Reliability - α

What It Is, Why, and How

Jason Nicholas, Ph.D.
November 13, 2008

Objective

- Introduction to reliability

- Meeting requirements of Body of Evidence guidelines for consistency

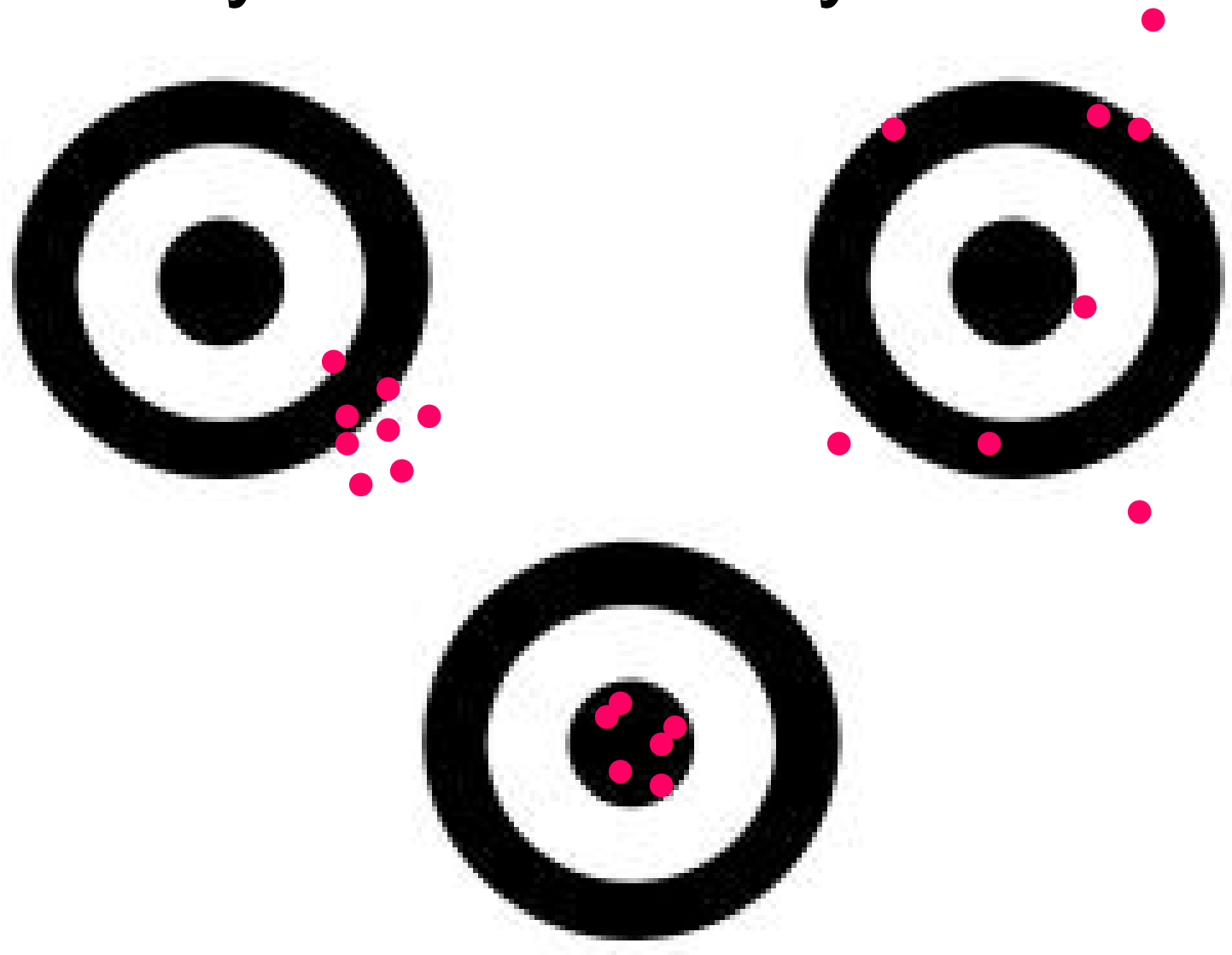
Evaluation Criteria for Body of Evidence Systems

1. Alignment
2. Consistency → ***Reliability***
3. Fairness
4. Standard Setting
5. Comparability

Validity and Reliability

- Bathroom Scale
- My Car

Validity and Reliability



Validity and Reliability

- Can I be reliable and not valid? ***Yes***
- Can I be valid and not reliable? ***No***
- Reliability is a necessary, but not a sufficient condition for validity

Validity

- Consider the following statement

  "*The assessment I created is valid*"

  - Correct or Incorrect?

- Incorrect

Validity

- An evaluation of the adequacy and appropriateness of the interpretations and uses of assessment results

- Example: An assessment of HSer's punctuation skills would not yield valid interpretations about 1st graders' abilities to add fractions

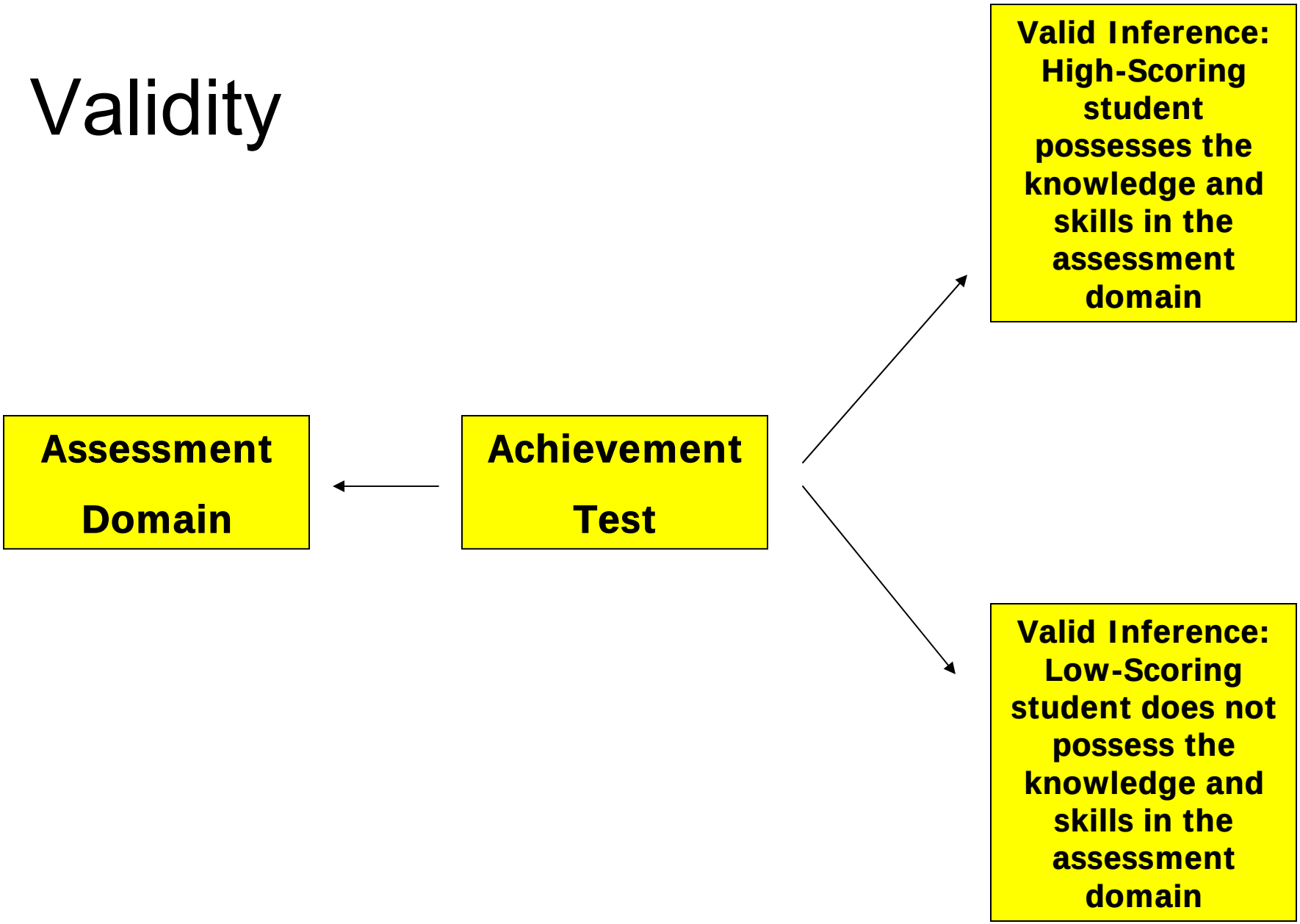
Validity

- Appropriateness of the interpretation of results of an assessment procedure for a given group of individuals
- Validity is a matter of degree; Not all or nothing
- Specific to some particular use or interpretation

Validity

- The interpretation of the assessment results or test scores is the operation that may or may not be valid

Validity


Valid Inference: High-Scoring student possesses the knowledge and skills in the assessment domain
Assessment Domain
Achievement Test
Valid Inference: Low-Scoring student does not possess the knowledge and skills in the assessment domain

Factors that Influence Validity

1. Unclear directions
2. Reading vocabulary and sentence structure too difficult
3. Ambiguity
4. Inadequate time limits
5. Overemphasis of easy-to-access aspects of domain at the expense of important, but hard-to-access aspects (construct under-representation)

Factors that Influence Validity

6. Test items inappropriate for the outcomes being measured (measure complex skills with low-level items)
7. Poorly constructed test items
8. Test too short to provide representative sample of domain being assessed
9. Improper arrangement of items (too hard of items too early)
10. Identifiable pattern of answers

Reliability

- The consistency of results produced by an assessment
- Reliability provides the consistency to make validity possible
- Reliability is the property of a set of test scores that indicates the amount of measurement error associated with the scores

Reliability

- Reliability describes how consistent or error-free the scores are
- Reliability is a property of a set of test scores, not a property of the test itself
- Most reliability measures are statistical in nature

Consistency from BOE

- The district presents evidence that it used **procedures for ensuring inter-rater reliability on open-ended assessments**. For assessments using **closed-ended items**, **measures of internal consistency** (or other forms of traditional reliability evidence) **indicate that the assessments comprising the system meet minimum reliability levels**.

Reliability

- Assessments in BOE systems are referred to as:
  - open-ended assessments
  - closed-ended assessments
- The focus of our discussion is on closed-ended assessments

Reliability

- From the Peer Review Scoring Guide
  - The procedures used to ensure reliability on closed-ended assessments are described
  - Desired, acceptable rates of reliability on closed-ended assessments are stated
  - Reliability data on closed-ended assessments (to meet or exceed average reliability coefficients greater than 0.85) is included

Let’s Get Technical

or actually Theoretical

(suspend all grasp of reality)

Reliability

- If student were to take an assessment again under similar circumstances, they would get the same score
- The property of a set of test scores that indicates the amount of measurement error associated with the scores
- How “error-free” the scores are

Reliability

- The degree to which a test's scores are free from various types of chance effects
- Reliability focuses on the error in students scores
- Can think of there being two types of errors associated with scores:
  - Random errors of measurement
  - Systematic errors of measurement

Reliability

- Random errors of measurement
  - Purely chance happenings
  - Positive or negative direction
  - Sources: guessing, distractions, administration errors, content sampling, scoring errors, fluctuations in the students state of being

Reliability

- Systematic errors of measurement
  - Do not result in inconsistent measurement, but affect utility of score
  - Consistently affect an individuals score because of some particular characteristics of the student or the test that has nothing to do with the construct
  - Hearing impaired child hears “bet” when examiner says “pet” → Score consistently depressed

Reliability

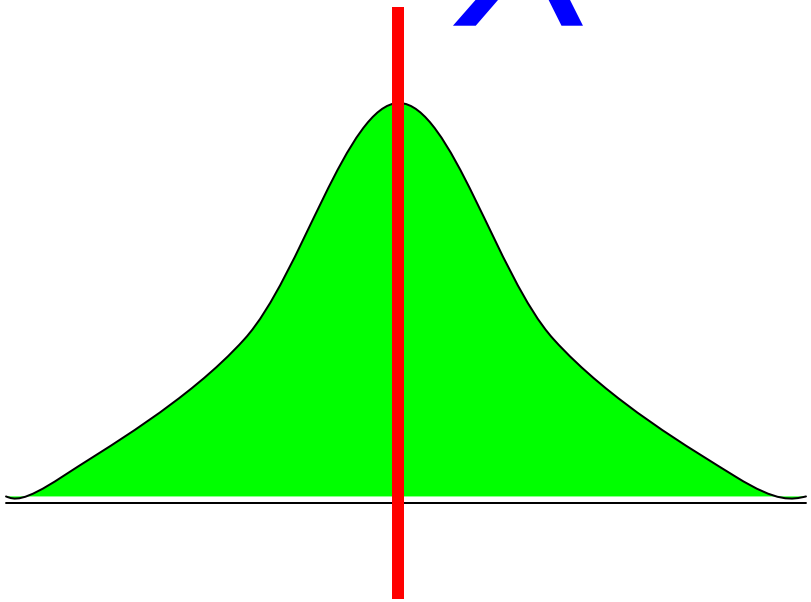
Observed Score = True Score + Error

$$X = T + E$$

Error = Observed Score – True Score

Reliability

$$X = T + E$$



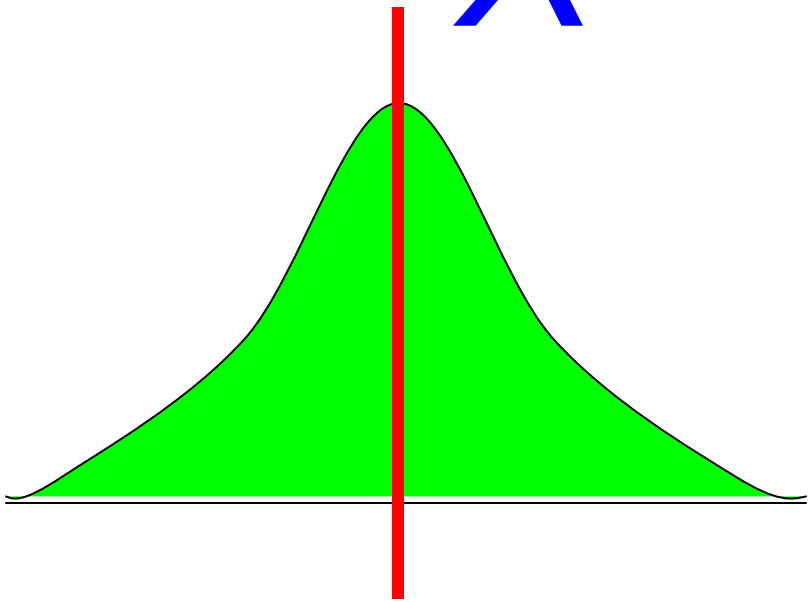
If were to give the assessment many times, we would assume the scores for the student would fall approximately normal

Where the center of the distribution would be the student's True Score

The scatter about the True Score is presumed to be due to errors of measurement

Reliability

$$X = T + E$$

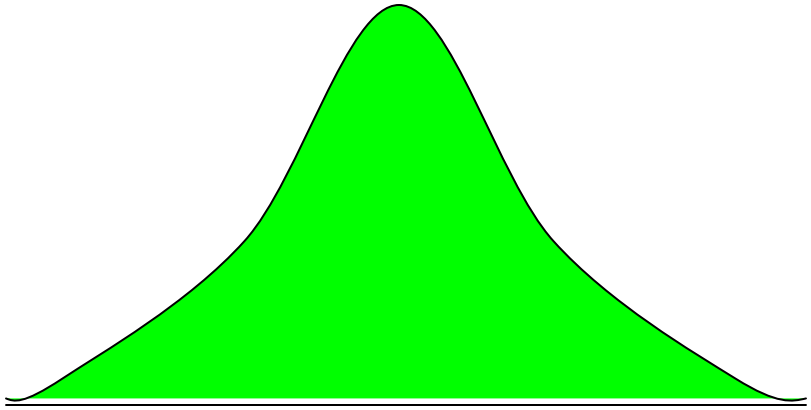


The smaller the standard deviation, the smaller the effect that errors of measurement have on test scores

So, over repeated testing we assume T is the same for an individual but we except that X will fluctuate due to the variation in E

Reliability

$$X = T + E$$



If we gave the assessment to lots of students, we would have the variability of the scores

$$\sigma_X^2 = \sigma_T^2 + Avg(\sigma_E^2)$$

Reliability

$$X = T + E$$

$$\sigma_X^2 = \sigma_T^2 + Avg(\sigma_E^2)$$

$$\text{Reliability} = \frac{\sigma_T^2}{\sigma_X^2}$$

Reliability

$$\text{Reliability} = \frac{\sigma_T^2}{\sigma_X^2}$$

Maximum = 1

All of the variance of the observed scores is attributable to the true scores

Minimum = 0

No true score variance and all of the variance of the observed scores is attributable to the errors of measurement

Greater reliability the closer to 1

Reliability

$$X = T + E$$

How closely related are the examinees Observed Scores and True Scores?

Correlation of two forms that measure the same construct (alternate forms)

Reliability

$$X = T + E$$

If we took two forms with the assumption they measure the same thing, students true score same on both (or linear) measurement errors truly random

The correlation between the two forms across students will be

$$\text{Reliability} = \frac{\sigma_T^2}{\sigma_X^2}$$

Let's Get Back to the Real World

So, how do we find out something about reliability since we don't know the student's True Score?

Estimate it

Reliability

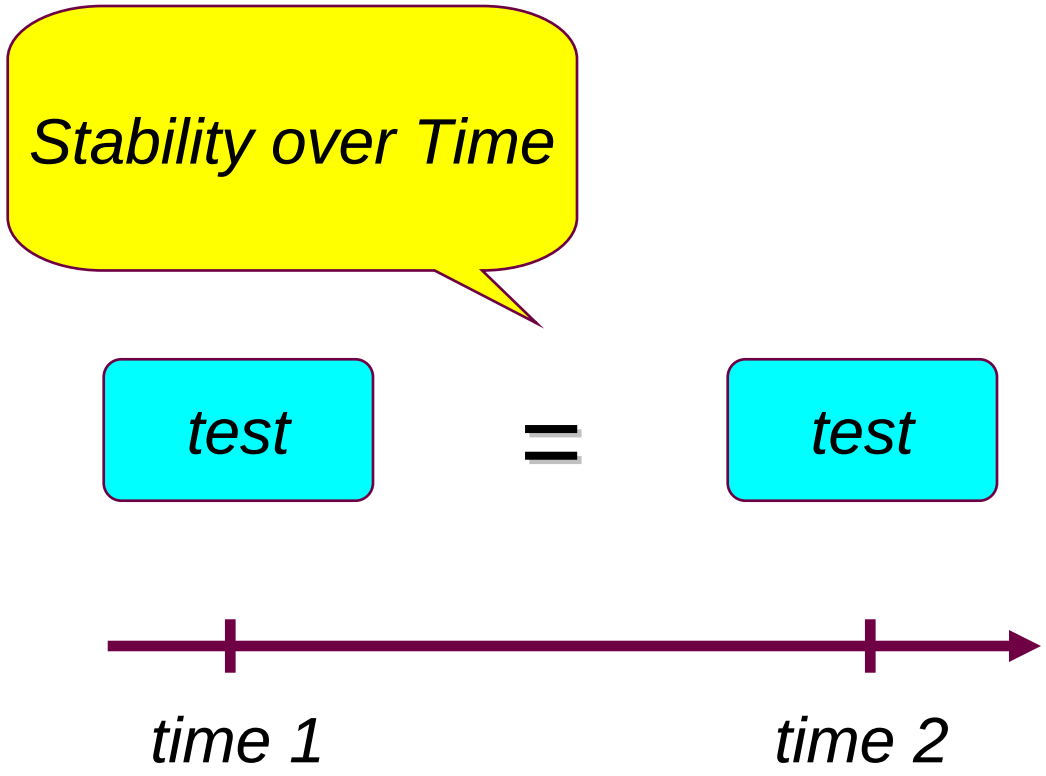
- Administer the test twice
  - Test-Retest Reliability
- Alternate form
  - Parallel Forms Reliability
- Internal consistency measures
  - Internal Consistency Reliability

Reliability

- Administer the test twice
  - measure instrument at two times for multiple persons
  - assumes there is no change in the underlying trait between time 1 and time 2
  - How long?
  - No learning going on?
  - Remember responses

  - Calculate correlation coefficient between test scores
  - *Coefficient of Stability*

Test-Retest Reliability



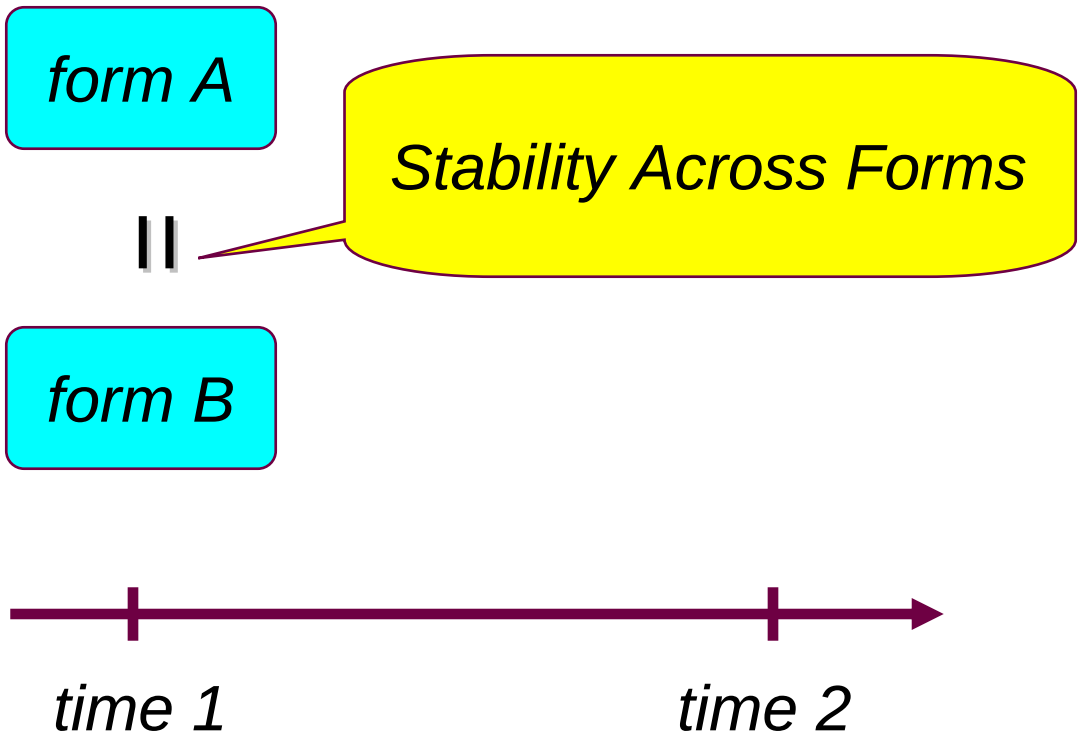
Reliability

- Alternate form
  - Forms similar
  - Short time period
  - Balance order of assessments
  - administer both forms to the same people
  - usually done in educational contexts where we need alternative forms because of the frequency of retesting and where you can sample from lots of equivalent questions

---

  - Calculate correlation coefficient between test scores from the two forms
  - *Coefficient of Equivalence*

Parallel-Forms Reliability



Reliability

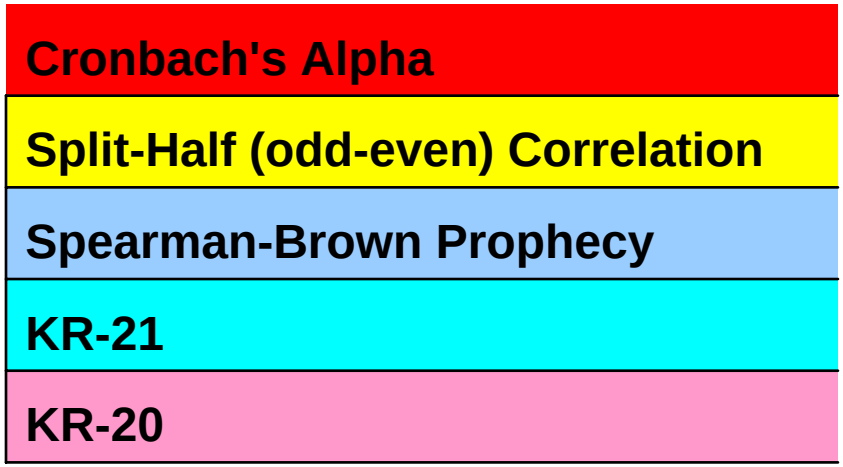
- Internal consistency measures
  - Statistical in nature
  - One administration
  - How well do students perform across subsets of items on one assessment
  - Students performance consistent across subsets of items, performance should generalize to the content domain
  - Main focus is on content sampling

Reliability

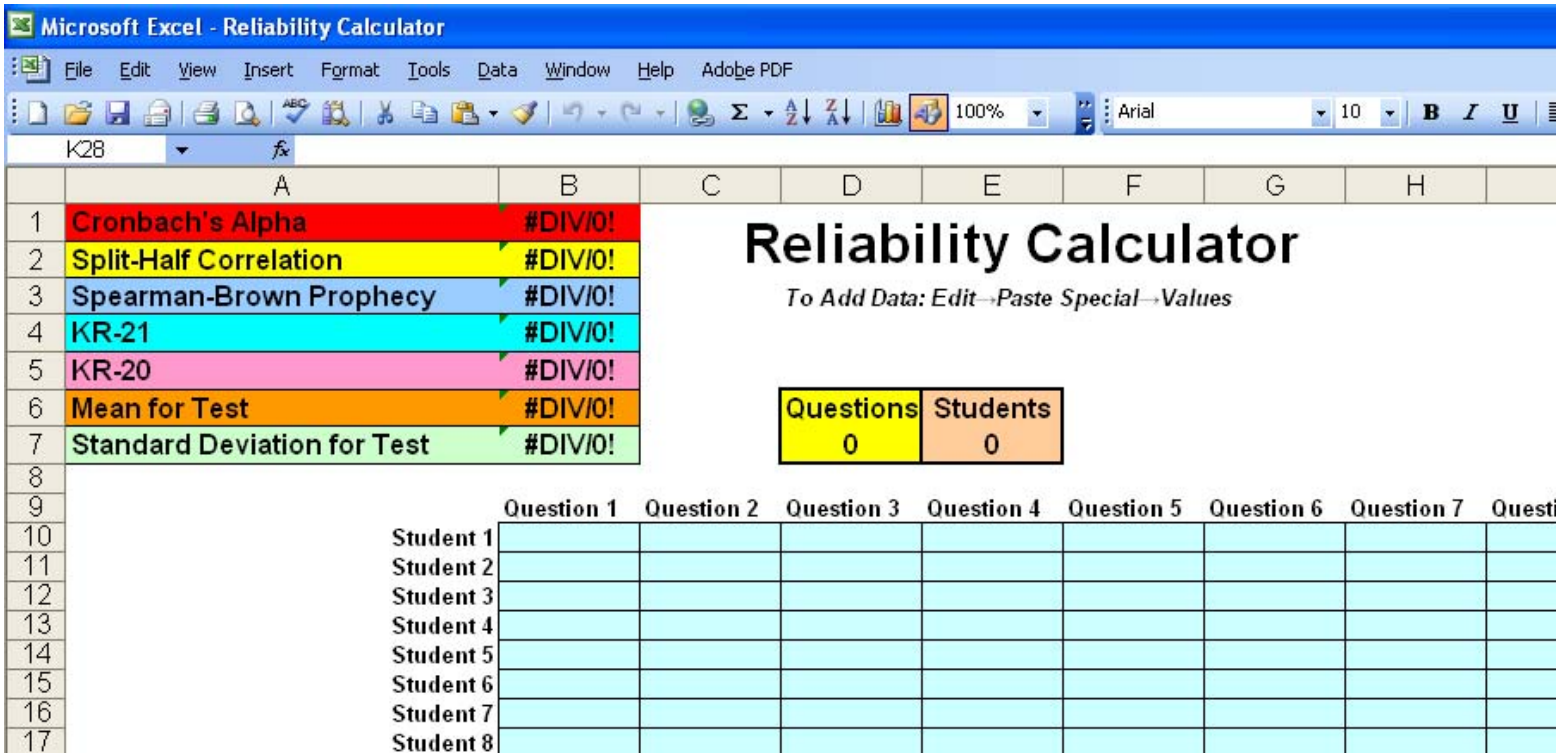
- Internal consistency measures
  - “Most appropriate to use with scores from classroom tests because these methods can detect errors due to content sampling and to differences among students in testwiseness, ability to follow instructions, scoring bias, and luck in guessing answers correctly.”
  - Two broad classes of internal consistency measures

Reliability

1. Split-Half
2. Variance Structure



Cronbach's Alpha
Split-Half (odd-even) Correlation
Spearman-Brown Prophecy
KR-21
KR-20



Split-Half

- Before scoring, split test up into two equal halves
- Create two half-tests that are as nearly parallel as possible
- The less parallel halves are, reduction in quality of reliability measure

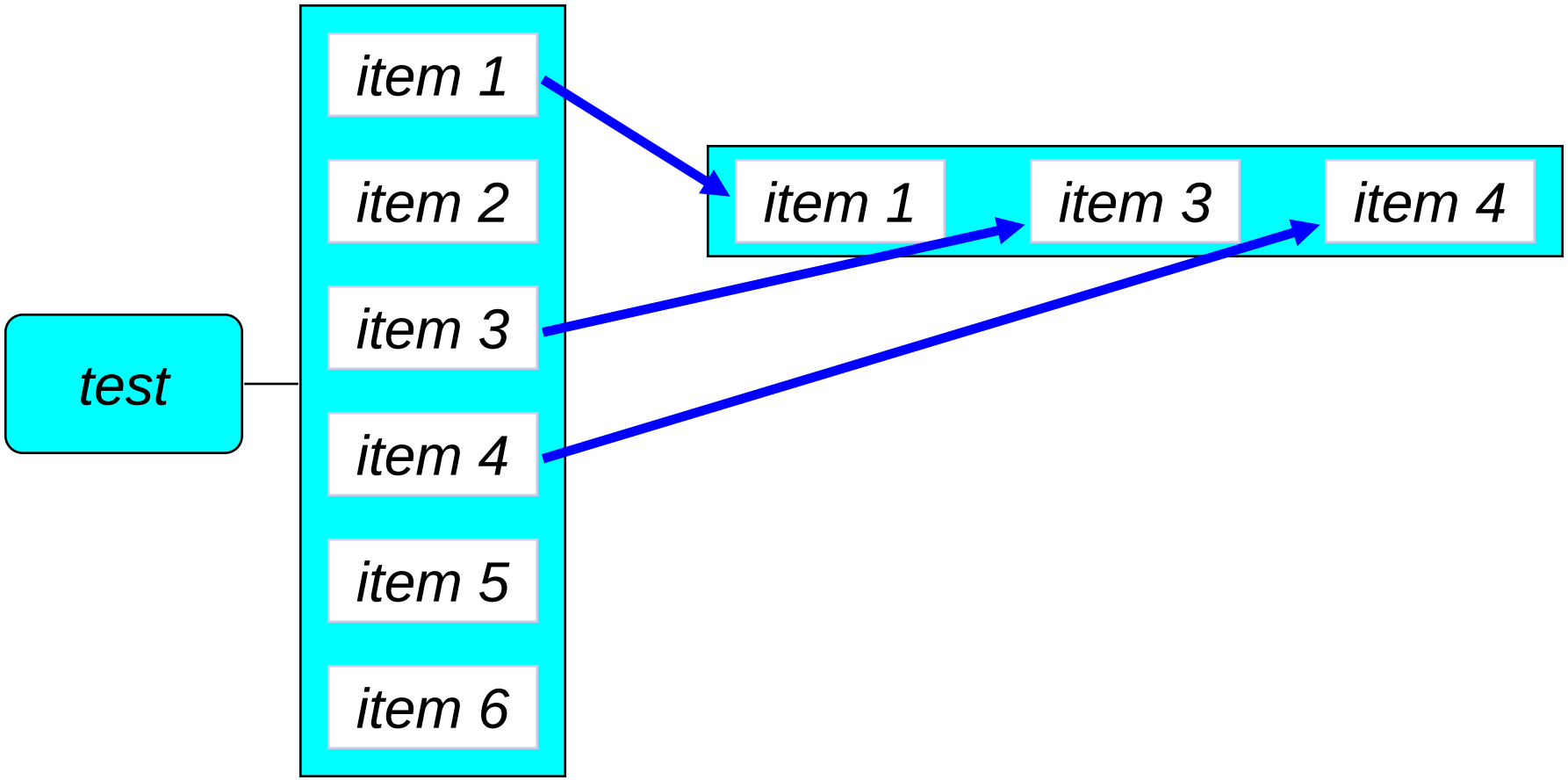
Split-Half

- Methods for splitting
  - Odd numbers to one form, even to another
  - Random assignment
  - Assign items so that forms are “matched” in content
  - Rank order items by difficulty values and then assign odd ranks to one form, even to another

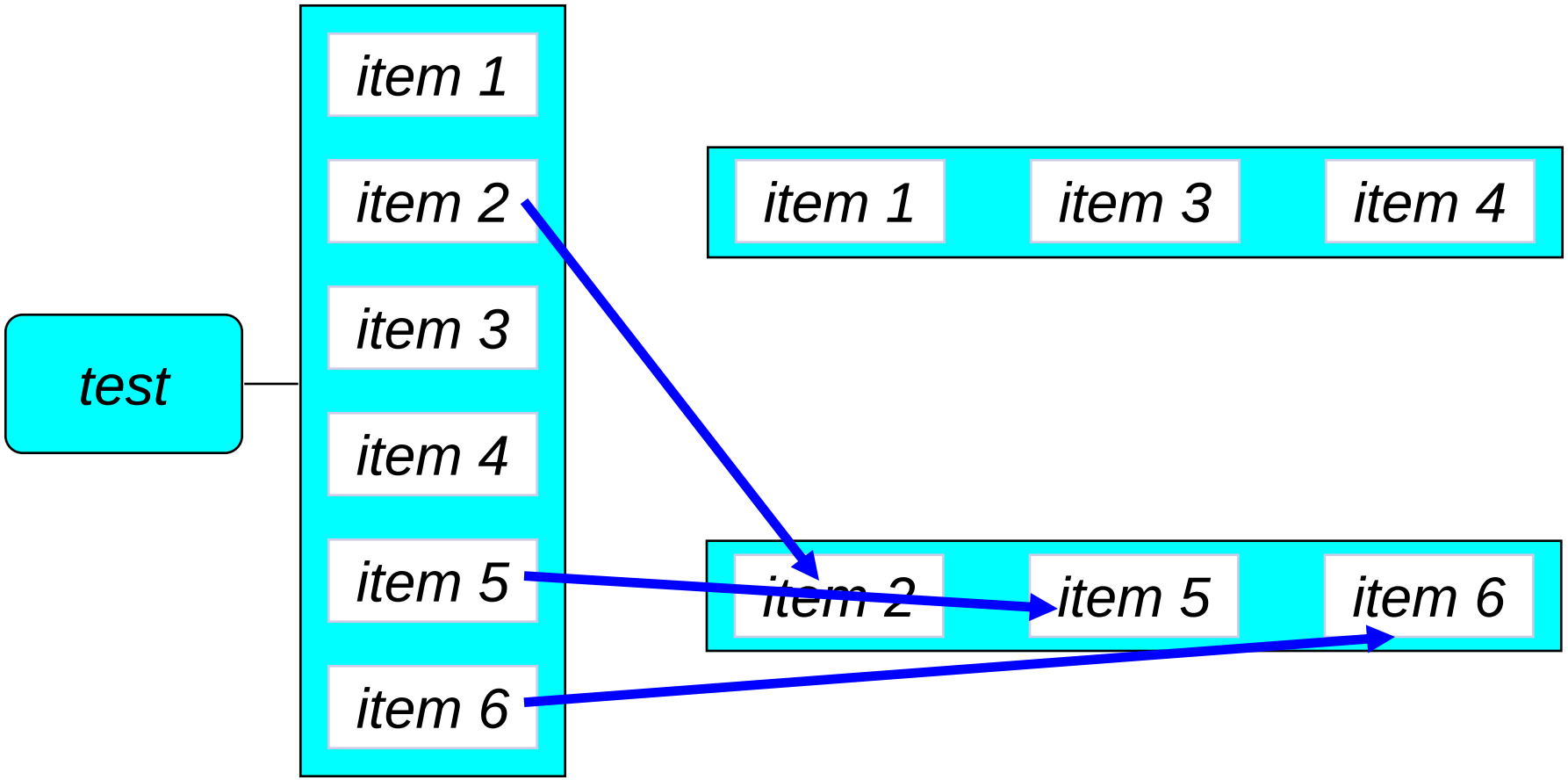
Split-Half

- Splitting completed
- Take student data from assessment
- Correlate Total student score on Form A with Total student score on Form B
- Correlation coefficient is the reliability measure

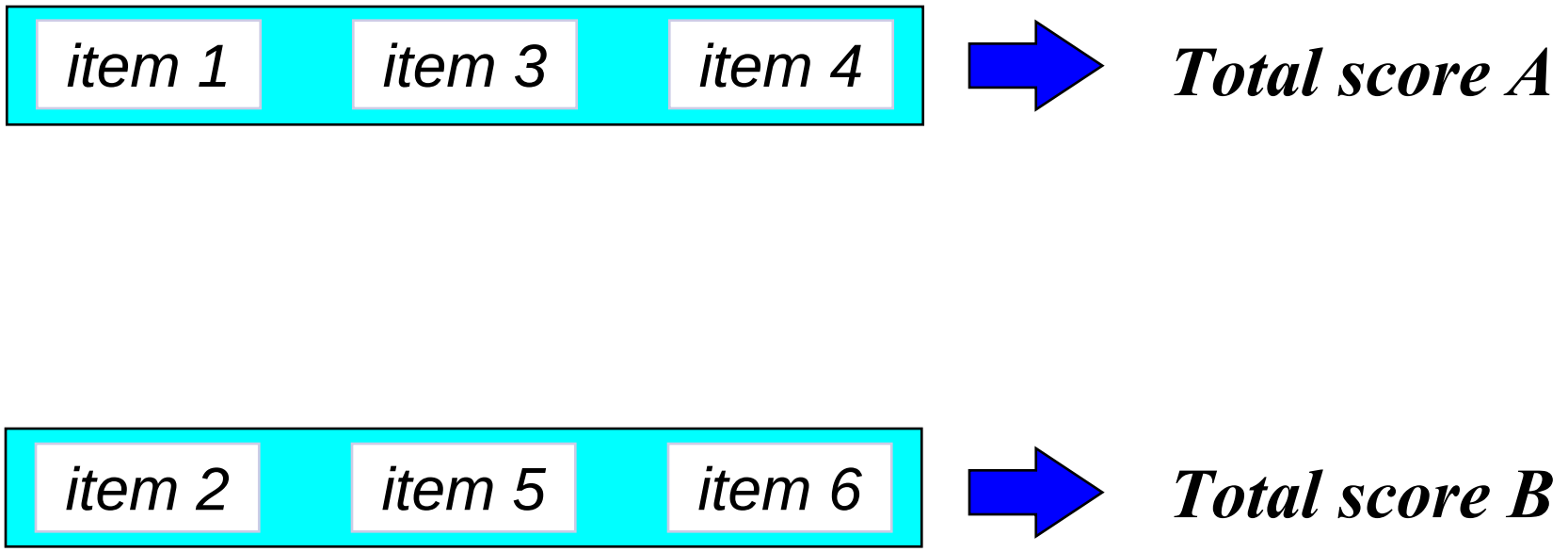
Split-Half



Split-Half


item 1
item 2
item 3
item 4
item 5
item 6
test
item 1
item 3
item 4
item 2
item 5
item 6

Split-Half


item 1
item 3
item 4
Total score A
item 2
item 5
item 6
Total score B

Split-Half

	Total score A	*Total score B*
Subject1	10	11
Subject2	13	14
Subject3	12	16
Subject4	11	15
Subject5	10	14
Subject6	17	13
Subject7	16	16
Subject8	14	15
Subject9	13	13
Subject10	12	13

Run correlation on the two lists of scores

Split-Half

- Likely to underestimate the reliability coefficient for the full-assessment
- Longer tests are generally more reliable than shorter tests since errors of measurement are reduced because of increased content sampling
- We can adjust for this

Spearman-Brown Prophecy

- Corrected estimate of the reliability coefficient of the full-length assessment

$$\text{SBPR} = \frac{2(split-half\ \ reliability)}{1+split-half\ \ reliability}$$

- Remember assumption that half-tests are strictly parallel. Less parallel, less accurate

Split-Half

- Split assessment, found correlation between students total scores across two splits → reliability = .34

Spearman-Brown Prophecy

$$\frac{2(.34)}{1+.34} = .51$$

Cronbach's Alpha

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum S_i^2}{S_T^2}\right)$$

k = number of items

S_i^2 = variance of item i

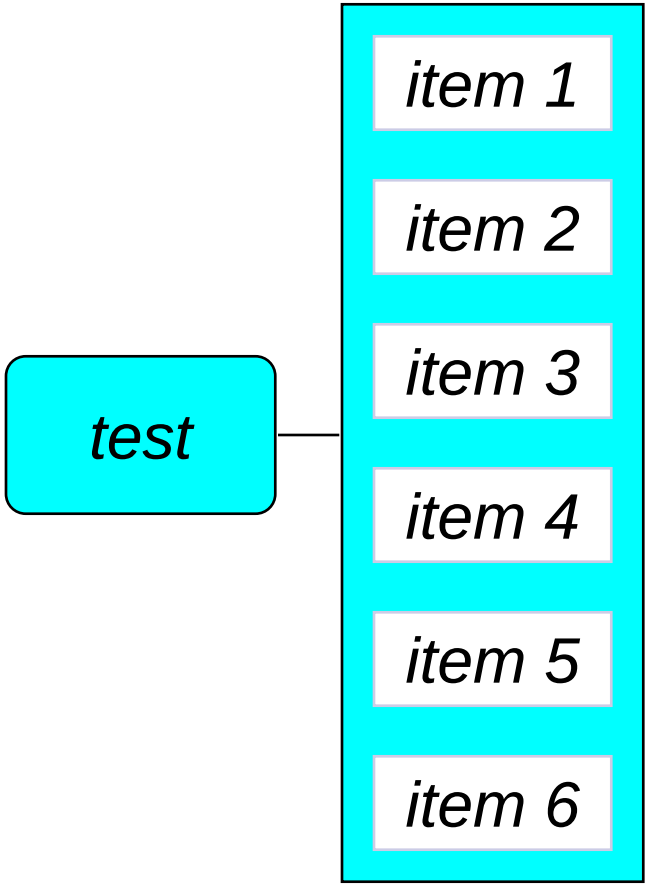
S_T^2 = variance of total test

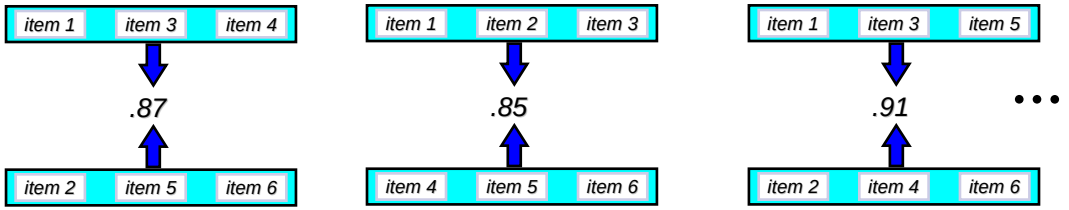
Cronbach's Alpha

- Can be used with multiple item types

- If were to get an Alpha = .80, we could say that at least 80% of the total score variance is due to true score variance

Cronbach's Alpha





SH_1 .87
SH_2 .85
SH_3 .91
SH_4 .83
SH_5 .86
...
SH_n .85

$\alpha = .85$

Like the average of all possible split half correlations

Kuder-Richardson 20

- Only used with dichotomous items

$$KR_{20} = \frac{k}{k-1}\left(1 - \frac{\sum p_i q_i}{S_T^2}\right)$$

k = number of items

p = proportion of group answering item i correctly

q = proportion of group answering item i incorrectly

S_T^2 = variance of total test

Kuder-Richardson 21

- Only used with dichotomous items

$$KR_{21} = \frac{k}{k-1}\left(1 - \frac{\bar{X}\left(k - \bar{X}\right)}{kS_T^2}\right)$$

k = number of items

p = proportion of group answering item i correctly

q = proportion of group answering item i incorrectly

S_T^2 = variance of total test

Kuder-Richardson 20

Kuder-Richardson 21

- When all items are of equal difficulty, KR20 and KR21 will be equal
- KR21 assumes equal difficulty of items, if not KR21 will be lower than KR20
- Publisher should not just report KR21
- KR21 easier to do by hand
- Sufficient lower bound for reliability

Interpretation of Reliability

- Reliability is based on a particular group of students on a certain day and under certain testing conditions

- Standards of Reliability
  - Published tests = .85-.95
  - For individual decisions = .85 minimum
  - For group decisions = .65 minimum
  - Teacher tests = .50 as long as we have other scores to be used in conjunction

Interpretation of Reliability

- Alpha and KR20 are focused towards assessments with homogenous content
- For assessments with heterogeneous content, Alpha and KR20 will be smaller than what is provided with split-half
- Alpha and KR20 not appropriate for speeded assessments
  - If speed is a factor, inflated reliability
  - Use test/retest or Alternate forms

What Affects Reliability?

- Under what circumstances do tests provide reliable scores?
- Consider
  - Assessment itself
  - Conditions under which assessment is given
  - Group of examinees being assessed
- Interaction of these that determines reliability

Assessment Itself

- Test Length
  - Longer = more reliable
  - Up to a certain point

- Item Type
  - Objectively scored items produce more reliable assessment
    - Eliminate scorer inconsistency
    - Cover more content

Assessment Itself

- Item Quality
  - Unclear items
  - Item too difficult for students
    - Skip or guess
  - Item too easy for students
    - Doesn’t hurt reliability, but doesn’t help
  - Best items are those that discriminate
    - Those students who possess the knowledge have a better chance of answering correct

Conditions of Administration

- Instructions
- Time limits
- Physical conditions

- Any factor that affects students differently will affect students test scores other than the difference in knowledge and skills

- These sources reduce reliability by introducing unwanted sources of random variation or measurement error into scores

Group of Examinees

- Reliablity depends on the range of ability in the group being tested
- A group that is narrow in its ability will produce a lower reliability (even though instrument the same)
- Example situation of improving instruction over time with the same instrument becoming less reliable

Group of Examinees

- With reliability, “we are looking at the capability of the test to make reliable distinctions among the group of examinees with respect to the ability measured by the test”

- If a big range of ability, a good test should be able to do this well. If small range, difficult to do.

Reliability

- From the Peer Review Scoring Guide
  - The procedures used to ensure reliability on closed-ended assessments are described
  - Desired, acceptable rates of reliability on closed-ended assessments are stated
  - Reliability data on closed-ended assessments (to meet or exceed average reliability coefficients greater than 0.85) is included