EDS 245: Psychology in the Schools

Stephen E. Brock, Ph.D., NCSP

Test Interpretation

Outline

- Types of test scores and what they mean
  - Raw Scores
  - Developmental (Age/Grade) Scores
  - Rank Scores
  - Standard Scores
  - Criterion Referenced Scores
- Standard error of measurement
- Activity

Raw Scores

- The number of correctly answered test items.
- Limits to interpretation?
  - Cannot be compared to other individuals (normative groups)
  - Cannot be compared to other test scores (relative strengths and weaknesses)
- Can show small changes over time

Developmental Scores

- Scores based on performance of others.
- Translates raw scores into age or grade equivalents.
  - Requires knowing the average raw score of children at different ages or grade levels.
  - Tells you that the child performs as well on the test as did the average _N_ year-old/grader who took the test.

Developmental Scores

- Strength
  - Easy to understand
- Weaknesses
  - Misleading
    - Is not the same as actually being at a given grade/age level.

Developmental Scores

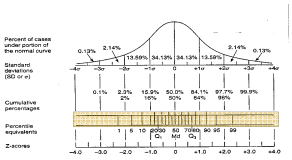
- "Steve obtained a math fact grade score of 6.9 (end of the sixth grade) on the WJIII. This means that even though he is in the 4th grade, he is ready for 7th grade math."
- True or False? Discuss. Interpret 6.9.

Rank Scores

- Scores based on performance of others.
- Percentile
  - The most commonly used rank score
  - The percentage of persons in the standardization sample who fall below a given raw score
    - You can't have a percentile score of 100 as you can exceed your own score
  - Scores of students are arranged in rank order from lowest to highest
  - The scores are divided into 100 groups.

Rank Scores

- Percentile Strength
  - Easy to understand
- Percentile Weakness
  - Unequal units, especially at the extreme ends of the distribution.

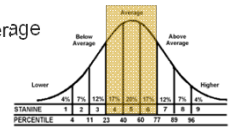


Rank Scores

- Decile ranks ("Deca" means "ten)
  - Scores of students are arranged in rank order from lowest to highest.
  - Scores are divided into 10 groups or bands (instead of 100 as with percentiles).
- Quartile ranks
  - Scores of students are arranged in rank order from lowest to highest.
  - Scores are divided into 4 equally sized groups or bands (instead of 10 as with deciles or 100 as with percentiles)

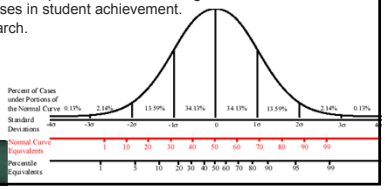
Rank Scores

- Stanine
  - Ranks scores into 9 bands within which fixed percentages of scores fall
    - 1, 2, 3 = below average
    - 4, 5, 6 = average
    - 7, 8, 9 = above average



Rank Scores

- Normal Curve Equivalent
  - Measures where a student falls along the normal curve.
  - Numbers on the NCE line run from 1 to 99, similar to percentile ranks.
  - NCE scores have a major advantage over percentiles in that they can be averaged.
    - This an important characteristic when studying overall school performance, and in particular, in measuring school-wide gains and losses in student achievement.
    - Used in research.

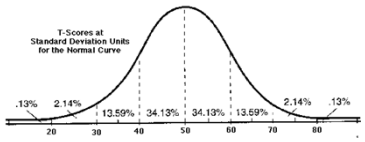


Standard Scores

- Scores based on performance of others.
- Express the how far a student's raw score is from the mean in terms of the standard deviation of the standardization samples score distribution.
- Z-scores
  - A measure of the distance in standard deviations of a score from a mean.

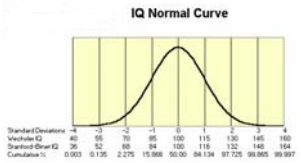
Standard Scores

- T-scores (t = 10**z** + 50)
  - A Z score with a mean of 50 and a standard deviation of 10.



Standard Scores

- Deviation IQ scores (IQ = 15**z** + 100)
  - A Z Score with a mean of 100 and a standard deviation of 15



Standard Scores

- Terman's Original IQ Classifications

IQ Range	Classification
140 and over	Genius or near genius
120-140	Very superior intelligence
110-120	Superior intelligence
90-110	Normal or average intelligence
80-90	Dullness
70-80	Borderline deficiency
Below 70	Feeble-mindedness

Standard Scores

- WISC IV Classifications

IQ Range	Classification
130 and above	Very Superior
120-129	Superior
110-119	High Average
90-110	Average
80-89	Low Average
70-79	Borderline
69 and below	Extremely Low

Standard Scores

- Woodcock's Classifications

IQ Range	Classification
Above 130	Very Superior
121-130	Superior
111-120	High Average
90-110	Average
80-89	Low Average
70-79	Low
Below 70	Very Low

Standard Scores

- Cautions when interpreting IQ test scores
  - http://www.wilderdom.com/intelligence/IQCautionsInterpretingIQ.html

Explaining Scores to Parents

- http://www.teachersandfamilies.com/open/parent/scores2.cfm

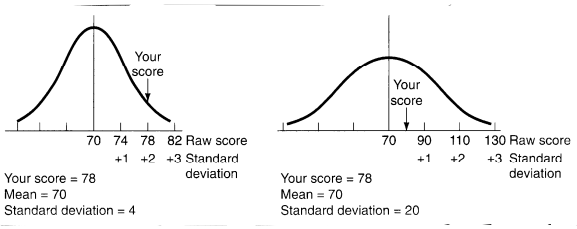
Criterion Referenced Scores

- Scores not based on performance of others.
- Percent correct
  - Percentile rank is often confused with these scores
- Grades
  - When not comparing to the performance of others

Error in Test Scores

- Error is always present in scores
  - Raw score = true score + error
- The goals are to
  - Reduce error
  - Account for error
- Reliability and validity reduce error
- Standard Error of Measure accounts for error

The Effect of Variance on Score Interpretation



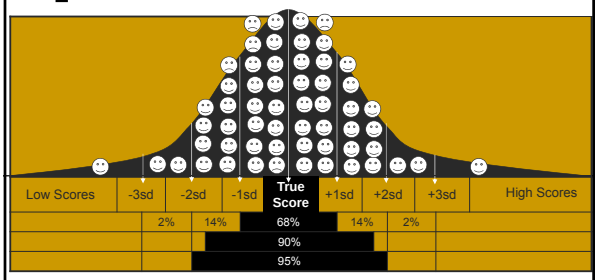
Standard Error of Measurement (SE_m)

- Estimate of the amount of variance in an obtained test score.
- How much is the person's score likely to differ from the true score?
- Formula
  - SEM=(SD) x square root of $(1 - r_{xx})$
  - r_{xx} a reliability coefficient

SE_m

- 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,
  - $15\sqrt{1 - .96} = 15\sqrt{0.04} = 15(0.2) = \mathbf{3} = SE_m$

Review

1. *Is it possible to construct a psycho-educational test that generates an individual's true score?*
2. *On a psycho-educational test, what is the "obtained" score?*
3. *Hypothetically, how would one determine the true test score for an individual?*
4. *What is the standard error of measurement* (SE_M)?
5. *How is the* SE_M *determined?*
6. *What can you say about a psycho-educational test that has a large* SE_M

Confidence Intervals

- 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.
  - It sets bands around obtained scores.
  - Helps to identify the range of scores within which the true score lies (vs. those results caused by factors other than the trait being assessed, e.g., good or bad luck).

Confidence Intervals

- 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.

Confidence Intervals

- The formula for a confidence interval is as follows:
  - CI = obtained test score $\pm$ $z(SE_m)$
    - The "z" in this formula refers to the z score obtained from a normal cure table.
    - Sample Normal Curve 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 $\pm$ 1.96(3).
    - 1.96 times 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 $\pm$6).

Confidence Intervals

- 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 test's standardization sample."

Wide Range Assessment of Memory and Learning

SUMMARY DATA

	SUM OF SCALED SCORES	INDEX	CONFIDENCE INTERVAL 68% 95%	PERCENTILE RANK
VERBAL MEMORY	23	103	97 - 113	58
VISUAL MEMORY	16	88	80 - 96	21
ATTENTION/CONCENTRATION	24	112	105 - 119	79
GENERAL MEMORY	22	101	— - —	53
SCREENING MEMORY			— - —	
WORKING MEMORY	20	100	92 - 108	50
VERBAL RECOGNITION	23	102	99 - 117	55
VISUAL RECOGNITION	23	109	101 - 117	73
GENERAL RECOGNITION	26	111	— - —	77

Bringing it all Together

- Let's start from the very beginning and bring it all together to ensure we understand SE_M and Confidence Intervals.

Bringing it all Together

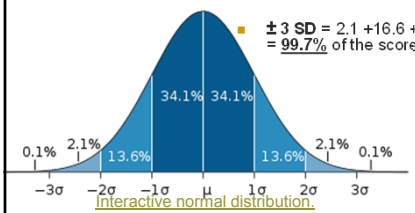
- Hypothetically, any psychological test variable measured enough times will be normally distributed.
- In other words, the obtained test scores will be normally distributed.
  - What does it mean to say that a variable (or test scores) is "normally distributed?
  - Answer is on the next slide

Bringing it all Together

- Mean, Median, and Mode are the same.
  - 50% of scores are above the mean
  - 50% of scores are below the mean
- ± 1 SD = 68% (most) of the obtained scores
- ± 2 SD = 95% (almost all) of the obtained scores
- ± 3 SD = 99.7% (virtually all) of the obtained scores

Bringing it all Together

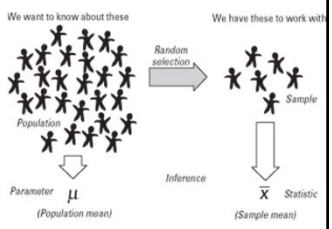
- **± 1 SD** = 34.1 + 34.1 = **68.2%** of the scores
- **± 2 SD** =1 3.6 + 34.1 + 34.1 +1 3.6 = **95.4%** of the scores
- **± 3 SD** = 2.1 +16.6 +34.1 + 34.1 +13.6 + 2.1 = **99.7%** of the scores



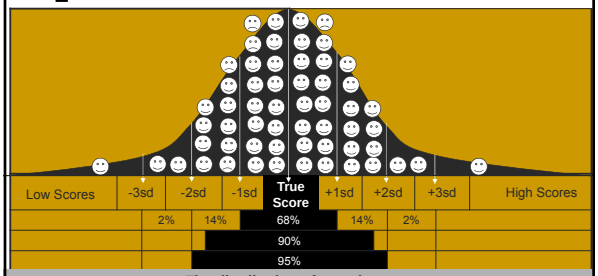
Interactive normal distribution.

Bringing it all Together

- With a large enough sample size the scores of a psychological test should be normally distributed.
  - We can use the test to tell us how far above or below the mean and individual student is (in terms of standard deviations) on the variable (e.g., IQ) measured by the test.

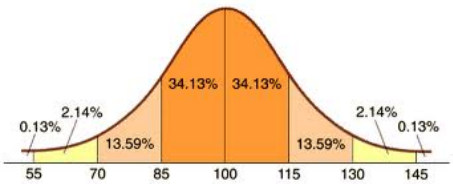


Bringing it all Together



The distribution of sample scores

Bringing it all Together

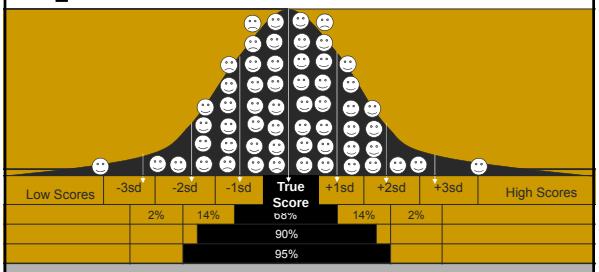


If the sample used to standardize the IQ test (and estimate population performance) is large enough we can expect the scores will be normally distributed. With most IQ tests the Z-scores are transformed such that Mean = 100 and SD = 15 standard score points.

Bringing it all Together

- If we could give the same test to an individual student over-and-over again the obtained scores should be normally distributed.
  - The hypothetical **true score is a constant**.
  - The **obtained test score** (the IQ score of the student you test) is **true score + error.**
  - The error associated with obtained scores is **random**.
  - Thus, the error scores should also be **normally distributed**.

Bringing it all Together



The Distribution of Error 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.**

Bringing it all Together

- Standard Error of Measurement (SE_M) allows the psychologist to account for the error always associated with a given obtained test score.
- It can be used to develop confidence intervals.

Bringing it all Together

- SE_M is obtained by multiplying the standard deviation (SD) of the test by the square root of 1 minus the reliability coefficient (r) of the test.

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

- What two statistics are needed to determine the SE_M?

Bringing it all Together

- $SE_M = SD\sqrt{1-r}$
  - Standard deviation for most IQ tests is 15.
  - But what is *r*?
    - SE_M is a reflection of the degree to which a test consistently yields the same results for an individual (i.e., that the test itself is reliable).
      - What is the Alternate Form Reliability correlation coefficient and how is it determined?

Bringing it all Together

- As part of the test development process, estimates of the tests reliability (or consistency) are determined.
  - For example, alternate form reliability would correlate individuals' scores on separate versions of the same IQ test.

Form A	Form B
101	102
100	100
98	96
115	114
88	88
120	122
61	65
99	98
95	95
…	…
106	106

r = .90

Bringing it all Together

- We now have the information we need to determine SE_M.
  - For an IQ test with a standard deviation of 15 and an alternate form reliability coefficient of .90
    - $15\sqrt{1-.9} = 15\sqrt{.1} = 15(0.316) = 4.743$
    - **5** = SE_m
  - We can now develop confidence intervals

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

Bringing it all Together

- A CI accounts for measurement error and allows a statement to be made about the range within which the testing subject's true score falls.
- It sets bands around obtained scores.

Bringing it all Together

- A CI helps to identify the range of scores within which the true score lies
  - vs. those results caused by factors other than the trait being assessed or error, e.g., sick or unusually attentive, good or bad luck.

Bringing it all Together

- The formula for a CI is as follows:
  - obtained test score ± z(SE_M)
    - The "z" in this formula refers to the *z* score obtained from a normal cure table.
    - Sample normal curve table.
      - 68% of error scores are accounted for by z = ±1
      - 90% of error scores are accounted for by z = ±1.65
      - 95% of error scores are accounted for by z = ±1.96
      - 99% of error scores are accounted for by z = ±2.58

Bringing it all Together

- CI = obtained test score ± z(SE_M)
  - For example, the 90% CI for an IQ test score of 105 for our test with a SE_M of 5 is 105 ± 1.65(5).
    - 1.65 times 5 equals 8.25.
    - Rounded down to 8, we can say that we are 90% confident that the student who obtained the IQ test score of 105 has a true IQ score falling in the range 98 to 112 (105±8).
    - For a test with a SE_M of 5 and a SD of 15, the range of IQ scores 98 to 112includes 90% of the obtained scores (i.e., true score + error) a student might be expected obtain on this test.

Test Bias

- Psychometric definition
  - A test with systematic differences in the meaning of test scores associated with group membership.
  - People from two groups who have the same observed score do not have the same standing on the trait of interest.
  - A test to predict some criterion of interest results in systematic over- or under- prediction based on group membership.

Possible Bias in Traditional Cognitive Measures

- Inadequate norms
- Testing formats
- Product-oriented response
- Unfamiliar with information tapped
- Lack of relationship between test and classroom
- Test-taking skills
- Trait measurement and stability

Cultural Considerations: Assessment

- Culture and value orientation
- Acculturation
- Culture and communication styles

QUIZ

1. The midpoint of a range of scores is the _____.
2. The _____ represents the most frequently achieved score within a range of scores.
3. The _____ _____ is a measure for variance in a set of scores.

QUIZ

4. Stanines break ordered scores into __ parts.
5. Most IQ scores have a mean of _____ and a standard deviation of _____.
6. A T-score has a mean of _____ and a standard deviation of _____.

QUIZ

7. What does a z score represent?
8. What is the standard error of measurement (SEM) and how is it helpful in score interpretation?
9. Cornelius' score on the Quick Draw Test placed him at the 35th percentile. What does this mean (in words a parent would understand)?

QUIZ

10. What is the main limitation in the use of percentile ranks?
11. What are primary limitation in the use of grade equivalent scores?
12. Henrietta Goodegg received a z-score of +1.5 on the speeded chicken plucking test. What is her T-score? Is she slower or faster than the average chicken plucker?

QUIZ

13. Red Island 's T-score on the chicken plucking test was 55. Legless Horn's score was 48. The test has a SEM of 5 and a 90% confidence interval of 8. Can we conclude that Red is definitely a faster chicken plucker than Legs? Explain your answer.

Questions?

Next Week (11/26/12): Measures of Validity)

No papers due

Reading ERIC Development Team (1999), and Messic (1990)