Lecture 4 Psychological Testing and Measurement

Sunthud Pornprasertmanit

Reliability

What is Reliability?

- Reliability refers to the consistency of scores obtained by the same persons when they are reexamined with the same test on
  - Different occasions
  - Different sets of equivalent items
  - Under other variable examining conditions

What is Reliability?

- The concept of reliability underlies the computation of the error of measurement
- We can predict the range of fluctuation likely to occur in a single individual's score as a result of irrelevant or unknown chance factors.

What is Reliability?

$$O = T + E$$

- O = Observed score (Individual differences by test
- T = True score (Real individual differences)
- E = Error of measurement

What is Reliability?

- Because *E* is chance factor, it does not correlate with *T*.

$$\sigma_O^2 = \sigma_T^2 + \sigma_E^2$$

- Therefore, variance of observed score is the sum of variance of true score and error of measurement variance

What is Reliability?

- Test reliability indicates the extent to which individual differences in test scores are attributable to "true" difference.

$$r_{xx} = \frac{\sigma_T^2}{\sigma_O^2} = 1 - \frac{\sigma_E^2}{\sigma_O^2}$$

What is Reliability?

- Such a measure of reliability characterizes the test when
  - it is administered under standard conditions
  - given to persons similar to those constitute the normative sample.

Type of Reliability

Type of Reliability Coefficient	Error variance
Test-retest Reliability	Time sampling
Alternate-Form (Immediate)	Content sampling
Alternate-Form (Delayed)	Time and content sampling
Split-Half	Content sampling
KR and Coefficient Alpha	Content Heterogeneity
Scorer Reliability	Interscorer differences

Test-retest Reliability

- Why the interval over which it was measured should always specified?
- What is the best interval to measure test-retest reliability?

Alternate-Form Reliability

- What is alternate form?
  - Same number of items
  - Same form
  - Cover same type of content
  - Range and level of difficulty should equal
  - Instruction, time limits, illustrative examples, format should be checked for equivalence.

Alternate-Form Reliability

- What are profits of alternate form test?
- Does alternate form affect from practice effect? If any, does practice effect affect alternate-form reliability?

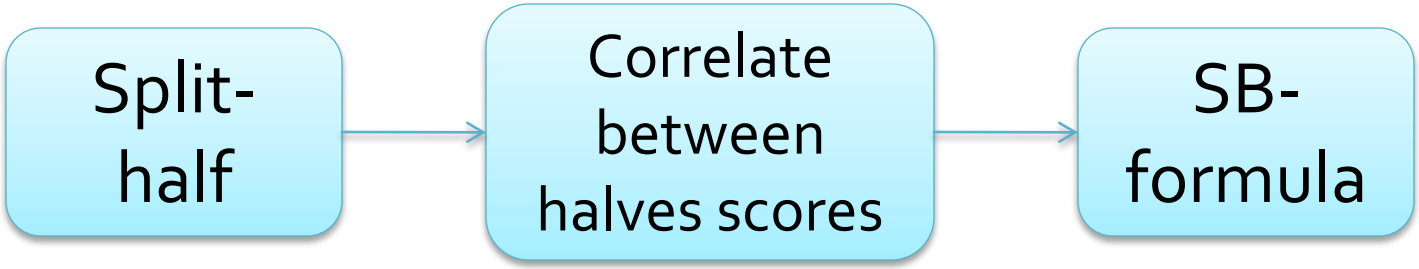
Split-Half Reliability

- The crucial step is to find equivalent halves.

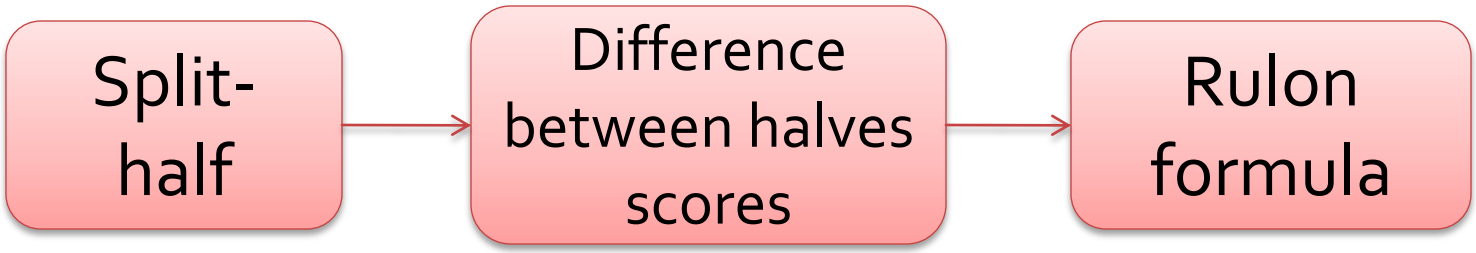
Split-Half Reliability

- Two popular ways for calculating

1)



2)



KR and Coefficient Alpha

- Interitem consistency is influenced by two sources of error variance
  - Content sampling
  - Heterogeneity of items

KR and Coefficient Alpha

- Is construct homogeneous in nature?
- Heterogeneous → unambiguous interpretation
- However, homogeneous → not adequately coverage contruct

KR and Coefficient Alpha

- For numerical scale items, a generalized formula has been derived, known as coefficient alpha:

$$r_{xx} = \left(\frac{n}{n-1} \right) \frac{SD_x^2 - \sum (SD_i^2)}{SD_x^2}$$

KR and Coefficient Alpha

- Coefficient alpha and split half reliability
- Coefficient alpha as a lower bound reliability
- High internal consistency = unidimensionality???
- Coefficient alpha and covariance among items

Scorer Reliability

- How to achieve high scorer reliability?

Reliability of Speeded Tests

- A pure speed test is one in which individual differences depend entirely on speed of performance.
- Such a test is constructed from items of uniformly low difficulty.
- The time limit is made so short that no one can finish all the items.

Reliability of Speeded Tests

- A pure power test has a time limit long enough to permit everyone to attempt all items.
- The difficulty of the items is steeply graded.

Reliability of Speeded Tests

- In actual practice, the distinction between speed and power tests is one of degree (varying in proportions).
- Why prevent perfect scores? (Except for criterion-referenced test)
- Truncated Distribution

Reliability of Speeded Tests

- All internal consistency (Split-half, KR and Alpha) is not suitable for estimating reliability of speeded tests, because it is spurious high.

Reliability of Speeded Tests

- Type of reliability that can be used
  - Test-retest reliability
  - Equivalent-form reliability
  - Split-half techniques made in terms of time by divide total time into quarters and counter-balance

Factors Affect NRT Reliability

- Sample Variance and Reliability (Range Restriction)
- Test Length
- Item Difficulty

Standard Error of Measurement

- Standard error of measurement is standard deviation of error scores.
- The more reliability coefficient, the less standard error of measurement.
- Computed by:

$$SEM = SD\sqrt{1 - r_{tt}}$$

Standard Error of Measurement

- Standard error of measurement can be used for true score estimate (by confidence interval)

$$CI_{1-\alpha} = X \pm z_{\alpha/2} SEM$$

Standard Error of Measurement

- Unlike the reliability coefficient, the error of measurement is independent of the variability of the group on which it is computed.
- However, SEM cannot be directly comparable from test to test.
- When consider SEM?

Reliability for Mastery Classifications

- What is concerned?
  - Decision Reliability
  - Score reliability

Reliability for Mastery Classifications

Decision Based on Form 1

Decision Based on Form 2	Nonmaster	Master	
Nonmaster	$p_{00} = .40$	$p_{01} = .10$	$p_{0.} = .50$
Master	$p_{10} = .30$	$p_{11} = .20$	$p_{1.} = .50$
	$p_{.0} = .70$	$p_{.1} = .30$	

The estimated probability of a consistent decision is

$$p = p_{11} + p_{00}$$

Reliability for Mastery Classifications

- Another formula is Cohen's Kappa:

$$\kappa = \frac{p - p_c}{1 - p_c}$$

- P_c is the chance probability of a consistent decision:

$$p_c = p_{1.}p_{.1} + p_{0.}p_{.0}$$

Reliability for Mastery Classifications

Decision Based on Form 1

Decision Based on Form 2	Nonmaster	Master	
Nonmaster	$p_{00} = .40$	$p_{01} = .10$	$p_{0.} = .50$
Master	$p_{10} = .30$	$p_{11} = .20$	$p_{1.} = .50$
	$p_{.0} = .70$	$p_{.1} = .30$	

$$p_c = p_{1.}p_{.1} + p_{0.}p_{.0} = (.7)(.5) + (.3)(.5) = .5$$

$$\kappa = \frac{p - p_c}{1 - p_c} = \frac{.6 - .5}{1 - .5} = .2$$

Reliability for Mastery Classifications

- Four factors may affect decision consistency
  - More Test length → More Reliability
  - Location of the cut score in the score distributions → At center, low reliability
  - High Test score generalizability → High Reliability
  - High Similarity of the score distributions for the two forms → High Reliability

Reliability for Mastery Classifications

- Example (5 items)
  - Test Difficulty = 40 %
  - Item = 5 items
  - Domain explained (ρ^2) = .40
  - Cutoff = 40 %
  - p = .66
- Example (10 items)
  - Test Difficulty = 40 %
  - Item = 10 items
  - Domain explained (ρ^2) = .57
  - Cutoff = 40 %
  - p = .71

Reliability for Mastery Classifications

- Four factors may affect decision consistency
  - More Test length → More Reliability
  - Location of the cut score in the score distributions → At center, low reliability
  - High Test score generalizability → High Reliability
  - High Similarity of the score distributions for the two forms → High Reliability

Reliability for Mastery Classifications

- Example (cutoff 40%)
  - Test Difficulty = 40 %
  - Item = 5 items
  - Domain explained (ρ^2) = .40
  - Cutoff = 40 %
  - p = .66
- Example (cutoff 20%)
  - Test Difficulty = 40 %
  - Item = 5 items
  - Domain explained (ρ^2) = .40
  - Cutoff = 20 %
  - p = .81

Reliability for Mastery Classifications

- Example (cutoff 40%)
  - Test Difficulty = 40 %
  - Item = 5 items
  - Domain explained (ρ^2) = .40
  - Cutoff = 40 %
  - p = .66
- Example (cutoff 80%)
  - Test Difficulty = 40 %
  - Item = 5 items
  - Domain explained (ρ^2) = .40
  - Cutoff = 80 %
  - p = .81

Reliability for Mastery Classifications

- Four factors may affect decision consistency
  - More Test length → More Reliability
  - Location of the cut score in the score distributions → At center, low reliability
  - High Test score generalizability → High Reliability
  - High Similarity of the score distributions for the two forms → High Reliability

Reliability for Mastery Classifications

- Example ($\rho^2 = .4$)
  - Test Difficulty = 40 %
  - Item = 5 items
  - Domain explained (ρ^2) = .40
  - Cutoff = 40 %
  - p = .66
- Example ($\rho^2 = .9$)
  - Test Difficulty = 40 %
  - Item = 5 items
  - Domain explained (ρ^2) = .90
  - Cutoff = 80 %
  - p = .90

Reliability for Mastery Classifications

- Four factors may affect decision consistency
  - More Test length → More Reliability
  - Location of the cut score in the score distributions → At center, low reliability
  - High Test score generalizability → High Reliability
  - High Similarity of the score distributions for the two forms → High Reliability

Class Assignment

Case 5.1 – 5.2

Case Feedback

3.2 Maid performance assessment sheet

Homework

Case 5.1-5.2

Exercise 5.1-5.2

Correction Performance Assessment Sheet

Next Lecture
Validity

Lecture 4 Psychological Testing and Measurement
Sunthud Pornprasertmanit

What is Validity?

Two Meaning of Validity

1) The validity of a test is the extent to which a test measures what it purports to measure.
2) Validity is an integrative evaluative judgment to degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment.

Types of Validity

- Content Validity
- Criterion-related Validity
- Construct Validity

Content Validity

- Content validity determine whether test content covers a representative sample of the behavior domain to be measured.
  - Like the process of job analysis
- The popular method is expert judgment.

Criterion-Related Validity

- Criterion-related validity indicate the effectiveness of a test in predicting an individual's performance in specified activities.
- Criterion Problem
  - Criterion Deficiency
  - Criterion Contamination

Criterion-Related Validity

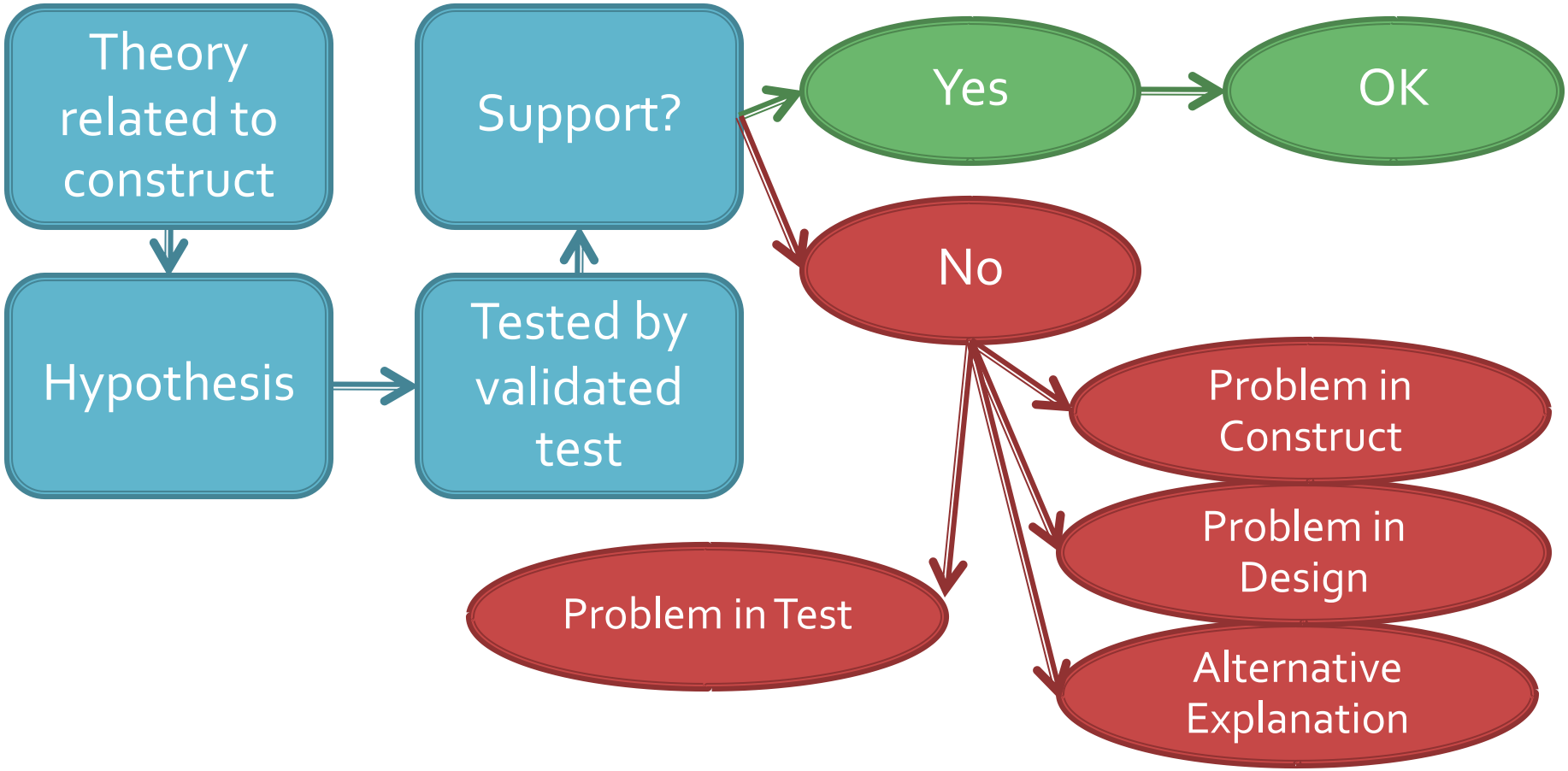
- Factors affect this validity
  - Inadequate Sample Size to achieve stat significant
  - Unreliability (Attenuation)
  - Restriction of Range
  - Differential Validity

Construct Validity

- Construct validity is the extent to which the test may be said to measure a theoretical construct or trait
- Construct validation requires the gradual accumulation of information from a variety of sources.

Construct Validity

- Process of Construct Validity



Construct Validity

- Example
  - Developmental Aspects
  - Nomological Network
  - Convergent and Discriminant Validity
    - Multitrait-multimethod matrix
  - Experimental Manipulation
  - Factor Analysis
    - Exploratory Factor Analysis
    - Confirmatory Factor Analysis