Introduction to Reliability

- What is reliability?
  - Reliability is an index that estimates dependability (consistency) of scores
- Why is it important?
  - Prerequisite to validity because if you are not measuring something accurately and consistently, you do not know if your inferences are valid
  - Should not base decisions on test scores that are not reliable

2 sources of measurement error

1) Random - individual fluctuation

not too serious

[use of large samples corrects for this]

2) Systematic - due to test itself

big problem

makes test unreliable

Reasons to be concerned with reliability

- *Provides a measure of the extent to which an examinee's score reflects random measurement error.*
  - *Measurement errors can be caused by examinee-specific factors.*
    - *motivation*
    - *concentration*
    - *fatigue*
    - *boredom*
    - *momentary lapses of memory*
    - *carelessness in marking answers*
    - *luck in guessing*
  - *Measurement errors can be caused by test-specific factors.*
    - *ambiguous or tricky items*
    - *poor directions*
  - *Measurement errors can be caused by scoring-specific factors.*
    - *nonuniform scoring guidelines*
    - *carelessness*
    - *counting or computational errors.*

Reliability

The extent to which the assessment instrument yields consistent results for each student

---

- **How much are students' scores affected by temporary conditions unrelated to the characteristic being measured (test-retest reliability)**
- **Do different parts of a single assessment instrument lead to similar conclusions about a student's achievement (internal consistency reliability)**
- **Do different people score students' performance similarly (inter-rater reliability)?**
- **Are instruments equivalent (alternate/equivalent/parallel forms reliability)?**

Internal consistency reliability

- Involve only one test administration
- Used to assess the consistency of results across items within a test (consistency of an individual's performance from item to item & item homogeneity)
- To determine the degree to which all items measure a common characteristic of the person
- Ways of assessing internal consistency:
  - Kuder-Richardson (KR20)/Coefficient alpha
  - Split-half reliability

Alternate-forms reliability

- Used to assess the consistency of the results of two tests constructed in the same way from the same content domain
- To determine whether scores will generalize across different sets of items or tasks
- The two forms of the test are correlated to yield a coefficient of equivalence

Test-retest reliability

- Used to assess the consistency of a measure from one time to another
- To determine if the score generalizes across time
- The same test form is given twice and the scores are correlated to yield a coefficient of stability
- High test-retest reliability tells us that if examinees would probably get similar scores if tested at different times
- Interval between test administrations is important—practice effects/learning effects

Internal Consistency Reliability for Objectively Scored Tests

- KR20 (Coefficient Alpha)
- KR21

Internal Consistency
Cronbach's Alpha

- ***1951 article: Estimates how consistently learners respond to the items within a scale***
- ***Alpha measures the extent to which item responses obtained at the same time correlate highly with each other***
- ***The widely-accepted social science cut-off is that alpha should be .70 or higher for a set of items to be considered a scale***
- ***Rule: more items, the more reliable a scale will be (alpha increases)***

KR20

- ***Dichotomously scored items with a range of difficulty:***
  - ***Multiple choice***
  - ***Short answer***
  - ***Fill in the blank***
- ***Formula:***

$$KR20 = [n/(n - 1)] \times [1 - (\Sigma pq)/Var]$$

KR20 = estimated reliability of the full-length test
n = number of items
Var = variance of the whole test (standard deviation squared)
Σpq = sum the product of pq for all n items
p = proportion of people passing the item
q = proportion of people failing the item (or 1-p)

Coefficient Alpha

- *Items that have more than dichotomous, right-wrong scores:*
  - *Likert scale (e.g rate 1 to 5)*
  - *Short answer*
  - *Partial credit*
- *Formula:*

$$\text{Alpha} = [n/(n-1)] \times [(Var_t - \Sigma Var_i)/Var_t]$$

Alpha	= estimated reliability of the full-length test
n	= number of items
Var_t	= variance of the whole test (standard deviation squared)
ΣVar_i	= sum the variance for all n items

KR21

- *Used for dichotomously scored items that are all about the same difficulty*
- *Formula:*

$$KR21 = [n/(n-1)] \times [1 - (M \times (n - M) / (n \times Var))]$$

KR21	= estimated reliability of the full-length test
n	= number of items
Var	= variance of the whole test (standard deviation squared)
M	= mean score on the test

Limitations of KR20 and KR21

1. Single moment in time
2. Generalization across domains
3. Speededness

Reliability for Subjectively Scored Tests

- Training and scoring
- Intra-rater reliability
- Inter-rater reliability

Intra-rater Reliability

- Used to assess each raters' consistency over time
- Agreement between scores on the same examinee at different times

Inter-rater Reliability

- Used to assess the degree to which different raters/observers give consistent estimates of the same phenomenon
- Agreement between the scores assigned by two raters (calculated as a percentage of agreement between the two or a correlation between the two)
  - Exact agreement for 5 points or less
  - Adjacent agreement for more than 5 points

Strategies to enhance reliability

- Objectively Scored Tests
  - Write "better" items
  - Lengthen test
  - Manage item difficulty
  - Manage item discrimination
- Subjectively Scored Tests
  - Training of scorers
  - Reasonable rating scale

Write better items

- Item writing checklist
  - General item writing
  - Stem construction
  - Response option development

Lengthen test

- ***Spearman-Brown Formula***

$$r_{kk} = k(r_{11}) / [1 + (k - 1)r_{11}]$$

r_{kk} = reliability of the test k times as long as the original test
r_{11} = reliability of original test
k = factor by which the length of the test is changed

Lengthen test

- ***Example using Spearman-Brown Formula:***

A test is made up of 10 items and has a reliability of .67. Will reliability improve if the number of items is doubled, assuming new items are just like the existing ones?

k = 20/10 = 2
r_{kk} = 2(.67)/[1 + (2 - 1).67] = 1.34/1.67 = .80

Lengthen test

- *Considerations:*
  - *Time available for testing*
  - *Fatigue of examinees*
  - *Ability to construct good test items*
  - *Point of diminishing returns - increasing test length by a lot will increase reliability but not enough to make it worth the testing time needed*

Item difficulty

- *Proportion of examinees who answered the item correctly:*

  Item difficulty = $\frac{\text{\# of people who answered correctly}}{\text{\# of total people taking the test}}$

- *Goal of .60 - .80*

Item difficulty

- Item is probably too easy:

Choices	#Selecting
A.	4
B.*	90
C.	4
D.	2

Difficulty = 90/100 = .90

Item difficulty

- Item is probably too difficult:

Choices	#Selecting
A.	16
B.	48
C.*	26
D.	10

Difficulty = 26/100 = .26

Item difficulty

- *Item is reasonably difficult:*

Choices	#Selecting
A.*	76
B.	7
C.	3
D.	14

Difficulty = 76/100 = .76

Assessment of Observation (Measurement)

Observed Score = True Score + Error

Standard Error of Measurement

- *Amount of variation to be expected in test scores*
- *SEM numbers given in tests are typically based upon 1 standard error*
- *Example– Score is 52 SEM is 2.5*

  68% of scores between 49.5 and 54.5 based upon repeated testing