

“Classical” or “Traditional” Test Theory

Basic concepts:

x is the observed summed-score

t is the “true score”

e is the “error”

$$x = t + e$$

The “true score” is a hypothetical entity with many definitions. The most commonly-used definition is that it is $E(x)$; that is, it is the expected value of x , or the mean of all the x s that would be obtained in (hypothetically) infinitely many retestings. The error component is defined by subtraction:

$$e = x - t$$

The Standard Results:

$$E(x) = E(t)$$

$$Var(x) = Var(t + e)$$

$$= Var(t) + 2Cov(t, e) + Var(e)$$

$$= Var(t) + Var(e)$$

Reliability

$$r_x = \text{Var}(t) / \text{Var}(x)$$

(by definition), and

$$\text{Var}(t) = \text{Var}(x) r_x$$

For two parallel measurements, x and x' :

$$\begin{aligned}r_{xx'} &= \frac{xx'}{x \ x'} \\ &= \frac{(t+e)(t+e')}{x \ x'} \\ &= \frac{t^2 + te + te' + ee'}{x \ x'} \\ &= \frac{t^2}{x} / \frac{x}{x'}\end{aligned}$$

Thus,

$$r_x = \frac{t^2}{x} / \frac{x}{x'} = r_{xx'}$$

Attenuation:

$$r_{xy} \quad r_{tx} = \sqrt{r_x} = \sqrt{r_{xx'}} ;$$

in other words, nothing can correlate more with x than its reliability (except by sampling error...).

Assessing reliability (or generalizability)

Test-retest (or generalizability over time):
Test twice, and $r_{xx'}$ is the estimate. This is the most intuitively appealing way to think about reliability, until you think about the psychology of testing twice. Then it almost never works very well.

Alternate forms (or generalizability over forms.

Test twice, using two different forms, essentially at the same time and $r_{xx'}$ is the estimate. It is frequently difficult to generate parallel alternate forms, except in major testing programs (where it is routine... and expensive).

Split halves—alternate forms within a single test:

Test once, split the test items into two halves. Score the two halves as x and x' and $r_{xx'}$ is an estimate of the reliability of tests half as long as yours. Unfortunately, that is not the reliability of your test.

Many years ago, Spearman (1910) and Brown (1910) did the algebra (dignified as the “Spearman Brown prophecy formula”) explaining that, under the assumptions of the traditional test theory plus the assumption that all of the items are equally correlated with each other,

$$r_{X^*X^{**}} = \frac{2r_{XX'}}{1 + r_{XX'}}$$

In general, the reliability of a test N times longer than the one giving the score x is

$$r_{X^*X^{**}} = \frac{Nr_{XX'}}{1 + (N-1)r_{XX'}}$$

Tables and graphs are available.

Internal Consistency Formulae:

A serious problem with split-half methods is, how do you split the halves? Different splits give different answers. Solution: Compute all possible split-half reliabilities, average them, and use the Spearman-Brown formula to correct for length. The result is called coefficient r_{tt} .

Coefficient

There are other derivations of this same number; and it is almost always computed using some computer program (it is not computed by splitting the test into all possible halves!). But the most informative formula is

$$= \frac{Nr}{1 + (N - 1)\bar{r}}$$

where N is the number of items and \bar{r} is an average interitem correlation.

Coefficient is just the Spearman-Brown formula taking the average of the interitem correlations as the reliability of little one-item tests.

Also note that as long as the average interitem correlation stays *positive*, the longer you make the test the more reliable it is. So what?

KR20 (so named because it is equation number 20 in Kuder & Richardson, 1937) is the special case of coefficient α for binary test items (α works for Likert type scales and such). KR21 is the next formula in the '37 paper, which is an over-simplified form assuming the items of the test are identical... it used to be used before computers, but has no sensible use now.

The Standard Error of Measurement:

$$SEM = SD \sqrt{1 - r_{xx'}} ;$$

The standard error of the *difference* between two scores is

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

E.g., rough numbers for the CEEB scales are

$$SEM = 100 \sqrt{1 - 0.92} = 30$$

$$SE_{diff} = \sqrt{(30)^2 + (30)^2} = 40$$

Test Reliability	Proportion of tests differing by more than			
	50 points	100 points	150 points	225 points
0.00	72%	48%	29%	11%
0.40	65%	36%	17%	4%
0.50	62%	32%	13%	2%
0.60	58%	26%	9%	1%
0.70	52%	20%	5%	0.4%
0.75	48%	16%	3%	0.2%
0.80	43%	11%	2%	0.1%
0.85	36%	7%	1%	0.01%
0.90	26%	3%	0.1%	0.001%
0.92	21%	1%	0.0%	0.0002%
0.94	15%	0.4%	0.0%	.00003%
1.00	0%	0%	0%	0%

Scatterplot of scores for two parallel forms; the reliability of each is

0.40

0.85

0.94





