Language test reliability

- On defining reliability
- Sources of unreliability
- Methods of estimating reliability
- Standard error of measurement

On defining reliability

- Non-technical definitions of reliability

A. Dictionary definition: something that is correct or true or someone who can be trusted

1. The incident cast doubt on her motives and reliability.
2. The reliability of these results has been questioned.

B. Stability or consistency of results or test scores

- Explanation of some key terms to understanding reliability

1. Observed score
2. True score
3. Error score
4. Mean
5. Variance
6. Correlation
7. Parallel tests: length of the test; item types; difficulty of item types; equivalence; independence

Correlation between parallel tests

A. Observed score = true score + error score
B. Uncorrelated nature of error scores with true scores

$$r = \frac{Vx - Ve}{Vx}$$

Sources of (error) variance

I. Meaningful variance
II. Measurement error or error variance
A. Systematic sources of error
B. Unsystematic sources of error

Potential sources of error variance

I. Variance due to environment

A. Location

B. Space

C. Ventilation

D. Noise

E. Lighting

F. weather

II. Variance due to administration procedures

A. Directions

B. Equipment

C. Timing

D. Mechanics of testing

III. Variance due to scoring procedures

A. Subjectivity

B. Evaluator biases

C. Evaluator idiosyncrasies

IV. Variance attributable to the test and test items

A. Test booklet clarity

B. Answer sheet format

C. Particular sample of items

D. Item types

E. Number of items

F. Item quality

G. Test security

V. Variance attributable to examinees
A. Health
B. Fatigue
C. Motivation
D. Emotion
E. Memory
F. Concentration
G. Forgetfulness
H. Carelessness
I. Testwiseness
J. Guessing
K. Chance knowledge of item content

Methods of estimating reliability

I. Some preliminaries

A. Reliability coefficient or reliability estimate

B. Range of reliability coefficient: 00.0 to 1.00

C. Correlation coefficient to estimate reliability

Theories of reliability

- Classical test score measurement theory (CTS model)
- Generalizability theory (G-theory)
- Item response theory (IRT)

Methods of estimating reliability

I. Internal-consistency reliability methods
A. Split-half method
1. The spearman-Brown split-half method
2. The Guttmann split-half method
B. Kuder-Richardson methods
1. KR-20
2. KR-21
C. Cronbach alpha
D. Rater estimates of reliability
1. Intrarater reliability
2. Interrater reliability

II. Stability methods
III. Equivalence methods

Internal-consistency reliability methods

- Split-half method
- Definition:

1. Dividing the same test into two parts and administering it to the same testees only once
2. Measurement of the same trait or ability of the two parts—homogeneity of items
3. Independence of the two parts
4. Equivalence of two parts
5. Importance of length
6. Difficulty of items

- Ways of splitting the test

1. Easy-to-difficult method
2. Odd-even method

- Estimating reliability from split-halves

1. Adjustment for full-test reliability (Spearman-Brown formula)

$$r = \frac{2\,(\text{r half})}{1 + (\text{r half})}$$

$$r = \frac{2\,(0.95)}{1 + (0.95)}$$

$$= \frac{1.90}{1.95} = 0.97$$

2. Guttman (no additional correction for length): independent but not equivalent

$$r_{xx'} = 2\,(1 - S^2_{h1} + S^2_{h2}/S^2_x)$$

$$r_{xx'} = 2\left(1 - \frac{.75 + .65}{.90}\right) = 2\left(\frac{0.4}{.90}\right) =$$

$$2\,(0.4444444) = 0.8888889$$

Advantages and disadvantages

I. Advantages
A. Practicality
1. No twice administration of the same test
2. No two different versions of the same test

II. Disadvantages
A. Insurance of homogeneity
B. Different subsections of the same test

KUDER-RICHARDSON METHODS

I. Kuder-Richardson formula 21

$$K - R21 = k/\ k - 1\ (1 - M\ (k - M)/kS^2$$

K-R21 =Kuder-Richardson formula 21

K = number of items

M = mean of test scores

S^2 = variance of test scores

II. Kuder-Richardson formula 20

$$r_{xx'}\ (K{-}R20) = k/k - 1\ (1 - \sum S_i^2/S_t^2)$$

K-R20 = Kuder-Richardson formula 20

K = number of items

ΣS_i^2 = item variance

S_t^2= test score variance

$$K-R21=\frac{k}{k-1}\left(1-\frac{M(k-M)}{kSD}\right)$$

$$K-R21=\frac{60}{60-1}\left(1-\frac{48(60-48)}{60(12.96)}\right)$$

$$=K-R21=\frac{60}{59}\left(1-\frac{576}{777.6}\right)$$

$$=K-R21=1.0169492\ (1-\ 0.7407407)$$

$$=K-R21=1.0169492\ x\ 0.2592593$$

$$=0.2636535$$

$$r_{xx'}\,(K\text{–}R20) = k/k - 1\,(1 - \sum S_i^2/S_t^2)$$

$$r_{xx'}(KR-20) = \frac{60}{59}\left(1 - \frac{1.55}{12.96}\right) =$$

$$1.0169492 \times 0.8804012 = 0.8953233$$

Advantages, assumptions and differences

I. Kuder-Richardson formulae
A. No administration of the same test twice
B. Lack of two different versions of the same test
C. No separate scoring of odd and even numbered items
D. No correlation coefficient calculation
E. No adjustment for length

II. Assumptions
A. Equality of items
B. Independence of items scored
C. Measurement of the same trait

III. Differences

A. K-R21 is simpler to calculate and more common in language testing than K-R20

B. K-R21 is more conservative than K-R20, yielding a lower reliability coefficient.

C. K-R20 is more accurate, yielding a higher reliability estimate.

Cronbach alpha or coefficient alpha

$$\alpha = k/k\text{ -1 }(1 - \sum s^2i/s^2x)$$

α = 40/40 -1(1 – 0.95/0.99) = 1. 025641 x 0.040404 = 0.04144 (based on item variances)

α = 2/2 -1 (6.25+ 1.21/12.96) = 2 (0.4243827) = 0.8487654 (based on two halves of the test)

Rater (scorer, mark-remark, or marker) reliability

I. Inter-rater reliability: reliability of ratings between two or more raters

A. Correlation between ratings of raters

B. Spearman-Brown prophecy formula to estimate full-rating reliability

Correlation coefficient for four raters:

Interrater correlations for a scored interview (N = 60)				
R1	1.00			
R2	.65	1.00		
R3	.75	.80	1.00	
R4	.55	.90	.70	1.00
	R1	R2	R3	R4

Spearman-Brown prophecy formula

$$r^{xx'} = n \times r / (n - 1)\, r + 1$$

$r^{xx'}$ = full-test reliability

r = correlation between the two test halves

n = number of times the test length is to be increased

= 4 x 0.55/ (4 – 1) .55 + 1

= 2.2/ 2.65

= 0.8301887

II. Intrarater reliability: reliability of ratings of the same rater over time

A. Use Pearson-product moment correlation coefficient when ratings are dealt with as parallel forms

B. Use Cronbach alpha when odd-even approach is adopted

Which one to choose?

- Conceptual clarity
- Ease of calculation
- Accuracy of results
- Frequency of appearance

Standard error of measurement

I. Definitions

A. Another more accurate way of looking at the consistency of a set of test scores

B. Interpretation of individually observed scores

C. Determination of a band around a testee's score within which that score to fall in case of repeated administrations of the same test to the same testee

D. Estimation of the probability with which the tester to expect scores to fall within one SEM, two SEMs, or more

E. Standard deviation of error scores across students on a given measure across various situations

$$Formula\ for\ SEM = S_x\sqrt{1-r}$$

$Formula\ for\ SEM = 4\sqrt{1-0.64}$

$= 4\sqrt{0.36}$

$= 4 \times 0.6$

$= 2.4$

Observed score minus or plus one SEM 15 ∓ 2.4

$= 15 - 2.4 = 12.60$

$= 15 + 2.4 = 17.40$

$Band = 12.60\ to\ 17.40$

$15 - 4.8 = 10..20$

$15 + 4.8 = 19.80$

$11.20\ to\ 19.80$

$15 - 7.20 = 7.80$

$15 + 7.20 = 22.20$

$7.80\ to\ 22.20$