RELIABILITY

I. RELIABILITY
   = the reproducibility of scores re:
     1. the same test on different occasions
     1. different sets of equivalent items
     2. the same test under different conditions

Reliability is a measure of variation in scores,
-> also measures amount of error in testing (vs. construct of interest)

Common sources of error in testing
1. The person taking the test
   - “organismic” factors

2. The situation
   “environmental” variability

3. The examiner/scorer
   differences in different people administering/scoring a test

4. Test content
   a. items are representative (not biased)
   b. items tap the construct of interest

5. Time
   - “temporal” reliability/variability
   - traits vs. states
No test is invulnerable
- statements of reliability an estimate of the extent to which scores are attributable to
  “true” differences among people, vs. chance error

Reliability & validity

1. **Reliability is a necessary (but not sufficient) condition for validity**
   a. A test cannot measure a construct of interest if the test is too error-prone
   b. A perfectly reliable test may be invalid, however

2. **Reliability places a direct ceiling on validity**

II. **TYPES OF RELIABILITY**

A. **Temporal stability/test-retest reliability**
   - does test measures the same thing over time

   Error variance = random fluctuations over time
   - from organismic variables
   - from environmental variables
   - other time-related effects

   Reliability coefficient = correlation of scores of individuals on both testing occasions

   Time intervals recommended = 1 week to 6 months
2 problems with test-retest approach

1. Too short an interval is likely to result in examinees recalling their previous answers
   - leads to inflation in reliability coefficient from tapping memory

2. An interval that is too long leaves the possibility that the person’s true score on the construct has changed
   - lowers the reliability coefficient

A. Alternate forms reliability
   - do 2 different versions of the same test measure the same thing?

   Constructed to be equivalent
   - same number of items, same format, same difficulty level
   - only difference = specific items
   & should have similar psychometric properties
     - ~ same mean, SD, reliabilities, validity
   -> VERY difficult to accomplish

Content sampling
   - techniques used to obtain a sample of items from universe of potential items
   & the representativeness of the test items compared to the item universe

Alternate form reliability may reflect
1. Content sampling alone

or 2. Content sampling + time variance

Low alternate forms reliability = problems in content sampling across the forms
- not testing the same domains
Alternate-forms reliability coefficient
- correlate scores on the 2 forms
- same procedure as test-retest

C. **Split-half reliability**
do all items on a test measure same thing
- measure of content sampling
- termed measures of internal consistency reliability

To measure split-half reliability
- are the 2 halves strongly related?
1. Correlate odd vs. even items
   - must be measuring same construct

   High split-half reliabilities = content sampling is adequate
   & consistent between the 2 halves

**Problem:**
Split-half techniques assess reliability of a half-test
- a test with more items usually is more reliable than 1 with fewer items
- more likely to measure construct

-> thus, split-half reliability underestimates a test’s internal consistency
2. Spearman-Brown prophecy formula

\[
S-Br. \quad r = \frac{(N)(r_{xy})}{1 + (N - 1)(r_{xy})}
\]

- \( r \) = estimated reliability of whole test
- \( N \) = factor by which test is lengthened (\( N = 2 \) if split-half)
- \( r_{xy} \) = Pearson correlation between the split halves

EX  \( r_{xy} = .9, N = 2 \)

\[
SBr. \quad r = \frac{(2)(.9)}{1 + (2-1)(.9)} = \frac{1.8}{1.9} = .947
\]

- original split-half reliability = .9
- est. reliability for whole test = .95

EX  \( r_{xy} = .85, N = 2 \)

\[
SBr. \quad r = \frac{(2)(.85)}{1 + (2-1)(.85)} = \frac{1.7}{1.85} = .919
\]

- original split-half reliability = .85
- est. reliability for whole test = .92
Cronbach’s Coefficient Alpha
- based on consistency of responses to test items
  - are items all testing the same thing?
- alpha = mean of all possible split-half reliabilities
- based on variance

Tests can focus on just 1 construct
- or can include several domains
- the more different domains that are included, the lower the overall consistency of the scores

High coefficient alpha = items measure the same construct
- a single dimension
Low coefficient alpha = the test taps multiple domains

Internal consistency coefficients are index of uni/multidimensionality
- test developed to measure a multifaceted domain will necessarily be heterogenous
- & will have a lower coefficient of internal consistency
- also, scores from such a test might be difficult to interpret

But, because several domains are tapped, test probably has good predictive power
- more confidence in that one score

vs. A unidimensional test (1 facet) will have higher measure of reliability
- also is very easy to interpret
- but perhaps not as real-world
- not as useful in making predictions about a domain that is multifaceted
Solution: Multifaceted test with several homogenous subtests
- each subtest taps different facet of the construct
- whole test is used for predictive purposes

4. KR-20
- related to Cronbach’s alpha
- formula also uses variances
  BUT used for tests with dichotomous items

D. Inter-rater reliability - do different examiners obtain the same score?
- especially for projective tests, observations, subjective scoring

To calculate:
1. Correlate scores by different raters
2. or Adjust with Spearman-Brown
3. Use Kappa — for nominal ratings

Summary: each reliability coefficient estimates the E component of a score
III. STANDARD ERROR OF MEASUREMENT (SEm)

68% = (34.13 x 2) = 1 SD around the mean
vs. 84th percentile = 1 SD above the mean

95% = (34.13 x 2) + (13.59 x 2) = 2 SD around the mean
vs. 98th percentile = 2 SD above mean

99% (99.74) = (34.13 x 2) + (13.59 x 2) + (2.15 x 2) = 3 SD around the mean
vs. 99th percentile (99.87) = 3 SD above the mean

Confidence intervals around scores
EX given a child whose “true” IQ = 100
- have obtained score, not true score
- obtained = estimate of the true score
obtained ≠ true score (error)

Error is random
- give person 100 IQ tests
- the scores would form a normal distribution around the true score
- thus, 68% would be ± 1 SD of the true score, 95% ± 2 SD, etc.

To estimate true score from observed score
- need SD of the error --> from the reliability coefficient

Standard error of measurement (Sem) = the SD of the error in a measure
- the variability of observed scores around the true score
- a special case of SEe
Formula
SEm = SD (of test) x √1 - reliab. coeff.

EX  WAIS IQ SD = 15, reliability of .89
SEm = 15 x √1 - .89 = 15 x √.11 = 5

As reliability increases, SEm decreases
- reliability = .90, SEm =
- reliability = .95, SEm =
- reliability = .99, SEm =

If reliability is perfect (1.00), SEm = 0 — no error
SEm = 15 x √1 - 1 = 15 x √0 = 0

To estimate a person’s true score from an obtained score:
- confidence interval using SEm

EX  test SD = 15, reliability = .89, & SEm = 5
- obtained score of 90, 68% time, interval of ± 1 SD from 90
  => 68% C.I. = 90 - 5 to 90 + 5 = 85 to 95
  => 95% C.I. = 90 - 2(5) to 90 + 2(5) = 80 to 100
  => 99% C.I. = 90 - 3(5) to 90 + 3(5) = 75 to 105

- reminder: always a band of error around an obtained score
IV. STANDARD ERROR OF THE DIFFERENCE (SEd)

- are 2 scores different from each other?

EX vocabulary score = 12, math = 16

Different factors could contribute

1. A true difference in ability
2. Reliability — error of measurement in each score
   - band of error around each score
   - are the bands overlapping?

2 formulas for computing SEd

1. \[ \text{SEd} = \sqrt{\text{SEm}_1^2 - \text{SEm}_2^2} \]
   - square the SEm of each test/subtest
   - subtract one squared SEm from other
   - obtain square root

2. \[ \text{SEd} = SD \sqrt{2 - r_1 - r_2} \]
   - SD times the square root of 2 minus the reliability of each subtest
   - if tests have same standard deviation (e.g., 2 IQ subtests, SD = 3)

Formulas are theoretically equivalent

- but don’t necessary give same SEd
- depends on the reliability estimate(s)
- Different sources of error & different types of reliability

EX for 1 test

- inter-rater = .95 error = 5%
- alternate forms = .90 error = 10%
- test-retest = .90 error = 10%

Thus, the aggregate reliability = .75
Information on a test may not give aggregate reliability, but just 1 or 2 types

Convention to use C.I.s of 95% (2 Sds)
- almost never use 1 SD
- use of 3 SDs is probably unnecessary

EX  WAIS-R Vocabulary subtest = 12, Math subtest = 16
    SEM for Vocabulary = .61
    SEM for Math = 1.14

    \[ \text{SEd} = \sqrt{\text{SEM}_{1}^2 - \text{SEM}_{2}^2} \]
    \[ \text{SEd} = \sqrt{1.14^2 - .61^2} = \sqrt{1.30 - .37} = .96 \]

To be 95% confident that the vocabulary/math difference is significant
- is difference in raw scores > 2 x Sem?
    16-12 > 2(.96)
- yes, 4 > 1.92
A patient
- 68yo Caucasian male
Jan, 1990 — FSIQ = 94
Jan, 1991 — FSIQ = 83

WAIS-R  FSIQ SD = 15  FSIQ r = .96
- use 95% confidence

$SE_d = SD\sqrt{2} - r_1 - r_2$