacher self-efficacy, and teacher self-efficacy scales have also been devised.

For a health psychology researcher, primarily two types of measures of self-efficacy are available. These instruments can be used to assess health related self-efficacy in two ways, including:

1. General perceived self-efficacy

2. Perceived self-efficacy pertaining to specific health behaviors.

The measures included in the above mentioned categories may be used in their original form as well as with alterations made according to the nature of problem under investigation.

1. General Perceived Self-efficacy (GSE) scale"

The most widely used measure of self-efficacy, General Self-Efficacy (GSE) Scale, was developed by Matthias Jerusalem, and Ralph Schwarzer in 1981. The original version, in German language, comprised 20 items, but the later version consisted of only 10 items (Schwarzer & Scholz, 2000; Schwarzer & Jerusalem, 1995). It is this 10 item version that is used by researchers studying self-efficacy in recent researches.

The GSE scale is available in at least 26 different languages. The scale was originally devised for predicting both coping and adaptation; how people coped with daily hassles, and how they adapted after having undergone stressful life events. This scale gauges a generalized sense of self-efficacy, indicating the overall global confidence that a person has about personal ability to cope with a wide range of situations that may be new, novel, taxing or demanding. GSE focuses upon a sense of competence that is broad and stable, rather than being only domain-specific (Schwarzer & Scholz, 2000).

The response format of GSE scale is uniform for all 10 items, consisting of a 4-point scale. The response options range from 'not at all' (definitely not) marked as 1, to 'exactly true' marked as 4. The final composite score is obtained by adding up the responses to all 10 items. The final score may range from 10 to 40. If the person marks 'not at all' in response to the entire range of items, he will be understood to be standing at the lowest possible level of self-efficacy. This can be taken to indicate a lack of self-efficacy. On the contrary a score of 40 will mean the person has the highest possible level of feeling self-efficacious. On average, the scale can be completed in 4 minutes. Some people may take longer, or lesser than the average time.

Alternate Approaches to Personality Assessment: Behavioral and Cognitive- Behavioral Testing

Behavioral Assessment:

At times mere psychological testing is not sufficient for making a good judgment about a person's personality. In fact when using intelligence, achievement, or aptitude tests as well one may not be sure whether to rely only on test results or not. There are occasions when one seeks other supportive evidence to support the information yielded by the test.

This is where the significance of behavioral assessment cannot be ignored.

Behavioral assessment provides us with first hand, direct, and tangible evidence about a person. What personality tests measure are traits, characteristics, or qualities and assumed to be underlying one's behavior. Behavioral assessment on the other hand yields information regarding a sample of behavior that is assumed to represent a person's behavior in various situations. It is always better and safer to supplement test results with behavioral information when some sort of decision making, diagnosis, or screening is involved. Behavioral assessment can be used as an independent approach as well. However, perhaps the best strategy would be to use a battery of procedures by employing psychometric tools along with behaviorism assessment.

Following is a brief description of approach/ procedures used for behavioral assessment.

Behavioral Observation:

The very first and the most basic tool/ method used for behavioral assessment is observation. Although all psychologists may use this procedure it is more commonly used by developmental, educational, child, and school psychologists. These psychologists observe the behavior of interest and keep a record of it.

The psychologists may employ various technical devices for recording the behavior of interest. Audio or video recordings add to the accuracy of observation by capturing the moments and those significant aspects of the subject's behavior that the observer may miss while taking down observational notes. The psychologist may employ trained staff for observational record keeping.

An off shoot of such observation is self- observation whereby a subject herself reports her own behavior.

A similar approach is 'self-monitoring' in which the subject keeps a record of her own behavior as it happens e.g. cigarettes smoked by a smoker during the day, binge eating episodes had by a bulimic, or seconds of spot running done by an overweight man in a day .

Recording Observed Behavior:

For recording the behavior in question a number of approaches may be used:

a. Narrative records; the observer may take notes while observing and record each and every thing that could be of interest. The advantage is that detailed notes are taken, but there are always chances that a lot might be missed while the observer is writing down the notes. A better approach is to take very short notes and fill out the gaps when the observation session is over.
b. As mentioned earlier, audio/ video recordings can be made.
c. Behavioral rating scales; one can record observed behavior in terms of codes rather than recording narratives. This procedure not only saves recording time but also makes possible inter rater uniformity when multiple observers are gathering information. Rating scales are designed in such a manner that one can record presence/absence, frequency, intensity and other aspects of behavior. Observers may develop their own scales or recording instruments, but may also choose from the available scales. Some of the available scales include The Play Performance Scale for Children (Lansky et al., 1985, 1987), Walker Problem Behavior Identification Checklist (Walker, 1983), Behavior Rating Profile(Brown & Hammil, 1978), and Social Skills Rating System (Gresham & Elliot, 1990) just to name a few.

Situational Performance Measure:

It is a form of observation in which a person's behavior is observed under specific circumstances. These circumstances can be real or simulated. For example a candidate for lecturer's position may be asked to deliver a model lecture in front of a real class; an applicant of heavy duty vehicle may be asked to drive a truck on a real

life busy road; a would be astronaut has to perform under simulated situation in a state of weightlessness. The subject's behavior under the chosen situation may be one or more observers.
One form of this type of observation is the use of situational stress tests when a person's behavior is observed under certain type, level or amount of stress, anxiety, or frustration. Such an approach may be used for jobs where the prospective candidate is expected to experience psychological pressure e.g. armed forces, bureaucracy. The U.S Office of strategic Services (OSS, 1948) has been reported to have used situational stress tests during World War- II for the selection of candidates for military intelligence and other positions.

Cognitive- Behavioral Assessment:
Verbal similar to the behavioral approach is the cognitive- behavioral approach to assessment. The only difference is in terms of the cognitive component.

Why Cognitive Behavioral Testing Is Needed?
The cognitive- behavioral testing is based on the cognitive- behavioral approach which is a relatively modern approach as compared to other approaches to testing.
In cognitive- behavioral testing, just like cognitive- behavioral therapy, an individual's own cognition, behavior, and related physiological responses are focused. The problem behavior itself is targeted for its understanding and treatment. Therefore the core of interest is the symptom behavior rather than the unconscious determinants or underlying causes that need to be explored, reached, and deciphered.
According to Kaplan & Saccuzzo, 2001, in comparison to traditionally used tests, the cognitive- behavioral tests target the 'disordered behavior' rather than the underlying cause. This approach is based on the behavior model and herein the symptom (reported as problem) is the focus of treatment. The analysis of disordered behavior is the goal of cognitive- behavioral assessment.
There are four steps involved in cognitive- behavioral assessment (Kaplan & Saccuzzo, 2001 :

1. Identification of critical behavior
2. Determining if the critical behaviors in question are in excess or deficits.
3. Evaluation of the frequency, duration, or intensity of the behavior being considered.
4. Based on step- 3, the frequency, duration, or intensity of the critical behavior is decreased or increased. If they were in excess then attempts to decrease would be made and if they were in deficit then an increase will be aimed for.

In order to give you a flavor of what cognitive- behavioral assessment is like, some of assessment methods/ procedures used for this purpose are being described here.

The Fear Survey Schedule (FSS):
It is a self-report procedure used for various clinical purposes. Primarily ratings on fear are taken on a rating scale. Initially introduced by Akutagawa (1956), the FSS is available in different versions after having undergone a number of revisions. Originally it had 50 items. Today the different versions have 50 to 122 items, employing either 5- point or 7- point scales.
The FSS items involve fear provoking situations and avoidance behaviors. The aim is to identify such situations and avoidance behavior in case of the subject being assessed. The items have been derived from clinical observations of actual cases (Wolpe & Lang, 1964) and laboratory experimental studies (Geer, 1965)

Irrational Beliefs Test (IBT):
People often hold irrational beliefs. Such beliefs do not have a logical or realistic basis but people simply cannot separate the belief from their cognitive system. A number of cognitive- behavioral tests are available for testing of irrational beliefs that people hold. One such test is the Irrational Beliefs Test or IBT developed by R. a. Jones (1968). The test contains 100 items. The test follows a 5- point scale format. The subject has to indicate level of agreement or disagreement with each item. Half of the items pertain to presence and the other half to the absence of particular irrational beliefs.

Kanfer and Saslow's Functional Approach:
Kanfer and Saslow (1969) were the ones who played one of the lead roles in the initiation of the cognitive-behavioral approach to assessment. In their functional approach, that is a behavior- analytic approach, excesses and deficits in peoples' behavior are focused. Rather than using traditional labeling of people into

psychopathological categories such as schizophrenic, psychotic, or neurotic their behavior is analyzed in view of excesses and deficits.

According to Kaplan & Saccuzzo (2001), "a behavioral excess is any behavior or class of behaviors described as problematic by an individual because of its inappropriateness or because of excesses in its frequency, intensity, or duration" (p. 480). On the other hand "behavioral deficits are classes of behavior described as problematic because they fail to occur with sufficient frequency, with adequate intensity, in appropriate form, or under socially expected conditions" (p.481).

The functional approach proposes that same laws operate in the development of normal and disordered behaviors. The difference occurs only in extremes. Therefore in the analysis the psychologist first identifies the excesses and deficits and then tries to help the client decrease or increase them accordingly.

Testing and Assessment in Health Psychology

Health psychology is one of the most popular areas of psychology. The main reason for its growing popularity is perhaps the fact that it has practical relevance to most people's life. With the growing body of research literature and evidence in health related issues, the number of available research and assessment tools is also increasing. This section discusses some of the tests and scales that a psychologist may use while assessing the subjects/ clients or when carrying out health psychological research.

The State- Trait Anxiety Inventory (STAI):
The STAI is based on the State- Trait Anxiety theory of Charles. D. Spielberger. The theory, and the inventory, assumes that anxiety is of two types; state anxiety and trait anxiety. State anxiety is an emotional reaction that may vary from situation to situation, whereas trait anxiety is a personality characteristic that may be found to be stable across situations. There are two scales and therefore two scores yielded by STAI, the A-State and A-Trait. The inventory has a 4- point scale format and the two scales have 20 items each.
The STAI has been, and may be, used with patients suffering from various health conditions and undergoing surgical or other stressful procedures for an assessment of anxiety .

The Ways of Coping Scale:
As the very name suggests, the Ways of Coping Scale(Lazarus, 1995; Folkman & Lazarus, 1980) assesses the way people cope with stress. It is one of the most po[pularly used tools in health psychology. It is a checklist in which the subject indicates the items/ thoughts and behaviors that apply to them. It contains 68 items and the following seven subscales:

a. Problem solving
b. Growth
c. Wishful thinking
d. Advice seeking
e. Minimizing threat
f. Seeking support
g. Self-blame

The subscales, research suggests, can be divided into two broad categories:

- Problem- focused strategies i.e., cognitive and behavior strategies for coping with stress. These strategies are attempts made to solve the problem.
- Emotion- focused strategies i.e., ways of dealing with the emotional response to stress. Such strategies do not help in resolving the problem.

Coping Inventory:
The Coping Inventory (Horowitz & Wilner, 1980) contains items that have been derived from clinical interview data. Its 33 items fall into three categories:

a. Activities and attitudes people adopt for avoiding stress.
b. Strategies for working through stressful events.
c. Socialization responses

The Social Support Questionnaire (SSQ):
The SSQ developed by I. g. Sarason and co- workers (1983) is an instrument that measures social support and related aspects. It contains 27 items of which each one has two parts. For every item the respondent has to endorse two things which ultimately culminate into two scores:

i. Listing the persons that the respondent can count on for support in given circumstances, these responses yield the number (N) score. The number of people listed in all 26 items is used to calculate an average (N) score.
ii. Indicating overall level of satisfaction with these supports. This leads to the satisfaction (S) score. This score may range from one to 6 for each item, one being very dissatisfied and 6 means very satisfied. The satisfaction score from all items is used to get an average (S) score.

Scales for Specific Health conditions Related Locus of Control

a. Drinking Locus of Control Scale:
A 25 item-scale measuring drinking locus of control is available. It follows a forced choice format that has been developed by Donovan, and O'Leary (1978). The items involve pairing of internal and external control alternatives.

b. Weight Locus of Control (WLOC) Scale:
This scale assesses the internal and external determinants of one's weight. The scale designed by Saltzer (1982), uses 6-point Likert scale format and is a 4-item measure.

c. Perceived Behavioral Control Measure:
Armitage, and Connor (1999) developed a measure to assess perceived behavioral control. The measure includes items like "Whether or not I eat a low fat diet is entirely up to me".

d. Desired Control Scale:
This 70-item rating scale was developed by Reid, and Zeigler (1981). It uses a 5-point response scale. The ratings range from 'strongly agree' to 'strongly disagree'.
The scale comprises two subscales, with 35 items each. The subscales include:

i. Desire of outcomes
ii. Beliefs and attitudes

e. Health Engagement Control Strategies / HECS
This scale, uses a 5-point rating scale, comprising 9 items and has been reported by Worsch, Schulz, and Heckhausen (2002). It contains items like "I invest as much time and energy as possible to improve my health". The rating options range from "almost never true" to "almost always true".

Assessment of Perceived Self-Efficacy Pertaining to Specific Health Behaviors:
Whereas the GSE scale has been found to be a good predictor of a general sense of personal competence across various situations, numerous studies have used the construct of self-efficacy for assessing its impact on specific health behaviors as well. Such studies have investigated the potential influence of self-efficacy on the initiation of health practices. Such practices include indulging in healthy lifestyles, avoiding or quitting unhealthy behaviors, and coping with specific health conditions and / or illnesses.

The main approach of assessment in this regard remains the same as that adopted in the measurement of GSE. However unlike the GSE measure, the specific health behavior measures focus upon the health condition in question alone, rather than on a global ability to handle a wide range of stressful situations in general. The researchers can simply replace the original items with items pertaining to specific health conditions, or devise similar measures on their own.
Many studies have used such very brief scales comprising only 4-5 items. In some cases even single item measures have been used. What needs to be kept in mind while devising and using such measures is the rule that the item or items should bear appropriate wording. The words used in the item / items should be theory-based and should convey exactly what the researcher wants to find out. The wording must very clearly include the mention of both the health action and the perceived barrier or condition for action. In this regard an 'if-then' sentence formation has been suggested.
The semantic structure recommended for health related research is as follows: 'I am certain that I can do XX, even if YY (barrier)' (Luszczynska, & Schwarzer, 2005). Scanning the available research literature, one can find the mention of at least around a dozen different measures of specific health condition- related perceived self-efficacy. Some of these measures are the altered forms of the scales developed by Schwarzer and colleagues, whereas the others have been designed and developed by independent researchers. Following is a brief description of specific health-condition related measures of self-efficacy that have been devised and used by researchers. These measures are available for those interested in exploring the relationship between self-efficacy and the health conditions and / or practices being investigated.

a. Measures for Assessing Exercise-Related Self-Efficacy:

Developed by Schwarzer, and Renner (2000) the exercise self-efficacy scale primarily focuses on the extent to which a person feels capable of overcoming barriers to adopting or maintaining the habit of exercising. The scale has a 4-point format, ranging from 'definitely not' to 'exactly true'.

A similar measure, self-efficacy for regular exercise, has been reported by Lorig et al (1996). The Exercise Regularly Scale assesses people's confidence in regularly doing certain physical activities such as gentle exercise, or aerobic exercise including walking, swimming, or bicycling. The subjects can choose from the ten response options for each item, starting from 'not at all confident' to 'totally confident'.

b. The Nutrition Self-Efficacy Scales:

Researchers have also developed and used self-efficacy measures specifically involving proper nutrition related behaviors. The Nutrition Self-efficacy Scale by Anderson, Winett, and Wojcik (2000) offers a choice from ten response options for each item, ranging from "very sure I cannot' to 'very sure I can'. The items assess the level of confidence of a person in terms of how certain he is that he can indulge in behavior involving the use of nutritious foods, such as taking a slice of bread containing fiber to school or work.

Schwarzer, and Renner (2000) have also reported the use of a Nutrition Self-efficacy Scale. The said scale once again involves the assessment of how certain a person is that she can overcome barriers to healthy eating as well as the extent to which the person can manage to stick to healthful foods despite personal or situational / social impediments. The response options, four in all per item, range from 'definitely not' to 'exactly true'.

c. Habit Cessation, And Abstinence Self-Efficacy:

Besides measures for gauging adoption of healthy behaviors, tools for self-efficacy related to assessing the confidence in the ability to refrain from unhealthy behavior, are also available. One of the earliest reports in this regard has been made by Annis (1987). The Situational Confidence Questionnaire is meant for examining alcohol abstinence self-efficacy. This instrument, with its 6-point scale format, provides response options in terms of percentages ranging from 0% to 100%. The mid-range response options include 20, 40, 60 and 80 percent. A response of 0% indicates 'not at all confident', while 100% means 'very confident' in resisting the urge to drink heavily even when circumstances were favorable for drinking a lot.

Schwarzer, and Renner (2000) report a similar scale that aims to assess the level of certainty with which a person feels in control of his own drinking behavior. The four response options range from 'definitely not' (1) to 'exactly true' (4).

Dijkstra, and De Vries (2000) have reported on a measure of self-efficacy that can be used with those trying to quit smoking. The Smoking Cessation Self-efficacy Scale is a 7- point scale in which the response options range from -3 to +3; from 'not at all sure I am able to', to 'very sure I am able to'. The scale assesses as to how much a person feels she can refrain from smoking in different situations.

d. Health-Protective Behaviors and Adherence To Medical Advice Self-Efficacy:

Luszczynska, and Schwarzer (2003) have reported on the use of two scales for measuring self-efficacy pertaining to breast self-examination (BSE). The first scale, Preaction BSE Self-Efficacy Scale can help examine the extent to which a woman feels able to perform regular BSE in spite of possible odds, besides a tendency to procrastinate and reschedule the plan. The scale offers five response options ranging from 'definitely not' to 'exactly true'. The other scale, Maintenance BSE Self-Efficacy Scale, also contains the same response options. This scale can be used to gauge the self-efficacy felt in maintaining the regular habit of BSE.

A scale for measuring adherence self-efficacy has been used by Mohr, Boudewyn, Likosky, Levine, and Goodkin (2001). The Adherence Self-efficacy Scale assesses self-efficacy related to self-injection. The response options go from 'I will not have any problems' (1), to 'I will not be able to tolerate it at all' (6).

Lesson 39

Measuring Personal Characteristics for Job Placement

Imagine if you had to advise a friend in job selection. Suppose that friend has two job options available with similar salary packages and located at same distance from home
What factors do we generally consider while taking such a decision?
Personal interest and aptitude? The skills and ability that the person has? The work place and the work setting? The prospective boss and colleagues? Or maybe all of these factors?
You are very well familiar with the role and significance of personal interests in career choice. We have discussed in detail the various tests and tools that can be used for the assessment of personal interests. But tests of interests are not the only measures that help in identifying whether a person is suitable for the job or not. In other words we have available a variety of other tests also for choosing the best person for the job.
Also, a number of assessment tools have been developed to assist you if you were to find out if you had the skills required for a job, or if the job or the work place was what you were made for.
Psychologists have developed a variety of measures that can gauge the suitability of individuals for a particular job by taking into account their personal characteristics as well as the features of the work setting. Different psychologists have proposed different theories in this regard. Based on these theories a number of assessment tools have been developed.

Osipow's Vocational Dimensions Approach: The Trait Factor Approach:
One psychologist who is best known for the use of trait factor approach for job decision making is Samuel Osipow. One can see that he has a global approach i.e., that considers a number of traits or let us say aspects of a person's personality. In this approach a number of tests, a battery of tests, are used for assessment.
The battery includes a variety of tests such as; the Kuder Occupational Interest Survey (Kuder, 1979), Strong-Campbell interest Inventory (Campbell, 1974), Seashore Measure of Musical Talents (Lezak, 1983), and Purdue Pegboard (Fleishman& Quaintance, 1984).
This approach gives quite comprehensive information regarding the traits and interests of the person. However it is criticized for not taking much into account the work environment Nevertheless this approach is found to be very useful in helping people make occupational decisions.

Roe's Career- choice Theory: The California Occupational Preference Survey:
The core feature of Roe's theory is its emphasis on 'person' or 'nonperson' orientation found in people. According to Roe, this orientation plays a significant role in people's career choice. In simpler terms maybe we can say that whether one likes to be with pother people or not affects one's career choice. The person/people - oriented people would be looking for jobs where they are in contact with other people e.g. Arts, entertainment, or other services.
The individuals who are not people- oriented would be seeking jobs that involve little interpersonal contact e.g. lab work, science and technology, field exploration etc. Roe drew some very interesting conclusions from extensive examination of the personalities of scientists. These scientists were working in different areas of study.
Roe proposes that career choices that people make in life are a result of their childhood experiences of relationship with their families. That is to say that people with different types of experiences of relationships with their family as a child will make different career choices.
According to Roe, whether people, as children, were reared in a warm family environment or a cold and aloof one determines if they are interested in other people or not. Children brought up in a family environment that is warm and accepting grow into people- oriented adults. On the other hand children who experienced a cold and aloof environment turn into adults who are interested in things rather than people (Roe & Klos,1969; Roe & Siegelman, 1964). Roe and Klos (1969) proposed the idea that occupational roles can be divided into two classes according to two independent continua.

The First Continuum: The extremes go from "orientation to purposeful communication" to "orientation to resource utilization"
The Second Continuum: The extremes go from "orientation to interpersonal relations" to "orientation to natural phenomena"
People make career choices according to where they stand on these two continua.

Lesson 40

Achievement and Educational Tests

Psychologists use tests in the educational set ups for a variety of reasons.
Tests may be used for an assessment of achievement, for diagnostic purpose, for onward transmission of test results to some other agency, for selection for a program, for entrance into an institution, or for screening before being chosen for specific skill acquisition training.
Tests may be used for evaluation of achievement or what students have learnt in a program of study. Tests may also be used for diagnostic purpose. This usually happens when the school teachers suspect some psychological/ behavioral problem, learning difficulty, some deficiency, or a similar problem. In such a case the child is referred to the school counselor or some other professional outside school.
At times the parents themselves might approach the school counselor for help. In such cases the child may be assessed for the identification of the problem.
The schools also use tests for selecting students with a certain level of intellectual ability, specific aptitudes, or skills. This is when the student is to be selected for a program of study or training that requires specific aptitude, orientation, or interest. Many institutions have their own admission tests that are used for selecting candidates for admission to their institution. On the other hand some agencies or institutions develop admission tests that are used by most institutions both nationally and internationally for admission purposes e.g. GRE, SAT, MAT etc.
However the tests most commonly used in academic settings are the achievement tests.

Achievement Tests:
Achievement tests are meant to assess if students or trainees have learnt whatever they were supposed to learn at the end of a course or program of instruction.
These tests measure the students' achievement alongside the effectiveness of a program.

What, How, and When of Achievement Tests:
The assessment of achievement involves three basic decisions.

What Is To Be Assessed?
This decision pertains to:
a. The course content to be covered by the assessment tool.
b. The instructional objectives that specify the expected and desired outcomes of the teaching-learning process.

How to Assess?
This decision pertains to:
a. The type of the assessment tool.
b. The administration procedure.
c. The number, format, and difficulty level of test items.

When to Assess?
This decision involves answers to these questions:
a. At what time during the academic session will the assessment take place?
b. Once in a term, or more than once?
This decision will affect the content area to be covered in assessment.
Of all the above mentioned issues and decisions, the most significant is to cover in the test the content area that the students have been taught keeping in mind the objectives specified for every component of the content.

Teacher Made Achievement Tests:
Teacher made achievement tests are the most common type of achievement tests.
Teachers, all over the world, and in all educational institutions are busy throughout the year either teaching or assessing their students. Teachers have a choice to design and develop their tests the way they like them to be.
A teacher made test can be either objective or subjective. On occasions it may be a combination of both. Objective and subjective type of items have their own advantages and disadvantages.

Objective or Forced Choice Type of Items:

- These are difficult to develop but easy to score.
- These allow the teacher to cover a wide range of content area.
- There is always a chance of selecting the right answer simply by guessing while actually not knowing the right answer. But this can be controlled. If the items are MCQs with 4-5 options per item, and the options are carefully developed then guessing can be controlled to a large extent.
- Another advantage of objective type items is uniformity of scoring across examiners. No matter who does the scoring the students will be receiving the same score.

The essential requirement for availing these benefits of objective tests is care in writing test items. The stem of every item should be clearly stated, should not be ambiguous, and should convey what the examiner wants to convey.

Even more important than this is the selection of appropriate response options. A good MCQ item is the one in which every option appears to be the right answer. Therefore only the ones who know the course content are able to select the right choice.

Subjective or Descriptive Tests:

On the other hand subjective or descriptive tests also have their advantages.

- The nature of the items is such that the examiner can test in depth knowledge of the students. However marking and evaluation of such examiner papers may be problematic.
- The inter examiner uniformity of scoring is doubtful in such tests.
- Examiners' personal or ideological biases may interfere with the objectivity in evaluation required of a just teacher.

It is ultimately for the examiners to decide as to what format they prefer to use and what would suit best to the course content.

Other Varieties of Achievement Tests:

Other than teacher made achievement tests, we have available a variety of standardized achievement tests that are used at national and international level.

We have discussed a number of such tests in earlier sections. Let us very briefly have a look at three of these tests which are most commonly used.

Standardized Achievement Tests:

The Scholastic Assessment Test (SAT-I):

The Scholastic Assessment Test or SAT-I, previously known as Scholastic Aptitude Test or SAT was first used in 1926, the test is the most commonly used college entrance test in the U.S.

SAT-I has two parts that contain the Verbal Reasoning and Mathematical Reasoning tests. These comprise further subtests.

SAT-II is also available.

Graduate Record Examination Aptitude Test (GRE):

GRE is one of the most well-known tests across the globe. It is the most commonly used graduate-school entrance test.

GRE measures general scholastic ability and contains three sections:

- Verbal (GRE-V),
- Quantitative (GRE-Q) and
- Analytic (GRE-A).

Miller Analogies Test (MAT):

MAT is the second major, widely used, scholastic aptitude test.

It is a verbal test that measures student's ability to find logical relationships for 100 different analogy problems.

Achievement versus Aptitude Tests:

Going through this brief description of these tests, you must have noticed that these tests are discussed as achievement tests whereas they have the term 'aptitude' attached to their name. If you have taken the indigenous GAT you must have realized that many of the items were more about your aptitude and ability rather than achievement in the conventional sense.

That is why one question commonly arises to the mind of most students of psychological testing i.e., what is the difference between achievement tests and aptitude tests?

In most situations these terms are used interchangeably.

If one analyzes logically, one can understand that it is not possible to cover in one test all of the content that students from different institutions, regions, and countries have studied.

Kaplan and Saccuzzo (2001, p. 343) have given a very good comparative description of the features of the two types of tests.

Grading, Percent Score, And Related Interpretive Issues:

School/ college tests usually use the grading system. Scores are also given in terms of percent. Grades make it easier to understand the relative position of students.

In large scale tests, like GRE or GAT, the results are communicated in terms of percentile ranks. These describe a candidate's position in relation with those scoring above as well as those scoring below him or her.

Lesson 41

Multicultural Testing

In the present and the following sections we will be discussing some specific issues that psychologists might come across when working in different situations and with different types of subjects.

Multicultural Testing:
At times psychologists are working with subjects or clients who come from a variety of cultural backgrounds and when their cultural background may interfere with their test performance. In such situations a need is felt for tests that can be used with all people and that are neither biased against or in favor of any specific cultural origin.
Multicultural testing refers to tests and testing procedures that are not affected by the cultural background of the test taker. Such situations may arise in testing scenarios where immigrants belonging to different cultural backgrounds settle in the developed countries and are to be tested on same variables using same tests. For example:

- When measurement of IQ or personality is to be done.
- When screening, short listing, or selection for jobs is to be done.
- When diagnosis of mal adjustment or mental illness is to be done.

Also, there are situations where the same tests are meant to be used with people based in different parts of the world and having different cultural origins and experiences.
Even people belonging to subcultures within a large society may experience cultural disadvantage.

The main idea is that the nature of many standardized tests is such that certain segments of the population may be at a disadvantage because of their origin or in other words, cultural disadvantage. In such situations there is a need to have tests that are free of cultural bias. Such tests may be called multicultural tests. The content, as well as the administration, scoring procedures, or scores are not affected by the cultural origin of the test taker. Multicultural testing is also known as transcultural testing or cross- cultural testing.

Factors That May Cause Cultural Bias:
The issue of cultural differences arises when people from one culture have to live in a culture very different from their own culture. There are a few factors that may put one culture at a disadvantage or advantage in comparison to another. These disadvantages become significant when people have to take psychological tests developed in cultures other than their own.
Anastasi and Urbina (2007) describe these as parameters along which cultural differences may be found. Such variables include:

a. Language: People are at a disadvantage if they can use only the language spoken in their own culture and not the one used in the culture wherein they have to adjust. As a consequence they will be handicapped if psychological tests administered to them are in the language that they are not familiar with.

b. Reading Ability: People will still be at a disadvantage if they cannot read. Most tests require certain level of reading ability. People may be familiar with the language that the test has been designed in, but they will remain handicapped if they cannot read the test items.

c. Speed: The speed required for completing a test may also cause problems. In some cultures life is very fast and people are familiar with a sense of urgency to meet deadlines. On the other hand the tempo of life is slower in some cultures and people are used to patiently waiting expected outcomes (e.g. rural and agricultural societies) rather than striving for immediate and rapid outputs.
Therefore persons coming from such cultures may find it difficult to cope with the demands of speed based tests.

d. Familiarity With The Format, Style, And Contents Of Tests: At times people may not be familiar with certain forms of test items and formats of tests. Consequently they find it hard to attempt certain types of items, may take longer than allocated time, and may also make mistakes because of not being able to understand what they were supposed to do.

For example people may find it difficult to attempt MCQ type questions if they have not seen such items previously. Even problems/items involving figures for assessing spatial reasoning may be a totally new experience for test takers who have never seen or made geometric drawings

Multicultural testing includes tests that are free of the bias that may arise out of cultural disadvantage stemming from variables such as language, reading ability, or speed.
In order to tackle the above mentioned issues, sources of bias, and disadvantages, certain measures are taken. Multicultural tests generally do not involve reading or writing, verbal ability, or test taking speed.
As far as familiarity with the format, style, and contents of tests is concerned it is an issue that is controlled by using performance and drawing based items. However the issue of familiarity with geometric drawings is concerned, it is controlled by avoiding the use of designs and patterns that most people cannot relate to.

Some Multicultural Tests

The Leiter International Performance Scale- Revised (LIPS-Revised): The LIPS-Revised (Roid & Miller, 1997) is an individually administered test whose first version was published undergone many revisions ever since. It measures intellectual ability.	
Developed in:	1940
Population:	Can be used with all age groups. Its 1997 version was standardized on a sample of 2000, aged 2 to 20 years, both atypical and normal subjects from the U.S.
Special features:	• This scale does not involve verbal instructions as such. • It is individually administered and follows a difficulty level sequence i.e., the easiest item is administered first. • There is no time limit. • Easels are used to present the graphic stimulus materials. The picture cards that the subject considers to be the appropriate response are placed in the provided response tray.
Measures:	The scale covers four domains: a) Reasoning b) Visualization c) Attention d) Memory
Tasks in domains:	The scale involves various tasks meant for various age levels Reasoning and Visualization: matching, form completion, design analogies, sequential ordering, paper folding, figure rotation, and classification. Attention and Memory: Sustained and divided attention measures; immediate and delayed memory tasks.

Raven Progressive Matrices:
One of the most popularly used nonverbal and culture free tests of general intelligence is the Raven Progressive Matrices (RPM).
As you are already familiar, it uses a multiple choice format. In each test item, the subject is asked to identify the missing element that completes a pattern. The test can be administered to groups or individuals of 5 years old to older adults. There are 60 matrices with a missing part presented in graded difficulty. The subject selects appropriate pattern from a group of eight options.

Goodenough-Harris Drawing Test:
The Goodenough-Harris Drawing Test is the quickest, easiest and less expensive nonverbal test for measuring intelligence. The subject is asked to draw a whole human figure. The test is scored for each item included in drawing. The subject gets credit for inclusion of elements such as individual body parts, proportion, perspective, clothing details etc.

The G-HDT scoring follows the age differentiation principle; older children tend to get more points because of greater accuracy. It is not a test of the subject's drawing or artistic skill. What is considered important is the development of conceptual thinking and accuracy of observation.

In the revised scale the test is not limited to the drawing of a man alone. The subject is asked to draw picture of a woman and of one's own self. The self-scale is used as a projective test of personality (Anastasi & Urbina, 2007).

The test has good psychometric properties. As previously mentioned, the scores on the G-HDT can be related to Wechsler IQ scores. The test can be more appropriately used in combination with other tests of intelligence.

IPAT Culture Fair Intelligence Test:

R. B. Cattell directed the development of this test. The IPAT Culture Fair Intelligence Test is a paper pencil test for three levels;

- Age levels 4-8 years and mentally disabled adults,
- Age levels 8-12 and randomly selected adults, and
- High-school age and above-average adults.

Lesson 42

Adaptive Testing and Other Issues: Computer Based Administration

Have you ever thought that different test takers have different attitudes toward test that they are taking?
They have different levels of motivation in taking the tests. Their abilities and aptitudes are different, and so is their test taking approach.
Subjects' response characteristics are also different and they affect their score or performance on a test .
Psychologists involved in test development have been working on the possibility of tests as well as test administration procedures that take such differences in consideration
In this section, we will discuss some of such issues.

Adaptive Testing:
As said earlier, psychologists have been working on the possibility of tailor- making the tests according to the individual response characteristics of the test takers.
The idea is that people should not be at a disadvantage because of their specific response characteristics. What happens usually is that people start taking the test and then they go on attempting all the items with increasing difficulty level no matter if the easier items were attempted correctly or not. As a consequence many people might score worse than what they could have achieved had the test been used according to individual response characteristics of the test takers.
Adaptive testing refers to the testing procedure whereby test item coverage is adjusted according to the response characteristics of individual subjects. There are different procedures available for this purpose. One such procedure described by Anastasi & Urbina (2007) involves **Two- Stage Adaptive Testing** with three measurement levels. The authors give the example of a hypothetical test that comprises 70 items in all. Ten items are placed in a routing test while the remaining 60 items are divided in three measurement tests of 20 items each.
These three measurement tests are of varying difficulty levels; easy, intermediate, and difficult. All subjects will attempt the ten items in the routing test but not all of the other 60 items. They will be taking any one of the three measurement tests depending on their performance on the routing test. Therefore everyone will be given 30 items in all, but the last 20 items that everyone will attempt will vary from person to person. So one can expect that if one could do the difficult items in the routing test then one will get the 'difficult' measurement test, a difference one could only do the easy items then one will be getting the 'easy' measurement test.
The authors (Anastasi & Urbina, 2007) have described an alternate to this two- stage model. This second model is the **Pyramidal Testing Model** since it progresses in the form of a pyramid. In this model every one begins with an item of intermediate difficulty level.
If one manages to answer the item correctly then one is given the next item of a higher level. If, on the contrary, one fails in the first item then one is routed downward to an item of lower difficulty level. This procedure is repeated until the test taker manages to answer the desired number of items. It can be seen in both of these models that the test takers are treated according to their response pattern. These procedures can be used as it is as well as in their varied forms. Although these procedures for adaptive testing can be done manually using simple paper and pencil, they are quite tedious. Computerized adaptive testing is a convenient option that provides facility to the psychologists.All that is required is the availability of the suitable software and the skill on part of the psychologist.

Computer Based Administration:
Like all other fields of life computers have been playing a very significant role in psychometrics. Availability of computers has facilitated the psychologists in a number of ways. They have made things possible and easier; whether it is computerized administration, scoring, item analysis, analysis of data obtained from large standardization samples, or adaptive testing. There are situations when group testing involves participants in large numbers or tests that involve tedious procedures.
In such situations computers can be, and are used for administration and scoring for example in case of use of multilevel batteries, different forms of educational testing, aptitude testing on its own or for career guidance, and achievement testing.
Today all major achievement and ability tests are administered and/or scored with the help of computers e.g. GRE, SAT, MAT, GAT, IELTS etc. Computers have made it possible to devise new ways and instruments of testing and assessment

Alternatives to Psychological Tests!!!

Can we use other ways of assessment rather than using psychological tests???

Interviews as Assessment Tools:

Psychological tests are just one form of instruments that a psychologist may use foe making assessment of people's personality, IQ, ability, aptitude or any other variable of interest. Interviews are another such instrument. Interviews provide an opportunity to have a direct, face to face, interaction with the person being examined. Interviews can be used as an alternative to psychological tests.

Kaplan & Saccuzzo, 2001, have highlighted **similarities between a test and an interview.**

According to them the two have these common features:

- Method for gathering data
- Used to make predictions
- Evaluated in terms of reliability
- Evaluated in terms of validity
- Group or individual
- Structured or un structured

Types of Interviews That Are Used For Assessment:

a. Evaluation interview: This interview helps the psychologists assess and understand why the student/ client/ individual has come to them.

b. Structured Clinical Interview: Structured interviews follow a fixed and set pattern of questions and procedures. This pattern may be decided by the clinic/hospital/institution or may be recommended by some other agency e.g. the use of DSM according to a sequence of steps.

c. Case history interview: This interview may be more detailed as compared to other types as it aims at in depth information. It usually takes a developmental approach. These are relatively flexible though focused.

d. Mental Status Examination: This interview is more fixed and focused and used more commonly in psychiatric settings. Usually it is used when some psychiatric, neurological, or emotional problem is suspected

d. Employment interview: These interviews are used by the employers for the selection of right people for the available jobs. Such interview may be both structured and/or unstructured, depending upon the nature of the organization, the employer, and the position for which interview is being made.

Interviewing Skills Required In Psychologists:

- Practice and training
- Command over language and vocabulary
- Overcoming personal complexes
- Empathy
- Flexibility and acceptance of the other person's opinion
- Control over own emotional reactions
- Cultural sensitivity
- Note taking skills and technological assistance

Lesson 43

Social and Ethical Considerations in Testing

The use of psychological tests may seem to be a simple and straightforward thing but it it may involve a variety of social and psychological issues pertaining to the use of these tests. Psychology being a very well organized discipline takes care of these issues. Psychologists have developed rules and regulations for carrying out research and all sorts of investigation, including psychological testing and assessment.

Psychologists are expected to follow a strict code of ethics in the endeavors they take up.
The most commonly followed ethical standards are the ones developed and published by the APA or the American Psychological Association
With the increasing application and popularity of Psychology there has been a growing concern about the way psychologists operate. This becomes more significant and relevant when psychological research and assessment are under discussion.

Although psychological research and assessment may be treated as two different areas, one can see that they overlap a lot. All psychological researches involve some form of instruments of data collection. These instruments are in fact most of the times some form of psychological tests. Therefore the ethical standards set for psychological research also apply to psychological testing. Before going into the details of ethics involved in psychological testing let us have a look at some agencies or sets of ethical standards available for psychologists involved in psychological testing:

APA Ethics Code:
This document by the APA covers most aspects of psychological testing ranging from confidentiality, development and use of psychological assessment techniques to legal and forensic contexts.

Principles for the Validation and Use of Personnel Selection Procedures:
This document containing guidelines for the validation and use of assessment procedures employed for personnel selection was developed in 1987 by Society for Industrial and Organizational Psychology (SIOP).

The RUST Statement/ Responsibilities of Users of Standardized tests:
The American Counseling Association (ACA) adopted this statement in 1989.

"Ability Testing: Uses, Consequences, and Controversies":
This book, by Wigdor and Garner (1982), covers all aspects of ability testing.
This publication, a two volume book, is actually the report of a project that investigated the use of ability tests in various settings. This four year project looked into the use of such tests in a variety of settings ranging from schools to job selection.

Board on Testing and Assessment / BoTA:
This board, established in 1993, primarily works on the use of psychological tests and other tools of assessment as tools of public policy. This board was created in the U.S. under support by the departments of Defense, Education, and Labor (Anastasi & Urbina (2007).

Ethical Issues in Psychological Testing

Ethical Issues and ethical Standards in Psychological Testing:
As mentioned earlier, a number of agencies have attempted to propose and set ethical standards for test use. These standards pertain to various aspects of testing, from test development to application of tests in a variety of situations for attaining a wide range of objectives.
In the forthcoming sections we will be discussing the general ethics that test developers and test users need to keep in mind while doing psychological assessment. Most of the times we refer to, and follow, the APA standards. On occasions some aspects of these standards or guidelines may be found in test manuals as well, particularly with reference to the person using or administering the test.

1. The Training and Eligibility of the Test User:

- The person using or administering the test should be properly trained and experienced.
- We know that the personality of the psychologist or the person responsible for test administration should have appropriate qualification and training.
- The standards of qualification may vary from place to place or institution to institution.
- However, generally the academic institutions or professional/psychological associations specify the minimum qualification for test use and administration.
- As far as training is concerned, the psychologist or the administrator should have completed sufficient number of hours of supervised training before carrying out independent test administration.
- Once again, the number of hours of training may vary from place to place and institution to institution.
- Training and qualification are specifically required in case of intelligence tests, particularly individually administered tests, and personality tests.
- This issue becomes even more significant when the interpretation of projective tests is in question.
- In case of achievement tests, especially objective tests, a compromise may be made on the qualification of the person responsible for test administration.
- In this regard we should consider what the Ethics Code states. According to the code the psychologists should "provide only those services and use only those techniques for which they are qualified by education, training, or experience" (APA, 1992, p. 1599).

2. Human Rights and Test Use:

- All individuals have a right to decide if they want to be tested or not.
- People should not be tested if they refuse to be tested.
- Legal or forensic scenarios may be an exception where testing is directed by a court of law.

3. Invasion of Privacy:

Confidentiality is an essential ingredient of counseling and clinical encounters
Unless allowed by the subject/participant/ test taker, the psychologist should not disclose and make public the following:

- That the person was tested
- The scores of the person
- The interpretation of the test results.
- The diagnosis, if any.

As said earlier, legal situations are an exception. Also, in case of achievement tests the test results are generally understood to be made public.

4. Confidentiality, Honesty, and Openness:

- The psychologist should explain the nature and purpose of the test to the subject.
- The subject also needs to be informed about the possible use of test results.
- The test results should not be used for any purpose other than those mentioned to the subject.
- As far as explaining the nature of tests before hand is concerned, it may be problematic in case of some tests.
- For example in case of projective tests like Rorschach, TAT, or WAT the test performance may be affected by information about the true nature of the test. In such situations the required information may be provided immediately after the test is over.

5. Care with Labeling:

- Unless essential, the psychologist should try to avoid labeling.
- People may be diagnosed with a certain problem, or their scores may be indicative of certain tendencies. However labeling should not be done unless that was the main objective of testing.
- Certain labels have a social stigma attached to them e.g. schizophrenic, or PWA (patient with AIDS). Such labels may be damaging the interests of the test taker in many ways; socially, psychologically, and even financially if the person is refused employment.

- Therefore the psychologist should try to avoid labeling if possible.

6. The Issue of Divided Loyalties:

At times the psychologists do assessment because it was requested and paid by an organization. In such situations the services of the psychologists are hired by the organization, but being their profession demands care and protection of the test taker as well. If there is a clash between the interests of the organization and those of the test taker then the psychologists have to take rational decisions.

In such situations the following steps may be taken (Kaplan & Saccuzzo, 2001):

- The psychologist must inform the clients beforehand about the purpose for which the test results will be used.
- The clients should also be informed about the limits of confidentiality.
- The results should be explained to the client or his representative.
- The organization may be provided only that much of information that was required.
- In case of adverse decisions the person's right to know the results should be preferred over test security.

7. Issues Pertaining To the Test Developers:

- The test developers should be fair and objective.
- They should use test content that is not gender biased or culturally biased.
- If the test is to be used with diverse populations then it should be culture free.

8. Issues Pertaining To the Test User In Diverse Populations:

- The test users or administrators should be fair to the test takers.
- Tests developed and standardized in other cultures should not be used blindly with subjects belonging to very different cultures.
- The tests should either be culture free, or translated, adapted, and standardized for indigenous culture.

Lesson 44

Assessment and Psychological Testing in Clinical & Counseling Settings

Psychological testing is a part of psychological assessment. Assessment is more than testing. It involves behavioral observation, interviews, and examination of case history. In the counseling and clinical settings psychological tests may be used as independent tools or as part of a complete assessment package.

Tests used in Clinical and Counseling Psychology:
In these settings psychological tests are used for diagnosis, induction in treatment groups or hospitals, for general assessment, and for gauging the rate of recovery. All intelligence and personality tests may be used. For example HTP can depict psychopathology.
Some tests are used for diagnosing specific learning disabilities e.g. Kaufman Test of Educational Achievement (K-TEA).

The counseling or clinical psychologists commonly use tests for the following purposes:

- General assessment of ability/ IQ
- General assessment of personality
- Diagnosis of intellectual deficits
- Diagnosis of mental disorders
- Assessment of aptitude
- Neuropsychological assessment
- Assessment of learning disabilities

You are familiar with many of the tests used for the above mentioned purposes. In addition to these tests behavioral assessment also proves to be an important tool.

Neuropsychological Testing:
In this section we will discuss areas that need special attention.
Neuropsychological testing is an area of psychological assessment that is quite complicated, particularly when it is with reference to the diagnosis of brain damage. The assessment may involve an extensive battery of tests. The person may be tested in a number of areas; cognitive ability, verbal ability, spatial relations etc.
A number of tests are required for making an exact diagnosis of the problem.

Some of the tests used for neuropsychological testing include the following:

Bender- Gestalt test and Benton Visual Retention Test: These two tests are quite commonly used for neuropsychological testing. However a single test may not prove to be an accurate instrument. Therefore batteries of tests are preferred for this purpose.

Halstead- Reitan Neuropsychological Test Battery (HRB- Reitan &Wolfson, 1993) and the **Luria-Nebraska Neuropsychological Battery:** Batteries like these are preferred over single tests because rather than providing information in one particular area they can provide information in a variety of areas.

According to Anastasi & Urbina (2007) these batteries are useful tools because:

- They provide measures of all significant neuropsychological skills.
- These and similar standardized batteries can detect brain damage with a high degree of success.
- Such a battery can help identify and localize the impaired brain areas.
- Differentiation between particular syndromes associated with cerebral pathology can be made.

Behavioral Assessment:

- The behavioral assessment procedures include the following:
- Self-report by the client
- Direct observation of behavior
- Physiological measures

Self-report By The Client:

- Self-reports can be made in various forms such as inventories and checklists.
- Clinical interviews can also be one of the procedures used for this purpose.
- One of the most commonly used such tools is the Beck Depression Inventory (BDI).
- The BDI involves self-ratings on 21 items that help assess the severity of depression.
- Alcohol Use Inventory (Horn, Wanberg, & Foster, 1990) is another such instrument.
- Some instruments involve multiple informants.

The Social Skills Rating System (SSRS):

Gresham &Elliott, 1990
Positive and problematic behaviors of students in educational and family settings can be evaluated.
There are separate forms for parents, teachers, and students themselves.

Behavior Assessment System for Children / BASC:

- This is one of the most comprehensive instruments of its kind.
- It includes behavior rating scales for teachers as well as parents.
- The children can be given a self-report questionnaire.
- This system also contains a form that can be used for coding and recording classroom behavior.
- In order to take developmental history from parents an additional structured interview is also available.

Direct Observation Of Behavior:

It may be recorded by the psychologist, parents, teachers or any other designated person
The observation takes place in naturalistic setting. The observation may be recorded in the form of narratives, checklists, rating scales, record forms or similar tools.

Physiological Measures:

Depending on the nature of problem a number of physiological measures may be used.
Such measures are employed in case of cases of anxiety, sleep disorders or similar cases.
These measures may include measures of cardiovascular activity, cerebral functioning, electrodermal, and electro-ocular activity. Behavioral assessment and clinical judgment

Evaluation of Various Assessment Techniques:

All techniques have their advantages and disadvantages. The psychologists may choose any method that best serves their purpose and has minimum limitations.

Lesson 45

Overview of the Course

This is the last lecture of this series. In this session we will go through, very briefly, whatever we have discussed so far. "Psychology is the scientific study of behavior and mental processes". Psychologists use carefully designed tools of data collection. Psychological tests are one of those tools. In all of the professional settings, where psychologists work, some form of testing and assessment is used.

"A test is a measurement device or technique used to quantify behavior or aid in the understanding and prediction of behavior" (Kaplan, & Saccuzzo, 2001).

Tests are a part of the assessment package and procedure

Types of Tests:

Tests can be categorized on the basis of:

- The purpose or the type of behavior/characteristics to be measured: personality, aptitude, intelligence, achievement etc.
- The administration procedure: individual versus group tests
- Speed versus ability tests
- Aptitude tests, achievement tests, or intelligence tests
- Ability versus personality tests
- Structured/objective tests versus projective tests
- Original versus translated and adapted tests
- Translated tests

Major Contexts of Current Test Use:

Psychological tests are designed for, and are used in numerous life situations

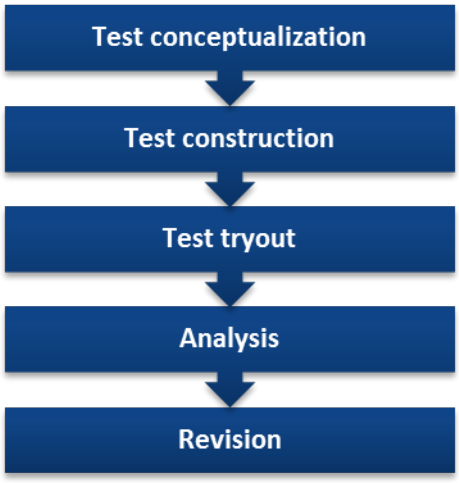
The three major context of test use:

- Educational testing
- Occupational testing
- Clinical and Counseling Psychology (Anastasi, and Utnins, 1997).

Test Construction:

Development of a good test takes place after going through a number of stages, keeping in view the established principles of test construction.

Test Development Process:



Essential Characteristics of Psychological Tests:

- A good psychological test should have these qualities:
- Validity: A test should measure what it is intended to measure.

- Reliability: A test should give consistent results. It should give same or similar results every time it is administered to the same subjects in same conditions.
- Norm development and standardization

Reliability:

According to the definition given by Anastasi & Urbina (2007) reliability refers to "the consistency of the scores obtained by the same persons when they are reexamined with the same test on different occasions, or with different sets of equivalent items, or under other variable examining conditions".

Reliability of a measure can be measured in a number of ways; the stability of scales over time, is it consistency between items, or something else.

a) Test- Retest Reliability:

- Test- retest reliability deals with two performances of the same test by the same persons on two different occasions.
- If reliability refers to the consistency and stability of scores over time then it will be measured using this method.
- The test- retest coefficient is also known as the 'coefficient of stability'.
- The test takers' scores on first administration of the test are correlated with their scores obtained on the second administration of the same test.

b). Alternate- Form Reliability

In this approach the test developer develops two alternate or parallel forms of the same test. Ideally speaking the two forms should be independently developed and completely *parallel* or *equivalent forms* of the same measure. They should match each other in all respects including:

- Same specifications
- Same instructions
- Same time limit
- Same content
- Same number of items
- Same item format
- Difficulty level

c). Split- Half Reliability

Alternate form reliability is a popular type of reliability but with some obvious limitation; Construction of two completely parallel forms is not easy and requires a lot of time and effort; even when to alternate forms are available, the practice effect and prior experience may affect performance on the second occasion.

The time gap between two administrations is another intervening variable.

To overcome these problems, in fact to remove these problems, split- half reliability is computed

e). Coefficient Alpha

- Another approach quite similar to the Kuder- Richardson technique is to calculate 'coefficient alpha'. Kuder- Richardson formula can be used only for tests in which items are scored as either zero or one.
- It is not applicable to tests where answers to items are assigned two or more scoring weights e.g. personality inventories where a number of response options with attached weights are available (never=0, occasionally=1, often=2 and so on).
- Coefficient alpha is a general formula that caters for such tests.

Validity

"Traditionally, the validity of a test has been defined as the extent to which a test measures what it was designed to measure" (Aiken, 1994, p.95)

"The extent to which a test measures the quality it purports to measure. Types of validity evidence include content validity, criterion validity, and construct validity evidence" (Kaplan & Saccuzzo, 2001, p.640). From these definitions one can realize that validity of a test is about the nature of a test with reference to the content

of the test as well as the content or domain in which it is rooted. It is about what the test measures and how well it measures. Validity is measured in several ways. These may be divided into three categories:

Content Validity Evidence:
"The evidence that the content of a test represents the conceptual domain it is designed to cover" (Kaplan & Saccuzzo, 2001, p.635).

Construct Validity Evidence:
"A process used to establish the meaning of a test through a series of studies. To evaluate evidence for construct validity, a researcher simultaneously defines some construct and develops the instrumentation to measure it. In the studies, observed correlations between the test and other measures provide evidence for the meaning of the test" (Kaplan & Saccuzzo, 2001, p.635).

Criterion Validity Evidence:
"The evidence that a test score corresponds to an accurate measure of interest. The measure of interest is called the criterion" (Kaplan & Saccuzzo, 2001, p.635).

Specific Procedures for Content Validity:
A number of careful decisions are taken at the time of test development when test specifications are made. *Test specifications* include:

- Instructional objectives, relative importance of topics/processes.
- It should clearly indicate the number of items for each topic.
- It may also provide sample material.

The process of content validation should include description of all procedures in the manual; that ensure that test is appropriate and representative.

Criterion-Related Validity:

- Whenever we plan and design a test we have a certain standard in mind that we want to meet.
- We relate score on our test with those achieved on the criterion or standard.
- One approach to assessment of validity is through comparing the test results with those on a criterion. A procedure where scores on a test being used are correlated with scores on a criterion.

Criterion Validity Evidence:
"The evidence that a test score corresponds to an accurate measure of interest. The measure of interest is called the criterion" (Kaplan & Saccuzzo, 2001, p.635).
The criteria used for this purpose may be in the form of scores on psychological tests, mental and behavioral measurement, classifications, grades, teacher's or supervisor's ratings etc.

Criterion Prediction Procedures:

- Two procedures may be adopted for this purpose:
- Predictive validity evidence
- Concurrent validity evidence

Predictive Validity Evidence:
"The evidence that a test forecasts score on the criterion at some future time" (Kaplan & Saccuzzo, 2001, p.638)

Concurrent Validity Evidence:
"evidence for criterion validity in which the test and the criterion are administered at the same point in time" (Kaplan & Saccuzzo, 2001, p.635).
When the validity of a test is measured with reference to a criterion, and both are administered at nearly the same time, then the resulting validity will be concurrent validity.
Anastasi & Urbina (2007) have provided an all-encompassing description of construct validity: "The construct validity of a test is the extent to which the test may be said to measure a theoretical construct or trait.

Assumptions Underlying Construct Validity:

Test scores will be highly correlated with tests measuring same construct/similar construct. This is convergent validity. Test scores will have weak/low (or at times may be negative) correlation with tests meant to measure construct which are different from the one measured by the main test. This is discriminant validity.

Norms:

The scores on any test become meaningful in the presence of norms that have been developed for that test.

Definition: "The test performance data of a particular group of test takers that are designed for use as a reference for evaluating or interpreting individual test scores".

In a more simple way it can be said that norms provide standards to which the results of the test takers on different measurements can be compared. Norms are the test performance of the standardization sample. Standardization sample is a group of people whose performance on a specific test is taken as a standard or norm for comparison.

All the other individuals' performance on this specific test is compared with the scores of this standardization sample.

A person's performance on a test is interpreted with reference to the distribution of scores in the sample used as a representative of the population.

Item Analysis:

Item analysis is "a set of methods used to evaluate test items. The most common techniques involve assessment of item difficulty and item discriminability" (Kaplan & Saccuzzo, 2001, p. 637)

But it may involve other things as well. For example item response valence, a content analysis etc.

Item Difficulty:

"A form of item analysis used to assess how difficult items are. The most common index of difficulty is the percentage of test takers who respond with the correct choice"

Item Discrimination:

A test is supposed to discriminate between those who know and those who do not know; those who score high and those who score low

"The item- discrimination index is a measure of the difference between the proportion of high scorers answering an item correctly and the proportion of low scorers answering the item correctly; the higher the value of d, the greater the number of high scorers answering the item correctly" (Cohen & Swerdlik, 1999).

Types of Tests:

We have discussed a variety of psychological tests in this course:

- Intelligence tests: individual and group tests, verbal and performance tests, speed and ability tests
- Personality tests: Personality inventories and projective tests
- Achievement tests: Ordinary school tests and standardized,
- Internationally used tests
- Tests used for occupational settings
- Tests for special populations
- Tests used in clinical and counseling settings
- Aptitude tests
- Tests of interests
- Variations of psychological tests: the Piagetian tasks.

Future of Psychological Testing:

- The application of psychological tests is becoming broader every day
- As the areas of application of psychology are also increasing
- BUT one should always keep the ethical standards in mind; these standards may have been set by the APA, the government, or any other professional and/or regulatory agency.