KAPLAN CONSULTING AND COUNSELING
(216) 766-5743

Fall 2005

Volume 1, Issue 2

Insights.... on Psychological Testing

Test Concepts Pertinent to Litigation

There are four significant test concepts pertinent to litigation: reliability, validity, standardization, and norms.

Reliability is the ability of a test to yield consistent results. In order to be reliable, both the test and the testing procedure must yield consistent results. Much like radar tracks a target, reliability refers to how accurately a test tracks changes in a subject.

Validity is a concept which refers to whether a test actually measures that which it intends to measure. Validity indicates the degree to which one is able to arrive at specific conclusions or make predictions about an individual based on one's test scores.

Standardization is the process of administering a test to a group of people in order to ascertain scores that are typically obtained. Standardization provides a mean (average) and standard deviation (spread) relative to a certain group. As a result of standardization, a test taker can determine how far above or below the average his/her score is relative to the normative group.

Norms are used as a reference against which psychological test data are interpreted. They represent the test performance of a standardization sample that reflects the general population under consideration. Norms provide a means of assessing a person's relative standing in comparison with others. Without normative data, the information obtained regarding a person is not empirically meaningful.

Significant Test Concepts

Reliability

Validity

Standardization

Norms

Predictive and Retrospective Accuracy

These two concepts are crucial to understanding testing that is based on standardized instruments. Predictive accuracy refers to the likelihood that a test is accurately classifying individuals or predicting whether they have a specific condition, characteristic, etc. Predictive accuracy addresses the likelihood that those who score positive on a test will fall into a specific group. Retrospective accuracy, however, begins not with the test but with the specific condition or characteristic that the instrument is designed to measure. Retrospective accuracy looks to the likelihood that those who fall into a particular group scored positively on a test. Confusing the directionality of the inference (the likelihood that those who score positive on a test will fall into a specific group versus the likelihood that those in a specific group will score positive on a test) is a cause of numerous assessment errors. For example, being a rape victim would more likely predict someone would score positively on a measure of Posttraumatic Stress Disorder (PTSD) than scoring positively on a PTSD test would predict that someone was a rape victim.

Inside this issue:

Prepared by

Kaplan Consulting and Counseling, Inc.
Robert G. Kaplan, Ph.D.
3401 Enterprise Parkway
Suite 340
Beachwood, Ohio 44122
Call: (216) 766-5743
RKapan@KaplanCC.com

Expertise in

Forensic Psychological Evaluation
Posttraumatic Stress Disorder
Workers' Compensation
Disability Evaluation
Workplace Violence
Sexual Harassment

Insights is a newsletter published by Kaplan Consulting and Counseling, Incorporated as a free service to legal professionals. Comments regarding this issue, suggestions for future issues, and requests for additional copies can be directed to the attention of Thomas A. Moran, J.C.D., B.C.E.T.S., Senior Litigation Analyst. Call (440) 225-4614 or e-mail thmoran@comcast.net.

Aspects of Validity

Validity is a critical issue in the arena of psychological testing. There are different types of validity including predictive validity, concurrent validity, content validity, and construct validity.

Predictive validity indicates the degree to which test results are accurate in forecasting some future outcome. For example, a psychological test may be administered to all individuals who seek to become police officers. The results may be used to predict those individuals who will be able to tolerate the stress of the position.

Concurrent validity is indicative of an instrument's ability to provide a basis for assessing accurately some current condition. For example, one might hypothesize that certain profiles of a psychological test are indicative of chemical dependency. To validate this hypothesis, test results could be compared with individuals who have known characteristics (i.e., chemically dependent individuals). If the test demonstrates adequate concurrent validity regarding this hypothesis, it could be substituted in some cases for the more elaborate and time consuming methods of assessment and diagnosis.

Content validity denotes the ability of a test to be generalized to the entire content of what is being measured. For example, the bar examination theoretically measures the basic knowledge, skills, and abilities necessary to practice as an attorney. The degree to which the bar examination accurately reflects or represents this larger domain is its content validity.

Construct validity indicates the general validity of a measurement device. It determines whether the instrument actually measures the concept under consideration. For example, individuals who coped well with a traumatic event should do better on a test of coping skills than individuals who did not cope well with that event.

The Different Types of Validity:

Predictive Validity
Concurrent Validity
Content Validity
Construct Validity

Measuring Truthfulness in Psychological Testing

Validity scales are used in many psychological tests to identify exaggeration, faking, equivocation, or deception by the test taker. For example, the Minnesota Multiphasic Personality Inventory-2 (MMPI-2) contains nine measures of validity that evaluate the truthfulness and the test taking attitude of the respondent. Among the most widely used are the F (Infrequency), L (Lying), and K (Defensiveness) scales. The F scale is made up of 60 items endorsed infrequently by the original MMPI normative sample. Elevated scores on this scale suggest that the respondent provided a large number of infrequent and therefore exaggerated answers to test items. The F_p scale helps differentiate elevations on the F scale that are the product of genuine psychopathology from those that are the result of over reporting. Providing the respondent with the opportunity to deny minor faults and flaws that most people acknowledge, the L Scale measures the likelihood that the test taker approached the test defensively, claiming excessive virtue. The K Scale measures less subtle forms of defensiveness. Highly correlated with the K Scale, the S Scale also measures the defensiveness of the test taker. Finally, the F-K Index is a useful measure of "faking good" when scores are positive or "faking bad" when scores are negative.

The Personality Assessment Inventory has eight measures of validity. Among the most widely used are the Negative Impression Scale (NIM) scale, the Positive Impression (PIM) scale and the Infrequency (INF) scale. A high score on the NIM scale suggests that the respondent attempted to portray himself or herself in an especially negative manner. An elevated score on the PIM scale indicates that the test taker attempted to portray himself or herself as exceptionally free of common human shortcomings. A high score on the INF scale suggests the test taker completed the instrument in an atypical way, possibly due to random responding, indifference, carelessness, confusion, or reading difficulties.

Validity Scales of the MMPI-2:

F Scale
F_p Scale
L Scale
K Scale

Validity Scales of the PAI:

NIM
PIM
INF

Plain English Psychological Testing Glossary

Achievement Test: An objective test that measures educationally relevant skills or knowledge.
Adaptive Testing: A sequential form of individual testing in which successive items in the test are chosen based primarily on the psychometric properties and content of the items and the participant's response to previous items.
Age Equivalent: The chronological age population for which a given score is the average score.
Age Norms: Values representing the typical or average performance of people in given age groups.
Alternate Forms: Two or more versions of a test that are considered interchangeable in that they measure the same characteristics, are intended for the same purposes, and are administered using the same directions.
Bias: Statistical bias is a systematic error in a test score; fairness bias may refer to the inappropriateness of content in the assessment instrument, either in terms of its irrelevance, overemphasis, exclusion, under representation, or irrelevant components in the test. Fairness bias usually favors one group of participants over another.
Calibration: The process of setting a test's scores, includes mean, standard deviation, so that the scores on a scale have the same relative meaning as scores on a related scale or test.
Ceiling: The upper limit of ability that can be measured by a test.
Classification Accuracy: The degree to which neither false positive nor false negative results occur when a test is used to classify an individual or event.
Composite Scores: A score that combines several scores by a specified formula.
Construct: Any characteristic that cannot be observed but rather is measured through indirect methods (e.g., intelligence, motivation).
Construct Equivalent: The extent to which the construct measured by one test is essentially the same as the construct measured by another test (see construct).
Construct Irrelevance: The extent to which test scores are influenced by factors that are irrelevant to the construct that the test is intended to measure. Such extraneous factors distort the meaning of test scores that are obtained.
Construct Under-Representation: The extent to which a test fails to capture important aspects of the construct that it is intended to measure.
Correlation: A statistical index used to represent the strength of a relationship between two factors, how much and in what way those factors vary, and how well one factor can predict the other.
Correlation Coefficient: The statistic or number representing the degree to which two or more characteristics are related ranging from +1 to –1. A correlation of 1 indicates that they are completely related while a correlation of –1 means that they are inversely related.
Criterion Referenced Test: A measure of specific criteria or skills in terms of absolute levels of mastery. The focus is on performance of an individual as measured against a standard or criteria rather than against performance of others who take the same test, as with norm-referenced tests.
Cut Score: A specific score on a scale, such that the scores at or above that point are interpreted differently from the scores below that point.
Domain Referenced Test: Test in which performance is measured against a well defined set of tasks or body of knowledge.
External Validity: The extent to which the results of a test can be generalized or extended to other situations.
Face Validity: The degree to which a test *appears* to measure that which it is intended to measure. Tests with high face validity are usually easy to fake.
False Negative: A situation in which an individual is assessed or predicted not to have a characteristic but, in truth, does have it.
False Positive: A situation in which an individual is assessed or predicted to have a characteristic but, in truth, does not have it.
Grade Equivalent: The estimated grade level that corresponds to a given score.
Internal Consistency: A measure of the degree to which a sample of data is representative of the whole.
Normal Distribution: The scores of a sample or population that, when graphed, have most scores in the middle and equal but fewer occurrences of high and low scores. The mean, median, and mode are all equal in perfect normal distribution.
Norms: Standards derived from giving a test to a sample of people similar to those who will take the test and that can be used to interpret the scores of future test takers.
Objective Testing: A style of testing that measures the participant's knowledge of objective facts.
Percentile: The percent of people in the norming sample whose scores were below a given score.
Probability of Error: The likelihood than an error caused the results that were obtained. This is usually small in well-designed tests.
Projective Test: Test that requires an individual to respond to ambiguous stimuli, revealing their tendencies and biases. The Rorschach Ink Blots is an example of a projective test.
Raw Score: The unadjusted score on a test, often determined by counting the number of correct answers, but more generally a sum or other combination of item scores.
Reliability: The extent to which a test is dependable, stable, and consistent.
Reliability Coefficient: A statistic used to determine or estimate reliability with 1 being perfect reliability and 0 being no reliability. Tests with reliability coefficients less than .7 are subject to great error.
Screening Measure: A fast, cost-efficient measurement for a large population to identify individuals who may deviate in a specified area. Screening measures often have high rates of false positive and false negative results.
Standard Deviation: A statistical measure of the variability of results; the higher the standard deviation, the greater the spread of data.
Standard Error of Measurement: The probability that a particular score is an accurate representation of an individual's ability or characteristic.
Standard Error of the Mean: An estimation of the amount of error in a test.
Standardization: The process of making a test the same for everyone so that the results can be compared to each other.
Standard Score: A score that indicates the amount of deviation from a population mean.

***MOST RELEVANT CONCEPTS HIGHLIGHTED IN RED**

(Continued on page 4)

Plain English Psychological Testing Glossary

(Continued from page 3)

Standardized Test: A form of measurement that has been normed against a specific population; standardization is obtained by administering the test to a given population and then calculating means, standard deviations, standardized scores, and percentiles. Equivalent scores are then produced for comparisons of an individual score to the norm group's performance.

Stanine: One of the steps in a nine point scale of standard scores with the mean being stanine 5.

T-Score: A standard score that sets the mean to 50 and the standard deviation to 10. Therefore, a T-score of 70 would be two standard deviations above the mean of the sample population.

Test Bias: An undesirable characteristic of tests in which item content discriminates against certain individuals on the basis of socioeconomic status, race, ethnicity, or gender.

Validity: The extent to which a test measures what it is purported to measure and the extent to which inferences and actions made on the basis of test scores are accurate.

Validity Coefficient: A statistic ranging from 0 to 1 which indicates how well a test measures a particular characteristic. Tests with validity below 0.3 are questionable.

Variability: The degree to which a distribution of scores varies around the mean. High variability means scores are spread wider apart and low variability means scores are relatively close together.

Variance: A measure of spread within a distribution (the square of the standard deviation).

Weighted Scoring: A method of scoring a test in which the number of points awarded for a correct (or diagnostically relevant) response is not the same for all items within the test.

Z Score: A standard score that sets the mean to 0 and the standard deviation to 1.

We're on the web: http://www.KaplanCC.com

Widely Used Psychological Tests

- **Beck Depression Inventory-II (BDI-II)**: Developed for the assessment of symptoms that correspond to DSM-IV criteria for Major Depressive disorder. This is a face-valid test on which 96% of normal individuals can fake depression.
- **Brief Battery for Health Improvement 2 (BBHI-2)**: A screening measure of psychological factors involved in pain and physical impairment. Useful for detecting malingered pain.
- **Halstead-Reitan Neuropsychological Test Battery**: Evaluates cognitive, intellectual, and perceptual processes.
- **Minnesota Multiphasic Personality Inventory-2 (MMPI-2)**: A broad-based test designed to assess a number of the major patterns of personality and psychological disorders. It is the most widely administered psychological test.
- **Millon Clinical Multiaxial Inventory-III (MCMI-III)**: Measures fourteen personality disorders and ten clinical syndromes for adults undergoing assessment or treatment. It has been thought to over-pathologize its subjects.
- **Personality Assessment Inventory (PAI)**: An objective measure of personality and psychological symptoms that uses the latest advances in test construction.
- **Rorschach Ink Blots**: Allows for an assessment of basic personality and psychopathology. Test scores are based on responses to ten ink blots. Only scoring based on the Exxner system has any empirical validity.
- **Structured Interview of Reported Symptoms (SIRS)**: The best test available for detecting malingered psychological symptoms.
- **Substance Abuse Subtle Screening Inventory-3 (SASSI-3)**: Designed to identify individuals who have a high probability of having a chemical dependency problem, particularly those trying to conceal it.
- **Test of Memory Malingering (TOMM)**: Aids in discriminating between those with legitimate memory impairment and malingerers.
- **Trauma Symptom Inventory (TSI)**: Designed to evaluate acute and chronic posttraumatic symptoms.
- **Wechsler Adult Intelligence Scale-Third Edition (WAIS-3)**: Assesses the intellectual ability of adults and gives IQ scores.