EDS 245: Psychology in the Schools

Stephen E. Brock, Ph.D., NCSP

Determining the Validity of Psychological Tests

Lecture Overview

- Introduction
- Reliability
  - Test-Retest
  - Parallel-Forms
  - Internal Consistency
  - Inter-rater
  - Standard Error of Measurement
  - Situation Specific Reliability

Lecture Overview

- Validity
  - Content
  - Criterion-Related
    - Concurrent
    - Predictive
  - Construct
  - Situation Specific Validity

Introduction

- A valid test necessarily generates reliable data.
- However, a reliable test is not necessarily valid.
- Why?

Introduction

- **Reliability**
  - The one word that best describes the concept of reliability is "*consistency*."
    - When used appropriately, the scores produced by a reliable psychological test will **consistently** tend to yield similar estimates of functioning.
    - Reliability coefficient (a correlation) indicates degree of relationship between two scores from the same test.
      - For most psycho-educational tests a correlation of **.80** or higher is acceptable.

Introduction

- **Reliability**
  - The one word that best describes the concept of reliability is "*consistency*."
    - Refers to the results (data obtained), not to the test itself.
    - Types of reliability statistics
      - Test-Retest
      - Parallel Forms
      - Internal Consistency
      - Inter-Rater

Introduction

- **Validity**
  - The one word that best describes the concept of validity is "*accuracy*."
    - When used appropriately, the scores produced by a valid psychological test will **accurately** reflect the psychological construct the test purports to measure.

Introduction

- **Validity**
  - The one word that best describes the concept of validity is "*accuracy*."
    - Tests are valid only for a specific purpose and within a particular context (e.g., as a measure of "achievement" for students who have received formal education).
    - Three principle types of validity are:
      - Content
      - Criterion-Related
      - Construct

Introduction

- A valid test necessarily generates reliable data.
- However, a reliable test is not necessarily valid.
- Why?

Test-Retest Reliability

- When separated by a period of time, two administrations of the same test yield consistent results.
  - Determined by correlating the two sets of scores.
  - Reflect stability of test scores over time.
  - Test-retest reliability coefficients that approach 1.0 (a perfect correlation) reflect a high degree of reliability.
  - Reliability coefficients tend to decrease as the interval between administrations increase. Thus, the interval between administrations should be considered when evaluating this type of reliability.

Parallel-Forms Reliability[1]

- When two equivalent forms of the test are available, administrations of the different forms of the test yield consistent results.
  - Determined by correlating the two sets of scores.
  - Reflect equivalence across forms.
  - Parallel-forms reliability coefficients (AKA "coefficient of equivalence") that approach 1.0 reflect a high degree of reliability.

[1]AKA Equivalent or Alternate Forms Reliability.

Internal Consistency Reliability

- When a measure is divided into parts, and the different parts of the test yield consistent results.
  - Determined by correlating sets of scores generated by different parts of the same test administration.
  - Reflect the degree to which items within the same test is measuring one construct.
  - Split-half reliability, K-R 20, or Cornbach's alpha correlation coefficients that approach 1.0 reflect a high degree of reliability.

Inter-rater Reliability

- When different examiners score the same test, and the different examiners' efforts yield consistent results.
  - Determined by correlating sets of scores generated by different examiners' scoring of the same test administration.
  - Reflect the degree to which different examiners can reliably obtain the same test results.
  - Inter-rater reliability correlation coefficients that approach 1.0 reflect a high degree of reliability.

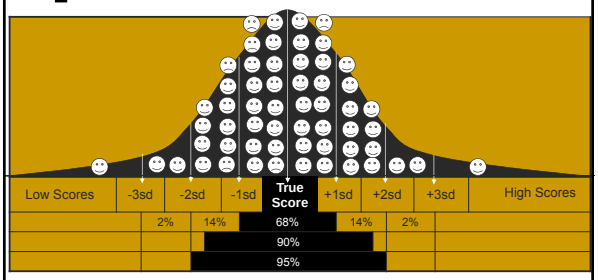
Situation Specific Reliability

- A test may be reliable for a population, but not for a specific testing subject.
  - Scores can be affected by idiosyncratic examiner/examinee factors.
    - Examinee is uncooperative, anxious, tired.
    - Examiner is incompetent
    - A Psycho-educational evaluation should always address this issue typically within a section labeled "Test-taking Behavior". For example…
      - "Cathi readily accompanied the examiner to the testing room and rapport appeared to be adequate. Rapport was... Level of activity and verbalizations were ... Her reaction to failure was ... Encouragement and praise resulted in ... Cathi's effort was consistent./inconsistent. Results are considered a reliable reflection of her present level of functioning."

Standard Error of Measurement

- No test is 100% reliable. All psychological test results are associated with some degree of measurement error.
- The standard error of measurement (SE_m) is an estimate of this error.
- SE_m is directly related to a test's reliability coefficients. Large SE_m scores are associated with relatively poor reliability and visa versa.
- SE_m is the standard deviation of the distribution of error scores.

SE_m: The Distribution of Obtained Scores



Psychometric theory suggests that if an individual was given the same test multiple times, these *obtained* scores would cluster around the true score. These *obtained* scores would be normally distributed and form a normal curve.

Standard Error of Measurement

- SE_m is obtained by multiplying the standard deviation of the test by the square root of 1 minus the reliability coefficient (r_{xx}) of the test.
  - $SE_m = SD \sqrt{1 - r_{xx}}$
    - For example, assume that a reading achievement test with a mean of 100 and a standard deviation of 15 has an internal consistency reliability coefficient of .96,
  - $15\sqrt{1 - .96} = 15\sqrt{0.04} = 15(0.2) = \mathbf{3} = SE_m$

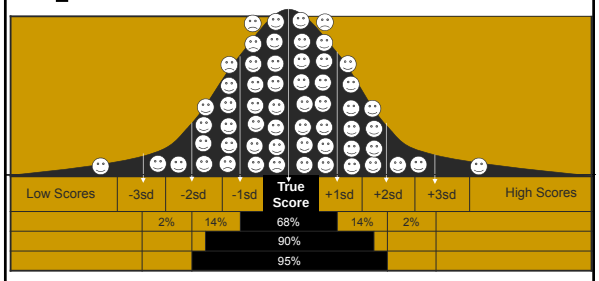
Standard Error of Measurement

- SE_m can be used to determine confidence intervals (CI).
  - A CI allows a statement to be made about the range within which the testing subject's true score falls.
- 68%, 90%, and 95% CI are typically used.
  - A 68% CI provides the range of scores within with a testing subject's true score lies 68% of the time. In other words, only 32 times out of 100 will the true score fall outside of this range.
  - A 90% CI provides the range of scores within with a testing subject's true score lies 90% of the time. In other words, only 10 times out of 100 will the true score fall outside of this range.
  - A 95% CI provides the range of scores within with a testing subject's true score lies 95% of the time. In other words, only 5 times out of 100 will the true score fall outside of this range.

Standard Error of Measurement

- The formula for a confidence interval is as follows:
  - CI = obtained test score ± $z(SE_m)$
    - The "z" in this formula refers to the *z* score obtained from a normal cure table.
    - For example, the 95% CI for an reading achievement test scaled score of 99 for our test with a SE_m of 3 is 99 ± 1.96(3).
    - 1.96 time 3 equals 5.88.
    - Rounded up to six, we can say that we are 95% confident that the student who obtained the test standard score of 99 has a true score falling in the range 93 to 105 (99 ±6).
    - In a psycho-educational report these data might be presented as follows:
      - "On this measure Jimmy obtained an standard score of 99±6. The chances are 95 out of 100 that Jimmy's true reading achievement falls in the range of scores 93 to 105. These data are well within the average range. Thus, it can be concluded that Jimmy's reading achievement is typical of children his age in this tests standardization sample."

SE_m: The Distribution of Obtained Scores



Psychometric theory suggests that if an individual was given the same test multiple times, these *obtained* scores would cluster around the true score. These *obtained* scores would be normally distributed and form a normal curve.

Activity

1. The *Excellent Intelligence Test* (EIT) has an internal consistency reliability coefficient of .95.
2. This test also has an standard deviation of 15
3. What is the SE_M for this measure?
4. Compute the 90, 95, and 99% CIs.
   1. 90%, z = 1.65
   2. 95%, z = 1.96
   3. 99%, z = 2.58

$SE_M = SD\sqrt{1-r}$

$CI = \pm Z(SE_M)$

Activity

1. The *Adequate Intelligence Test* (AIT) has an internal consistency reliability coefficient of .90.
2. This test also has an standard deviation of 15
3. What is the SE_M for this measure?
4. Compute the 90, 95, and 99% CIs.
   1. 90%, z = 1.65
   2. 95%, z = 1.96
   3. 99%, z = 2.58

$SE_M = SD\sqrt{1-r}$

$CI = \pm Z(SE_M)$

Activity

1. The *Marginal Intelligence Test* (MIT) has test-retest reliability coefficient of .80.
2. This test also has an standard deviation of 15
3. What is the SE_M for this measure?
4. Compute the 90, 95, and 99% CIs.
   1. 90%, z = 1.65
   2. 95%, z = 1.96
   3. 99%, z = 2.58

$SE_M = SD\sqrt{1-r}$

$CI = \pm Z(SE_M)$

Activity

1. Interpret the IQ score of 91 for all three IQ tests
   - EIT = 90% CI of ± 5 (SEM = 3)
   - AIT = 90% CI of ± 8 (SEM = 5)
   - MIT = 90% CI of ± 12 (SEM = 7)

Interactive normal distribution

$SE_M = SD\sqrt{1-r}$

$CI = \pm Z(SE_M)$

Validity

- The degree to which a test **accurately** measures what it is supposed to measure and, **consequently**, permits appropriate interpretations.
- Valid (or *accurate*) tests are always reliable (or *consistent*) tests.
- Valid for specific purposes and populations.
- A matter of degree.

Validity

- A psycho-educational report should always include a statement regarding validity. Typically this statement precedes the listing of psychological procedures.
- For example:
  - "Children with Cathi's characteristics are represented in the tests standardization samples. In addition, the measures administered have been validated for the purposes for which they were used."

Content Validity

- Degree to which the test measures the intended content area.
- Includes both item (item relevance to content area) and sampling (sample of total content area) validity.
- Determined by logic/expert judgment.
  - e.g., the content validity of a science test would be determined by a group of experienced science teachers

Criterion-Related Validity

- The degree to which a test (the predictor) correlates with a second measure (the criterion).
  - **Concurrent Validity**. Both measures are administered in the same time frame. How well the measure reflects current functioning.
    - e.g., the correlation between 7th science test results and student grades given by their 7th grade science teacher.
  - **Predictive Validity**. Both tests area administered at different times. How well the measure predicts future performance.
    - Should be at least .60 for IQ tests.
    - e.g., the correlation between 7th grade science test results and student grades given by their 8th grade science teacher.

Construct Validity

- The extent to which the test reflects the construct it is intended to measure. It requires a series of studies (including studies to determine content and criterion-related validity research), and is the most important form of validity.
- Does the test measure what it is supposed to measure.
  - e.g., the 7th grade science test positively correlates with other 7th grade science achievement test results (*convergent validity*). In addition, the science test correlates to a higher degree with other science tests than it does with tests of other academic areas (*discriminate validity*).

Situation Specific Validity

- A test may be valid for a population, but not for a specific testing subject.
  - Test related factors
    - Test taking skill
    - Anxiety
    - Motivation
    - Speed
    - Understanding of directions
    - Item/format novelty
    - Examiner/examinee rapport
    - Physical handicaps
    - Bilingualism
    - Educational exposure
    - Important examinee characteristics not present in the test normative sample

Evaluating Tests

- Test coverage and use
  - There must be a clear statement of recommended **uses**.
  - There must be a clear description of the **population** for which the test is intended.

From Rudner, L. M. (1994). Questions to ask when evaluating tests. *Practical Assessment, Research & Evaluation, 4*(2). Retrieved November 21, 2002, from http://pareonline.net/getvn.asp?v=4&n=2

Evaluating Tests

- Appropriate samples for test validation and norming
  - The samples used for test validation and norming must be of **adequate size** and must be sufficiently representative to substantiate validity statements, to establish appropriate norms, and to support conclusions regarding the use of the instrument for the intended purpose.

From Rudner, L. M. (1994). Questions to ask when evaluating tests. *Practical Assessment, Research & Evaluation, 4*(2). Retrieved November 21, 2002, from http://pareonline.net/getvn.asp?v=4&n=2

The Importance of Checking Norms

- The original norms for the Halstead-Reitan tests are not well founded. Halstead's "normal" population consisted of 29 subjects (8 women) and 30 sets of scores. Ten of these subjects were servicemen who became available for Halstead's study because the were under care for 'minor' psychiatric disturbances. One was awaiting sentencing for a capital crime (in the state at that time it could have been either life imprisonment or execution. Halstead notes that the subject appeared "anxious"). Four were awaiting lobotomies because of behavior threatening their own life and/or that of others. Two sets of scores were made by one subject, a young man, since he was still waiting at the hospital after two months and so took the test again. This is the group whose test performance defined the unimpaired range for the cutting scores in general use with the Halstead tests. (Bolls 1981)

Evaluating Tests

- Reliability
  - The test is sufficiently reliable to permit stable estimates of the ability levels of individuals in the target group.
- Content Validity
  - Content validity refers to the extent to which the test questions represent the skills in the specified area.

Adapted from Rudner, L. M. (1994). Questions to ask when evaluating tests. *Practical Assessment, Research & Evaluation, 4*(2). Retrieved November 21, 2002, from http://pareonline.net/getvn.asp?v=4&n=2

Evaluating Tests

- Criterion Validity
  - The test adequately predicts performance.
- Construct Validity
  - The test measures the "right" psychological constructs.

Adapted from Rudner, L. M. (1994). Questions to ask when evaluating tests. *Practical Assessment, Research & Evaluation, 4*(2). Retrieved November 21, 2002, from http://pareonline.net/getvn.asp?v=4&n=2

Evaluating Tests

- Test Administration
  - Detailed and clear instructions outline appropriate test administration procedures.
- Test Reporting
  - The methods used to report test results, including scaled scores, subtests results and combined test results, are described fully along with the rationale for each method.

From Rudner, L. M. (1994). Questions to ask when evaluating tests. *Practical Assessment, Research & Evaluation, 4*(2). Retrieved November 21, 2002, from http://pareonline.net/getvn.asp?v=4&n=2

Evaluating Tests

- Test and Item Bias
  - The test is not biased or offensive with regard to race, sex, native language, ethnic origin, geographic region or other factors.

From Rudner, L. M. (1994). Questions to ask when evaluating tests. *Practical Assessment, Research & Evaluation*, *4*(2). Retrieved November 21, 2002, from http://pareonline.net/getvn.asp?v=4&n=2

Questions?

Next Meeting (12/7/16): The Future of School Psychology

Readings:
1. Canter (2007,June)

Activity:
Review final exam study guide