Eulogio "Amang" Rodriguez Institute of
Science and Technology - Manila Campus
College of Arts and Sciences

Psychological Assessment

INTEGRATIVE COURSE IN PSYCHOLOGY (ICOPSCYH)

EARIST CAS | BS PSYCHOLOGY REVIEW CLASS

Academic Year 2023-2024

INTRODUCTION: CONCEPTS, APPLICATIONS AND LIMITATIONS

Test -a measurement device or technique used to quantify behavior or aid in the understanding and prediction of behavior.

Instruments - checklists, scales, surveys, and inventories to provide information

Item - specific stimulus to which a person responds overtly; this response can be scored or evaluated. Items are specific questions or problems that make up a test.

Psychological Test - set of items that are designed to measure characteristics of human beings that pertain to behavior

P**sychological Testing** - a systematic procedure for measuring a sample of be- havior by posing questions in a uniform manner.

- a systematic procedure for observing behavior and describing it with the aid of numerical scales or fixed categories (Cronbach, 1951).
- Objective and standardized measure of sample of behavior (Anastasi & Urbina, 1997)

- Classical Test Theory: X = T + e
- Standardized/uniform conditions
- Established scoring and interpretation
-

Psychological Assessment - Assessment is the gathering and integration of psychology-related data for the purpose of making psychological evalua- tion that is accomplished through the use of tools such as tests, interviews, case studies, behavioral observation, and especially designed apparatuses and measurement procedures (Cohen, Swerdlik, & Sturman 2013)
- An extremely complex... process of solving problems (answering questions) in which psychological tests are often used as one of the methods of collecting relevant data (Maloney & Ward 1976)

ASSESSMENT VS. TESTING (Cohen, et. al., 2013)

	Assessment	Testing
Objective	To answer referral question, solve a problem, or arrive at a decision through the use of tools of evaluations.	To obtain some gauge, usually numerical in nature, with regard to the ability or attribute.
Process	Typically individualized and focuses on how the individual processes rather than simply the results of that processing.	May be individual or group in nature with little or less regards with the mechanics of contents and processes.
Role of Evaluator	The assessor is key to the process of selecting tests and/or other tools of evaluation as well as drawing conclusion from the entire evaluation	The tester is not key to the process; practically speaking, one tester may be substituted for another tester without appreciably affecting the evaluation
Skill of the Evaluator	Typically requires an educated selection of tools of evaluation, skill in evaluation, and thoughtful organization and integration of data.	Typically requires technician-like skills in terms of administering and scoring a test as well as in interpreting a test result.
Outcome	Entails a logical problem-solving approach that brings to bear many sources of data designed to shed light on a referral question	Typically yields a test score or series of test scores.

PSYCHOLOGICAL ASSESSMENT TIMELINE (Erford, 2013)

220 BCE - Chinese set up civil service exams to select Mandarins
500 BCE - Greeks may have used assessments for educational purposes
1219 AD - English University administers first oral examination
1510 - Fiteherbert proposes the first measure of mental ability (identification of one's age, counting 20 pence, etc.)
1540 - Jesuit Universities administer first written exam
1575 - Spanish Physician Juan Huarte de San Juan defines intelligence in Examen de Ingenius (independent judgment, meek compliance when learning)
1692 - German Philosopher Thomasius advocates for gaining knowledge of the mind through objective, quantitative methods

1799 - In working with the "Wild Boy of Aveyron", Itard differentiates between normal and abnormal cognitive abilities

1837 - Seguin develops the Seguin Form Board Test and opens school for mentally retarded children

1838 - Esquirol advocates differences between mental retardation and mental illness, proposes that mental retardation has several levels of severity

1869 - Galton, of Individual Psychology, authors Hereditary Genius, sparking study on individual differences and cognitive heritability.

1879 - Wundt establishes world's first psychological laboratory at the University of Leipzig in Germany

1888 - J.M. Catell establishes assessment laboratory at the University of Pennsylvania, stimulating the study of mental measurements

1890 - Catell coins the term "Mental Test"

1897 - Ebbinghouse develops and experiments with tests of sentence completion, short term memory and arithmetic

1904 - Spearman espouses two-dimensional theory of intelligence (g=general factor, s = specific factor).

- Pearson develops the theory of correlation.

1905 - E.L Thorndike writes about test development principles, later introduces one of the first textbooks on the use of measurement in education.

- First standardized group test of achievement published.
- Jung's Word Association Test published.
- Binet and Simon introduce first "intelligence tests" to screen French public school children for mental retardation

1912	Stern introduces term "Mental Quotient"
1916	Terman publishes the Stanford Revision and Extension of the Binet-Simon Intelligence Scale
1917	Yerkes and Colleagues publishes the Army Alpha and Army Beta tests, designed for the intellectual assessment and screening of the U.S. military recruits
1919 – 1931	Multiple standardized tests were published including, but not limited to Rorschach Inkblot Test,
	Draw-A-Person Test, Stanford Achievement Test, etc.
1933	Thurstone introduces multiple factor analysis.
	Johnson develops a test scoring machine.
1935	Murray and Morgan develop the Thematic Apperception Test
1936	Piaget publishes *Origins of Intelligence*
1937	Terman and Merill revised their earlier work as the Stanford-Binet Intelligence Scale
1938	Bender publishes the Bender Visual Motor Gestalt Test
1939	Wechsler introduces the Wechsler-Bellevue Intelligence Scale
1940	Hathaway and McKinley publish the Minnesota Multiphasic Personality Inventory (MMPI)
1955	Wechsler revises Wechsler-Bellevue Intelligence Scale to Wechsler Adult Intelligence Scale (WAIS)
1956	Bloom publishes Taxonomy of Educational Objectives

1980s	Dramatic improvements on the published standardized tests and introduction of the Computer Assisted Psychological Assessments
1990s	Revisions of tests were published; WISC-III (1991), Wechsler Individual Achievement Test (WAIT – 1992), WAIS-III (1997)
2000s	Revisions of tests were published; Scholastic Achievement Test (SAT – 2002), WPPSI-III (2002), WISC-IV (2003), Stanford-Binet Intelligence Scale-Fifth Ed. (2003), WAIS-IV (2008), WAIT-II (2009), DSM-5 (2013)

CURRENT DEVELOPMENT 1980'S TO PRESENT (Bernardo, 2011)

- A new interest in psychological testing emerged in the 1980's; some psychologists have developed entirely new approaches.
- Psychological testing remains one of the most important yet controversial topics within Psychology.
- Testing is indeed one of the essential elements of Psychology. Several areas of Psychology depend upon measurement.

- An understanding of the basic principles of measurement is necessary for the effective study of any area of human behavior.
- The use of specific psychological tests is one of the basic skills of the applied psychologists. Training in Psychological testing is one of the psychologists' most distinguishing marks.
- In today's complex society, the relevance of the principles, applications and issue of psychological testing extends far beyond the field of psychology. Attorneys, physicians, business managers, education and many other professionals are frequently confronted with reports based on psychological tests.
- The future of testing may very well depend upon you and people like you. The more you know about testing, the better you will be able to base your decisions on facts and to ensure that tests are used for the most beneficial and constructive purposes.
- In the Philippine setting, the current trend in testing is towards indigenization, that is, development of locally-made tests that are attuned to the culture and behavior of the Filipino.

PHILOSOPHIES OF PSYCHOLOGICAL ASSESSMENT

Psychometric Approach

- Originated in the America which yield numerical estimate of the aspect of performance, personality factors, and other psychological variables
- Psychometric theory involves several distinct areas of study. First, psychometricians have developed a large body of theory used in the development of mental tests and analysis of data collected from these tests. This work can be roughly divided into classical test theory (CTT) and the more recent item response theory (IRT) (Rasch, 1960/1980).
- A currently widespread definition, proposed by Stanley Smith Stevens (1946), is that measurement is "the assignment of numerals to objects or events according to some rule".

Impressionistic Approach

- Originated in Germany which aims to come up with a comprehensive picture of a person's overall psychological functioning as oppose to providing quantitative figure in describing a person
- Impressionism attempts to capture/represent psychological perception of experience; experience as perceived by the mind, not just the eye or an objective observer ("Impressionism", n.d).

ASSUMPTIONS ABOUT PSYCHOLOGICAL TESTING AND ASSESS- MENT

(Cohen, et. al., 2013)

1. ***Psychological traits and states exist.*** A trait has been defined as "any distinguishable, relatively enduring way in which one individual varies from another" (Guilford, 1959, p. 6). States also distinguish one person from another but are relatively less enduring (Chaplin et al., 1988). The word distinguishable conveys the idea that behavior labeled with one trait term can be differentiated from behavior that is labeled with an- other trait term. The phrase "relatively enduring way" in the definition serves as a reminder that a trait cannot be expected to be manifest in an individual 100% of the time,

2. ***Psychological traits and states can be quantified and measured.*** Test authors, much like people in general, have many different ways of look- ing at and defining the same

phenomenon. Just think, for example, of the wide range of ways a term such as "aggressive" is used. We speak of an "aggressive salesperson," an "aggressive killer," and an "aggressive dancer," and in each of those different contexts "aggressive" carries with it a different meaning. If a personality test yields a score purporting to provide information about how aggressive a test taker is, a first step in understanding the meaning of that score is understanding how "aggressive" was defined by the test developer. More specifically, what types of behaviors are presumed to be indicative of someone who is aggressive as defined by the test?

3. ***Test-related behavior predicts non-test-related behaviors.*** Many tests involve tasks such as blackening little grids with a number 2 pencil or simply pressing keys on a computer keyboard. The objective of such tests typically has little to do with predicting future grid-blackening or key-pressing behavior. Rather, the objective of the test is more typically to provide some indication of other aspects of the examinee's behavior. For example, patterns of answers to true-false questions on the MMPI are used as indicators of the presence of mental disorders.

4. ***Tests and other measurement techniques have strengths and weak- nesses.*** Competent test users understand a great deal about the tests they use. They understand, among other things, how a test they use was developed, the circumstances under which it is appropriate to adminis- ter the test, how the test should be administered and to whom, how the test results should be interpreted and to whom, and what the meaning of the test score is. Competent test users understand and appreciate the limitations of the tests they use, as well as how those limitations might be compensated for by data from other sources. All of this may sound quite commonsensical. It probably is. Yet this deceptively simple assumption that test users know the tests they use and are aware of the tests' limitations-is emphasized repeatedly in the codes of ethics of associations of assessment professionals.

5. ***Various sources of error are part of the assessment process.*** In the context of the assessment enterprise, "error" need not refer to a deviation, an oversight, or something that otherwise violates what might have been expected. To the contrary, "error" in the context of psycho- logical testing and assessment traditionally refers to something that is not only expected but actually considered a component of the measure- ment process. In this context, error refers to a long-standing assumption that factors other than what a test attempts to measure will influence performance on the test. Because error is a variable in any psycholog- ical assessment process, we often speak of error variance. Test scores earned by examinees are typically subject to questions concerning the degree to which the measurement process includes error. For example, a score on an intelligence test could be subject to debate concerning the degree to which the obtained score truly reflects the examinee's IQ, and the degree to which it was due to factors other than intelligence.

6. ***Testing and assessment can be conducted in a fair and unbiased manner.*** All major test publishers strive to develop instruments that, when used in strict accordance with guidelines in the test manual, are fair. One source of fairness-related problems is the test user who attempts to use a particular test with people whose background and experience are different from the background and experience of people for whom the test was intended. In such instances, it is useful to empha- size that tests are tools that, like other, more familiar tools (hammers, ice picks, shovels, and so on), can be used properly or abused. Some potential problems related to test fairness are more political than psychometrician nature, such as the use of tests in various social programs For

example, heated debateoften surrounds affirmative action programs in selection, hiring, and access or denial of access to various opportunities. In many cases, the real question to be debated is, "What do we, as a society, wish to accomplish?" not "Is this test fair?"

7. ***Testing and assessment benefit society.*** At first glance, the prospect of a world devoid of testing and assessment might seem very appealing, especially from the perspective of a harried student preparing for a week of midterm examinations. Yet a world without tests would most likely turn out to be more of a nightmare than a dream. In such a world, people could hold themselves out to the public as surgeons, bridge builders, or airline pilots regardless of their background, ability, or professional credentials. In a world without tests, teachers and school administrators could arbitrarily place children in different types of special classes simply because that is where they believed the children belonged. Considering the many critical decisions that are based on testing and assessment procedures, as well as the possible alternatives (including decision making on the basis of human judgment, nepotism. and the like), we can readily appreciate the need for the assessment enterprise and be thankful for its existence.

TYPES OF ASSESSMENT TOOLS

1. Standardized vs. Non-standardized
Any test in which the same test is given in the same manner to all test takers, and graded in the same manner for everyone, is a standardized test. Standardized tests are perceived as being fairer than non-standardized tests, because everyone gets the same test and the same grading system. Principles of Standardized Tests

I. Reliability
II. Validity
III. Test Administration, Scoring, and Interpretation

2. Verbal vs. nonverbal
Verbal tests are not necessarily spoken but may be written and nonverbal is considered performance tests,

3. Paper and pencil vs. Performance
The use of paper and pencil/ballpoint pen compared to a performance test

4. Individual vs. group testing
Deals with the number of test takers during a testing procedure

5. Speed vs. Power
Deals with the time component of a test and difficulty level

6. Objective vs. Subjective
Reflects the methods used to score the assessment tool.

7. Cognitive vs. Affective or Human Ability vs. Personality

- Cognitive instruments are those that assess cognition: perceiv- ing, processing, concrete and abstract thinking, and remembering.
- Affective instruments assess interests, attitudes, values, mo- tives, temperaments, and non-cognitive aspects of personality. There are also tests that measure psychomotor skills.

TYPES OF PSYCHOLOGICAL TESTS

Intelligence tests.
These are used to measure intelligence, or your ability to understand your envi- ronment, interact with it and learn from it
Examples:

- Wechsler Adult Intelligence Scale (WAIS)
- Wechsler Intelligence Scale for Children (WISC)

- Stanford-Binet Intelligence Scale (SB)

Personality tests.
These are used to measure personality style and traits. Personality tests are commonly used in research or to assist with clinical diagnoses.
Examples:
- Minnesota Multiphasic Personality Inventory (MMPI)
- Thematic Apperception Test (TAT)
- Rorschach, also known as the 'inkblot test'
-

Attitude tests.
Such as the Likert Scale or the Thurstone Scale, are used to measure how an individual feels about a particular event, place, person or object.

Achievement tests.
These are used to measure how well you understand a particular topic fe, (mathematics achievement tests).
Examples:
- American College Test (ACT)
- kowa Test of Basic Skills.
- STAR Early Assessment.

Aptitude tests.
These are used to measure your abilities in a specific area (ie. clerical skills).
Examples:
- Differential Aptitude Test (DAT)
- Wide Range Achievement Test

Neuropsychological tests.
It is used to find out how damage to your brain is affecting your ability to reason,concentrate, solve problems, or remember.
Examples:
- Barcelona Neuropsychological Test (BNT)
- Cambridge Neuropsychological Test Automated Battery (CANTAB)
- Cognistat (The Neurobehavioral Cognitive Status Examination)

Vocational tests.
It is the process of determining an individual's interests, abilities and aptitudes and skills to identify vocational strengths, needs and career potential.
Examples:
- Career Occupational Preference Survey (COPES)
- High School Placement Test (HSPT)

Direct observation tests.
It involves the observation of people as they complete activities. This type of assessment is usually conducted with families in a laboratory, home or with children in a classroom.
Examples:
- The Parent-Child Interaction Assessment-II (PCIA)
- The MacArthur Story Stem Battery (MSSB)
- Dyadic Parent-Child Interaction Coding System-II

Biographical Information Blank
The Biographical Information Blanks or BIB is a paper-and-pencil form that includes items that ask about detailed personal and work history. It is used in the hiring of employees by matching the backgrounds of individuals to requirements of the job.

PURPOSES OF PSYCHOLOGICAL ASSESSMENT (Erford, 2013)

1. **Screening** is a quick procedure, usually involving a single measure, done for the purpose of determining whether deeper diagnostic assessment is necessary or warranted. A screening process is by no means comprehensive, and the instruments used for this purpose are sometimes held to lower standards of psycho- metric accuracy, although this is not always a desirable practice. Accuracy in screening is just as critical as accuracy in

diagnosis because both procedures, done correctly, save students and clients emotional pain, time, and money

2. **Diagnosis** entails "a detailed analysis of an individual's strengths and weaknesses, with the general goal of arriving at a classification decision" (Erford, 2013). Diagnosis always involves more than one measure and often includes a battery of tests. Such a battery is usually composed of a series of tests that are integrated to yield specific information or identification decisions

3. **Treatment Planning and Goal Identification.** Assessment helps clients and students to understand where they are and where they want to go, a key facet of developing a client's goals and objectives. Thus, a primary purpose of assessment is to help establish goals, often through a combination of assessment methods, including interviewing and standardized testing.In addition, the information gathered from an initial assessment can be helpful in planning a client's treatment. Frequently, student or client strengths, weaknesses. es, challenges, and resiliency factors and resources are confirmed or better understood through assessment procedures. "Treatment planning must flow logically from assessment results, fit the given environmental context of the client, and be individualized to mesh with the client's strengths and weaknesses" (Erford, 2013).

4. **Progress Evaluation.** One cannot overemphasize the importance of assessing and evaluating progress to determine whether the treatment or intervention is working. Failure to periodically evaluate treatment progress is unethical and unprofessional, not to mention inefficient. If a treatment is having no positive effects, the client is wasting time and money while continuing to experience discomfort. Tests and inventories can be very helpful aids in assessing treatment outcomes.

USES OF PSYCHOLOGICAL TESTS (Mote, 2017)

Assessment
Psychologists use tests during one of your fist few sessions to assess your problem. A psychologist tests at this point to supplement his clinical interview and to determine the severity, duration and extent of your problems

Setting Goals
Psychologists use test results to help you set goals for improvement. Psychologists use unusual results, such as a high occurrence of depression, to develop specific and measurable goals. Goals such as "produce the frequently of depression to half of that initially discovered" are clear and can be measured to show improvement.

Determining Interventions
Psychologists also use tests to identify the most effective interventions for you. Personality tests, such as the Myers-Briggs Type Indicator, can reveal much about how you think and the way you relate to other people. These tests reveal your strengths as well as your weaknesses. For example, if your test results reveal that you are a highly analytical person, interventions such as reading and rational analysis of problems may be effective in helping you make desired changes.

Reviewing Progress
Most psychologists use tests as a way of reviewing what you have accomplished in treatment. If you scored high on the Beck Anxiety Inventory in your initial assessment, re-taking the test three months later may reveal lower anxiety and provide you with momentum to keep up the work

Closure
A psychologist does not want to keep you dependent on her. Her goal is to build your competence and confidence so you can manage

your problems on your own. Psychologists often use tests as a way of ending treatment. Test results are used as evidence in closing discussions about the progress you have achieved on

LIMITATIONS OF PSYCHOLOGICAL TESTS

1. Scores can't reveal how or why the individual obtained a certain score
2. May seem to give favorable responses
3. Doesn't measure the ability or potential to apply and appreciate information gained
4. Results can't make decisions for the examinee
5. Chance error on individual interpretation of scores
    a. SEM reasonable limits of scores and yet maintain its reli- ability
    b. b. SE Difference between two scores for test of significance
    c. c. SEMargin of error expected an individual's predicted criterion
6. It only infer from a sample of behavior
7. It uses limited scale only
8. Easily affected by extraneous variables
9. Many measure hypothetical constructs and theoretical abstracts only

STATISTICAL FOUNDATIONS

According to Orense & Reyes (2014)

Descriptive Statistics - Methods used to provide a concise description of a collection of quantitative information
Inferential Statistics - Methods used to make inferences from observations of a small group (n) to a larger group (N).
Measurement - application of rules for assigning numbers to objects.
Objects - Qualities or attitudes that are transformed into numbers.

Scales of Measurement and their Properties

Nominal Scale.
The word nominal is derived from nomen, the Latin word for name. Nominal scales merely name differences and are used most often for qualitative variables in which observations are classified into discrete groups. The key attribute for a nominal scale is that there is no inherent quantitative difference among the categories. Variables on a nominal scale are often called categorical variables.

Examples:
Not really a "scale" because it does not scale objects along any dimension. It simply labels objects

Gender is a nominal scale
Male-1
Female-2
Religious Affiliation
Catholic-1
Protestant-2
Jewish-3
Muslim-4
Other-5

Strength of Nominal Scales
Easy to generate from closed questions, large amounts of questions can be collected quickly; increasing reliability

Weakness of Nominal Sales
Without linear scale participants may be unable to express degrees of response, can only use the mode as a measure of spread

Ordinal Scale
Ordinal scales rank-order observations. Class rank and horse race results are examples. There are two salient attributes of an ordinal scale. First, there is an underlying quantitative measure on which the observations differ. For class rank, this underlying quantitative attribute might be composite grade point average, and for horse race results it would be time to the finish line.

The second attribute is that individual differences on the underlying quantitative measure are either unavailable or ignored. As a result, ranking the horses in a race as 1st, 2nd, 3rd, etc. hides the information about whether the first-place horse won by several lengths or by a nose

Strength of Ordinal Scale
Indicates relative values on a linear scale instead of just totals; more informative than nominal data

Weakness of Ordinal Scale
Gaps between the values aren't equal so a mean cannot be used to assess central tendency

Interval Scales.
An interval scale has a constant interval but lacks a true 0 point. As a result, one can add and subtract values on an interval scale, but one cannot multiply or divide units. Interval relationships are meaningful
Examples:

Temperature used in day-to-day weather reports is the classic example of an interval scale. The assignment of the number to a particular height in a column of mercury is an arbitrary convenience apparent to everyone familiar with the difference between the Celsius and Fahrenheit scales. As a result, one cannot say that 30o C is twice as warm as 15o C because that statement involved implied multiplication. To convince yourself, translate these two into Fahrenheit and ask whether $60 F is twice as warm as 50o F.

Nevertheless, temperature has constant intervals between numbers, permitting one to add and subtract. The difference between 28o C and 21o Cis 7 Celcius units as is the difference between 53o C and 46o C. Again, convert these to Fahrenheit and ask whether the difference between 82.4o F and 69.8o F is the same in Fahrenheit units as the difference between 127.4o F and 114.5o F

Strength of Interval Scale
More informative than ordinal and nominal as the points are directly comparable because they are all of equal value; scientific measures used to record the distance between values are highly reliable

Weakness of Interval Scale
In interval scales that do not contain scientific measurements there is no absolute baseline to the scale so scoring 0 may not mean the participant doesn't demonstrate the variable but that the scale doesn't measure it

Ratio Scales
A ratio scale has the property of equal intervals but also has a true point. As a result, one can multiply and divide as well as add and subtract using ratio scales. Units of time (msec, hours), distance and length (cm, kilometers), weight (mg, kilos), and volume (cc) are all ratio scales. Scales involving division of two ratio scales are also themselves ratio scales. Hence, rates (miler per hour) and adjusted volumetric measures (mg/dL) are ratio scales. Note that even though a ratio scale has a true 0 point, it is possible that the nature of the variable is such that a value of 0 will never be observed. Human height is measured on a ratio scale but every human has a height greater than 0. Because of the multiplicative property of ratio scales, it is possible to make statements that 60 mg of fluoxetine is three times as great as 20 mg.

With these, there is practically no ratio scale data in psychology and other social sciences. For instance, there is no such thing as ""intelligence or "O" aggression as personality trait

MEASUREMENT SCALES: OTHER VIEWS (Sheskin, 2011)

Continuous and Discrete Variables
A *continuous variable* has an infinite number of possible values between any two points on the measurement scale. For example, mouse weight

will have an infinite number of possible values between 25 grams and 26 grams because one could always add extra decimal places to the measurement.

A *discrete variable* on the other hand can only take on a limited number of values. By their nature, all categorical variables are discrete, but so are many variables measured on ratio scales. One very important type of discrete variable measured on a ratio scale is a count such as the number of pups in a rat litter or number of correct responses on a memory task. Counts are always positive integers.

Bounded Variables

All variables probably have mathematical bounds imposed by nature the weight of an adult human brain, for example, has a lower and upper bound even though the exact numbers for these bounds may be unclear. The term bounded variables, on the other hand, refers to measurement scales with a mathematical boundary. Counts, for example, have a lower bound of 0 while percent have a lower bound of 0 as well as an upper bound of 100.

Bounded variables may not do not have to present problems for statistics. For example, many statistics apply the normal curve to data. The equa tion for the normal curve assumes that scores are symmetric around the mean and can range from negative infinity to positive infinity (although as scores deviate more and more from the mean, the probability of observing them becomes less and less). Many variables measured as percent, however, may have values that cluster close to 0 (or 100). In such cases, mathematical transformations of the original variables are used and the statistical analysis is performed on the transformed values

Categorical versus Categorized Variables

True *categorical variables* place observations into groups in which there is no implication about differences in magnitude between the groups. A dichotomous variable has two mutually exclusive categories with no implication of a difference in magnitude between the categories. Sex (male versus female) is dichotomous as is virginity (either you are or you were). The term binary variable is synonymous with dichotomous variable. The phrase polychotomous refers to a categorical variable with more than two mutually exclusive classes.

Categorized variables, on the other hand, develop arbitrary but often meaningful classes from a variable that is inherently quantitative. A dichotomized variable is one that has an underlying quantitative scale in which an arbitrary cut point is used to divide observations into two classes. Classic examples are the cutoffs for cholesterol levels or blood pressures that are used in medicine to make treatment decisions about hypercholesterolemia and hypertension. The sem polychotomous applies to two or more cut points that result in more than two ordered classes-eg. blood pressure could give hypotensive, normotensive, and hypertensive groups.

As in most attempts at classification, there are large gray areas that are not well captured by distinguishing categorical from categorized variables. Can- cer is clearly categorical but within cancer victims it is important to quantify the stage of the cancer. Psychiatric disorders provide another gray area. Many regard a diagnosis of antisocial personality disorder as a categorized variable reflecting an underlying continuum of prosocial to antisocial behavior, but the same consensus will not be found for schizophrenia.

A further gray area is the ordered group variable. An ordered group variable has one (or possibly more) underlying quantitative dimension(s) but lacks clear cut-points to separates the groups. Division of professorial ranks into assistant, associate, and full professors is an example of an ordered group vari- able. Some classification

systems in psychopathology also use ordered groups. For example, the "schizophrenia spectrum" may be defined as not affected, schizotypal personality disorder, and schizophrenia.

Type of Scale	*Magnitude*	*Equal Intervals*	*Absolute 0*
Nominal	No	No	No
Ordinal	Yes	No	No
Interval	Yes	Yes	No
Ratio	Yes	Yes	Yes

DESCRIBING DISTRIBUTIONS

1. **Frequency Distribution.** The number of times that a certain value or range of values (score interval) occurs in a distribution of scores. Distribution , on the other hand is a tabulation of scores (ordered either in ascending or descending value) showing the number of individuals in a group obtaining each score or contained within a specified fixed range of scores (score interval).

2. **Frequency Polygon**

   a. **Skewness.** A measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

      i. *Positive Skew-* In positively skewed distributions, the mean is usually greater than the median, which is always greater than the mode.

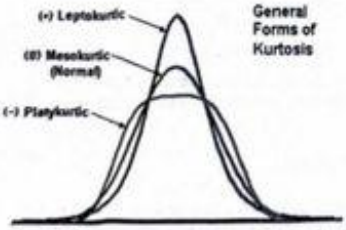
      ii. *Negative Skew-* A distribution is negatively skewed, or skewed to the left, if the scores fall toward the higher side of the scale and there are very few low scores.

   b. **Kurtosis** - is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.

      i. *Platykurtic*

      ii. *Leptokurtic*

      iii. *Mesokurtic*



General Forms of Kurtosis

IMAGE SOURCE: https://mvpprograms.com/help/mvpstats/distributions/

   c. **Normal Curve / Normal Distribution.** Also known as the bell-shaped curve because of its distinctive appearance in that scores are distributed symmetrically about the middle, such that there are an equal number of scores above as below the mean, with more scores concentrated near the middle than at the extremes. The normal distribution is a theoretical distribution defined by specific mathematical properties that many human traits and psychological characteristics appear to closely approximate (e.g., height, weight, intelligence, etc.).

Some features of the normal distribution are:

1.) The mean, median, and mode are identical in value.

2.) The scores are distributed symmetrical about the mean (50.0% above the mean and 50.0% below the mean).

3.) 68.26% of the scores are within 1 standard deviation of the mean (34.13% above the mean and 34.13% below the mean).

4.) 95.44% of the scores are within 2 standard deviations of the mean (47.72% above the mean and 47.72% below the mean)

5.) 99.72% of the scores are within 3 standard deviations of the mean (49.86% above the mean ad 49.86% below the mean).

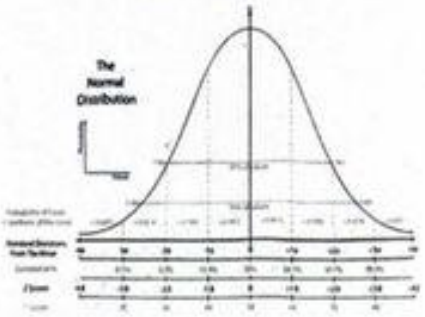


IMAGE SOURCE: https://math.stackexchange.com/questions/2148839/why-do-depictions-of-the-normal-distribution-in-textbooks-often-not-look-normal/2149119

3. **Measures of Central Tendencies**

   a. **Mean.** The average of a set of scores obtained by adding all scores in the set and dividing by the total number of scores. For example, the mean of the set {15, 13, 18, 16, 14, 17, and 12} is 15

$$\text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}}$$

$$= \frac{15+13+18+16+14+17+12}{7}$$

$$= \frac{105}{7}$$

$$= 15$$

b. **Median**. The middle point (score) is a distribution of ranked-ordered scores that divides the group into two equal parts, each part containing 50% of the data. The median corresponds to the 50th percentile. Half of the scores are below the median and half are above it, except when the median itself is one of the obtained scores. For example, the median of the set (4, 7, 8, 9, 10, 10, 12) is 9.

c. **Mode**. The score that occurs most frequently in a distribution of scores. For example, the mode for the set (65, 72, 85, 85, 70, 90) is 85. A distribution can have more than one mode. For example, the modes for the set (65, 72, 72, 72, 85, 85, 85, 70, 70, 90) are 72 and 85, since both these scores have a frequen- cy of three, which is the highest frequency for the distribution.

4. Measures of Variability

a. **Range**. The range of a distribution of scores is defined as the difference between the two extremes (maximum score minus minimum score), and is a rough indication of the spread or variability of the scores. The range of the set (65, 72, 85, 85, 85, 70, 90) is calculated as (90-65), which is equal to 25.

b. **Semi-interquartile Range (Quartile Deviation).** A measure of spread or dispersion. It is computed as one half the differ- ences between the 75th percentile (often called (Q3)] and the 25th percentile (Q1). The formula for semi-interquartile range is therefore: (Q3-Q1)2.

c. **Standard Deviation.** A statistic that measures the degree of spread or dispersion of a set of scores. The value of this statis tic is always greater than or equal to zero. If all of the scores in a distribution are identical, the standard deviation is equal to zero. The further the scores are away from each other in value, the greater the standard deviation. This statistic is calculated from using the information about the deviations (distances) between each score and the distribution's mean. It is equiva- lent to the square root of the variance statistic. The standard deviation is often the preferred m thod of examining a distri- bution's variability since the standard deviation is expressed in the same units as the data.

5. Standardized Scores

a. **Individual Score**

i. *Percentile Score/Rank.* The percentage of scores in a specified distribution that fall at or below the point of a given score. Percentile Ranks range in value from 1 to 99, and indicate the status or relative standing of an individual within a specified group (eg, norms group), by indicating the percent of individuals in that group who obtained lower scores. For example, if a student earned a 72nd Percentile Rank in Language, this would mean he or she scored better than 72 percent of the students in a particular norm group who were administered that same test of Language. This also implies that only 28 percent (100-72) of the norm group scored the same or high- er than this student. Note however, an individual's percentile rank can vary depending on which group is used to determine the ranking. A student is simultaneously a member of many groups: classroom, grade, building, school district, state, and nation. Test developers typically publish different sets of percentile ranks to permit schools to make the most relevant comparisons possible.

b. **Linear Standard Score**

i. *Z-Score.* A type of standard scores such that the distribution of the scores for a specified population has a mean of 0.0 and a standard deviation of 1.0. The z-score indicates the amount a student's score (X) deviates from the mean in relation to the stan- dard deviation (SD) of the group. For example, if a student's score is 25, and the group's mean and standard deviation are 15 and 5, respectively, then the student's z-score would be 2.0, indicating that the student scored 2 standard deviations above the group mean.

$$Z = \frac{X - \bar{X}}{SD} = \frac{(25 - 15)}{5} = 2$$

ii. *T-Score.* A normalized standard score, having a mean of 50 and a standard deviation of 10. T-Scores are a direct transformation of z-scores and range (rough- ly) from 20 to 80 (corresponding to approximately 3 standard deviations above and below the mean). Example: T-Score for a score 3 standard deviations below the mean:

$$\begin{aligned} \text{T-Score} &= (z)(10) + 50 \\ \text{T-Score} &= (-3.0)(10) + 50 \\ &= (-30) + 50 \\ &= 20 \end{aligned}$$

c. **Normalized Standard Scores**

1. *Stanine*. The name stanine is simply a derivation of the term "standard-nine" scale. Stanines are normalized standard scores, ranging in value from 1-9, whose distribution has a mean of 5 and a standard deviation of 2. Stanines 2 through 8, are equal to a 1⁄2 standard deviation unit in width, with the middle stanine of 5 defined as the range of scores % of a

standard deviation below to % of a standard deviation above the mean. Stanines can, more easily, be thought of as coarse groupings of percentile ranks (see below), and like percentile ranks indicate the status or relative rank of a score within a particular group. Due to their coarseness, stanines are less precise indicators than percentile ranks, and at times be misleading (e.g., similar PR's can be grouped into different stanines [e.g., PR-23 and PR-24] and dissimilar PR's can be grouped into the same stanine [e.g., PR-24 and PR-40]). However, some find that using stanines tends to minimize the apparent importance of minor score fluctuations, and are often helpful in the determination of areas of strength and weakness.

ii. *Sten.* A sten score indicates an individual's approximate position (as a range of values) with respect to the population of values and, therefore, to other people in that population. The individual sten scores are defined by reference to a standard normal distribution. Unlike stanine scores, which have a midpoint of five, sten scores have no midpoint (the midpoint is the value 5.5). Like stanines, individual sten scores are demarcated by half standard deviations. Thus, a sten score of 5 includes all standard scores from -.5 to zero and is centered at -0.25 and a sten score of 4 includes all standard scores from -1.0 to -0.5 and is centered at -0.75. A sten score of 1 includes all stan- dard scores below -2.0. Sten scores of 6-10 "mirror" scores 5-1.

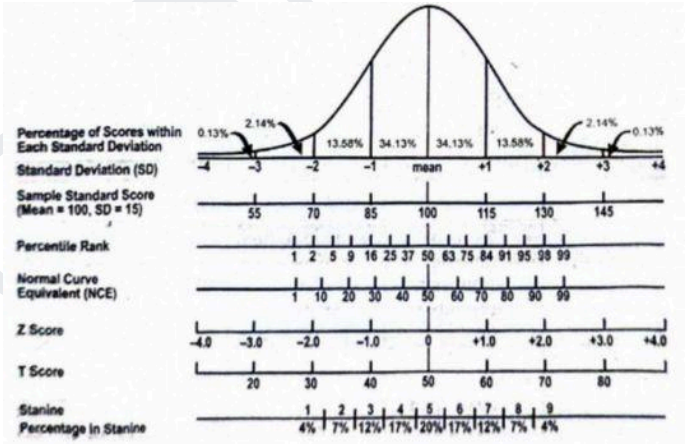


IMAGE SOURCE: http://www.wikiwand.com/en/standard_score

CORRELATIONAL STATISTICS – A statistic that indicates the degree of relationship between any two sets of scores obtained from the same group of individuals. The degree of association is computed and measured through correlation coefficient (Singh, 2007).

CORRELATION COEFFICIENT – This refers to a mathematical index that describes the direction and magnitude of a relationship. This may be usually either Positive (Direct) or Negative (Inverse) (Singh, 2007).

+/- 0.0 to 0.19	Very weak, negligible correlation
+/- 0.20 to 0.39	Weak, low correlation
+/- 0.40 to 0.59	Moderate correlation
+/- 0.60 to 0.79	Strong, high correlation
+/- 0.80 to 1.0	Very strong correlation

Coefficient of Determination – squared value of the correlation coefficient. It is the proportion of the total variation in scores on Y that we know as a function of information about X (Singh,2007).

CORRELATIONAL TECHNIQUES(Orense & Reyes, 2014)

Statistical Tool	*Set of Scores*	*Type of Measure*	*Interpretation*
Pearson r	2 sets of scores from same respondents	Both are Interval or Ratio (Scale data)	-1 to +1 = perfect .91 to .99 = very high .71 to .90 = high .41 to .70 = moderate .21 to .40 = slight/ weak .00 to .21 = no corr.
Spearman's rho	• 2 sets of ranking from same respondents. • N is not more than 30	Both are ordinal	-1 to +1 = perfect .91 to .99 = very high .71 to .90 = high .41 to .70 = moderate .21 to .40 = slight/ weak .00 to .21 = no corr.
Kendall's Tau	• 2 sets of ranking from same respondents. • N can be more than 30 • Tau coefficient is smaller than spearman's rho, when the tools are used.	Both are ordinal	-1 to +1 = perfect .91 to .99 = very high .71 to .90 = high .41 to .70 = moderate .21 to .40 = slight/ weak .00 to .21 = no corr.
Kendall's W	More than two sets of ranking	All are ordinal, ranking from several judges or raters	0 to 1 0 = no agreement between judges 1 = perfect agreement between judges
Phi Coefficient	2 or more sets of frequencies	All are Nominal Data	-1 to +1 = perfect .91 to .99 = very high .71 to .90 = high .41 to .70 = moderate .21 to .40 = slight/ weak .00 to .21 = no corr.
Multiple correlation	3 or more sets of Pearson	All are ratio or Interval (Scale Data)	Relationship between one variable and a combination of two other variables.

PARAMETIC TESTS (Singh, 2007)

1. **Z-test** is any statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution. (NOTE: $n \geq 30$; single parameter)

   Example: Suppose that in a particular geographic region, the mean and standard deviation of scores on a reading test are 100

points, and 12 points, respectively. Our interest is in the scores of 55 students in a particular school who received a mean score of 96. We can ask whether this mean score is significantly lower than the regional mean, that is, are the students in this school comparable to a simple random sample of 55 students from the region as a whole, or are their scores surprisingly low?

2. **T-test** assesses whether the means of two relatively small groups of normal distributions are statistically different from each other.

3. **One-way analysis of variance (ANOVA)** is used to determine whether there are any significant differences between the means of three or more independent (unrelated) groups.
Example: a researcher wishes to know whether different pacing strategies affect the time to complete a marathon. The researcher randomly assigns a group of volunteers to either a group that (a) starts slow and then increases their speed, (b) starts fast and slows down or (c) runs at a steady pace throughout. The time to complete the marathon is the outcome (dependent) variable.

4. **Two-way analysis of variance (ANOVA)** compares the mean differences between groups that have been split on two independent variables (called factors) on the dependent variable.
Example: you may want to determine whether there is an interaction between physical activity level and gender on blood cholesterol concentration in children, where physical activity (low/ moderate/high) and gender (male/female) are your independent vari- ables, and cholesterol concentration is your dependent variable.

NONPARAMETRIC TEST (Singh, 2007)
Situations in which nonparametric test are used:

A. The data involve measurements on nominal or ordinal scales. In these situations, you cannot compute the means and variances that are essential part of the parametric tests.
B. The data do not satisfy the assumptions underlying parametric tests.
C. The data have extremely high variance, which can undermine the likeli- hood of significance for a parametric test. In this case, the scores can be converted to categories or ranks, and a nonparametric test can be used as an alternative.

Some Nonparametric Tests (Singh, 2007)

1. **Chi Square**statistic is used to investigate whether distributions of categorical variables differ from one another.

   ❖ **chi-square test for independence.** The test is applied when you have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables.

Example: in an election survey, voters might be classified by gender (male or female) and voting preference (Democrat, Republican, or Independent). We could use a chi-square test for independence to determine whether gender is related to voting preference.

   ❖ **chi-square goodness of fit test.** The test is applied when you have one categorical variable from a single population. It is used to determine whether sample data are consistent with a hypothesized distribution.

Example: suppose a company printed baseball cards. It claimed that 30% of its cards were rookies; 60%, veterans; and 10%, All-Stars. We could gather a random sample of baseball cards and use a chi-square goodness of fit test to see whether our sample distribution differed significantly from the distribution claimed by the company.

2. **Binomial Test** is an exact test of the statistical significance of deviations from a theoretically expected distribution of observations into two categories. One common use of the binomial test is in the case where the null hypothesis is that two categories are equally likely to occur (such as a coin toss).

3. **Sign-test** is the alternative test to the Wilcoxon test for dependent data. One requirement of the **Wilcoxon test** is that the data needs to be at least interval scaled. For the sign test the data needs to be at least ordinal scaled.

PARAMETRIC VS. NONPARAMETRIC (Orense& Reyes, 2014)

	Parametric	Non-parametric
Assumed distribution	Normal	Any
Assumed variance	Homogeneous	Any
Typical data	Ratio or Inter-val	Ordinal or Nominal
Data set relationships	Independent	Any
Usual central measure	Mean	Median

— 362 —

Benefits	Can draw more conclusions	Simplicity; Less affected by outliers
Tests		
Choosing	Choosing parametric test	Choosing a non-parametric test
Correlation test	Pearson	Spearman
Independent measures, 2 groups	Independent-measures t-test	Mann-Whitney test
Independent measures, >2 groups	One-way, independent-measures ANOVA	Kruskal-Wallis test
Repeated measures, 2 conditions	Matched-pair t-test	Wilcoxon test
Repeated measures, >2 conditions	One-way, repeated measures ANOVA	Friedman's test

RELIABILITY (McLeod, 2007)

Reliability is defined as the repeatability of measurement. If one gave the sameinterview, questionnaire, or test on two occasions, then how well wou the results on thefirst administration predict those of the second administratior If the level of predictionis high, then the psychometric instruments is said to b reliable. If predictability is poor,then the instrument is unreliable.

The time interval between the administrations of the measuring instr mentsdepends on the nature of the trait under study and on the type of reliabil One type ofreliability, termed internal consistency reliability, actually measur reliability for asingle administration. Internal consistency reliability is approp ate only for scales whereeach item measures the same trait. Here, the reliabili consists of how well scores on,say, a subset of items predict scores on other subsets of items. When all the items on asingle scale—e.g., the word in a voc ulary test—are purported to measure the sametrait—vocabulary ability—then the scale should have high internal consistency reliability.

Some traits—mood, for instance—can change in a matter of hours. Hence, aneuroscientist interested in affect might examine the reliability of a mood measure usinga time interval of an hour or two. Other traits—e.g., intel ligence—might be expected tochange little over the course of a week or two. Hence, the time interval required to assessthe reliability of a measure of intel ligence could be much greater than that for mood.Often, psychometricians wil administer a different form of the test on each occasion to control for memory effects.

— 363 —

It is important to distinguish reliability from trait stability, although in practice the two concepts blend into each other rather than being discrete entities. The time interval for reliability testing depends on a theoretical implication and/ or common-sense judgment about how quickly individual differences in trait may change for natural reasons. Stability, on the other hand, is an empirical question about how well one can predict trait scores over a long period of time. Reliability of measurement is a prerequisite for assessing trait stability. If the measure has poor reliability, then and assessment of stability must be questioned. High reli- ability, on the other hand, does not imply high stability. A measure of mood may be reliable, but the correlation between scores assessed now and a year from now may be low.

TYPES OF RELIABILITY (McLeod, 2007)

The split-half method

It assesses the internal consistency of a test, such as psychometric tests and questionnaires. There, it measures the extent to which all parts of the test contribute equally to what is being measured.

This is done by comparing the results of one half of a test with the results from the other half. A test can be split in half in several ways, e.g. first half and second half, or by odd and even numbers. If the two halves of the test provide similar results this would suggest that the test has internal reliability.

The reliability of a test could be improved through using this method. For example any items on separate halves of a test which have a low correlation (e.g. r = .25) should either be removed or re-written.

The split-half method is a quick and easy way to establish reliability. However it can only be effective with large questionnaires in which all questions measure the same construct. This means it would not be appropriate for tests which measure different constructs.

For example, the Minnesota Multiphasic Personality Inventory has sub scales measuring behaviors such depression, schizophrenia, social introversion. Therefore the split-half method was not an appropriate method to assess reliability for this personality test.

Inter-Rater or Inter-Observer Reliability

Used to assess the degree to which different raters/observers give con- sistent estimates of the same phenomenon. The test-retest method assesses the external consistency of a test. This refers to the degree to which different raters give

consistent estimates of the same behavior. Inter-rater reliability can be used for interviews. Note, it can also be called inter-observer reliability when referring to observational research. Here researchers observe the same behavior inde- pendently (to avoided bias) and compare their data. If the data is similar then it is reliable.

Where observer scores do not significantly correlate then reliability can be improved by:

- Training observers in the observation techniques being used and making sure everyone agrees with them.
- Ensuring behavior categories have been operationalized. This means that they have been objectively defined.

For example, if two researchers are observing 'aggressive behavior' of children at nursery they would both have their own subjective opinion regarding what aggression comprises. In this scenario it would be unlikely they would record aggressive behavior the same and the data would be unreliable.

However, if they were to operationalize the behavior category of aggression this would be more objective and make it easier to identify when a specific behavior occurs.

For example, while "aggressive behavior" is subjective and not operationalised, "pushing" is objective and operationalized. Thus researchers could simply count how many times children push each other over a certain duration of time.

Test-Retest Reliability

Used to assess the consistency of a measure from one time to another. The test-retest method assesses the external consistency of a test. Examples of appropriate tests include questionnaires and psychometric tests. It measures the stability of a test over time.

A typical assessment would involve giving participants the same test on two separate occasions. If the same or similar results are obtained then external reliability is established. The disadvantages of the test-retest method are that it takes a long time for results to be obtained.

The timing of the test is important; if the duration is to brief then participants may recall information from the first test which could bias the results. Alternatively, if the duration is too long it is feasible that the participants could have changed in some important way which could also bias the results.

METHODS OF RELIABILITY ASSESSMENT AND SOURCES OF MEASUREMENT ERROR (Orense & Reyes, 2014)

Source of Error	Example	Method	How Assessed
Time sampling	Same test given at two points in time	Test-retest	Correlation between scores obtained on the two occasions
Item Sampling	Different items used to assess the same attribute	Alternate forms or parallel forms	Correlation between equivalent forms of the test that have different items
Internal consistency	Consistency of items within the same test	1. Split half	1. Corrected correlation between two halves of the test
		2. Inter-item analysis for Intelligence test	2. KR_{20}
		3. Inter-item analysis for personality test	3. Cronbach's alpha
Observer differences	Different observers recording	Inter-rater or Inter-scorer reliability	Kappa Statistic

— 365 —

STANDARD ERROR OF MEASUREMENT

Standard Error of Measurement

[Error that exists in an individual's test score]

$$SEM = \sigma\sqrt{1 - r}$$

Standard Deviation

Reliability

Examples: $\sigma = 10; r = .90 \longrightarrow SEM = 3.16$

$\sigma = 10; r = .60 \longrightarrow SEM = 6.32$

Image Source: http://slideplayer.com/slide/3587762/

— 366 —

The standard error of measurement (SEm) estimates how repeated measures of a person on the same instrument tend to be distributed around his or her "true" score. The true score is always unknown because no measure can be constructed that provides a perfect reflection of the true score(Harvill, 1991). SEm is directly related to the reliability of a test; that is, the larger the SEm, the lower the reliability of the test and the less precision there is in the measures taken and scores obtained. Since all

measurements contain some error, it is high- ly unlikely that any test will yield the same scores for a given person each time they are retested.

SEm's Confidence Interval(Harvill, 1991)

Statements about an examinee's obtained score (the actual score that is received on a test) are couched in terms of a confidence interval -- a band, interval, or range of scores that has a high probability of including the examin- ee's "true" score. Depending on the level of confidence one may want to have about where the "true" score may lie, the confidence band may be small or large. Most typical confidence intervals are 68%, 90%, or 95%. Respectively, these bands may be interpreted as the range within which a person's "true" score can be found 68%, 90%, or 95% of the time. It is not possible to construct a confidence interval within which an examinee's true score is absolutely certain to lie. It is important to report the confidence interval associated with a child's obtained score so that the reader can be informed of the probability that the examinee's true score lies within a given range of scores.

Selection of the confidence interval will depend on the level of certain- ty that one wishes to have about where the examinee's true score may lie given their obtained score. The 68% confidence level is the one most typically reported in psychoeducational evaluation reports. This is often reported in the following manner; "Given the student's obtained score of there are two out of three chances that the individual's true score would fall between_ (low score in range) and (high score in range)." Most test authors, however, suggest that a higher level of confidence statement be made. These would be stated as "nine out of ten chances" for the 90percent confidence band or "95 out of 100 chances" for the 95 percent confidence band. Obviously the increased levels of confidence would expand the range of scores included in the probabil- ity statements. Test publisher recommendations should be followed in reporting confidence levels.

Standard Error of Estimation (Harvill, 1991)

Some tests, such as the Wechsler Intelligence Scale for Children, 4th Edition (WISC-IV), provide the user with the standard error of estimation (SEest) , another form of standard error of measurement. This statistic takes into account regression toward the mean and the fact that scores at the extreme end (very high or very low scores) of the distribution are more prone to error than scores near the average. Because of this fact, the standard error of estimation is not equivalent around the obtained score. This would be reflected in a statement such as, "Given an obtained IQ score of 125, there are two out of three chances that this student's true score lies between 122 and 127 (-3 and +2 of the obtained score)." The WISC-IV manual provides standard error of estimation for the 90 and 95 percent levels of confidence.

Applications of the SEm and SEest (Harvill, 1991)

How should the SEm be used in program eligibility determination?

The SEM is a characteristic of the test that reflects the probability that an examinee's true score falls within a given range of scores. No score within the range of scores (except the obtained score) has a higher probability of occurring than any other score within that range. Using the 68% confidence level, for example, if a child receives an intelligence test score of 115 with a SEm of three (3) points, there is a 68% probability that the child's true score falls within the range of 112 to 118. It would not be appropriate to select the highest or lowest numbers within that range as the best estimate of the child's true score. In fact, the best estimate of any child's true score on a given test is the obtained score given, appropriate test administration procedures are followed, there is good effort and motivation on the part of the exam- inee, and there are no conditions within the testing situation that would unduly influence test scores. In the sample case just cited, that would be 115. The SEm should be treated as

information one has about a test to be considered by the examiner and/or eligibility committee in determin- ing the presence or not of a disability or giftedness.

How should the SEm be used in program eligibility determination?

The SEM is a characteristic of the test that reflects the probability that an examinee's true score falls within a given range of scores. No score within the range of scores (except the obtained score) has a higher probability of occurring than any other score within that range. Using the 68% confidence level, for example, if a child receives an intelligence test score of 115 with a SEm of three (3) points, there is a 68% probability that the child's true score falls within the range of 112 to 118. It would not be appropriate to select the highest or lowest numbers within that range as the best estimate of the child's true score. In fact, the best estimate of any child's true score on a given test is the obtained score given, appropriate test administration procedures are followed, there is good effort and motivation on the part of the exam- inee, and there are no conditions within the testing situation that would unduly influence test scores. In the sample case just cited, that would be 115. The SEm should be treated as information one has about a test to be considered by the examiner and/or eligibility committee in determin- ing the presence or not of a disability or giftedness.

Are there specific actions that should be taken by the evaluator in deciding how to obtain a measure of or to estimate the child's ability if the obtained score is judged to not be the best estimate?

The examiner should have sufficient basis for dismissing the obtained score as the best estimate of a child's true score on a given test. The standard error of measurement should not be used to unilat- erally extend or restrict the definition of giftedness or a disability such as mental retardation. That is, scores that fall within a given confidence interval of two standard deviations above the mean or two standard deviations below the mean are not "automatically" interpreted as "gift- ed" or "mentally handicapped" respectively. The eligibility committee should guard against promoting such actions as, there is just as much of a chance of inaccurately placing a student, as there is in inaccurately not placing a student in a program. As stated earlier, all things being equal, the best estimate of the true score for a given individual is always the obtained score. Naturally, additional testing is indicated if very questionable results are achieved.

How should test scores be provided in interpretative reports?

In order to accurately reflect the measurement error inherent in each test, it is appropriate to reflect in test interpretation reports the confidence that one has regarding the range within which an examin- ee's true score lies given their obtained score. Although the obtained score is the best estimate of the true score, one can be more confident in stating that the true score lies within a given range of the obtained score. Recommended confidence intervals are generally provided in test manuals.

Must the standard error of measurement be reported for all test scores?

No. This is but one of many statistical features of a test and does not have to be reported by the examiner in an interpretive report. Because of the greater weight that is placed on scores obtained on tests of intelligence, it is strongly recommended that appropriate confidence levels be provided to the reader of the psychological report. For other types of tests, the examiner should be prepared to provide this information, particularly when it is critical to decisions that must be made about the examinee.

\

HOW TO IMPROVE RELIABILITY

1. Quality of test items (Concise statements, homogeneous words (some sort of uniformity)
2. Adequate sampling of content domains (Comprehensiveness of items)
3. Longer assessment (More test items)
4. Developing a scoring plan (Especially for subjective tests = Rubrics)
5. Ensure Validity

VALIDITY

Test validity is an indicator of how much meaning can be placed upon a set of test results. In psychological and educational testing, where the im- portance and accuracy of tests is paramount, test validity is crucial (Weiner & Braun, 1988).

TYPES OF VALIDITY (Shuttleworth, 2009)

1. Face Validity -content of the test reflects the materials it is supposed to measure, according to the test takers.Face validity is a measure of how representative a research project is 'at face value,' and whether it appears to be a good project.
Example: The RPm Board Exam should reflect the materials provided in the table of specification.

2. Content Validity - sometimes called logical or rational validity, is the estimate of how much a measure represents every single element of a construct.

Content validity is often seen as a prerequisite to criterion validity, because it is a good indicator of whether the desired trait is measured. If elements of the test are irrelevant to the main construct, then they are measuring something else completely, creating potential bias. In addition, criterion validity derives quantitative correlations from test scores.

Content validity is qualitative in nature, and asks whether a specific element enhances or detracts from a test or research program.

Content validity is related to face validity, but differs widely in how it is evaluated.

Face validity requires a personal judgment, such as asking participants whether they thought that a test was well constructed and useful. Content validity arrives at the same answers, but uses an approach based in statistics, ensuring that it is regarded as a strong type of validity.

For surveys and tests, each question is given to a panel of expert analysts, and they rate it. They give their opinion about whether the question is essential, useful or irrelevant to measuring the construct under study.
Their results are statistically analyzed and the test modified to improve the rational validity.
Example: The RPm Board Exam should comprise a wide range of subjects.

3. Criterion-Related Validity. It assesses whether a test reflects a certain set of abilities.
To measure the criterion validity of a test, researchers must calibrate it against a known standard or against itself. Comparing the test with an established measure is known as concurrent validity; testing it over a period of time is known as predictive validity. It is not necessary to use both of these methods, and one is regarded as sufficient if the experimental design is strong. One of the simplest ways to assess criterion related validity is to compare it to a known standard.

A new intelligence test, for example, could be statistically analyzed against a standard IQ test; if there is a high correlation between the two data sets, then the criterion validity is high. This is a good example of concurrent validity, but this type of analysis can be much more subtle.

o Concurrent Validity -measures the test against a benchmark test and high correlation indicates that the test has strong crite- rion validity. The tests are for the same, or very closely related, constructs and allow a researcher

to validate new methods against a tried and tested stalwart.

Example: The scores on the Mechanical Aptitude Test correlated significantly with supervisory ratings of the worker's performance conducted at the same time

The Weaknesses of Concurrent Validity

Concurrent validity is regarded as a fairly weak type of validity and is rarely accepted on its own. The problem is that the benchmark test may have some inaccuracies and, if the new test shows a correlation, it merely shows that the new test contains the same problems.

For example, IQ tests are often criticized, because they are often used beyond the scope of the original intention and are not the strongest indicator of all round intelligence. Any new intelligence test that showed strong concurrent validity with IQ tests would, presumably, contain the same inherent weaknesses.

Despite this weakness, concurrent validity is a stal- wart of education and employment testing, where it can be a good guide for new testing procedures. Ideally, researchers initially test concurrent validity and then follow up with a predictive validity based experiment, to give a strong foundation to their findings.

Predictive Validity - is a measure of how well a test predicts abilities. It involves testing a group of subjects for a certain construct and then comparing them with results obtained at some point in the future.

It is an important sub-type of criterion validity, and is regarded as a stalwart of behavioral science, education and psychology,

Most educational and employment tests are used to predict future performance, so predictive validity is regarded as essential in these fields.

Example: High score in RPM Board Exam should predict high performance in psychometric work

Weaknesses of Predictive Validity

Predictive validity is regarded as a very strong mea- sure of statistical validity, but it does contain a few weakness- es that statisticians and researchers need to take into consider- ation. Predictive validity does not test all of the available data, and individuals who are not selected cannot, by definition, go on to produce a score on that particular criterion.

In the university selection example, this approach does not test the students who failed to attend university, due to low grades, personal preference or financial concerns. This leaves a hole in the data, and the predictive validity relies upon this incomplete data set, so the researchers must always make some assumptions. If the students with the highest grade point averages score higher after their first year at university, and the students who just scraped in get the lowest, researchers assume that non-attendees would score lower still. This down- wards extrapolation might be incorrect, but predictive validity has to incorporate such assumptions.

Despite this weakness, predictive validity is still

regarded as an extremely powerful measure of statistical accuracy. In many fields of research, it is regarded as the most important measure of quality, and researchers constantly seek ways to maintain high predictive validity.

4. Construct Validity - It defines how well a test or experiment measures up to its claims. It refers to whether the operational definition of a vari- able actually reflects the true theoretical meaning of a concept.

The simple way of thinking about it is as a test of general- ization, like external validity, but it assesses whether the variable that you are testing for is addressed by the experiment.Construct validity is a device used almost exclusively in social sciences, psychology and education.

Example: A score in an IQ test should reflect one's intelligence

Threats to Construct Validity

Hypothesis Guessing

This threat is when the subject guesses the intent of the test and consciously, or subconsciously, alters their behavior.

For example, many psychology departments expect students to volunteer as research subjects for course credits. The danger is that the students may realize what the aims of the research are, potentially evaluating the result. It does not matter whether they guess the hypoth- esis correctly, only that their behavior changes.

Evaluation Apprehension

This particular threat is based upon the tendency of humans to act differently when under pressure. Individual testing is notorious for bringing on an adrenaline rush, and this can improve or hinder performance. In this respect, evaluation apprehension is related to ecological external validity, where it affects the process of generalization.

Researcher Expectancies and Bias

Researchers are only human and may give cues that influence the behavior of the subject. Humans give cues through body language, and subconsciously smiling when the subject gives a correct answer, or frowning at an undesirable response, all have an effect. This effect can lower construct validity by clouding the effect of the actual research variable. To reduce this effect, interaction should be kept to a minimum, and assistants should be unaware of the overall aims of the project.

Poor Construct Definition

Construct validity is all about semantics and labeling. Defining a construct in too broad or too narrow terms can invalidate the entire experiment.

For example, a researcher might try to use job satisfaction to define overall happiness. This is too narrow, as somebody may love their job but have an unhappy life outside the workplace. Equally, using general happiness to measure happiness at work is too broad. Many people enjoy life but still hate their work!

Mislabeling is another common definition error: stating that you intend to measure depression, when you actually measure anxiety, compromises the research. The best way to avoid this particular threat is with good planning and seeking advice before you start your research program.

Construct Confounding

This threat to construct validity occurs when other constructs mask the effects of the measured construct.

For example, self-esteem is affected by self-confidence and self-worth. The effect of these constructs needs to be incorporated into the research.

Interaction of Different Treatments

This particular threat is where more than one treatment influ- ences the final outcome.

For example, a researcher tests an intensive counseling pro- gram as a way of helping smokers give up cigarettes. At the end of the study, the results show that 64% of the subjects successfully gave up. Sadly, the researcher then finds that some of the subjects also used nicotine patches and gum, or electronic cigarettes. The construct validity is now too low for the results to have any meaning. Only good planning and monitoring of the subjects can prevent this.

Unreliable Scores

Variance in scores is a very easy trap to fall into.

For example, an educational researcher devises an intelligence test that provides excellent results in the UK, and shows high construct validity. However, when the test is used upon immigrant children, with English as a second language, the scores are lower. The test measures their language ability rather than intelligence.

Mono-Operation Bias

This threat involves the independent variable, and is a situa- tion where a single manipulation is used to influence a construct.

For example, a researcher may want to find out whether an anti-depression drug works. They divide patients into two groups, one given the drug and a control given a placebo. The problem with this is that it is limited (e.g. random sampling error), and a solid design would use multi-groups given different doses. The other option is to conduct a pre-study that calculates the optimum dose, an equally acceptable way to preserve construct validity.

Mono-Method Bias

This threat to construct validity involves the dependent vari- able, and occurs when only a single method of measurement is used. Using a variety of methods, such as questionnaires, self-rating, physiological tests, and observation minimizes the chances of this particular threat affecting construct validity.

For example, in an experiment to measure self-esteem, the researcher uses a single method to determine the level of that construct, but then discovers that it actually measures self-confidence.

Convergent Validity - measures of constructs that theoretically should be related to each other are, in fact, observed to be related to each other

o Discriminant Validity - measures of constructs that theoretically should not be related to each other are, in fact, observed to not be related to each other

If a research program is shown to possess both of these types of validity, it can also be regarded as having excel- lent construct validity.

In many areas of research, mainly the social sciences, psychology, education and medicine, researchers need to analyze non-quantitative and abstract concepts, such as level of pain, anxiety or educational achievement. A researcher needs to define exactly what trait they are measuring if they are to maintain good construct validity.

Constructs very rarely exist independently, because the human brain is not a simple machine and is made up of an interlinked web of emotions, reasoning and senses. Any research program must untangle these complex interactions and establish that you are only testing the desired construct.

This is practically impossible to prove beyond doubt, so researchers gather enough evidence to defend their findings from criticism.

The basic difference between convergent and discriminant validity is that convergent validity tests whether constructs that should be related, are related. Discriminant validity tests whether believed unrelated constructs are, in fact, unrelated.

Example: Imagine that a researcher wants to measure self-esteem, but she also knows that the other four constructs are related to self-esteem and have some overlap. The ultimate goal is to make an attempt to isolate self-esteem.

In this example, convergent validity would test that the four other constructs are, in fact, related to self-esteem in the study. The researcher would also check that self-worth and confidence, and social skills and self-appraisal, are also related.

Discriminant validity would ensure that, in the study, the non-overlapping factors do not overlap. For example, self-esteem and intelligence should not relate (too much) in most research projects.

As you can see, separating and isolating constructs is difficult, and it is one of the factors that make social science extremely difficult.

Social science rarely produces research that gives a yes or no answer, and the process of gathering knowledge is slow and steady, building on top of what is already known.

FACTORS THAT CAN LOWER VALIDITY

1. Unclear Directions
2. Difficult reading vocabulary
3. Ambiguity in statements

4. Inadequate time limits
5. Inappropriate level of difficulty
6. Poorly constructed test items

TEST DEVELOPMENT

Characteristics of a Good Test (Doverspike, 2017)

1. Reliable

Reliability refers to the accuracy of the obtained test score or to how close the obtained scores for individuals are to what would be their "true" score, if we could ever know their true score. Thus, reli- ability is the lack of measurement error, the less measurement error the better. The reliability coefficient, similar to a correlation coefficient, is used as the indicator of the reliability of a test. The reliability coefficient can range from 0 to 1, and the closer to 1 the better. Generally, experts tend to look for a reliability coefficient in excess of .70. However, many tests used in public safety screening are what is referred to as multi-dimensional. Interpreting the meaning of a reliability coefficient for a knowledge test based on a variety of sources requires a great deal of experience and even experts are often fooled or offer incorrect interpretations. There are a number of types of reliability, but the type usually reported is internal consistency or coefficient alpha. All things being equal, one should look for an assessment with strong evidence of reliability, where information is offered on the degree of confidence you can have in the reported test score.

2. Valid

Validity will be the topic of our third primer in the series. In the selection context, the term "validity" refers to whether there is an expectation that scores on the test have a demonstrable relationship to job performance, or other important job-related criteria. Validity may also be used interchangeably with related terms such as "job related" or "business necessity." For now, we will state that there are a number of ways of evaluating validity including:

Content
Criterion-related
Construct
Transfer or transportability
Validity generalization

A good test will offer extensive documentation of the validity of the test.

3. Practical

A good test should be practical. What defines or constitutes a practical test? Well, this would be a balancing of a number of factors including: Length - a shorter test is generally preferred

- *Time* - a test that takes less time is generally preferred
- Low cost speaks for itself
- Easy to administer
- Easy to score
- *Differentiates between candidates* - a test is of little value if all the applicants obtain the same score
- *Adequate test manual* - provides a test manual offering ade- quate information and documentation
- *Professionalism* - is produced by test developers possessing high levels of expertise

The issue of the practicality of a test is a subjective judgment, which will be impacted by the constraints facing the public-sector jurisdiction. A test that may be practical for a large city with 10,000 ap- plicants and a large budget, may not be practical for a small town with 10 applicants and a miniscule testing budget.

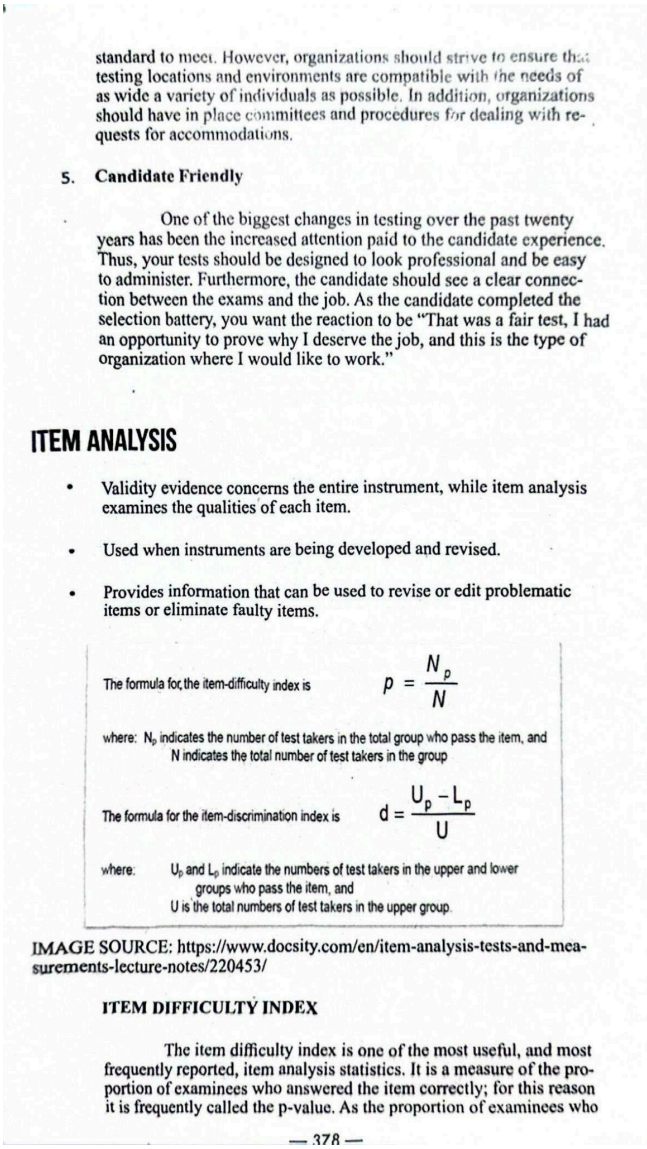
4. Socially Sensitive

A consideration of the social implications and effects of the use of a test is critical in public sector, especially for high stakes jobs such as public safety occupations. The public safety assessment professional must be considerate of and responsive to multiple group of stakeholders. In addition, in evaluating a test, it is critical that attention be given to:

Avoiding adverse impact - Recent events have highlighted the importance of balance in the demographics of safety force

personnel. Adverse impact refers to differences in the passing rates on exams between males and females, or minorities and majority group members. Tests should be designed with an eye toward the minimization of adverse im- pact. A complicated topic, I addressed adverse impact in greater depth in previous blog posts here and here.

Universal Testing - The concept behind universal testing is that your exams should be able to be taken by the most diverse set of applicants possible, including those with disabilities and by those who speak other languages. Having a truly universal test is a difficult, if not impossible,



standard to meet. However, organizations should strive to ensure that testing locations and environments are compatible with the needs of as wide a variety of individuals as possible. In addition, organizations should have in place committees and procedures for dealing with requests for accommodations.

5. **Candidate Friendly**

One of the biggest changes in testing over the past twenty years has been the increased attention paid to the candidate experience. Thus, your tests should be designed to look professional and be easy to administer. Furthermore, the candidate should see a clear connection between the exams and the job. As the candidate completed the selection battery, you want the reaction to be "That was a fair test, I had an opportunity to prove why I deserve the job, and this is the type of organization where I would like to work."

ITEM ANALYSIS

- Validity evidence concerns the entire instrument, while item analysis examines the qualities of each item.
- Used when instruments are being developed and revised.
- Provides information that can be used to revise or edit problematic items or eliminate faulty items.

The formula for the item-difficulty index is $p = \frac{N_p}{N}$

where: N_p indicates the number of test takers in the total group who pass the item, and
N indicates the total number of test takers in the group

The formula for the item-discrimination index is $d = \frac{U_p - L_p}{U}$

where: U_p and L_p indicate the numbers of test takers in the upper and lower groups who pass the item, and
U is the total numbers of test takers in the upper group.

IMAGE SOURCE: https://www.docsity.com/en/item-analysis-tests-and-measurements-lecture-notes/220453/

ITEM DIFFICULTY INDEX

The item difficulty index is one of the most useful, and most frequently reported, item analysis statistics. It is a measure of the proportion of examinees who answered the item correctly; for this reason it is frequently called the p-value. As the proportion of examinees who

— 378 —

got the item right, the p-value might more properly be called the item easiness index, rather than the item difficulty. It can range between 0.0 and 1.0, with a higher value indicating that a greater proportion of examinees responded to the item correctly, and it was thus an easier item. For criterion-referenced tests (CRTs), with their emphasis on mas- tery-testing, many items on an exam form will have p-values of .9 or above. Norm-referenced tests (NRTS), on the other hand, are designed to be harder overall and to spread out the examinees' scores. Thus, many of the items on an NRT will have difficulty indexes between 4 and .6.

ITEM DISCRIMINATION INDEX

The degree to which an item differentiates correctly among the examinees on the behavior domain (i.e. the high scorers and the low scorers).

METHODS OF ITEM DISCRIMINATION

- ***Extreme Group Method***
  - Examinees are divided into two groups based on high and low scores.
  - Item discrimination index is then calculated by subtract- ing the proportion of examinees in the lower group from the proportion of examinees in the upper group who got the item correct or who endorsed the item in the expected manner.
  - Item discrimination indices can range from + 1.00 (all of the upper group got it right and none of the lower group got it right) to 1.00 (none of the upper group got it right and all of the lower group got it right)
  - The determination of the upper and lower group will depend on the distribution of scores. If there is a normal distribution optimal dispersion is accomplished by using the upper 27% for the upper group and lower 27% for the lower group (Kelly, 1939). For small groups Anastasi and Urbina (1997)

suggest the range of upper and lower 25% to 33%.

- The interpretation of an item discrimination index de- pends on the instrument, the purpose it was used for, and the group taking the test.
- In general, negative item discrimination indices, particularly small positive indices are indicators that the item needs to be eliminated or revised.

Correlational Method

- This approach reports the relationship between the performance on the item and a criterion.
- Often the criterion is performance on the overall test.
- Commonly used methods are the point biserial and phi coefficient.
- The result is a correlation coefficient that ranges between - 1.00 and +1.00, with the positive and larger coefficients reflecting items that are better discriminators.

ITEM RESPONSE THEORY (IRT) OR LATENT TRAIT THEORY(Em- bretson & Reise, 2000)

- Its basis is different from classical test theory, where the focus is on arriving at one true score and attempting to control error in measurement.
- In IRT, the focus is on each item and establishing those items that actu- ally measure one's ability or the respondent's level of latent trait.
- It rests on the assumption that the performance of an examinee on a test item can be predicted by a set of factors called traits, latent traits or abilities.
- Using IRT, we get an indication of an individual's performance based not on the total score, but on the precise items the person answers correctly.
- It suggests that the relationship between examinees' item performance and the underlying trait being measured can be described by a mono- tonically increasing function called an item characteristic function or an item characteristic curve.

STANDARDS FOR QUALIFICATIONS OF TEST USERS (ACA, 2003)

The professional qualifications essential to the use of tests in counsel- ing arise from a synthesis of knowledge, skills, and ethics. While some professional groups are seeking to control and restrict the use of psychological tests, the American Counseling Association believes firmly that one's right to use tests in counseling practice is directly related to competence. This competence is achieved through education, training, and experience in the field of testing. Thus, professional counselors with a master's degree or higher and appropriate coursework in appraisal/assessment, supervision, and experience are qualified to use objective tests. With additional training and experience, professional counselors are also able to administer projective tests, individual intelligence tests, and clinical diagnostic tests. This training may occur in graduate school, in post-graduate professional development instruction, or in supervised training in use of the test. Professional counselors are qualified to use tests and assessments in counseling practice to the degree that they possess the appropriate knowledge and skills, including the following areas:

1. Skill in practice and knowledge of theory relevant to the testing context and type of counseling specialty.

Assessment and testing must be integrated into the context of the theory and knowledge of a specialty area, not as a separate act, role, or entity. In addition, professional counselors should be skilled in treatment practice with the population being served.

2. A thorough understanding of testing theory, techniques of test construction, and test reliability and validity.

Included in this knowledge base are methods of item selec- tion, theories of human nature that underlie a given test, reliability, and validity. Knowledge of reliability includes, at a minimum: methods by which it is determined, such as domain sampling, test-retest, parallel forms, split-half, and inter-item consistency, the strengths and limita- tions of each of these methods; the standard error of measurement, which indicates how accurately a person's test score reflects their true score of the trait being measured; and true score

3. A working knowledge of sampling techniques, norms, and descriptive, correlational and predictive statistics.

Important topics in sampling include sample size, sampling techniques, and the relationship between sampling and test accuracy. A working knowledge of descriptive statistics includes, at a mini- mum: probability theory, measures of central tendency; multi-modal and skewed distributions, measures of variability, including variance and standard deviation; and standard scores, including deviation IQ's, z-scores, T-scores, percentile ranks, stanines/stens, normal curve equivalents, grade- and age-equivalents. Knowledge of correlation and prediction includes, at a minimum: the principle of least squares; the direction and magnitude of relationship between two sets of scores; deriving a regression equation; the relationship between regression and correlation; and the most common procedures and formulas used to calculate correlations.

4. Ability to review, select, and administer tests appropriate for clients or students and the context of the counseling practice.

Professional counselors using tests should be able to describe the purpose and use of different types of tests, including the most wide- ly used tests for their setting and purposes. Professional counselors use their understanding of sampling, norms, test construction, validity and reliability to accurately assess the strengths, limitations, and appropriate applications of a test for the clients being served. Professional counselors using tests also should be aware of the potential for error when relying on computer printouts of test interpretation. For accuracy of interpretation, technological resources must be augmented by a coun- selor's firsthand knowledge of the client and the test-taking context.

5. Skill in administration of tests and interpretation of test scores.

Competent test users implement appropriate and standardized administration procedures. This requirement enables professional counselors to provide consultation and training to others who assist with test administration and scoring. In addition to standardized procedures, test users provide testing environments that are comfortable and free of distraction. Skilled interpretation requires a strong working knowledge of the theory underlying the test, test's purpose, statistical meaning of test scores, and norms used in test construction. Skilled interpretation also requires an understanding of the similarities and differences between the client or student and the norm samples used in test construction. Finally, it is essential that clear and accurate communica- tion of test score meaning in oral or written form to clients, students or appropriate others be provided.

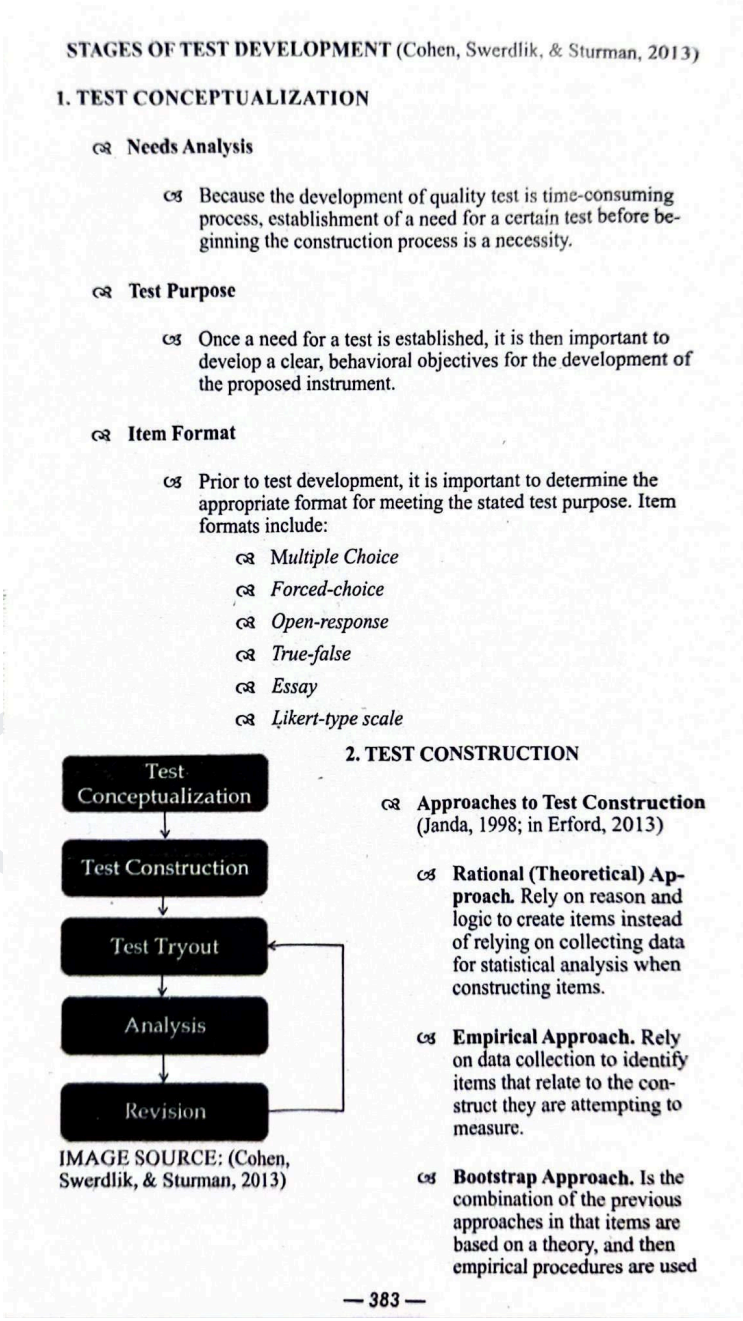
6. Knowledge of the impact of diversity on testing accuracy, including age, gender, ethnicity, race, disability, and linguistic differences.

Professional counselors using tests should be committed to fairness in every aspect of testing. Information gained and decisions made about the client or student are valid only to the degree that the test accurately and fairly assesses the client's or student's characteris- tics.

Test selection and interpretation are done with an awareness of the degree to which items may be culturally biased or the norming sample not reflective or inclusive of the client's or student's diversity. Test us- ers understand that age and physical disability differences may impact the client's ability to perceive and respond to test items. Test scores are interpreted in light of the cultural, ethnic, disability, or linguistic factors that may impact an individual's score. These include visual, auditory, and mobility disabilities that may require appropriate accommodation in test administration and scoring. Test users understand that certain types of norms and test score interpretation may be inappropriate, de- pending on the nature and purpose of the testing.

7. Knowledge and skill in the professionally responsible use of assessment and evaluation practice.

Professional counselors who use tests act in accordance with ACA's Code of Ethics and Standards of Practice (1997), Responsibil- ities of Users of Standardized Tests (RUST) (AAC, 2003), Code of Fair Testing Practices in Education (JCTP, 2002), Rights and Respon- sibilities of Test Takers: Guidelines and Expectations (JCTP, 2000), and Standards for Educational and Psychological Testing (AERA/ APA/NCME, 1999). In addition, professional school counselors act in accordance with the American School Counselor Association's (AS- CA's) Ethical Standards for School Counselors (ASCA, 1992). Test users should understand the legal and ethical principles and practices regarding test security, using copyrighted materials, and unsupervised use of assessment instruments that are not intended for self-administra- tion. When using and supervising the use of tests, qualified test users demonstrate an acute understanding of the paramount importance of the well-being of clients and the confidentiality of test scores. Test users seek on-going educational and training opportunities to maintain competence and acquire new skills in assessment and evaluation.



STAGES OF TEST DEVELOPMENT (Cohen, Swerdlik, & Sturman, 2013)

1. TEST CONCEPTUALIZATION

- **Needs Analysis**
  - Because the development of quality test is time-consuming process, establishment of a need for a certain test before be- ginning the construction process is a necessity.
- **Test Purpose**
  - Once a need for a test is established, it is then important to develop a clear, behavioral objectives for the development of the proposed instrument.
- **Item Format**
  - Prior to test development, it is important to determine the appropriate format for meeting the stated test purpose. Item formats include:
    - *Multiple Choice*
    - *Forced-choice*
    - *Open-response*
    - *True-false*
    - *Essay*
    - *Likert-type scale*

IMAGE SOURCE: (Cohen, Swerdlik, & Sturman, 2013)

2. TEST CONSTRUCTION

- **Approaches to Test Construction** (Janda, 1998; in Erford, 2013)
  - **Rational (Theoretical) Approach.** Rely on reason and logic to create items instead of relying on collecting data for statistical analysis when constructing items.
  - **Empirical Approach.** Rely on data collection to identify items that relate to the construct they are attempting to measure.
  - **Bootstrap Approach.** Is the combination of the previous approaches in that items are based on a theory, and then empirical procedures are used

— 383 —

to verify that the items actually measure or are highly related to the construct to measure.

Scaling. The process of setting rules for assigning numbers in measure- ment.

- What scale of measurement *(nominal, ordinal, interval) t*o use?

- Will it be a function of age *(age-based scale)* or a function of grade (grade-based)?
- Would the scores measure a single factor *(unidimensional)* or multiple ones *(multidimensional)*?
- Are the scores compared to all other stimuli in the scale (comparative) or the scores differed quantitatively of each other *(categorical)*?

Writing Items.

What range of content should the items cover?

Item Pool/Bank - is the reservoir or well from which items will or will not be drawn for the final version of the test.

Which of the many different item formats should be em- ployed?

Item Format - variables such as form, plan, struc-ture, arrangement, and layout of individual test items.

Selected-Response Format - require test-takers to select a response from a set of alternative responses

Examples of Selected - Response Format
Multiple-choice format
Matching
True-false

Constructed - Response Format - require test-takers to supply or create the correct answer
Examples of Constructed - Response Format
Completion Item - requires the examinee to provide a word or phrase

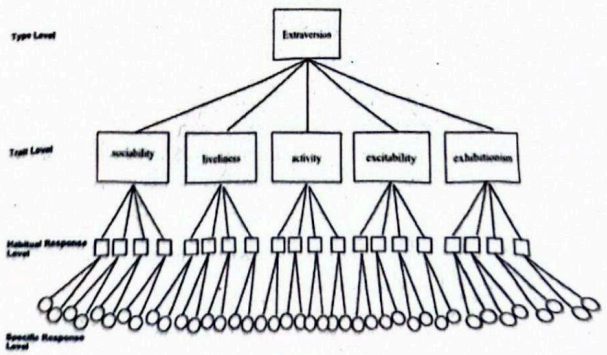
that completes a sentence.

- *Short-Answer Item* – requires the examinee to respond to a question succinctly (with a "short" answer, that is not beyond a paragraph or two)
- *Essay* – requires the examinee to respond to a question that demonstrates recall of facts, understanding, analysis and/or interpretation.

- **Scoring Items**
  - **Cumulative** – this model assumes that the number of items endorsed or responded to match the key which represent the degree of the construct of trait the test measured; the higher the score, the greater the degree of the construct present.
  - **Class** – use to categorize individuals for the purpose of description or prediction
  - **Ipsative** – indicates how an individual has performed on a set of variables or scales

3. TEST TRYOUT

- **Pilot Testing.** Subjecting the test to pre-testing for factor and item analysis.
  - Rule of Thumb in Selecting Participants for Pilot Testing
    A x 5 to 10 = n
    where A is the items of the questionnaire
    n is the total number of sample participants
  - Rule of Thumb in Selecting Participants for Validation:
    - 20 participants per Item



4. ANALYSIS

- **Index of Item's Difficulty** -refers to the 'difficulty level' of each item in the given test.
- **Index of Item's Reliability** - provides an indication of the internal consistency

of the test (factor analysis and inter-item consistency).

- **Index of Item's Validity** - is a statistic designed to provide an indi- cation of the degree to which a test is measuring what it purports to measure.
- **Index of Item Discrimination** - indicate how adequately an item separates or discriminates between high scores and low scores on an entire test
- **Qualitative Item Analysis** - various non-statistical procedures de- signed to explore how individual test items work. Some topics than can be considered are cultural sensibility, 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 and during the test, tests taker's overall impressions, test taker's preparation, and test taker's preferences.

5. REVISION

- **Test revision as a stage of new test development.** This is a judicious act on making the information pertinent to the test and molds it into its final form and further improvement.
- **Test revision in the life cycle of an existing test.** An existing test must keep its present from as long as it remains "useful" but it should be revised when significant changes in the domain represented, or new condition of test use and interpretation, make the test inappropriate for its intended use
- **Cross-validation** - refers to the revalidation of a test on a sample of test takers other than those on whom test performance was originally found to be a valid prediction of some criterion.
- **Co-validation** - defined as a test validation process conducted on two or more tests using the same sample of test takers. When used in con- junction with the criterion of norms or the revision of existing norms, the process is referred to as ***co-norming.***

SCALE CONSTRUCTION: PRINCIPLES AND PROCEDURES (del Pilar, 2015)

OUTLINE OF THE SCALE CONSTRUCTION PROCESS

1. Definition of the target construct

• Make a "Working definition" of the construct being constructed a test for

2. Item writing to generate the item pool

- Best done by a team of writers
- Can be supplemented by interviews with people who appear to be high and low on the target construct
- Item writing model based on Eysenck's (1947) behaviorist model
- Using the "Affect-Behavior-Cognition" as an Item writing model. Include self-observation and other people's.

3. Item review: qualitative item analysis

• *Item writing* - imaginative

• *Item review* - critical

• *Characteristics of good items:*

- simple in form, affirmative in direction
- free of excessive levels of social desirability
- likely to result in large variance in responses
- suitable for all target respondents
- phrased in a manner that produces immediate identifica- tion or rejection
- as a set, keyed in both directions

• *Further considerations during item review:*

- Rule of thumb on minimum number of items for item testing: twice the target length of the scale
- If the number of acceptable items, in terms of item char- acteristics, exceeds the maximum length that can be test- ed, items should be reviewed for

prototypicality. Those most prototypical are selected for item testing.

4. Item testing

- Items should be tested on a sample similar to those on which the scale will be used.
- For many scales for use by the general population, heterogeneous student samples are often appropriate.
- Rule of thumb for minimum sample size for item testing N=100

5. Item selection: quantitative item analysis

• Family of methods that aim to select items that will contribute to the reliability and validity of the scale

• Reliability analysis:

- Repeatability; the extent to which a scale is free of measurement error
- Sources of unreliability: environmental factors, person factors, scale factors
- Observed Score (X) = True Score (T) + Error (E)
- The smaller the average size of the error scores across individuals, the more reliable the scale.
- The smaller the average size of the error scores across different items in a scale, the more reliable the scale.

• How do we pick out the good items?

- Item-total correlation:
  - Good items are those with high itc.
  - Poor items are those with low itc.

• Procedure for reliability analysis (for high internal consistency)

- Compute the reliability for all trial items.
- Compute the correlation of each of the trial items with the total on the trial items.
- Identify the item with the lowest item-total correlation.
- Remove the item from the list, recompute total scores.
- Recompute the reliability for the remaining items.
- Repeat Steps 2 to 5 until the reliability begins to go down.

• Caveat:

- The preceding method is appropriate for scales that mea- sure only a single construct (unifactorial).
- When used for a construct that has facets or subcomponents, it could result in selecting only the largest component.

6. Estimating the reliability of the scale

• The correlation of trial items with the total may be due to chance.

• Consequently, the reliability of the sample on which the item analysis was done will often be higher than the reliability to be obtained from other samples.

• Thus, the estimate of the reliability of the scale should be based on a sample other than the item-selection sample.

7. Validation studies (optional)

• Validation by its nature is a lengthy process.

• For a scale that is proposed by its author as a research tool for general use, the author may conduct a few validation studies.

• Initial validation studies often take the form of demonstrating convergent validity with other scales measuring the same or similar constructs, and mean differences between a group expected to be high or low on the construct, and a group from the general popula- tion.

8. Norming, if for applied purposes (optional)

• If the scale is to be used for applied purposes, norms are developed so that scores may be interpreted.

• A meaningful norm group should be identified and sampled so as to obtain a representative sample of the group.

• Percentile ranks indicate the approximate position of a score in the norm group. When interpreting scores, the normative sample should be clearly specified in terms of demographic

characteristics, as well as possible peculiarities it might have, eg,type and location of participating schools, or organizations.

Other Important Factors necessary for Test Construction : Orense & Reyes (2014) identify the following

Coefficient of Determination - this value tells us the proportion of the total cation in scores on Y that we know as a function of information about X.

Multiple Regression - the goal is to find the linear combination of the 2 or more variables (X, Z, K...) that provides the best prediction of Y

Factor Analysis - used to study the interrelationships among a set of variables without reference to a criterion. It is also a form of data-reduction technique.

Factors that Contribute to Test Scores

1. Innate factors
2. Background and Environment
3. Personality
4. Situation
5. Test Demands
6. Random Variations

NORMING

Norm-Referenced Tests (or NRTs)

Compare an examinee's performance to that of other examinees. It is any standardized test or evaluative instrument for which the resulting scores are interpreted or acquire additional meaning in terms of com- parisons made to a specified group (Le... reference group) for which the individual or group belongs (e.g., age or grade). Tests can be considered both norm-referenced and criterion-referenced. It depends only on the use and interpretation of the scores (Cohen, et. al, 2013).

Norm-Referenced Interpretation

A score interpretation based on a comparison of a test taker's performance to the performance of other people in a specified reference pop- alation (e.g., age groups, grade groups, etc.). Norm-referenced interpretations can be for individuals (ie, student norms) or for institutions (eg, school norms) and could involve converting scores to scale scores (or standard scores), percentile ranks. stanines, grade equivalents, etc.. depending on use of the test and the information provided by the test publisher. Norm-referenced interpretations allow educators to get an "external" look at the performance of their students in relation to rest of the nation(Cohen, et al, 2013).