Reliability

Lecture 4 Psychological Testing and Measurement

Sunthud Pornpresertmanit

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 correlate with T.

$$\boxed{\sigma_O^2 = \sigma_T^2 + \sigma_E^2}$$

- 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.

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

What is Reliability?

- Any condition that is irrelevant to the purpose of the test represents error variance: test taking time, rapport, instructions etc.
- Factors that might be considered error variance for one purpose would be classified under true variance for another.

What is Reliability?

- Such a measure of reliability characterizes the test when it is administered under standard conditions and 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

- It is the correlation between the scores obtained by the same persons on the two administrations of the test
- The error variance corresponds to the random fluctuations of performance from one test session to the other
- It shows the extent to which scores on the test can be generalized over different occasions.

Discussing Question?

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

Alternate-Form Reliability

- The correlation between the scores obtained on the two forms represents the reliability coefficient of the test.
- It is a measure of both temporal stability and consistency of response to different item samples (or test forms).

Alternate-Form Reliability

- Error from content sampling is the fluctuation from item random sampling from population pools.
- To what extent do scores on this test depend on factors specific to the particular selection of items?
- Sometimes, this form of reliability can be administered in immediate succession or delayed taking test.

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.

Discussing Question?

- 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

- Two scores are obtained for each person by dividing the test into equivalent halves.
- Split-half reliability provides a measure of consistency with regard to content sampling.
- Temporal stability of the scores does not enter into such reliability because only one test session is involved.

Split-Half Reliability

- Sometimes, this type of reliability is called a coefficient of internal consistency.
- How to split the test in order to obtain the most nearly equivalent class?

Split-Half Reliability

- The correlation of two halves scores actually give the reliability of only a half-test.
- Other things equal, the longer a test, the more reliable it will be, because of large content sampling

Split-Half Reliability

- The effect that lengthening or shortening a test will have on its coefficient can be estimated by means of Spearman-Brown formula:

$$r_{xx} = \frac{nr_{hh}}{1+(n-1)r_{hh}}$$

$$n = \frac{number\ of\ new\ test}{number\ of\ old\ test}$$

Split-Half Reliability

- Alternate method for finding split-half reliability is Rulon formula:

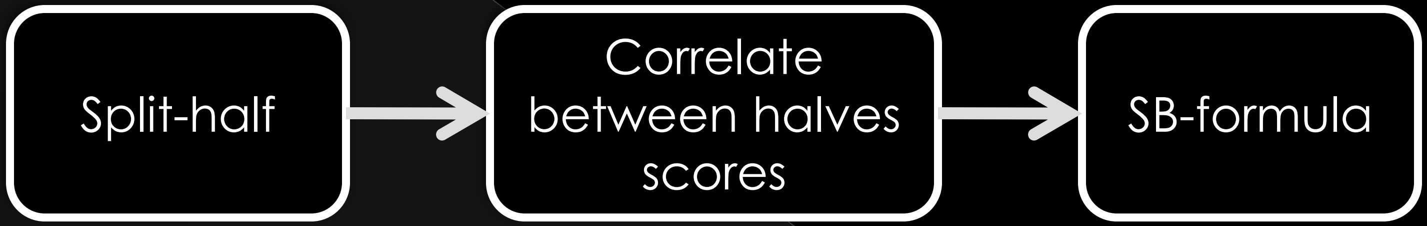
$$r_{xx} = 1 - \frac{SD_d^2}{SD_x^2}$$

d = different of scores between two halves

Split-Half Reliability

- Summary

1)

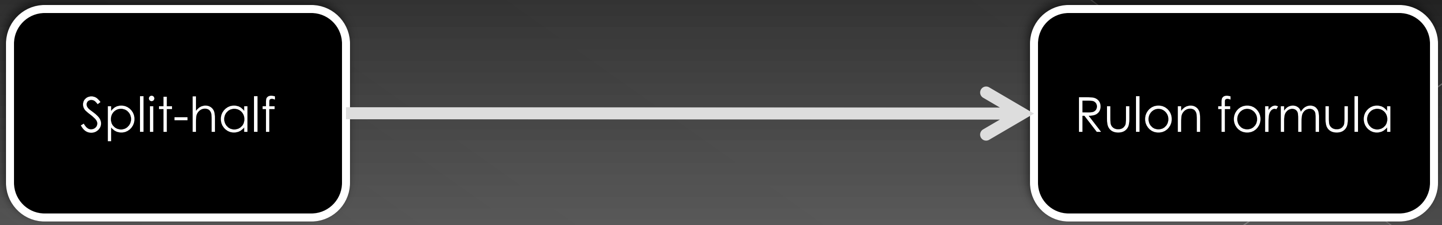


Split-half

Correlate between halves scores

SB-formula

2)



Split-half

Rulon formula

KR and Coefficient Alpha

- This method is based on the consistency of responses to all items in the test.
- Interitem consistency is influenced by two sources of error variance
  - Content sampling
  - Heterogeneity of behavior domain sampled

KR and Coefficient Alpha

- It is apparent that test scores will be less ambiguous when derived from relatively homogeneous tests.
- The question whether the criterion that the test is trying to predict is itself relatively homogeneous or heterogeneous is relevant to utility of homogeneous test.

KR and Coefficient Alpha

- Unambiguous interpretation of test scores could be combined with adequate criterion coverage.

KR and Coefficient Alpha

- The most common procedure for finding interitem consistency is “Kuder-Richardson formula 20”

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

KR and Coefficient Alpha

- KR-20 can be used only for dichotomous items.
- Cronbach (1951) showed that KR-20 is actually the mean of all split-half coefficients (by Rulon formula) resulting from different splitting of a test.
- The difference between KR and split-half reliability coefficients may be serve as a rough index of the heterogeneity of a test.

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 can be considered as the lower bound to a theoretical reliability coefficient known as the coefficient of precision.
- One common interpretation of coefficient alpha is that a relatively high value of alpha indicates that the test items are unidimensional (measuring only one trait).

KR and Coefficient Alpha

- Because alpha is a function of item covariances, and high covariance between items can be result of more than one common factor.
  - For example, scores to items on an essay test in social studies may be determined both by examinees' writing abilities and by their knowledge of the content.

Scorer Reliability

- In individual test, there is evidence of considerable examiner variance.
- Scorer reliability can be found by
  1) Having a sample of test papers independently scored by two examiners.
  2) Two scores obtained by each test taker are then correlated in the usual way

Discussing Question?

- How to achieve high scorer reliability?
- Can reliability coefficient be interpreted as percentage of true variance and error variance?
- Does variance of score affect reliability coefficient?

Administration errors

- It is not necessary to report reliability for administration errors, exclude scorer reliability, because it can experimentally controlled.

Generalizability Theory

- Experimental designs that yield more than one type of reliability coefficient for the same group permit the analysis of total score variance into different components.
- The statistical analysis developed by Cronbach, Glaser, & Rajaratnam (1972) called generalizability theory use ANOVA theory to partition source of variance.

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, and the test includes some items too difficult for anyone to solve, so that no one can get a perfect score.

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

Dependence of Reliability Coefficients on the Sample Tested

- When a test is to be used to discriminate individual differences within a more homogeneous sample than the standardized group, the reliability coefficient should be redetermined on such a sample.
- Reliability vary between groups differing in average ability level. (may be affected by floor or ceiling effect)

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:

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

Standard Error of Measurement

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

$$CI_{\%} = X \pm z_{\%/2} SEM$$

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.

Standard Error of Measurement

- Neither reliability coefficients nor errors of measurement can be assumed to remain constant when ability level varies widely.

Standard Error of Measurement

- It is particularly important to consider test reliability and errors of measurement when evaluating the different between two scores.
- Unless considering SEM, small differences may be overemphasized.

Standard Error of Measurement

- Standard error of different between two scores

$$\boxed{SE_{diff} = \sqrt{(SEM_1)^2 + (SEM_2)^2}}$$

- This standard error can be used to create confidence interval.

Reliability for Mastery Classifications

- The accuracy of test scores as domain score estimates is of less interest when the test is used to make mastery classifications.
- Decision consistency concerns the extent to which the same decisions are made from two different sets of measurements.

Reliability for Mastery Classifications

		Decision Based on Form 1		
		Nonmaster	**Master**	
Decision Based on Form 2	**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

- Four factors may affect decision consistency
  - Test length
  - Location of the cut score in the score distributions
  - Test score generalizability
  - Similarity of the score distributions for the two forms

Reliability for Mastery Classifications

- The more test length, the more probability of consistent decision.
- Decision consistency tends to be lowest when the cut score is close to the center of the test score distribution.
- Increasing generalizability tends to increase decision consistency
- P tends to be smaller for the group with a mean score close to cut score.

Reliability for Mastery Classifications

Number of Items	ρ^2	Cut score (Percent-Correct Scale)			
		.20	.40	.60	.80
5	.40	.81	.66	.68	.81
10	.57	.83	.71	.77	.90

Mean percent-score is .40 for all exams

Test Mean	ρ^2				
	.10	.30	.50	.70	.90
3.0	.57	.63	.69	.78	.90
4.8	.96	.93	.91	.91	.94

The cut score, expressed on the total score scale, is 3 for all entries in the table.

Reliability for Mastery Classifications

- Two forms of a test will tend to yield more-consistent decisions for a group characterized by heterogeneous domain scores than for a group characterized by homogeneous domain scores
- Substantial decision consistency can occur even when test score generalizability is low.

Reliability for Mastery Classifications

- Other things being equal, decision consistency tends to be smaller when test score distributions are dissimilar.

Reliability for Mastery Classifications

- When two tests have the same distributions, are statistical independent and have cutoffs at the median of score, P = .50.
- Corrected formula of P is

$$\boxed{P^* = 2P - 1}$$

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

- These formula are affected by these factors as same as *P*.
  - Test length
  - Location of the cut score in the score distributions
  - Test score generalizability
  - Similarity of the score distributions for the two forms

Reliability for Mastery Classifications

- Otherwise, *P*, *P** and κ can be computed in test-retest reliability, criterion-related validity and convergent validity.