## Glossary of Terms

### Terms

**Clinical Cut-off Score**: A test score that is used to classify test-takers who are *likely* to possess the attribute being measured to a clinically significant degree (such as major depressive disorder or posttraumatic stress disorder). Assuming that the test is scored such that higher scores indicate higher levels of the attribute, test-takers who score at or above the clinical cutoff score are classified as "test positives," whereas test-takers whose scores fall below the cutoff score are classified as "test negatives." The setting of clinical cutoff scores typically involves evaluating rates of *sensitivity* and *specificity*, *negative predictive power* and *positive predictive power*, and *false negatives* and *false positives*, associated with a range of possible cutoff scores. These rates are each calculated by comparing the decisions made by the test against decisions made by a "gold standard" authoritative measure, such as a structured clinical interview. The "test cutoff score" judged to be optimal for a given application is then chosen.
*Sensitivity and specificity* are two important indices used in evaluating the accuracy with which a given diagnostic measure classifies test-takers who *have* versus *do not have* a given clinical condition. *Sensitivity* deals with *inclusion*, and refers to the test's ability to correctly classify test-takers who actually *have* the condition. This index answers the question, "How *sensitive* is the test at identifying actual positives?" It is calculated by dividing the number of actual positives who score at or above the clinical cutoff score (that is, the true positives) by the total number of test-takers who have the condition (all actual positives). Values for sensitivity range between 0 and *1.0*; the higher the value, the better the test is at accurately classifying actual positives. Tests with high sensitivity have a low *false negative rate*, meaning that they rarely misdiagnose people who actually have the condition.
In contrast, *specificity* deals with *exclusion*, and refers to the test's ability to correctly classify test-takers who do not have the condition. It answers the question, "How well does the test *rule out* actual negatives?" It is calculated by dividing the number of actual negatives who score below the clinical cutoff score (the true negatives) by the total number of test-takers who do not have the condition (all actual negatives). Values for specificity range between *0* and *1.0*; the higher the value, the better the test is at excluding actual negatives. Tests with high specificity have a low *false positive rate*, meaning that they rarely misdiagnose people who actually do not have the condition.

**Confidence Interval for True Scores:** A range of test scores within which one can be certain, with a specified level of confidence (say, with 68%, 95%, or 99% certainty) that a person's "true score" falls. (The true score is the average test score that a given test-taker would receive if she took the same test an infinite number of times.) Confidence intervals for true scores are created using the standard error of measurement. Specifically, confidence intervals can be formed because errors in measurement are presumed to be *normally distributed*, therefore allowing 68%, 95%, and 99% confidence intervals to be affixed around the observed score by inserting +/- 1, +/-2, or +/- 3 standard errors of measurement around the observed score, respectively.

**Continuous Scale:** A scale that measures the quantity of something on a *continuum* that represents gradually increasing amounts of the trait or attribute being measured (such as height, weight, or age). Many psychological concepts, such as optimism, intelligence, and level of mental distress, are measured in such a way that they are considered to be *continuous* variables.

**Correlation Coefficient:** An index describing the degree of *linear association* between two variables. The correlation coefficient varies between -1.0 and +1.0. Positive correlations

(between 0 and +1.0) indicate that high values in one variable are related to high values in another variable. Negative correlations, (between -1.0 and 0) indicate that high values in one variable are related to low values in another variable. Correlation coefficients closer to -1.0 or 1.0 indicate stronger associations, whereas correlation coefficients closer to 0 indicate weaker associations. Correlation does not accurately capture *nonlinear* associations, such as curvilinear (e.g., U-shaped) relationships.

**Cronbach's Alpha:** A commonly used index of the *internal consistency reliability* of a test or measure which, based on the average of the inter-item correlations, helps test users to judge whether the items are measuring a single underlying dimension or characteristic. Cronbach's Alpha measures the extent to which the individual test items *cohere* or "stick together," such that test takers consistently respond to items measuring the same thing in the same ways. Use of Cronbach's Alpha is based on the assumption that all the test items are measuring the *same* underlying attribute (not a mixture of different attributes) with the same degree of sensitivity.

**Criterion-Referenced Test:** Criterion-referenced tests are designed to predict, with maximum accuracy, scores on another test or standard external to the test itself (termed a *criterion*) given the test-taker's observed score on the test. (Criterion-referenced tests are also sometimes termed *criterion-keyed tests*.) Criterion-referenced tests can be used to generate such predictions as "How well, given her performance on the Scholastic Assessment Test (SAT) that she took while in high school, will this student do if she enters college?" "What is the likelihood, given his score on this test, that this juvenile offender will commit another serious criminal offense if he is released?" and "What is the likelihood, given her test score, that this applicant will succeed at this job?"

**Cross-Validation:** The evaluation of whether the psychometric properties of a test developed and validated in one sample of a given population can be repeated or *replicated* in a new sample from that same population. If a test fails to replicate in the new sample, there is a significant likelihood that the original results may have occurred as a result of chance factors (such as unique characteristics of the original sample, mode of test administration, or random error). Conversely, if the test's properties successfully replicate, it can be inferred that the test's psychometric properties reflect substantial characteristics about the underlying trait as measured by the test in that population. Successful cross-replication strengthens test users' confidence in the validity of the test as used in that population. Tests whose psychometric properties *replicate* across repeated samples from a given population are referred to as cross *validated* within that population. Cross-validation also refers to circumstances in which a test originally developed and replicated within one population is successfully cross-validated in one or more samples from a *different* population.

**Cumulative Frequency Scores:** Scores that indicate the proportion of test-takers whose scores fall *at or below* a given test score value.

**Dichotomous Variable:** A variable that is divided into two (and only two) discrete categories, such as male versus female, or "completed treatment" versus "dropped out of treatment."

**Discontinuous Scale:** A scale on which the categories are discrete and separate but ordered according to level or magnitude. Individuals clearly fall into one of the distinct categories (such as grade level in school).

**Frequency Distribution:** A summary table or graph that lists the range of scores for a given test, and the number of test-takers who obtained each specific test score. A frequency distribution records the frequency with which each test score is observed.

**Factor Analysis:** Factor analysis is a "data reducing" statistical tool that examines the interrelations among a set of variables (such as test items) in order to identify their underlying

structure. Factor analysis "extracts" clusters of strongly correlated variables and groups them into factors. Generally, one factor is extracted per cluster of strongly correlated variables. Factor analysis assists in reducing a large number of observed variables (such as 30 individual test items) into a much smaller number of variables (such as 5 test subscale scores). The subscales reflect the *factors*, which define the structure that underlies the set of variables. If a test is factor analyzed and generates one factor, the test is usually interpreted as being *unidimensional* (measuring one thing). Conversely, if a test generates multiple factors, the test is interpreted as being *multidimensional* and is subdivided into multiple subscales in accordance with its factor structure. Intelligence tests are one type of psychological test whose subscale structure is developed through the use of factor analysis. Tests derived from *multiple correlated* factors may be subdivided into multiple subscales that can also be combined to form a total-scale score (such as a test of PTSD symptoms that can be scored to create B-Symptom, C-Symptom, and D-Symptom *subscale* scores, as well as a *total-scale* score).

**Likert-type Scale:** As originally developed in 1932, the Likert Scale is a type of summative rating scale used to measure attitudes, such as asking test-takers to indicate the extent to which they agree or disagree with a given statement. Likert scales typically have five to seven possible response choices, the most common ranging from 1 to 5 (e.g., 1 = strongly disagree, 2 = disagree, 3 = not sure, 4 = agree, 5 = strongly agree). Because they are simple to understand and usually reliable, Likert scales have been adapted for a wide variety of applications, including clinical assessment instruments. These adaptations of the original scale are termed *Likert-type scales*. Examples include frequency scales (e.g., 0 = not at all, 1 = infrequently, 2 = sometimes, 3 = often, 4 = most or all the time) and intensity scales (e.g., 0 = not at all, 1 = a little, 2 = a moderate amount, 3 = a lot, 4 = a great deal). The scale is scored by calculating the sum (or alternatively, the average) of the test items to form a composite test score.

**Longitudinal/Maturational Effects:** Refer to changes in test-takers' scores that are caused by natural maturational processes as they take place over time (i.e., longitudinally). Examples include reaching puberty, increasing in one's ability to think abstractly, and increasing in stages of moral development as a youth matures from childhood into adolescence.

**Mean:** The *arithmetic average* of a frequency distribution of test scores. The mean is created by summing all test scores together and dividing by the number of test scores. It is a summary measure or index of the *central tendency* of a set of data.

**Median:** The middle score in a frequency distribution of test scores. The median is created by identifying the specific score at which 50% of test takers scored above, and the other 50% scored below. It is a summary measure or index of the *central tendency* of a set of data.

**Mode:** The most frequently observed score in the frequency distribution. The mode is created by identifying the test scores that were most commonly obtained. It is a summary measure or index of the *central tendency* of a set of data.

**Test Norms:** A statistical description of the test performance of a well-defined group (the normative sample) that serves as a reference by which to gauge the performance of other test-takers who subsequently take the test. Most norms tables show, in descending order, various test scores and the percentage of people in the reference group who scored below each score level (i.e., the cumulative percentile). Thus, knowing an individual's score, you can quickly determine how he or she compares in relation to the reference group. Potential types of test norms include *gender norms* (to permit comparisons of boys to boys only, and girls to girls only), *age norms* (to permit comparisons to same-age peers), *grade norms* (to permit comparisons to pupils in the same grade), *race or ethnic norms* (to permit within-group comparisons), *national norms* (to permit comparisons to a nationally representative sample), and *local norms* (to permit comparisons to other test-takers who live within the same geographic region).

**Test Norming:** To norm a test is to administer (in a *standardized manner*) the test to one or more *normative samples* for the purpose of creating test *norms* (such as age norms, grade norms, ethnic group norms, gender norms, national norms, local norms, and so forth). A test should be standardized before it is validated, and validated (to a reasonable degree) before it is normed.

**Normative Sample:** A selected sample of test-takers who are assembled together to take the test for the purposes of creating test norms. Members of the normative sample are typically selected on the basis of some common characteristic or characteristics, depending on the type of norms that the test developers wish to create. These may be grade norms, age norms, sex norms, racial or cultural norms, nationally representative norms, local norms, or some combination thereof.

**Norm-Referenced Test:** A norm-referenced test is designed to furnish test users with information about the meaning of a given test score by comparing that score to a distribution of scores (called *test norms*) from other representative test-takers (called the *normative sample*). The use of norm-referenced tests permits answering such questions as "How has this test-taker performed relative to a comparison group of test-takers (made up of members of the normative sample)? Most norm-referenced tests calculate test-taker's standing in terms of "percentile rank" (i.e., cumulative percentage, or the percentage of test-takers in the normative sample who scored at or below a given test score).

**Operational Definition:** Instruments and procedures that have been selected to measure the attribute under study. This could be weighing a child with a bathroom scale (an operational definition) to measure his or her weight, measuring her height with a ruler (another operational definition), or measuring her school attendance (operationally defined by counting the number of unexcused absences in her school attendance records).

Operational definitions become more complex when they are used to measure phenomena that cannot be directly seen (like measuring a boy's level of perceived social support on a 5-point Likert-type frequency scale). In particular, much of what is measured in psychological assessment are *hypothetical constructs* (like resilience, anxiety, intelligence, motivation, or perceived social support) which, because they are not physical entities that have a physical size, shape, and weight, cannot be directly measured. Thus, operational definitions are *always* one degree removed from the hypothetical construct. Instead, operational definitions are measuring (with some degree of imprecision and hence error) the "measurable phenomena" to which the hypothetical construct gives rise (including responses to test questions)-but *never the actual construct itself*. Because hypothetical constructs are *measured* by operational definitions, the constructs themselves are essentially *defined* by those operations. Thus, it is important to remember that a *hypothetical construct (such as resilience) will "behave" no better than the operational definitions used to measure it will allow.* Therefore, even if resilience as a hypothetical construct is meaningful and influential in the natural world, if the test used to measure it is of poor quality, "resilience" scores will perform poorly and make the construct appear inconsequential. Evaluating how well an operational definition measures the hypothetical construct it is intended to measure is one of the most important tasks of psychometrics. Evaluating the psychometric properties of a test involves asking such questions as: Are these operational definitions reliable and valid? Free from significant bias? Culturally appropriate? Developmentally appropriate? Clinically relevant?

**Percentiles:** Scores that denote the proportion of test-takers whose scores fall *at or below* a given test score value, expressed *in percentage units*. For example, a test score that corresponds with the 5th percentile is one at which 5% of the test-takers scored at or below. A test score that corresponds with the 95th percentile is one at which 95% of the test takers scored at or below.

**Periodicity:** Refers to the clinical course of a given variable, typically a psychological

symptom or sign, that is, its tendency to intensify, decrease, go into remission, or fluctuate over time. Periodicity is a particularly useful concept when evaluating whether a given condition is cyclical (such as in cyclothymia or manic-depressive disorder), whether it fluctuates as a function of the presence or absence of risk or protective factors (such as life stresses or social support), or how the symptoms respond to treatment.

**Psychological Assessment:** The gathering and integration of psychology-related data for the purpose of conducting a psychological evaluation. Psychological assessment uses such instruments as psychological tests, interviews, case studies, behavioral observations, and specially-designed apparatuses and measurement procedures.

**Psychometrics:** The specialized branch of psychology dealing with the properties of psychological tests, such as reliability, validity, and test bias. Psychometrics can also be viewed as a specialized branch of statistics that is dedicated to describing, evaluating, and improving the properties of psychological tests.

**Range:** An index of scale variability, as measured by the distance between the highest observed score and the lowest observed score in a frequency distribution.

**Raw Scores:** Test scores that *have not* been transformed or converted to any other form. Raw scores are in the metric of the original test, whatever that metric is (such as kilograms, or agreement/disagreement on a 5-point Likert-type scale). Raw scores are also referred to as *observed test scores.*

**Standard Scores:** Standard scores are raw scores that have been transformed or converted from their original metric to another metric that has an *arbitrarily set mean and standard deviation.* For example, converting raw scores into a standard score with a mean of 0 and a standard deviation of 1 transforms them into *Z scores,* whereas converting raw scores into a standard score with a mean of 50 and a standard deviation of 10 transforms them into *T scores.* Standard scores are generally considered to be easier to interpret, more meaningful, and more clinically useful than the untransformed "raw" scores from which they were derived.

**Test Standardization:** To *standardize* a test is to develop specific (i.e., *standardized*) procedures for *consistently* administering, scoring, and interpreting a test, so that the test procedure is similar for each test-taker. Tests that are not standardized are vulnerable to introducing error into observed test scores due to potential differences in the ways in which the test is administered, scored, or interpreted across test-takers.

**Variability:** A general term used to describe the degree of dispersion or "spread-outness" of test-taker's scores in reference to the mean. There are three key measures of variability: the *range* (the distance between the highest and lowest score), the *variance* (the average squared deviation score), and the *standard deviation* (the square root of the variance). A distribution in which test scores are tightly compressed around the mean will have limited variability; whereas a distribution in which scores are dispersed far from the mean will have much variability.

## Reliability

**Test Reliability (general definition):** The extent to which a measurement instrument yields consistent, stable, and uniform results over repeated observations or measurements when the test is administered under the same testing conditions. In a general sense, test reliability refers to the *degree to which observed test scores are free from measurement error.* As test reliability increases, the more the observed test score variance is free from error. That is, observed differences are presumed to reflect "true" differences in the amount of the attribute being measured, and not error. For example, to what extent does an observed test score of, say, 100 reflect a given test-taker's *"true" score*, as opposed to reflecting the influence of

random error (such as guessing) or systematic error (e.g., whether she was rated by a conservative versus a liberal judge)? Another way to view test reliability involves understanding how *consistently* a test performs in its measurement operations. There are a variety of methods for estimating test reliability. These including gauging the degree of *consistency* in test-taker's responses over (a) repeated test administrations (*test-retest* reliability), (b) different versions of the same test (*alternate-forms or parallel-forms reliability*), (c) different sections of the same test (*split-half* reliability), (d) other items within the same test (*internal consistently* reliability), or (e) other raters (*inter-rater* reliability).

**Internal Consistency Reliability:** Refers to the extent to which the items within a test "cohere" or inter-correlate. The higher the index of internal consistency reliability, the greater the confidence that the test items are, collectively, measuring the same underlying trait or dimension. A commonly used index of internal consistency reliability is *Cronbach's Alpha.* Alpha values of .70 or higher are considered to be minimally acceptable evidence that the items collectively measure the same thing for research applications; whereas Alpha values near to .90 are considered optimal for clinical applications.

**Inter-Rater Reliability:** Inter-rater reliability refers to the degree of agreement between ratings made by two or more independent judges who are rating the same person or attribute. For example, two psychologists with expertise in clinical assessment may administer a structured clinical interview to the same client. A commonly used index of inter-rater reliability, the Kappa Coefficient, could be used to evaluate the degree of agreement in their respective ratings of whether a given symptom is present versus absent, or whether the person meets criteria for the disorder, after correcting for chance agreement.

**Test Re-Test Reliability:** Test-retest reliability refers to the degree to which an assessment yields similar results from one testing occasion to another *in the absence of intervening maturation, learning, or other life experiences* that may influence the actual value of the attribute being measured. Test-retest reliability is traditionally calculated by administering the same test to the same test-takers at two separate points in time, the correlating their "Time 1" test score with their "Time 2" test score. The duration of time separating the two test administrations depends on the presumed stability of the attribute being measured. Many clinical instruments use a 2-week test-reliability interval, whereas some intelligence tests use intervals spanning several months or longer.

**Alternate-Forms Reliability:** A form of reliability in which (typically two) different forms of the same test are administered to the same subjects on separate occasions. The test-taker's scores on "Form A" of the test are then correlated with their respective scores on "Form B" to calculate the alternate-forms reliability coefficient. The "alternate forms" used in evaluating this form of reliability should be equivalent in the *content coverage* and *level of difficulty* of their respective test items. An advantage of alternate forms reliability over test-retest reliability is that the use of alternate forms reduces "carryover" effects into the second test administration that are produced by test-takers' memories of how they answered the test at the first administration. Alternate-forms reliability is easier to achieve than parallel-forms reliability, which carries more stringent assumptions about the equivalence of the tests. (Note, different forms of the test created by simply changing the order of the *same* set of test items do *not* qualify as alternate forms of one another. Rather, Forms A and B must consist of different sets of items that are presumed to measure the attribute under study equally well.)

**Parallel-Forms Reliability:** A form of reliability in which (typically two) different forms of the same test are administered to the same subjects on separate occasions. The test-taker's scores on "Form A" of the test are then correlated with their respective scores on "Form B" to calculate the parallel-forms reliability coefficient. The "parallel forms" used in evaluating this form of reliability should be equivalent in the *content coverage, level of difficulty, means, variances, and measurement error* in their respective items. Another way of defining parallel-forms tests is that each test-taker's total-test scores on the two parallel forms will correlate equally with that test-taker's "true score." Parallel-forms reliability is a very difficult form of

reliability to achieve, because the test items must be so similar to each other that they are essentially *interchangeable*-that is, they measure the same attribute with the same degree of sensitivity, and with the same amount of measurement error. An advantage of parallel forms reliability over test-retest reliability is that the use of parallel forms reduces "carryover" effects into the second test administration that are produced by test-takers' memories of how they answered the test at the first administration.

**Response Format:** Refers to the scaling method, or metric, used to record test-taker's responses to test questions or items. Examples include true/false scaling, Likert-type scaling, multiple choice scaling, open-ended, essay, fill in the blank, and matching.

**Sensitivity to Clinical Change:** Refers to the extent to which a given test item score, or total-test score, changes in its value as a result of treatment (such as psychotherapy or pharmacotherapy). Tests that show good sensitivity to clinical change generally use scales with multiple small increments. It is generally more difficult to detect clinical change with a 2-point (e.g., True/False) or a 3-point (e.g., Some/A Little/A Lot) scale than with a scale with more response options, such as Likert-type scales. For example, some test items measuring delinquent behavior capture "lifetime incidence" behaviors (such as "Have you ever stolen anything valued over $1,000?") that will not change even if the youth's behavior has improved during treatment. Such items are generally not sensitive to clinical change. In contrast, an item such as "How many days did you skip school during the past month?" may be much more sensitive to detecting changes in behavior that occurred during the course of treatment.

**Standard Deviation:** An index of the variability or spread of data around their mean, indicating the degree to which data or test scores are compressed versus dispersed around the mean. Mathematically, the standard deviation is computed by taking the square root of the variance, another index of variability.

**Standard Error of Measurement:** A theoretical concept consisting of the standard deviation of the distribution of observed test scores that a given test-taker would theoretically generate by taking a given test an infinite number of times. The random fluctuations in test scores across repeated administrations of the test is presumed to reflect random error, and to be randomly distributed (thus creating a *standard normal distribution of test scores*). Wider fluctuations indicate lower test reliability, whereas smaller fluctuations indicate high test reliability.

Estimates of the standard error of measurement depend heavily on the reliability of the test-as test reliability approaches its upper limit of 1.0, the standard error of measurement shrinks in size. A perfectly reliable test will have a standard error of measurement of 0.0. The standard error of measurement is used to calculate *confidence intervals* around a test-taker's observed test score that allow the test administrator to calculate the likelihood that the interval contains the test-taker's "true" score. (The true score is a theoretical concept, consisting of the mean of the test scores that the test-taker would generate if she took the test an infinite number of times.) These confidence intervals can be formed because errors in measurement across repeated test administrations are presumed to be *normally distributed*, therefore allowing 68%, 95%, and 99% confidence intervals to be affixed around the observed score by inserting +/- 1, +/-2, or +/- 3 standard errors of measurement around the observed score, respectively.

**Standard Normal Distribution:** The standard normal distribution is a theoretical data distribution that is bell shaped, with symmetrical tails that taper off in both directions. This distribution has unique properties that greatly assist with test interpretation. Specifically, in the standard normal distribution:

- the mean, the median, and the mode are exactly the same value;
- approximately 68.26% of scores fall within +/- (plus and minus) 1 standard deviation unit above and below the mean;

- approximately 95.44% of scores fall within +/- 2 standard deviation units above and below the mean;
- approximately 99.72% of scores fall within +/- 3 standard deviation units above and below the mean.

Knowing what proportion of scores fall within certain intervals around the mean allows test users to create *confidence intervals*, wherein they can estimate, with a certain probability level, that a test-taker's *true score* (that is, the test score that she would obtain if she were measured with perfect accuracy) falls within a certain range around her *observed test score.*

**T-Score:** T-scores are a type of standard score that have been converted or transformed from raw scores to have a mean of 50 and a standard deviation of 10. Many commonly used psychological tests convert raw scores into T scores to assist with interpretation. If the frequency distribution of T scores calculated from the raw scores conforms to a standard normal (bell) shape (that, is, if they are *normalized T scores*), then a T score of 20 falls approximately at the 0.1st percentile, a score of 30 falls at the 2.3rd percentile, a score of 40 falls at the 15.9th percentile, a score of 50 falls at the 50th percentile, a score of 60 falls at the 84th percentile, a score of 70 falls at the 97.7th percentile, and a score of 80 falls at the 99.9th percentile. The practice of assigning a "clinical cutoff score" of T = 70 or higher is using a statistical approach to identifying "clinical cases" by selecting test takers whose scores fall at or above the 98th percentile of the normative sample's test scores (that is, within approximately the upper 2 percent).

**Test Validation:** Refers to a set of procedures designed to systematically evaluate whether a given test or procedure measures what it purports to measure. This entails evaluating one or more specific types of test validity, including content validity, criterion-referenced validity, predictive validity, or construct validity.

**Variance:** An index of the variability or spread of data around their mean, indicating the degree to which data or test scores are compressed versus dispersed around the mean. Distributions with little variance are bunched tightly around the mean, whereas distributions with much variance are spread out far from the mean. Outliers (extreme data points) can artificially inflate the variance and standard deviation of the distribution in misleading ways, making it appear more spread out than it generally is.

## Validity

**Test Validity (general definition):** Test validity refers to *the degree to which a test actually measures what it claims or purports to measure, rather than something else.* There are a variety of methods for estimating validity, which are defined below. It is generally accepted that a test must first be accepted as reliable (that is, low in measurement error) before it is considered valid. A test can be reliable but still not be valid for a given testing application, but it cannot be considered valid if it is unreliable.

**Concurrent Validity:** A form of criterion-referenced validity that is determined by the degree to which the scores on a given test (such as a test of the degree of a child's exposure to sexual abuse) correlate with scores on another, already established test administered at the same time (such as a test of the frequency of the child's posttraumatic stress reactions). Concurrent validity is gauged by the strength of the resulting correlation, sometimes termed a *validity coefficient.* Tests that show strong correlations with external criterion variables that are measured at the same point in time are considered to show good concurrent validity.

**Content Validity:** A method of establishing validity based on expert judgment that the content of the measure is consistent with the *content domain* of the attribute under study. For example, does a test of positive parenting skills cover all relevant dimensions of positive parenting? Does a test of self-concept adequately cover all relevant dimensions of a child's

self-concept?

**Construct Validity:** refers to the degree to which a set of operational definitions measures the attribute that the test purports to measure, and not other attributes or measurement error. (That is, does it measure what it claims to measure, and does it not measure what it claims not to measure?) Construct validity may be considered the "parent" of all validities because it *encompasses* all other types of validity. Construct validity can be evaluated in many different ways. One popular method involves administering several tests *that purport to measure the same or related attributes* to the same group of test-takers and then calculating the patterns of relationships among the test scores. Tests that measure the same attribute should correlate strongly, and tests that measure theoretically related attributes should correlate to a moderate degree.

**Convergent Validity:** A type of criterion-referenced test validity that is typically evaluated in combination with *discriminant validity* (also defined below). Evidence for convergent validity is found when tests that purport to measure the same attribute correlate strongly, and when tests that purport to measure theoretically related attributes correlate to a moderate degree. Conversely, evidence for discriminant validity is found when tests that purport to measure theoretically unrelated attributes correlate negligibly (with small correlation coefficients that do not significantly differ from 0.0). Taken together, tests that show convergent and discriminant validity will correlate differentially with measures of related, versus unrelated, attributes.

**Criterion Referenced Validity:** Refers to the degree to which a test correlates with other tests (or other criteria that are external to the test) that measure a theoretically-related attribute. This attribute may be the same attribute that the to-be-validated test measures, or a related attribute, and may be measured either concurrently with the test, prior to the test, or at a future point in time.

**Discriminant Validity** (also known as divergent validity): A type of criterion-referenced test validity that is typically evaluated in combination with *convergent validity.* Evidence for discriminant validity is found when tests that purport to measure theoretically unrelated attributes correlate negligibly (with small correlation coefficients that do not significantly differ from 0). Conversely, evidence for convergent validity is found when tests that purport to measure the same attributes correlate strongly, and when tests that purport to measure theoretically related attributes correlate to a moderate degree. Taken together, tests that show convergent and discriminant validity correlate *differentially* with measures of related, versus unrelated, attributes.

**Face Validity:** A very superficial form of test validity, considered by many test experts to not constitute a "true" form of validity at all. Evaluating face validity involves gauging, based on a visual inspection alone, whether the test items *appear to measure what they claim to measure.* A test (such as the Minnesota Multiphasic Personality Inventory, or MMPI), may be a "valid" measure of an attribute, yet show low face validity. Low face validity is a desirable property in some applications, such as forensic assessment, given that it makes "faking" test scores more difficult for test-takers who do not wish to give an accurate response. However, in the other applications, high face validity may be valuable, such as when creating take-home questionnaires for caregivers for which a high response rate is desired.

**Postdictive Validity:** A form of criterion-referenced validity that is determined by the degree to which the scores on a given test (such as a test of the degree of a child's exposure to physical abuse) are related to the scores on another, already established test or criterion administered at a previous point in time (such as an index of the child's school functioning for the year before, as gathered from school records). The degree of test postdictive validity is gauged by the magnitude of the resulting correlation, a *postdictive validity coefficient*.

**Predictive Validity:** A form of criterion-referenced validity that is determined by the degree to

which the scores on a given test (such as a test of the degree of an adolescent's exposure to community violence) correlate with scores on another, already established test or criterion that is administered at a future point in time (such as a test of the frequency of the adolescent's somatic complaints). The degree of test predictive validity is gauged by the magnitude of the resulting correlation, termed a *predictive validity coefficient.*

**Discriminant-Groups Validity** refers to the degree to which a test intended to measure a given trait can discriminate between two or more groups that are theorized to differ in their levels of that trait. For example, a trauma researcher could administer a test of posttraumatic stress disorder (PTSD) to groups of physically abused, sexually abused, and "non-traumatized" children. Evidence for discriminant-groups validity would be found if the non-traumatized groups scored significantly lower than the other two "trauma-exposed" groups.

**Factorial Validity:** Evidence for the factorial validity of a test is found when the test's factor structure in a factor analysis corresponds with the theorized structure of the hypothetical construct that the test is purported to measure. For example, if a test developer's working theory of positive parenting practices proposes that there are three basic dimensions of parenting, and a factor analysis of a test of positive parenting practices generates three factors that correspond with those three proposed dimensions, then the test is inferred to show *factorial validity.*

**Hypothetical Construct:** A theoretical concept that is not directly measurable, but which gives rise to phenomena that can themselves be directly measured using operational definitions. Most "psychological" concepts are hypothetical constructs, because they are not physical entities that can be direct measured physically. Examples include intelligence, resilience, depression, anxiety, attitudes, motivation, optimism, hope, fear, faith, guilt, group cohesion, and marital satisfaction. Operationally defining hypothetical constructs always involves some degree of measurement error, because one is not directly measuring the construct, but instead, the measurable phenomena to which it gives rise. For example, one cannot directly measure resilience, but one can measure *behavioral manifestations* of resilience, such as doing well in school amidst severe adversity.

**Z Score:** Z-scores are a type of standard score that have been transformed or converted from raw scores to have a mean of 0 and a standard deviation of 1. Note that this transformation will not automatically *normalize* the resulting distribution of Z scores such that they resemble a standard normal distribution. That is, after transformation, the resulting distribution of Z scores will retain the same overall shape as its "parent" raw scores-whether it is normally shaped, positively skewed, and so forth.

**Test Standardization:** A *standardized* test is a test that has a standard or consistently applied set of procedures for test administration, scoring, and interpretation. Test standardization refers to the process of developing those procedures. *Test standardization*, test validation, and test norming are related, but not identical, procedures. A test should be standardized before it is validated, and validated (to a reasonable degree) before it is normed.

**Test Validation:** A validated test is a test for which specific types of reliability and validity data considered relevant for a specific testing purpose have been systematically examined and judged to be acceptable for that purpose. *Test validation* refers to the process of collecting specific types of test reliability and validity data in preparation for a given testing application. A test may be validated for certain measurement applications, but not be validated for other applications.

## References

Aiken, L. R., & Groth-Marnat, G. (2006). *Psychological testing and assessment* (12th ed.).

New York: Allyn & Bacon.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Prospect Heights, IL, USA: Waveland Press.

Cohen, R. J., & Swerdlick, M. E. (2005). *Psychological testing and measurement: An introduction to tests and measurement* (6th ed.). New York: McGraw-Hill.

Dawson, B., & Trapp, R. G. (2001). *Basic and Clinical Biostatistics.* New York: Lange/McGraw-Hill.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* New York: Freeman.