

Lesson 1

Difference Between Testing And Assessment:

What is a test?

“A test is a measurement device used to in the understanding and prediction of behavior ”.

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.

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:

- Confidentiality
- No hurt or damage
- Respect the viewpoint of subject

Lesson 2

Researchers and historians agree that China was perhaps the first country to develop and use tests in the formal sense of mental measurement. We should also know that the Chinese had developed a system of mental measurement even in 2200 years BC.

Chinese had developed a civil service testing program. Under the Chinese emperors, oral examinations were arranged every third year in order to determine work evaluations and promotion decisions.

Han Dynasty (206 B.C.E. to 220 C.E). These early tests pertained a variety of topics e.g. revenue, civil law, military affairs, agriculture, and geography.

Join VU Group: <https://chat.whatsapp.com/HuGWSMfRr7AFCZZ33kmJyg>

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 bow shooting, 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.

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.

In 1883, the American Civil Service Commission was established. This commission developed and administered competitive examinations for certain government job

Charles Darwin wrote On the Origins of Species in 1859. 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)

British: Francis Galton

France: Alfred binet

America: James McKean Catell

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.

Join VU Group: <https://chat.whatsapp.com/HuGWSMfRr7AFCZZ33kmJyg>

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 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”.

". 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.

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.

Cattell and Galton held the common understanding that intellectual function could be measured by measuring sensory discrimination and reaction time

He also developed some sensorimotor tests. He established his anthropometric lab at the International exposition of 1884. 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

James McKeen Cattell. 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

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.

- Binet preferred to stick to the term ‘mental level’ whereas many other psychologists used ‘mental age’ instead. L. M. Terman and associates at Stanford University. This revision is called Stanford- Binet.. 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

As a result of the work of Binet and his associates the Ministerial Commission for the Study of Retarded Children was established

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.

Lesson 3

Esquirol: 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. He felt that language use by the person being diagnosed for mental retardation was the most reliable criterion for assessing intellectual level.

Seguin: 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.

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.

Lecture # 4

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

1. Individual differences exist
2. Psychological traits and states can be quantified and measured
3. Various approaches to measuring aspects of the same thing can be useful
4. Assessment can provide answers to some of life's most important questions: (test can form on spot for different things)
5. Assessment can pin point phenomena that require further attention or study: (diagnosis)
6. Various source of data is required for assessment process
7. Various sources of error are part of the assessment process.
8. Tests and other measurement techniques have strengths and weaknesses.
9. Test-related behaviors predict non-test related behaviors: for example in HTP (house, tree, person)
10. Present day behavior sampling predicts future: (useful for future too)

Join VU Group: <https://chat.whatsapp.com/HuGWSMfRr7AFCZZ33kmJyg>

11. Testing and assessment can be conducted in a fair and unbiased manner.
12. Testing and assessment benefit society.

Types of Tests:

- Individual versus group tests
- Ability, achievement, or aptitude tests
- Intelligence versus personality tests
- Speed test versus ability tests
- Structured personality tests versus projective tests
- Verbal versus non-verbal/performance tests.
- 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: WAIS and WISC, Stanford-Binet Scales, and Kaufman Scales are tests that are to be administered individually.

Group test : 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.

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**.

Achievement tests on the other hand are meant and designed for measuring the effect of educational programs or trainings.. **SAT is an example of such tests**.

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. **Differential aptitude tests (DAT) are the most widely used aptitude test batteries.**

3. Intelligence verses Personality tests:

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**. **Clerical Speed and Accuracy Test in DAT.**

Join VU Group: <https://chat.whatsapp.com/HuGWSMfRr7AFCZZ33kmJyg>

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:

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. That is why these are called objective tests. EPPS or Edwards Personal Preference Schedule is an example of this variety of personality tests.

An unclear or abstruse 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. HTP, WAT word association test, Rotter's incomplete sentence blank.

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 true 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.

However some tests are 'available to all'.. For example Multidimensional Health Locus of Control Scale and Self Efficacy Scale have placed their scales on websites and anybody can borrow and use them.

Major Contexts of Current Test Use:

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

i) Educational testing:

Join VU Group: <https://chat.whatsapp.com/HuGWSMfRr7AFCZZ33kmJyg>

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
- California Achievement Test

b. Tests of Minimum Competency in Basic Skills:

These tests measure mastery of basic skills in children as well as adults. These areas include reading, language and applied mathematics.

EXAMPLE: TABE battery or Tests of Adult Basic Education.

c. Teacher-made class room Tests:.

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.

ii) Occupational Testing:

Psychological tests are used for a variety of objectives to be attained in occupational or workplace settings.

- 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.

Join VU Group: <https://chat.whatsapp.com/HuGWSMfRr7AFCZZ33kmJyg>

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).

Lecture # 5

The following are potential confounding variables that need to be looked into, watched, and thoroughly controlled.

- testing process itself,
- examiner related variables
- the subject/examinee related variables

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

A good testing room

- Noise free
- Unnecessary traffic free
- Enough light
- Control temperature
- Comfortable seat
- Examiner neither too close nor too far

These components of anxiety can be measured with the help of several test inventories like Test Anxiety Inventory developed by Spielberger and his coworkers.

Join VU Group: <https://chat.whatsapp.com/HuGWSMfRr7AFCZZ33kmJyg>

Treatment of the test anxiety includes the behavior therapy procedures with the combination of the cognitive therapies

Lecture 6

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. Standardization sample is a group of people whose performance on a specific test is taken as a standard or norm for comparison.

Norms perform two functions or the purpose of norms

- Normative function(relative standing position)
- Comparative function(is it good or bad)

Statistical Concepts Used In Psychological Testing:

The aim of the statistical method is to organize and summarize the quantitative data.

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 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:

A diagram consisting of rectangles whose y-axis represents frequency of a variable and whose x-axis represent the class interval

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 on x-axis and frequencies are plotted on y-axis.

2-Normal Bell-Shaped Curve:

Join VU Group: <https://chat.whatsapp.com/HuGWSMfRr7AFCZZ33kmJyg>

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 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 $M = \Sigma X/N$. For example, children obtained the following marks in an 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 4th 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 curve 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:

Standard Deviation is a measure which shows how much variation (such as spread, dispersion, spread,) from the mean exists denotes by S.D

Variance: Variance is known as the “average squared deviation” around the mean. The formula of variance is $\Sigma(X-X)^2/n$. 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.

The more individual differences reveal the larger S.D while small individual differences indicate the low values of S.D

Lecture # 7

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**.

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)

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.

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.

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: n

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

Join VU Group: <https://chat.whatsapp.com/HuGWSMfRr7AFCZZ33kmJyg>

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)

- 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:

Any phenomena explain in rank order or sequence.

Gesell Developmental Schedules

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
- b) Uniformities of developmental sequence.

For example, a chronological sequence can be observed in visual fixation and in hand and finger movements 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

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.

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.

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:

$$IQ = \frac{MA}{CA} \times 100$$

If the child's MA equals to CA then the child's IQ will be exactly 100. IQ of 100 represents the average or means 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.

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.

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.

Lecture # 8

Standardization:

Norms are established for the sake of standardization of any test. Standardization is the process whereby a test is given to a large number of people than the raw score of those people are gathered which are statically analyzed and their mean and standard deviation is calculated and used as a standard for future. A standardized test is the one that has normative data, as well as clearly specified administration and scoring procedures.

The Normative Sample:

Normative sample should be a representative of a particular population for which the normative sample has been 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.

- 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.

- In order to obtain a true representative sample, careful sampling procedures need to be followed. Most populations comprise sub groups. The sample therefore should include members from all strata. **Proportionate stratified random** sampling is the best approach for selecting a representative sample
- 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.
- 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, then 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

National anchor norm

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 ‘equi-percentile method’. Test scores are compared and their equivalence is determined with reference to their corresponding percentile scores.

National anchor norms are very helpful in 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:

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. 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 basis of specific purpose of a test or subtests.

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 & Swardlik). 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.

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 group. SAT

Item Response Theory:

Item response theory can be understood in terms of the ‘latent trait models’. 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.

Lesson 9

Nature and Uses:

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.

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”

Glaser (1963) was first one to use the term ‘criterion-referenced testing’. 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. 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.

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

Domain referenced tests are commonly used in education and job proficiency

Less individual difference and can improve with learning.

Content Meaning:

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

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. .

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. 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 exist.

Minimum Qualification and Cutoff scores:

As we said earlier, mastery tests may adopt an all-or-none approach or make a three-way distinction. Minimum acceptable standard. 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. 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.
- b) 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.
- c) Whenever possible, cutoff scores should be established and verified with the support of experiential information.
- d) Test scores should be obtained from groups that are clearly different from each other on the relevant criterion behavior.

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

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.

The issue of norm versus criterion referenced tests.

Norm-referenced is a type of test that assesses the test taker's ability and performance against other test takers. It could also include a group of test takers against another group of test takers. This is done to differentiate high and low achievers

Criterion-Reference is a type of test that assesses the test taker's ability to understand a set curriculum. In this test, a curriculum is set in the beginning of the class, which is then explained by the instructor. At the end of the lesson, the test is used to determine how much did the test taker understand. This test is commonly used to measure the level of understanding of a test taker before and after an instruction is given

Pilot Work:

In Pilot testing, a selected group of end users try the system under test and provide the feedback before the full deployment of the system. In other words, it means to conduct a dress rehearsal for the usability test that follows. The process may involve the creation, revision, and deletion of many test items.

2. Test Construction:

Scale

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
- 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:

- **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.

- **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 of judges.
- **Sorting tasks (ordinal scale)** 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** we will only deal with this scales, where the number of possible values is finite. These possible values are often referred to as the *levels* of the scale.

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.

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** (count correct answers). 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** (near to correct or the concept is right)

A third scoring model is **ipsative scoring (sub 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 item 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

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. Matching item
- d. The Likert format
- e. The category format
- f. 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 and true or 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 $\frac{1}{2}$. 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

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”.

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.

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

$$\text{Corrected score} = R - \frac{W}{n-1}$$

Where, R= the number of right responses

W= the number of wrong responses

n= the number of choices for each item

Guessing should be avoided. 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.

c. 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.

d. 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.

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

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.

e. Checklists and O –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.

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.

Lesson 12

DeVellis (1991) has given a number of guidelines for writing items. Some of those are as follows:

- Clear definitions and item specificity: (objective and item should be clear)
- Developing an item pool: (make lot of items and then work on them)
- Appropriate level of reading difficulty: (readable and understandable by the test taker)
- Reasonable length of items:
- Avoid 'double- barreled' items: (more than one ideas in same statement)
- 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'.
- Avoid double negatives: e.g.
- Avoid biases

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

- Remembering
- Understanding
- Thinking

Lesson 13

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

Join VU Group: <https://chat.whatsapp.com/HuGWSMfRr7AFCZZ33kmJyg>

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. 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:

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)

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 size of correlation coefficient generally acceptable for the purpose of reliability is around .80 or .90.. However a coefficient less than .60 is usually not acceptable.

Lesson 14

Types of test 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
- Many changes may take place in the test takers.. 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 and they can be replaceable . They should match each other in all respects including:

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 series, 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'

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 randomly 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. But here fatigue and difficulty level can intervene
- A better option can be to consider odd numbered items in one half, and the even numbered items in the other half.
- 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:

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

r_{tt} = 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- 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.

= Coefficient of reliability of whole 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

$\sum pq$ = 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:

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

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

Lecture 15

Relationship between the Sample Tested and Reliability:

Two things affecting the reliability of sample

- Variability
- Ability level

Variability: Variability of score, or in other words individual differences, reliability index is better with sample having individual differences. 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. (Test is taken many times add their score average is calculated and standard deviation is calculated)

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

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

A standard error of measurement, estimates the variation around a “true” score for an individual when repeated measures are taken. About 68% people fall between -1 and +1 standard deviations from the mean.

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

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.

Lecture # 16

Definition:

- “Traditionally, the validity of a test has been defined as the extent to which a test measures what it was designed to measure”

Join VU Group: <https://chat.whatsapp.com/HuGWSMfRr7AFCZZ33kmJyg>

Validity is not, and should not be measured and reported as “highly” valid or “low” validity. It is reported in specific terms.

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.

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”

Content-Description Procedures:

Nature:

“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”

- Certain considerations are needed for content validation measure:
 - The test should 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.

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.

- **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 is 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 and inference are not involved e.g. questionnaire or scale for a survey.

Lesson 17

The validity of a test can be assessed in more than one way:

- a. Content validity: 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 lay down at the time of test planning, then the test is considered valid.
- b. Criterion related: you calculate the correlation between the results of your measurement and the results of criterion measurement. High correlation shows validity
- c. Construct validity: it is about ensuring that the method of measurement matches the construct you want to measure

According to Aiken

“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 set on the basis of test)

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”. 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

Concurrent Validity Evidence:

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”. 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 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:

- Academic achievement ; used for intellectual tests and similar tests
- Matriculation marks at the time of admission to higher classes; used for selection and screening.
- For those who have completed, or discontinued education, years of education can be good criteria.
- Performance in specialized training; used for selection, admission, achievement.
- Instructor’s ratings

- 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} 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**. (things should not overlap in criterion and test, concept can be same but not the elements)