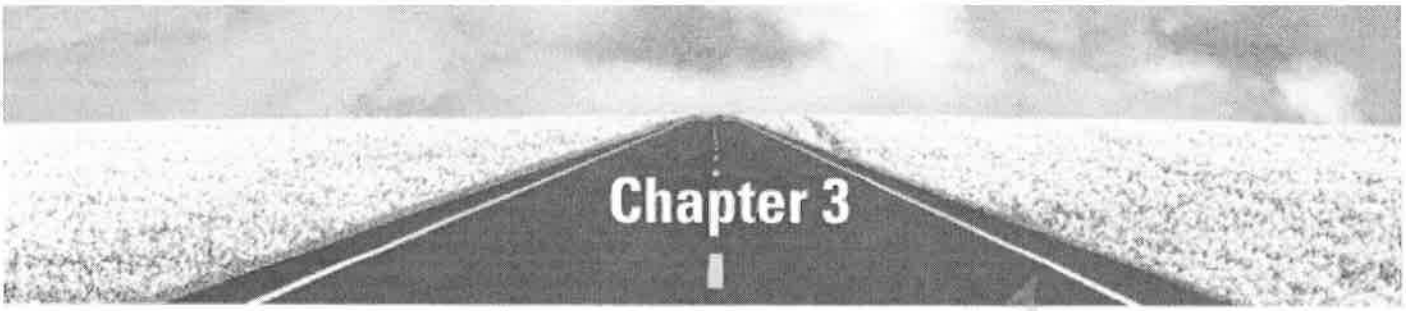

Chapter 3

Foundations of Standardized Assessment

Matthew E. Sprong

Noel A. Ysasi

Key Terms: psychometric properties, reliability, validity, descriptive statistics, substance abuse assessment

Key Objectives: Students should be better prepared to identify and gather needed client assessment data, and interpret assessment information and use it in substance abuse treatment planning.

Introduction and the Nature of Tests

Assessment is a crucial factor in the treatment of individuals who have drug and alcohol challenges. Often times, individuals requiring treatment have several barriers that may decrease the likelihood of having a perceived high quality of life. Such barriers may include loss of employment, difficulty maintaining healthy relationships, and legal challenges. Counseling professionals are tasked with identifying these barriers for several reasons, including the proper placement of the individual in the appropriate level of care (see Chapter 7), and appropriate counseling and behavioral interventions can be included (see Chapter 8) in the development of the treatment plan (see Chapter 10). Although these topics are of important interest for a counseling professional, the current chapter will focus on standardized assessment. In Chapter 2, diagnostic criteria of substance-related disorders were discussed. Often times, instruments are developed to assist in this diagnostic process. However, it is important to identify valid and reliable instruments to aid in this process. If invalid and unreliable testing occurs, the client and counseling professional may be unable to make informed decisions related to treatment.

Methods and Sources of Assessment Information

Assessment is a systematic procedure for collecting information that is used to make inferences and decisions about an individual and involves several data collection methods from multiple sources to yield relevant, accurate, and reliable information (Drummond & Jones, 2010). There are several methods to obtain information from the individual with the drug and/or alcohol barrier, such as an interview, direct observation and reviewing collateral data, and psychological testing. A comprehensive interview allows for the counseling professional to gather a great amount of information in a short period of time. Gathering information related to a person's hobbies, employment, and personal relationships are some examples of information that can be used during counseling treatment. Additionally, an interview will assist a counseling professional in understanding the individual's current situation.

Clinical observation is one method where counseling professionals can obtain valuable information related to the client's current level of functioning (e.g., interaction with others, mood patterns), observe behaviors from clients that are used in order to determine a diagnosis, and assist in the treatment plan development. Furthermore, clinical observations are the basis of therapy and treatment and are a means by which a counseling professional can learn about his or her client (Sommers-Flanagan & Sommers-Flanagan, 2017). A counseling professional is also able to obtain collateral information from people associated with the client, such as family members, friends, medical professionals, previous treatment records (if available), and significant others. The benefit of direct observation and reviewing collateral information is that a counseling professional can compare information from what he or she is observing to determine whether the individual is making progress and completing short- and long-term goals, or if the client is not making progress. The information briefly described above (i.e., interview, direct observation, reviewing collateral information) is essential in the treatment plan development. However, the purpose of the current chapter is a focus on testing, while discussing the process of developing a test. By the end of the chapter, counseling professionals or future counseling professionals should be able to review a test and evaluate the feasibility of using that test within clinical practice.

Psychological Testing

Psychological testing is a valuable approach toward assisting a counseling professional in (a) properly diagnosing the individual, (b) determining other psychosocial limitations, and (c) evaluating the appropriate level of care for the individual in need of drug/alcohol treatment. A test is defined as an objective and standardized measure of a sample of behavior (Trochim, Donnelly, & Arora, 2014) and provides a sample of behavior that is used to predict how a person will behave in other situations of interest. A test has two primary functions, including (1) diagnosis or an estimate of a person's current status (behavior) and (2) prognosis or a prediction of a person's future status (behavior). To properly diagnose and create a prognosis of a person's future status, there are several factors to consider when selecting an instrument to use. An important factor is whether the test is standardized or implies uniformity of procedure (rules) for administration, scoring, and interpretation; and makes it possible to compare scores of different people because they have been exposed to the same stimuli. In Chapter 2, several different criteria were provided in terms of diagnosing an individual with a substance-related disorder. Although a counseling professional may be able to read the criteria and determine beyond a reasonable degree of certainty whether the individual meets the diagnosis for a mental health disorder, some instruments provide the counseling professional the ability to determine whether the behavior falls above or below an average range. Therefore, when selecting standardized assessments, the counseling professional should evaluate the test manual (all credible tests/instruments should have one) and determine if the following eight categories are discussed (Drummond & Jones, 2010):

1. Uniform test materials
2. Time limits
3. Exact oral instructions
4. Preliminary demonstrations or items, examples of which are provided
5. How to handle questions that arise from participants
6. Details of test environments
7. Norms for comparing test performance
8. Guidelines for test performance interpretation

These aforementioned categories are essential to selecting a test that will provide accurate information that can be used in treatment planning. If these features are not provided, a counseling professional

Table 1 National Certification Addiction Counselor Levels

Level of Credential	Educational Requirements	Work Experience
Level 1 (NCAC I)	GED, high school diploma, or higher	Three years' work experience or 6,000 hours supervised experience
Level 2 (NCAC II)	Bachelor's degree or higher in Substance Use Disorder/Addiction and/or related counseling subject	At least 5 years full-time experience or 10,000 hours of supervised experience
Master Addiction Counselor	Master's degree or higher in Substance Use Disorder and/or related counseling subject	At least 3 years full-time or 6,000 hours of supervised experience as a Substance Use Disorder/Addiction counselor

Adapted from "The National Certification Commission for Addiction Professionals."

may administer the test in an inappropriate manner, and the results of the test may not be objective and may provide inaccurate results. For a test to be objective, it must first avoid any subjective bias of the examiner. Specifically, the individual providing the assessment should refrain from increasing or decreasing the scoring of the test based off personal and/or professional biases. For example, teachers within a public school system often discuss with school psychologists the behavioral issues among their students. If a teacher provides the licensed specialist in school psychology (LSSP) their informal diagnosis of the student (i.e., attention deficit disorder; ADD), the LSSP can inadvertently look for behavioral signs for this particular mental health disorder and thus diagnose them with ADD. Second, the creation of test items is determined by empirical data, and test items are only administered after they have been assessed for reliability and validity.

For example, the National Certification Commission for Addiction Professionals (NCCAP) offers a standardized certification examination for individuals desiring to demonstrate they meet the minimum competencies necessary to work with clients with substance use disorder/addiction. The purpose of this credential is to standardize the quality of addiction prevention, intervention, treatment, and continuing care services (NCCAP). The certification provides three levels of credentialing which are outlined in Table 1.

Despite the requirements for sitting for an examination, each test taker should receive a subjective and reliable test which measures knowledge directly related to the area of specialty, and test conditions should be consistent for all.

Scenario 1

Suppose two individuals (Mandy from Illinois and Roger from Wisconsin) are both registered to sit for the NCCAP level-2 certification examination. When Mandy registers for the examination online, she receives no information regarding accommodations, test environment, and policies/procedures. After arriving at the testing site, the test administrator provides a hardcopy of the examination whereby Mandy must circle the correct answer choice and write out the answers for each essay question. Mandy is given a time limit of 1.5 hours. Roger, on the other hand, registers online and is provided with test site information which includes basic requirements for requesting accommodations, time limit for examinations, is ensured a noise-reduced environment, and will be provided with an electronic test format.

Once Mandy and Roger complete the examination, both are notified within 2 weeks whether they will pass the test. As the reviewers for the examination are scoring the test, they come across the

essay questions and score them differently despite the answer choices for both Mandy and Roger being nearly identical.

As is noticeable from the aforementioned hypothetical scenario, the test administration was not uniform (both Mandy and Roger received different test formats), the time limit was different (while both were given 1.5 hours, Mandy had to write out her answer choices which requires more time), details of the test environment were only provided to Roger, and reviewers were not provided with how to interpret the essay questions. in line with the categories required. In order to ensure the examination is credible, the eight categories must be established. In addition to determining a test has a standardized procedure established, it is essential for a counseling professional to understand scale development and statistical concepts, reliability and validity of tests, and how to interpret the results of these types of tests.

Scale Development and Statistical Concepts

Aside from evaluating an instrument for uniformity of test administration, additional concepts counseling professionals should become familiar with prior to selecting a test to use in practice consist of the following: (a) understanding types of sampling, (b) validity, (c) reliability, (d) statistical concepts, (e) standardization process, (f) understanding assessment scores, and (g) how to interpret assessment results. A counseling professional familiar with each of these concepts will likely have a better arsenal when selecting the best test to use in clinical practice. As is likely discussed in other classes associated with training future counseling professionals, there is limited time available to meet with clients and it is essential to gather as much information in as little time as possible. Therefore, the selection of a test that can gather large amounts of information is important. However, if the test lacks "essential ingredients," the results may not be accurate.

Sampling

Suppose a researcher was interested in conducting a study where you survey individuals on their honest opinion toward the President of the United States. If the researcher surveys all of the residents in a metropolitan area that is known to be Republican, and the President is Republican, the results are likely to be skewed in favor of the President. Although it is nearly impossible to survey every person in the population to receive a generalizable result which reflects the viewpoints of most, there are ways in which researchers can obtain reliable results by selecting a subset of the population. When surveying a sample of individuals within a population, this is called sampling.

Probability Sampling. There are two types of ***sampling*** in research, including probability sampling and nonprobability sampling (Fitzpatrick, Sanders, & Worthen, 2011). **Probability sampling** is the process of utilizing some form of random selection of participants from the population of interest. For example, if there are 100 students in an introduction to drugs and alcohol course, a researcher may randomly select a specific number of students (e.g., 20 out of 100) from the class for a study (simple random sampling). In theory, this sample of the population will be representative of the entire student body associated with that program, because each student had an equal chance of being selected. There are several other types of probability sampling other than simple random sampling, including but not limited to:

1. ***Stratified Random Sampling***: dividing the population into homogeneous subgroups and then taking a simple random sample in each subgroup.
2. ***Systematic Random Sampling***: after determining the sample size desired, all of the students in the introduction to drugs and alcohol course are assigned a number (we will say there are 100 students in the course, and our desired sample size is 25 students). We would then divide the population (100 students) by our desired sample size (25 students), which would equal the number 4. We

would randomly select a number between 1 and 4. If we chose the number 3, we would then select every third student to participate in the study (e.g., 3, 6, 9, etc.).

3. ***Cluster Sampling:*** Although both stratified and systemic random sampling provide an accurate picture of the population, they both have their drawbacks. Both require for the researcher to have a sampling frame of the population. In other words, you must have all identifying information of each person in the population you are studying to carry out that form of sampling method. For example, if you wanted to survey all students at your university, that would require having the contact information (i.e., emails) for each student.

In such instances, cluster sampling is often used. To carry out cluster sampling, the researcher would first sample groups or clusters of participants rather than all participants. In other words, consider your home state had a total of 500 K–12 schools and you wanted to survey students to determine whether they were being exposed to illegal substances and if so, at what age. Rather than trying to survey all students among the 500 schools, you would randomly select a chosen number (e.g., 20) and then survey those specific schools. This method of sampling often involves a multistage process whereby you sample large clusters, then sample smaller clusters.

Nonprobability Sampling. However, most researchers in the behavioral sciences do not carry out probability sampling. Rather, researchers often rely on **nonprobability sampling**. With this approach, researchers have no way of knowing whether a particular case "Jane Doe" will be chosen from the sample.

- ***Convenience Sampling*** includes participants which are readily available. For example, as a student, how many times have you been given the opportunity to participate in a study by your professor? When your professor sends out the evaluation online, she/he has no way of knowing who will complete the survey but is relying on a group of participants available to them.
- ***Snowball Sampling*** is the process of identifying someone who meets the criteria for inclusion in a study, and then asking this individual if he or she was able to refer or recommend others that would also meet the inclusion criteria so the researcher can contact them.

There are several other examples of nonprobability sampling procedures available, which include modal instance sampling, quota sampling, and expert sampling. Although these methods of sampling procedures are often used in research studies, it is important to be familiar with sampling procedures as often times, scales should be developed on prior peer-reviewed articles and literature. Additionally, scale questions (items) are generally developed based on the literature relevant to the construct. Furthermore, determining if an adequate sample size of participants included in the development of the test/assessment is crucial to determining if you have the ability to generalize the results to the entire population, called **external validity**. There are several types of internal validation methods which scale developers will often use to validate an instrument (test).

Validity

Validity deals with how well a test measures what it claims to measure rather than measuring something else. A well-designed instrument that has strong validity assists a counseling professional in determining whether the findings are genuine and not the result of something else. The validity of a test is related to two questions, including (1) what does the test measure and (2) how well does the test measure the trait/characteristic?

Face validity is the first type of validity that a counseling professional can use to determine whether an instrument could be beneficial to use in practice. Specifically, face validity refers to whether a measure appears to measure what it's intended to. For example, if a counseling professional was interested in using

an instrument that measured drinking/drug behavior, but also measured criminal activity, and potential for psychological diagnoses, the test would likely include questions that are not just related to drinking behavior. There are three additional types of validity by which a counseling professional should also pay attention to when evaluating a test manual of an instrument, including content validity, criterion-related validity (concurrent and predictive), and construct validity.

Content validity refers to the extent to which a test represents all components of a given construct. More specifically, content validity involves "the systematic examination of the test content to determine whether it covers a representative sample of the behavior domain to be measured" (Anastasi & Urbina, 1997, p. 114). The process of scale development usually involves individuals who are content experts in the construct or domain of interest. For example, if a test developer was interested in creating a fifth-grade math test, a panel of fifth-grade math teachers would be appropriate because they are able to determine whether the questions are appropriate for students enrolled at that level. Another example could be applied to the scenario regarding the national certification examination discussed in scenario 1. It is important for experts in drug and alcohol counseling (e.g., researchers, clinicians, and treatment administrators) to be included in the development of test items, because they can create questions that measure the knowledge of future drug and alcohol counseling professionals. Using fifth-grade math teachers to create test questions may not fully grasp the knowledge, skills, and abilities that a future certified alcohol and drug counselor must have in order to be a successful clinician. There are four factors that are considered when evaluating content validity:

1. Does the test measure the content domain of interest?
2. Do the test items constitute a representative sample of the behavior we want to measure?
3. Did the method involve defining the content/behavior domain being measured and determine whether test items adequately sample that domain?
4. Were a panel of experts used in the development of test items?

It is also important to consider the panel of experts and evaluate their professional background. Within the test manual, often times the content experts used in the development of test items will be provided with their professional affiliations. It is essential to determine how many experts were used, while also determining if a variety of professional background were used (e.g., drug/alcohol counseling training educators, researchers in this counseling specialty, clinicians).

Criterion-related validity involves demonstrating how test scores are correlated with a particular aspect of a client's behavior outside of the test. This "outside behavior" of interest correlates with performance on the test, such as the general record examination (GRE) which is often used to determine the predictability of success among the applicant if accepted into a graduate program. There are two types of criterion-related validity, including concurrent validity and predictive validity. Concurrent validity refers to establishing the relationship between test score(s) and some aspect of the client's current behavior (concurrent validity deals with diagnosis). Suppose we compared the scores of the national certification examination for addiction counselors and the grades in the addiction courses within their academic training program. We could determine the degree that scores on the certification examination are related to the performance in the addiction-related courses. If we compared the GRE examination to first semester performance in graduate school, we could establish predictive validity. Predictive validity refers to establishing the relationship between test score(s) and some future performance of the client (predictive validity deals with prognosis). Using that GRE example, some researchers have predicted that the higher a student performs on the GRE, the greater the likelihood that he or she will be successful in graduate school. However, there are other factors that are generally included in graduate school success (e.g., social relationships, financial considerations). Predictive validity is basically does the test predict what it is supposed to predict.

Construct validity refers to the extent that the test measures the theoretical construct that it claims to measure (Drummond & Jones, 2010). This is established by examining the correlation coefficient between scores on the test in question and scores on another test that measures the same construct of interest (convergent validity) or a different construct of interest (discriminant validity). Convergent validity should produce positive correlations between tests measuring the same or similar trait/construct. For example, an eighth-grade spelling test should be correlated with other eighth-grade spelling tests. Discriminant validity should produce correlations reflecting weak relationships or negative relationships between tests measuring different traits/constructs (e.g., an eighth-grade spelling test should not be correlated with an eighth-grade math test). Let's pretend a counseling professional was working with a client who uses drugs to escape discomfort associated with large crowds; we might administer the Liebowitz Social Anxiety Scale (Liebowitz, 1987). Earlier research by Liebowitz suggested there are two levels of social anxiety, including fear of social situations and avoidance of social situations. Now suppose that we developed an instrument that included a section that measured both fear and avoidance (social anxiety). We would expect to find a strong positive correlation between the items related to social anxiety of our scale and the Liebowitz Social Anxiety Scale, assuming we administered both scales to the same group of individuals. This process is called convergent validity.

In addition to establishing convergent validity, test developers may be interested in demonstrating how the construct for which the scale measures is not related to another construct (discriminant validity). Using the social anxiety example, if we provided the same client a confidence of public speaking in a large crowd scale, we would expect there to be a negative relationship between both scales. In theory, it would be difficult for a person who has severe social anxiety to be able to speak in front of large audiences. Although this may not be true in every situation, we would expect a negative relationship (as social anxiety increases, confidence in speaking in front of large crowds decreases).

To establish convergent validity, a scale developer should produce positive correlations between tests measuring the same or similar trait/constructs. For example, if a test developer created an instrument to measure psychiatric symptoms that would related to the diagnosis of Bipolar in the *DSM-5*, if this test along with an already established test were given to the same group of individuals, then there should be positive correlation between the items of the two scales. Likewise, discriminant validity should produce correlations reflecting weak relationships or negative relationships between tests measuring different trait/constructs. Remember, the closer a correlation coefficient is to 1 or negative 1, the stronger the relationship. The closer the correlation coefficient is to zero, the weaker the relationship. For example, there is usually a positive relationship between studying for an examination and the grade on the examination. If the correlation coefficient is closer to 1, that means the relationship is strong. If the correlation coefficient was closer to zero, that means the relationship between study time and grade on the examination is weaker in nature.

A ***correlation coefficient*** is important to understand moving forward in this chapter. A correlation is the measure between two variables and quantifies the degree of association (covariation) between two variables. A correlation describes in numerical terms the degree or strength of the relationship in addition to the direction (i.e., positive or negative as illustrated in Figure 1). The negative or positive portion of the correlation represents the direction of the relationship, but NOT the strength. As one variable increases, the other variable decreases (e.g., as drinking behavior increases, cognitive functioning decreases). A positive correlation indicates that as one variable increases, the other variable increases as well (e.g., as study time increases, the grade on the examination increases). In terms of the strength of the relationship, the closer the coefficient is to negative 1 or positive 1, the stronger the relationship. There is no difference in a negative 1 or positive 1 when examining the strength of the relationship. The closer the coefficient score is to zero, the weaker the relationship (zero = no correlation).

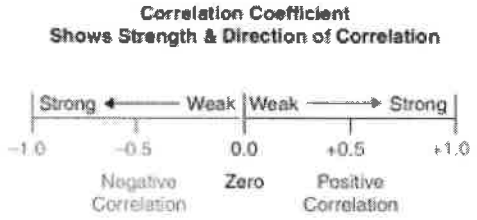

Figure 1.

Reliability

Reliability refers to the extent to which a scale produces consistent results, if the measurements are repeated (Howell, 2009). Another way to think about reliability is whether the instrument consistently measures what it is intended to measure or if you measure the same thing would you get the same score? Consider stepping on a weight scale and you weigh yourself every evening for the past week. You would expect for the scale to be within at least a five-pound range, but what if Monday you weighed 125 pounds and Friday you weighed 180 pounds? Would your scale be reliable? We will be discussing three ways that a test developer may measure reliability, including (1) internal consistency, (2) test–retest reliability, and (3) parallel-forms reliability. Each type of reliability analyses produces a score ranging from 0.00 to 1.00, with higher scores indicating higher accuracy. The observed score produced from each of these types of analyses is composed of a person's true score and error. An instrument is said to be reliable if it accurately reflects the true score and minimizes the error component. The closer the observed score is to the numerical value of 1, it is presumed that the error decreases. Several scholars will argue that there is a relationship between reliability and validity. Specifically, it is important for a valid test to accurately measure what it intends to measure. As can be ascertained from the image below, a test can be reliable but not valid. In the first image of the bull's-eye, one can have high accuracy and precision, but may not hit the target in the location that is desired. Another way to view concept is to consider the scores on the national drug/alcohol certification examination and how the test may be reliable, but it might not measure all of the content areas that it claims to. Specifically, if the examination consists of questions which focus on a clinical diagnosis of schizophrenia rather than substance abuse, then the test does not measure what it is intended to measure. The second bull's-eye demonstrates how a scale can measure a construct of interest, but the results may not be consistent. A test may not be valid or reliable as the construct has not been accurately measured, and the questions may be poorly written, which could increase error associated with the test (decreasing the reliability coefficient). The final bull's-eye represents what the desired instrument should be, in terms of being both reliable and valid.

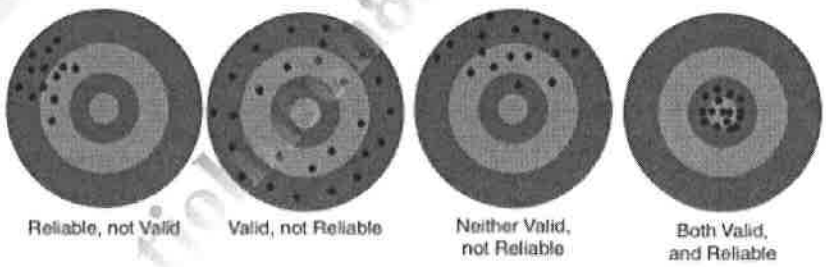

Figure 2.

In addition to evaluating the validity of a test when selecting an instrument to use in practice, additional considerations consist of the following: (a) determine if the test/user manual discusses reliability and (b) examine the reliability score to determine how consistent the test is. There are other types of reliability (e.g., inter-rater/observer reliability) that may be used in other types of research; however, these will not be discussed in this chapter.

Internal consistency is used to assess the consistency of results across items within a test, or how well the items on a test measure the same construct or idea. Test developers will determine the appropriate type of statistical analysis appropriate to measure reliability depending on the level of measurement used and the type of reliability being measured. There are three primary types of statistical procedures that are primarily used by instrument developers to measure internal consistency, including Cronbach's alpha and the split-half procedure (measured by Spearman and Brown formula, and/or Guttmann's formula). It may be difficult for a counseling professional to determine if each of the reliability procedures immediately above was conducted properly, but a counseling professional should still be able to determine if the reliability of the instrument is adequate. The table below is a visual depiction of the internal consistency categories based on the Cronbach alpha score.

Researchers have long argued what score is necessary for an instrument to be reliable instrument. However, the closer the score is to 1, the higher the reliability. A minimum of 0.7 should be the cutoff point in most instances, but a counseling professional should aim at a 0.8 or higher. Remember, the lower the reliability score, the greater the error (or the less likely you are to get the same score if you are administered the test again). Also, if the test manual of an instrument indicates an instrument's reliability was 1.0 (perfect reliability), a counseling professional should become skeptical as there is usually always error associated in statistics and a perfect reliability score is near impossible.

The **split-half** procedure is the process of dividing the test into equivalent halves in order to form two sets of items. The entire test is administered to the group of individuals, and the total score for each set of items is computed. The split-half reliability score is obtained by determining the correlation between the two sets. This is beneficial because the test is only being administered once, thus lessening the potential of time sampling error (as discussed in the test–retest reliability section below). It must be noted that splitting the test into the "first" and "second" half is not advised because of time sampling error (fatigue) and content sampling error (difficulty of items). Therefore, an odd-even split of items is one method to produce scores for each participant (Drummond & Jones, 2010). Longer tests generally produce stronger reliability, but the split-half reliability function reduces the number of items by 50 percent. Therefore, the Spearman-Brown formula for calculating the reliability correlation coefficient is used because it adjusts for the decrease in the number of items (Huck, 2011). Other types of internal consistency analyses include the Kuder–Richardson procedure (commonly used) or kappa (k), which are also used to analyze nominal level data. Weighted

Table 2 Internal Consistency Categories

Cronbach's Alpha	Internal Consistency
$\alpha \geq 0.9$	Excellent
$0.9 > \alpha \geq 0.8$	Good
$0.8 > \alpha \geq 0.7$	Acceptable
$0.7 > \alpha \geq 0.6$	Questionable
$0.6 > \alpha \geq 0.5$	Poor
$0.5 > \alpha$	Unacceptable

Kappa techniques are used to analyze ordinal level data, and intraclass correlation coefficients and internal consistency reliability coefficients are used to analyze interval or ratio level data (Fitzpatrick et al., 2011).

Test–retest reliability is a mechanism where a group of people are administered a test, and then at a later point, readministered the same test. The two test administrations can occur over a short period of time such as a day, or as long as a year or even more. The researcher's primary goal is correlate the two sets of scores to assess for consistency (Drummond & Jones, 2010). The resulting correlation coefficient will examine the two sets of scores with the final score describing the degree of reliability. Moreover, one would expect the score on each test to be similar since the same individuals are being tested. However, some problems may arise from this type of reliability analyses, including (a) differences in performance on the second test may actually change due to the first test [e.g., the test taker learns new information and their responses change]; (b) constructs of interest may change over time independent of the stability of the measure; (c) the interval between the two administrations of the test may be too long and the construct you are attempting to measure may have changed; or (d) the interval may be too short, which causes the reliability to be inflated because of memory. Additionally, the test developer may have difficulty in making sure the testing environment is identical between both samples.

Parallel or alternate forms are used to assess the consistency of results of two tests constructed in the same way from the same content domain (Fitzpatrick et al., 2011). Both tests must have the same rules for administration, the same number of items, items should be in the same format, and the range and level of difficulty should be equal for both tests. Essentially, parallel or alternate forms means that there are two versions of a test where questions for each form measure the construct of interest. For example, certifications examinations (e.g., NCCAP level-2 certification examination) oftentimes have more than one version of the examination with questions from each version being similar to each other (e.g., difficulty of the questions, construct being measured). This is beneficial because if a test taker needs to retake a certification examination, the questions will not be identical to the previous version, thus eliminating some learning of specific answers to test items (learning may have occurred due to additional studying, but not for specific test items). In terms of reliability, an instrument developer would provide each version of the examination to the same sample of people within a short time period. The correlation coefficient is then observed by the test developer to determine how reliable each form is (remember, the closer to 1.00, the more reliable).

Statistical Concepts

Statistics is a branch of mathematics that provides the tools for analyzing numerical data in a method (Howell, 2009). Descriptive statistics are used to organize, summarize, and present data in a systematic fashion (Trochim et al., 2014), while inferential statistics include procedures for reaching tentative conclusions from data using probability theory. Descriptive statistics include frequency distributions, measures of central tendency, and variability. A frequency distribution is to accomplish the task of organizing, summarizing, and describing data (collections of measurements). For example, when evaluating the test manual, the data should be organized and summarized in a manner that the counseling professional can observe the characteristics (e.g., gender, race, age, educational background) of the sample of the population that participated in taking the survey.

The ***measures of central tendency*** represent an average or typical value of the data (mean, mode, median), and described the central point of the distribution, while variability describes how the scores are scattered around the central point. The mean is the average of the scores of the participants. This is accomplished by adding up all of the numbers (scores), and then dividing by how many numbers there are.

Example: 2, 4, 5, 3, 4, 2, 1, 5, 6, 7, 4, 3
Step 1: Add up the numbers (2+4+5+3+4+2+1+5+6+7+4+ 3) = 46
Step 2: Divide by the count (how many numbers) = 46/12 = 3.83

The mode is the most frequently occurring score. If there is one mode, it is considered unimodal. Whereas, having two modes is bimodal. Having more than two modes is multimodal. Below are two examples as represented in Table 3:

The median is the score that divides the distribution into two parts such that an equal number of scores falls above and below that point (Trochim et al., 2014). The formula to calculate the median is $(N + 1) / 2 = n^{th}$ digit. The first step is to rewrite the number so they are in order from smallest to largest.

Numbers: 10, 11, 5, 14, 22, 32, 15, 17, 13
Rewrite: 5, 10, 11, 13, 14, 15, 17, 22, 32

Note: There are nine numbers in the list, so the middle one will be the $(9 + 1) \div 2 = 10 \div 2 = 5$th number.

There is statistical software that a test developer will use to calculate these numbers. However, it may be beneficial to counseling professionals to know what these numbers mean. The ***measure of variability*** represents the dispersion of scores about the average value (Howell, 2009). The goal is to obtain a measure of how spread out the scores are in a distribution. A measure of variability usually accompanies a measure of central tendency as basic descriptive statistics for a set of scores. This allows a scale developer to measure how accurately any individual score represents the entire population. The smaller the variability, the closer the scores are clustered together and will provide a good representation of the entire data set (Drummond & Jones, 2010). The range, interquartile range, and variance and standard deviation is how variability of the data can be measured.

The ***range*** is the maximum score minus the minimum score and is affected by extreme values. This represents the total distance covered by the distribution. The ***interquartile range*** is the distance covered by the middle 50% of the distribution (difference between quartile 1 and quartile 3). The figure below represents a normal distribution. The range would represent the highest score on the test minus the lowest score on the test. The normal distribution represents the population, where 68% of the population would be between −1 and 1 standard deviations, 95% of the population would be between −2 and 2 standard deviations, and 99.7% of the population would be between 3 standard deviations. Standard deviations represent the distance from the mean, median, and mode (in a normal distribution). In a normal distribution, most cases cluster in the center of the range and the number of cases decreases as extremes are approached, and the curve is bilaterally symmetrical with a single peak in the center (Howell, 2009).

Most human traits approximate the normal curve (e.g., height, weight, intelligence, academic achievement, aptitudes, and personality). For example, suppose researchers were attempting to establish guidelines to what appropriate weight is based on a person's biological sex, age, and race. These researchers would need to have a representative sample for each of these categories.

Table 3 Measures of Central Tendency

Value	Frequency	Total
13	1, 1, 1, 1	4
14	1, 1	2
16	1	1
18	1	1
21	1	1

Note: One Mode (Unimodal): 13, 13, 13, 13, 14, 14, 16, 18, 21 = **Mode = 13**

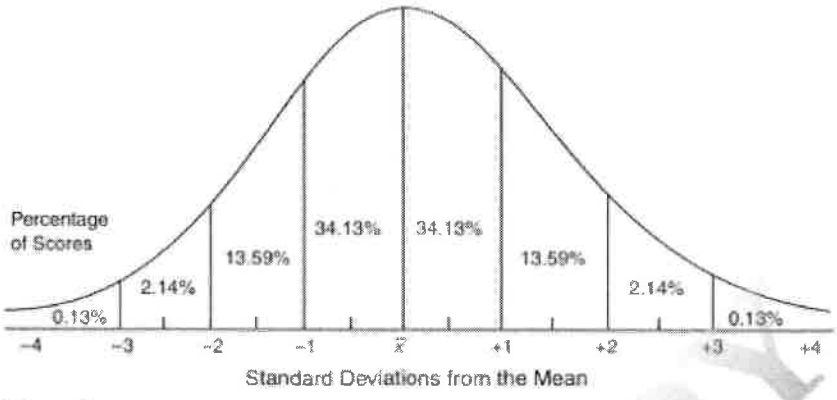

Figure 3.
Adapted from Mark Leary–*Introduction to Behavioral Research Methods, 6th ed,* 2012.

Standard scores allow for comparing an individual's raw score to others who have taken the same test. There are several types of standard scores, including Z-scores, T-scores, Stanines, CEEB scores, Deviation IQs, and percentiles. In norm reference testing, we compare a person to others, as described above and represented in the standard scores mentioned in the prior sentence. A "norm" is the average performance and is established by administering the test to a large representative sample of people for whom the test was designed. After collecting the data, researchers would create a normal distribution (evaluating the measures of central tendency and dispersion) to determine what a normal range is. As displayed in the figure above, 68% would fall between −1 and 1 standard deviations. For test development, establishing a normal distribution is helpful as we can observe whether an individual is considered "normal" on the construct that is being measured. Percentile rankings or scores are another method of evaluating an individual. A percentile rank informs how well an individual has performed compared to other people. For example, a percentile rank score of 99 on the national certification examination would indicate a test taker as scoring higher than 98 percent of the people that have taken the test. Without standardizing the scores of the sample, it would be difficult to make comparisons. A normal distribution, as mentioned above, will result in a sample of test subjects as representative of the population, and with a relatively large sample. Standard scores assist in the interpretation of an individual's test (raw) score, show how an individual's score fits within the total distribution of scores, and create a technique for comparison (Howell, 2009).

Chapter Activity

Identify an assessment that is used in the treatment of people with drug or alcohol challenges (e.g., Global Appraisal of Individual Needs [GAIN], Addiction Severity Index [ASI]). Locate the test manual for the assessment that you have chosen. Using the checklist below, answer the following categories: Basic Information, Background Information, Psychometric Properties of the Test. Determine based on the information provided in the chapter whether this is an instrument that is reliable and valid, and could be used in practice. With a classmate, practice administering your instrument by following the specific instructions in the test manual. After administering the test, answer the questions in the category of Administration of the Test, and Scoring and Interpretation of the Test. Once this is complete, reevaluate all of the completed categories and provide your overall impression of the test you have selected. Although this process may seem time-consuming, it will provide for the appropriate selection of a test to be used within clinical practice.