Advanced Tests and Measurement

Reliability

What We Will Cover

- What reliability is.
- How a test's reliability is estimated.
- How to interpret and use reliability estimates.
- How to enhance reliability.

What Is It?

An estimate of the consistency of a test score.

- Permits an estimate of the amount of error in a score.
- The more error, the less stable (reliable) the score is

1

Definitions

True Score *The stable characteristics of the individual being tested or the attribute being measured.*

Error *Features of the individual, test content, and situation that influence a score but which have nothing to do with the attribute being measured.*
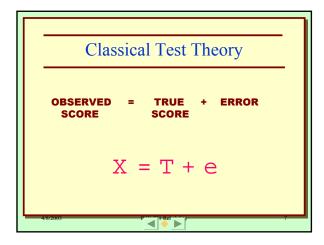

Sources of Variability #1

1. **Person.**
   - **True level of the trait or construct being measured.**
   - **Variability in the person not connected with the trait (error).**
2. **Test.**
   - **Content that is related to the trait being measured.**
   - **Errors in content sampling (error).**
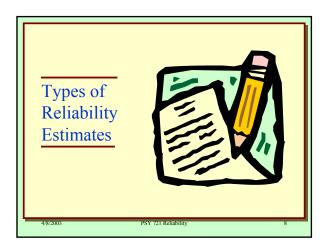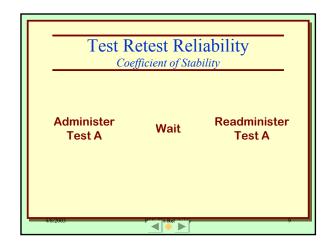   - **Errors in item construction (error).**

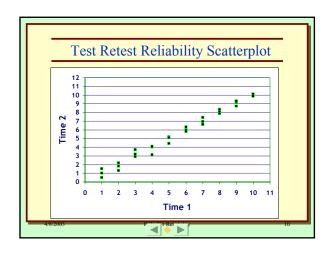## Sources of Variability #2

3. **Test administration.**
   - **Inconsistent test administration (error).**

4. **Scorer error.**
   - **One scorer is inconsistent (error).**
   - **Two scorers don't give the same assessment to the same behavior (error).**

Classical Test Theory

OBSERVED = TRUE + ERROR
SCORE   SCORE

$$X = T + e$$

Types of Reliability Estimates

Test Retest Reliability
Coefficient of Stability

Administer   Wait   Readminister
Test A   Test A

Test Retest Reliability Scatterplot

What good is it?

What sources of error are detected with test-retest reliability?
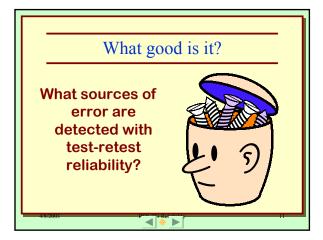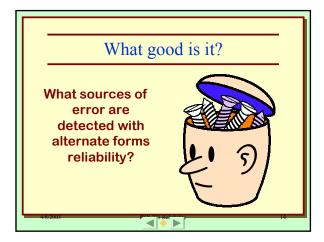

Issues with r_{tt}

- **The interval between time 1 and time 2 is important.**
- **Subject reactivity.**
- **Carry over effect.**
- **Time consuming.**
- **Assumes no change in the individual.**

Alternate Forms Reliability

Administer Test A   **No significant wait**   **Administer Test B**

What good is it?

What sources of error are detected with alternate forms reliability?

Issues

- **Practice effect.**
- **Fatigue effect.**
- **Time delay.**
  - **Back-to-back.**
  - **Interval.**

Sweeney's Measure of Verbal Fluency

USE EACH OF THE FOLLOWING WORDS CORRECTLY IN A SENTENCE.

1. Cat
2. House
3. Automobile
4. Phrenologize
5. Coat

6. Marble
7. Dog-flogger
8. Variance
9. Beetle
10. Crayon

---

3. Internal Consistency

1. Split half reliability.

2. Kuder Richardson (KR-20).

3. Coefficient Alpha.

Split Half Reliability

1. Divide the test into two sub-tests.

2. Correlate the scores on the sub-tests.

Issues With Split Half

- **Which halves?**

- **Correcting for length.**

- **Speeded tests.**

- **Only a single analysis.**

Kuder-Richardson (KR-20)

- **Used with test items that can be scored pass-fail.**
- **Represents the mean of all possible split-half coefficients.**
- **Expressed in terms of a correlation coefficient.**

Coefficient Alpha

- **Used when**
  - **There is no 'pass-fail'.**
  - **Multiple responses to an item.**
- **Represents the mean of all possible split-half coefficients.**
- **Expressed in terms of a correlation coefficient.**

Extroversion Scale (α = .66)

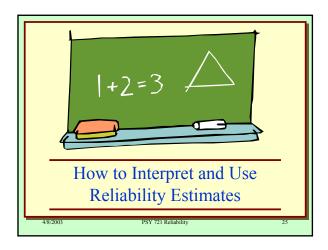| Item	Corrected Item-Total Correlation	Alpha if Deleted
Energetic	.238	.657
Bashful	.508	.587
Quiet	.451	.603
Shy	.541	.580
Withdrawn	.327	.644
Bold	.263	.655
Talkative	.307	.643
Extraverted	.251	.664

What you can learn.

- Errors due to content sampling.

- Errors due to heterogeneity of the content domain.

- Scoring errors.

Interrater(Scorer) Reliability

1. Do different scorers give the same evaluation of the same test?

2. Does the same scorer give the same evaluation of the same test?

How to Interpret and Use Reliability Estimates

Interpretation of r_{tt}

Can be interpreted as the % of variance attributable to TRUE SCORE.

r_{tt} = **percent of TRUE SCORE variability in a score.**

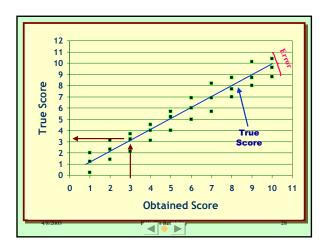$1 - r_{tt}$ = **percent of ERROR variability in a score.**

Mathematically Speaking

$$\sigma^2_{Total} = \sigma^2_{True\ Score} + \sigma^2_{Error}$$

$$r_{tt} = \frac{\sigma^2_{True\ Score}}{\sigma^2_{True\ Score} + \sigma^2_{Error}}$$

Standard Error of Measurement (SEM)

An index of the amount of error (inconsistency) in an individual's test score.

An estimate of the standard deviation of the error in a test.

How to Calculate SEM

$$SEM = SD_{test} \sqrt{1 - r_{tt}}$$
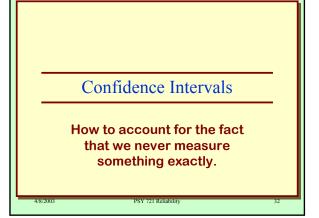
Example Calculation

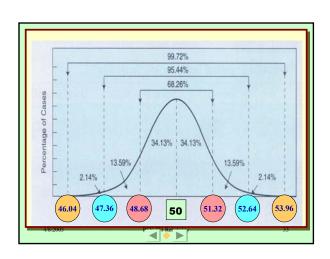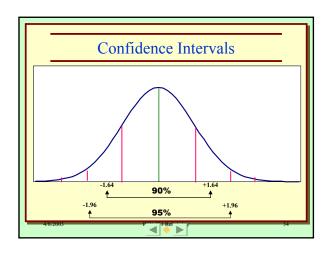| Mean = 50	SD = 4	$r_{tt} = .89$

$$SEM = 4\sqrt{1 - .89}$$
$$SEM = 4\sqrt{.11}$$
$$SEM = 4 \; x \; .33$$
$$SEM = 1.32$$

Confidence Intervals

How to account for the fact that we never measure something exactly.

11

Confidence Intervals

-1.64   90%   +1.64

-1.96   95%   +1.96

Magic Numbers

Confidence Interval	z-score limits
90th	+/- 1.64
95th	+/- 1.96
99th	+/- 2.58

Reliability of Difference Scores

- **Issue**
  - **Both tests have random error.**
  - **The difference between the two test scores does not take into account the SEM for each test.**
  - **The Standard Error of Difference (SE_{diff}) is the estimate of error in difference scores.**
  - **SE_{diff} is greater than either SEM.**

12

SE$_{diff}$ Example

PDQ Test of Conceptual Flexibility

Mean	SD	r_{tt}	SEM
40	8	.90	2.53

$$SE_{diff} = SD\sqrt{2 - r_{11} - r_{22}}$$

$$SE_{diff} = 8\sqrt{.2}$$

$$SE_{diff} = 3.58$$

How to Enhance Reliability

- **Increase test length.**
- **Remove inconsistent items.**
- **Correct for attenuation.**
- **Standardize the scoring system.**
- **Live with it.**

THE

END

13