

## Basic Test & Measurement Concepts: Scales

- What are the different types of measurement scales?
  - Nominal
  - Ordinal
  - Interval
  - Ratio

5/22/06

These scales are listed on this slide from least to most sophisticated.

**Nominal**—qualitative—least sophisticated—scores used for grouping purposes—group things into categories. For example, department stores—Sears, Dillards, Burdines, Macy’s, Nordstrom.

**Ordinal**—rank ordering—like places in a race—equal intervals usually do NOT exist, meaning that the distance between point 1 and 2 may not be the same as the distance between point 3 and 4. (For example, places in a race, like first place, second place, third place).

**Interval**—equal differences between scores—zero point is arbitrary. No absolute zero. An example is temperature (except Kelvin scale where zero indicates the complete absence of heat)..

**Ratio**—equal distance between increments—and a real or true zero point exists—can apply all math operations with this kind of scaling. An example is weight.

## Basic Test & Measurement Concepts: Central Tendency

- What are the three types of central tendency statistics?
  - Mean
  - Median
  - Mode
- Which one of these three is used most commonly? When would it be most likely for the other two to be reported also?

5/22/06

Mean— mathematical average

Median— score or # in the middle of the distribution of scores

Mode— most common score (the one that occurs most frequently)—have you ever heard someone say, “that was the modal response?”

Mean is most common. The other two are reported often when the data are ordinal or nominal.

## Basic Test & Measurement Concepts: Variability & Distributions

- What are the two common indices of variability? How are they calculated or defined?
  - Range
  - Standard Deviation
- A normal distribution is completely specified when the mean & variance have been given.

5/22/06

Range—subtract the lowest from the highest (not the best reflection of variability however)

SD—an index of the extent to which scores deviate from the mean in a distribution. It is the square root of the average of the sum of the squared distances of each score from the mean.

**You should be able to describe the relation between the standard deviation and means for various distributions when shown pictures of distributions.**

## Basic Test & Measurement Concepts: Variability & Distributions

- In a normal distribution of scores on some human trait, most people earn scores clustered near the \_\_\_\_\_ and few people earn scores near the \_\_\_\_\_ of the distribution.
- Distributions also may be skewed. What does a positively skewed distribution look like? Negatively skewed?

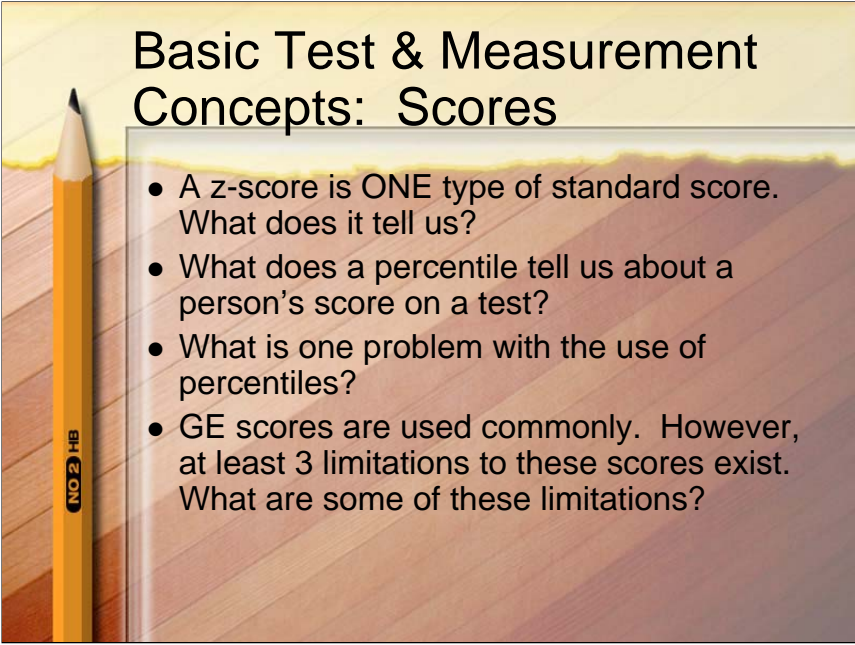
In a normal distribution of scores on some human trait, most people earn scores clustered near the **mean or middle of the distribution** and few people earn scores near the **tails of the distribution**.

Distributions also may be skewed. This is when the data are distributed unevenly—or the distributions are not symmetrical (scores bunch up at one end or the other).

What does a positively skewed distribution look like? The scores are bunched up at the low end of the distribution with just a few high scores.

What does a Negatively skewed look like? The scores are bunched up at the high end of the distribution and there are just a few low scores.

To help you remember, think of the curved shape of the distribution forming an arrow or a pointer (points to the type of skew).



## Basic Test & Measurement Concepts: Scores

- A z-score is ONE type of standard score. What does it tell us?
- What does a percentile tell us about a person's score on a test?
- What is one problem with the use of percentiles?
- GE scores are used commonly. However, at least 3 limitations to these scores exist. What are some of these limitations?

A **z-score** tells us the # of standard deviation units a particular score is above or below the mean (for any normal distribution).

**Percentiles** are the most commonly used and many consider them to be more understandable than standard scores. They tell us the percentage of cases to which a person's score is equal to or better than. Unfortunately, percentiles are often confused with percentages.

**Grade equivalent scores** have a number of limitations despite their common usage. They have unequal units of measurement (ordinal data), but this is often forgotten when they are interpreted. They should not be added or subtracted.

Also, the same score may have different meaning for students of different ages (and on different tests).

They are misinterpreted often!

## Basic Test & Measurement Concepts: Correlations & Reliability

- What does the correlation between two variables tell us?
- What is reliability?
- Why do tests need to be reliable?
- For what is reliability a necessary, but not sufficient, condition?
- What are some of the factors affecting reliability of a test?

5/22/06

**Correlation** describes the degree to which variables are related (e.g., IQ and achievement). Does NOT imply causation.

You should be able to examine graphs of data scatterplots and describe the relation between two variables as either positive, negative, none, or curvilinear.

**Reliability** is the degree of consistency in measurement—same result is given consistently.

Tests need to be reliable to increase our confidence that the obtained score is as close to the “true” score as possible—limit error.

Reliability is a necessary but not sufficient condition for validity.

Factors affecting reliability include: characteristics of the individual (temporary ones, like hunger or lasting—experience taking tests); test length; test-retest interval; guessing; variation in testing situation (break pencil).

## Basic Test & Measurement Concepts: Reliability

- What 4 types of reliability are often calculated with respect to tests? Any disadvantages to these?
- Reliability Standards (Salvia & Ysseldyke)
  - Administrative purposes for groups--a minimum of .60 is suggested.
  - Screening decisions for one student--a minimum of .80 is suggested.
  - Important educational decisions for one student--a minimum of .90 is suggested.

### The Types of Reliability are:

**Test-Retest**—may overestimate if the second test is given too soon, may underestimate if the test retest interval is too long. Inefficient use of time.

**Equivalent form**—for A, B (like versions of the measurement exam we have). Drawbacks of this process— time and interval length decisions and difficulty in getting a truly equivalent form.

**Split-half**—advantage of this approach is that you only have one test session (no retest interval issues), BUT estimates can be inflated because it does not account for any changes in the person from one measurement session to the next.

**Internal consistency**—item to item consistency.

## Basic Test & Measurement Concepts: SEM

- The standard error of measurement is the standard deviation of what?
- Test A & B have identical means and SDs. Test A has a SEM of 4.8 and test B has a SEM of 16.3. Which test is more reliable? Why?
- What is the SEM often used for with respect to test results?

**SEM** is the standard deviation of the distribution of an individual scores around their true score if that individual were tested again and again. **Read the example from the book that is copied into your notes!**

2nd bullet question—Test A. Tests with higher reliability have a lower SEM.

3rd bullet question—often used to construct confidence intervals (68% and 95% are probably the most common).

## Estimated True Score & Confidence Intervals

- The less reliable the test, the greater the difference between "true score" and the obtained score.
- Estimated True Score (ETS) = Mean + ( $r_{xx}$ ) (Obtained Score - Mean).

For the 2nd bullet the  $r_{xx}$  is the tests reliability.

## Estimated True Score & Confidence Intervals

- When the obtained score is below the test mean and the reliability coefficient is less than 1.00, the ETS is *always* larger than the obtained score.
- When the obtained score is above the test mean and the reliability coefficient is less than 1.00, the ETS is *always* less than the obtained score.

## Estimated True Score & Confidence Intervals



- ETS often is used to calculate confidence intervals. Follow these steps:
  - 1. Select the degree of confidence
  - 2. Find the z-score associated with that degree of confidence
  - 3. Multiply the z-score by the SEM (if SEM not given it can be calculated from the reliability)
  - 4. Find the ETS
  - 5. Take the product of the z-score and the SEM, and both add it to & subtract it from the ETS.

Step #5 is  $ETS \pm (z\text{-score})(SEM)$

This is information you will want to keep handy for the future because, while many tests provide you with confidence bands in tables in the test manual, many others do not and you will want to know how to calculate them yourself!

## Basic Test & Measurement Concepts: Validity

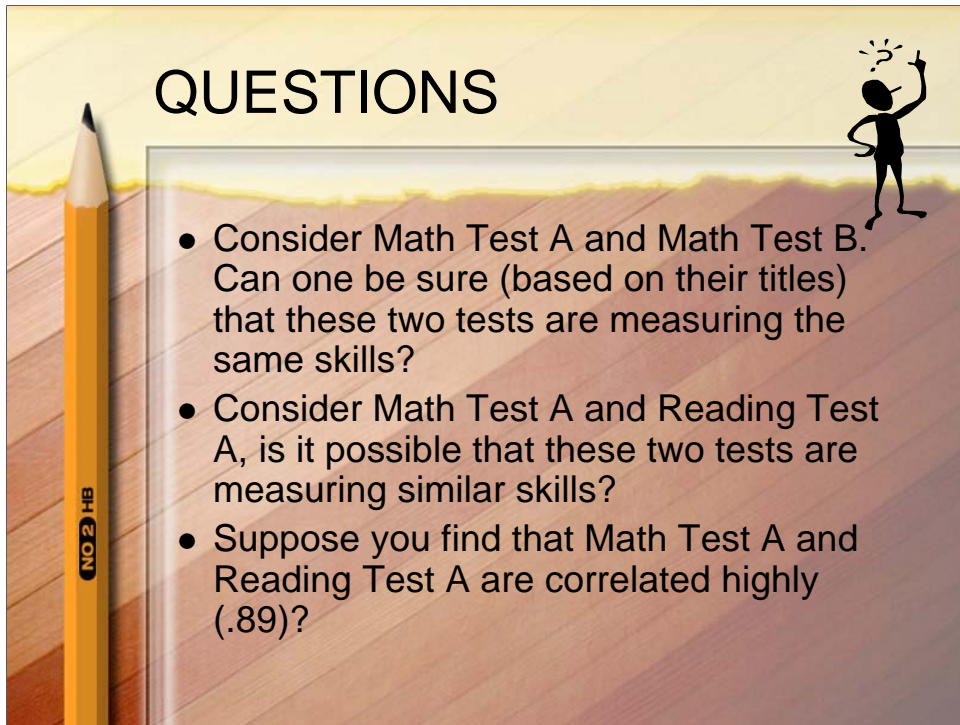
- What does it mean to say that a test is valid? What must validity be tied to?
- What are the different types of validity?
- Who's responsibility is it to make sure that a test is reliable & valid for any specific purpose?

What does it mean to say that a test is valid? **That it measures what it is supposed to measure.** What must validity be tied to? **The purpose for which the test is intended (what question you are trying to answer or decision you are trying to make).**

What are the different types of validity? **Content—adequate sample. Criterion-related (either predictive or concurrent). And Construct—capability of the test in measuring some trait or construct.**

Who's responsibility is it to make sure that a test is reliable & valid for any specific purpose? **The Test USER!**

# QUESTIONS



- Consider Math Test A and Math Test B. Can one be sure (based on their titles) that these two tests are measuring the same skills?
- Consider Math Test A and Reading Test A, is it possible that these two tests are measuring similar skills?
- Suppose you find that Math Test A and Reading Test A are correlated highly (.89)?

## Answers to Questions...

1. No, this cannot be guaranteed.
2. Yes
3. Still could be measuring different things—there could be a third intermediary variable—not reading or math, but visual acuity for example.

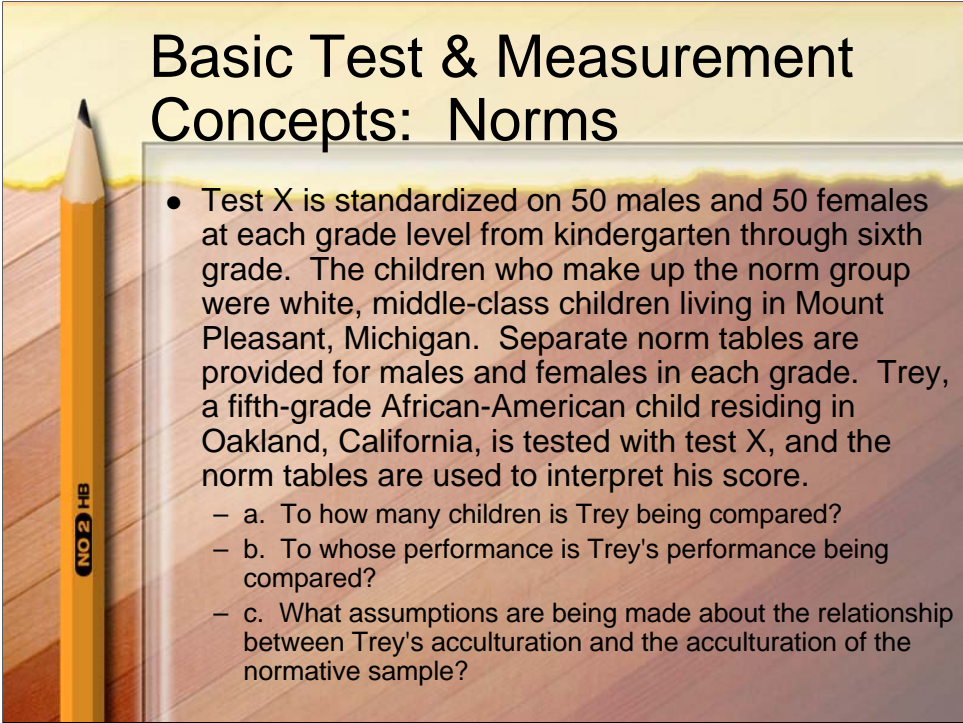
## Basic Test & Measurement Concepts: Norms

- What are the criteria for determining if a test's norms are adequate?
- In a the normative information provided in test manuals, what information often is missing? Why would this information be important to know?
- What would your primary concern be with the use of a poorly normed test?

What are the criteria for determining if a test's norms are adequate?  
**Representativeness and how recent they are (this idea is tied to representativeness).**

In a the normative information provided in test manuals, what information often is missing? **Ages of the sample.** Why would this information be important to know? **Want to know if norms are extrapolated or interpolated or not.**

What would your primary concern be with the use of a poorly normed test?  
**Misinterpretation of the results. Cannot make accurate or meaningful interpretations.**

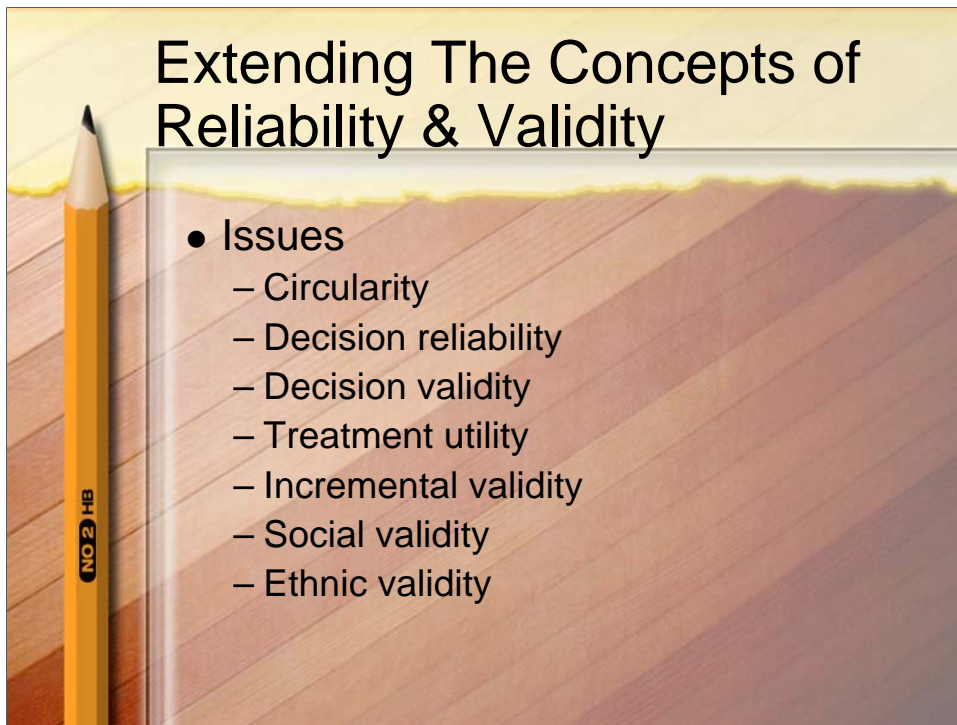


## Basic Test & Measurement Concepts: Norms

- Test X is standardized on 50 males and 50 females at each grade level from kindergarten through sixth grade. The children who make up the norm group were white, middle-class children living in Mount Pleasant, Michigan. Separate norm tables are provided for males and females in each grade. Trey, a fifth-grade African-American child residing in Oakland, California, is tested with test X, and the norm tables are used to interpret his score.
  - a. To how many children is Trey being compared?
  - b. To whose performance is Trey's performance being compared?
  - c. What assumptions are being made about the relationship between Trey's acculturation and the acculturation of the normative sample?

Answers to questions...

- A. 50 males
- B. White children from Michigan
- C. That they have the same learning experiences and learning opportunities



-**Circularity**--using correlation between 2 scales as measure of validity--the reliability & validity of the criterion measure has to be established first (are constructs measured adequately, can they be interpreted with individual children, are they useful for intervention design, etc.).

-**Decision reliability**--consistency of decision outcomes across instruments, methods, raters, & times of measurement. If WISC-IV given along with Vineland to establish mental retardation would SB-5 & Vineland give the same results?

-**Decision validity**--appropriateness of using assessment data for a specific decision purpose. Evaluation of intervention outcomes over time--to what degree did the assessment lead to positive outcomes for the child.

-**Treatment utility**--subset of decision validity--the degree to which assessment leads to beneficial treatment outcomes--largely untested and often unexamined.

-**Incremental**--contributions beyond what is already known--improved predictions--does the assessment tell us something we don't already know or provide information leading to intervention that we could not otherwise obtain.

-**Social validity**--defining socially significant problems for behavior change & designing interventions that are acceptable to significant others.