RELIABILITY

Contents

If viewing on screen you can click on a contents item above to jump to the page the item is on

1. Definition and basic formulas

Reliability theory is concerned with the measurement, nature, and control of error in measurement by psychological (and other) tests.

Some knowledge of reliability theory and its associated calculations is essential to the informed selection and interpretation of test results.

This module will provide an introduction to classical reliability theory, and the formulas based on it, which are likely to prove useful in clinical practice.

Reliability theory assumes that the variance of a test can be split into two components:

1. true variance;
2. error variance

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2$$

where :

σ^2_x = total test variance

σ^2_t = true variance

σ^2_e = error variance

It is assumed that:

1. Mean error = zero
2. The correlation between error scores and (1) true scores; and (2) other error scores = zero

Definition - the **reliability coefficient** of a test - r_{xx} - is the proportion of its variance which is true variance.

$$r_{xx} = \frac{\sigma_t^2}{\sigma_x^2}$$

The proportion of variance which is true variance can be estimated from the correlation between 2 parallel tests. Parallel tests are hypothetical tests with different items, which nevertheless have the same mean; the same variance; and the same correlations with other tests and variables. In real life we have to use approximations to them. Most of the best known methods for assessing reliability are attempts to approximate to parallel tests (but see also Section 14 below on Coefficient Alpha):

1. **test-retest** – the same test is given on two separate occasions, and the reliability coefficient is estimated from the correlation between scores on first and second administrations.
2. **split-half** – the test is divided up into two halves. The reliability coefficient is estimated by correlating one half with the other half. An adjustment has to be made to compensate for the fact that the half tests will inevitably be less reliable than the longer full one (see 3.9 below)
3. **parallel form** – a second test with the features of the first but with different test items is constructed. The reliability coefficient is assessed by correlating the two tests

Why does the correlation between two parallel tests indicate the proportion of variance which is true variance?

Before giving the proof, let's spell out some assumptions again.

- An individual's score on a test is made up of two components - the true score for that person and an error score.
- The error score can be positive or negative.
- The mean error score, for an individual tested repeatedly or for a group of individuals tested simultaneously, is zero
- The correlation between error scores and other error scores and true scores is zero.

All of this means that the following are assumed, or follow from these assumptions.

$$X = T + E$$

$$M_x = \overline{T} + \overline{E} = \overline{T}$$

$$x = (T + E) - \left(\frac{\sum (T + E)}{N} \right) = T - \overline{T} + E - \overline{E} = t + e$$

where:

X = test score

T = true score

E = error score

$x = X - Mean\ X$

$t = T - Mean\ T$

$e = E - Mean\ E$

In what follows, remember that:

$$r_{xy} = \frac{\sum xy}{N\sigma_x\sigma_y}$$

$$\therefore r_{xy}\sigma_x\sigma_y = \frac{\sum xy}{N}$$

If we call the first test X_1 and the second X_2, the correlation between them will be:

$$r_{x_1x_2} = \frac{\Sigma(t+e_1)(t+e_2)}{N\sigma_{x_1}\sigma_{x_2}} = \frac{\sum t^2 + \sum te_1 + \sum te_2 + \sum e_1e_2}{N\sigma_{x_1}\sigma_{x_2}}$$

$$= \frac{\sum t^2/N + \sum te_1/N + \sum te_2/N + \sum e_1e_2/N}{\sigma_{x_1}\sigma_{x_2}}$$

$$= \frac{\sigma_t^2 + r_{te_1} + r_{te_2} + r_{e_1e_2}}{\sigma_{x_1}\sigma_{x_2}}$$

because any correlation between error and error, and error and true scores = 0, and because $\sigma_{x_1} = \sigma_{x_2}$ this will equal:

$$r_{x_1x_2} = \frac{\sigma_t^2 + 0 + 0 + 0}{\sigma_x^2}$$

$$\therefore r_{x_1x_2} = \frac{\sigma_t^2}{\sigma_x^2} = r_{xx}$$

2. Correlation between obtained scores and true scores

The reliability coefficient is in fact a coefficient of determination which tells us the proportion of variance accounted for by the correlation between true scores and obtained or actual scores.

Just as the correlation coefficient in general is the square root of the coefficient of determination, the correlation between true scores and obtained scores will equal the square root of the reliability coefficient.

$$r_{tx} = \sqrt{r_{xx}}$$

3. Correlation between obtained scores and error scores

Error scores have a positive correlation with obtained scores. The correlation is:

$$r_{xe} = \sqrt{1 - r_{xx}}$$

The diagram below shows how mean error scores increase with:

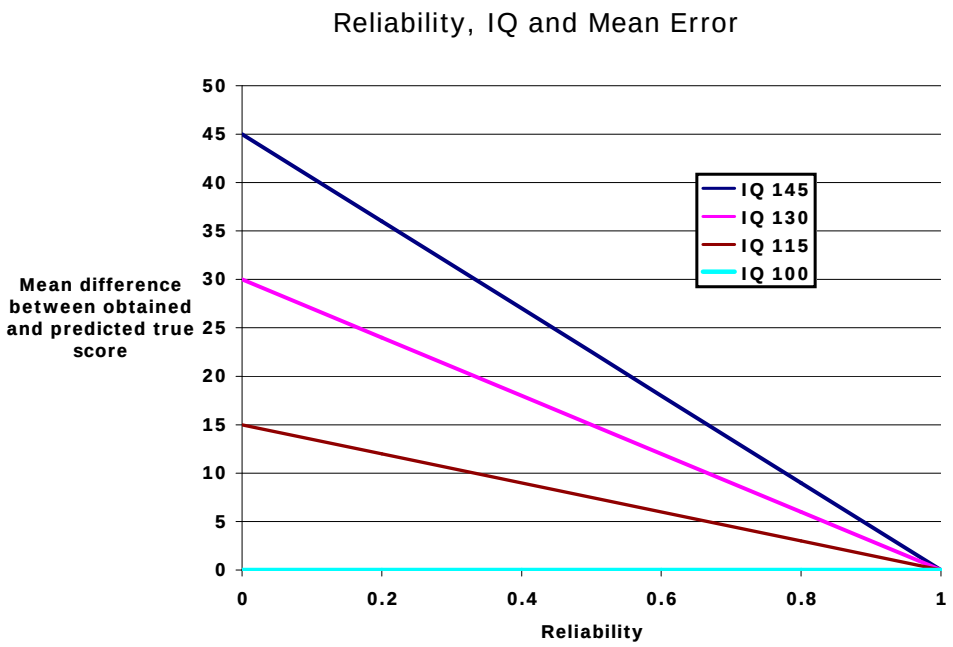
(a) distance of obtained score from the mean;
and
(b) reductions in reliability.

(The scores are on a hypothetical intelligence test with mean = 100, sd = 15)

You can see that error **increases**

1. with increasing distance of the obtained score from the mean
2. as reliability decreases.

Thus with a reliability of 0.7, the difference between obtained and true scores at IQ 55 or IQ 145 would be 13.5 points on average, while at IQ 100 it would be zero on average.



4. Predicting the true score from the obtained score

The predicted true score will be:

$$\hat{T} = r_{xx}(X - M_x) + M_x$$

This might look a little peculiar at first glance as the reliability coefficient is a coefficient of determination, and usually in prediction we use the correlation coefficient NOT the coefficient of determination. And we know that the correlation between obtained scores and true scores is the square root of the reliability coefficient.

What has happened here is this. The raw score equation for predicting Y from X is:

$$\hat{T} = \sqrt{r_{xx}}\frac{\sigma_t}{\sigma_x}(X - M_x) + M_t$$

but

$$\sqrt{r_{xx}} = \sqrt{\frac{\sigma_t^2}{\sigma_x^2}} = \frac{\sigma_t}{\sigma_x}$$

$$so \sqrt{\frac{\sigma_t^2}{\sigma_x^2}} \times \frac{\sigma_t}{\sigma_x} = \frac{\sigma_t}{\sigma_x} \times \frac{\sigma_t}{\sigma_x} = \frac{\sigma_t^2}{\sigma_x^2} = r_{xx}$$

Actual true scores will be normally distributed around this predicted true score with a standard deviation of: $\sigma_x\sqrt{r_{xx}(1-r_{xx})}$. So the 95% confidence limits for the range within which the true score will lie will be: $r_{xx}(X-M_x)+M_x$ plus or minus 1.96 x $\sigma_x\sqrt{r_{xx}(1-r_{xx})}$ (But see later discussion in Confidence Interval module)

5. Predicting the obtained score from the true score.

The equation for predicting the obtained score from the true score is:

$$\hat{X} = T \ .$$

This is because:

$$\hat{X} = \sqrt{r_{xx}}\left(\frac{\sigma_x}{\sigma_t}\right)(T-M_t)+M_x$$

$$\because M_t = M_x \text{ and } \because \sqrt{r_{xx}} = \frac{\sigma_t}{\sigma_x}$$

$$\hat{X} = \left(\frac{\sigma_t}{\sigma_x}\right)\left(\frac{\sigma_x}{\sigma_t}\right)(T-M_x)+M_x \text{ so}$$

$$\hat{X} = T$$

Actual obtained scores will be normally distributed around the predicted obtained score with a standard deviation of: $\sigma_x\sqrt{1-r_{xx}}$. The 95% confidence limits for the obtained score predicted from the true score are: T plus or minus $1.96 \times \sigma_x\sqrt{1-r_{xx}}$ (See also discussion in Confidence Interval module)

The following table summarises the various predicted scores and their standard errors in relation to reliability.

You will note that predicting a retest score from the score on first testing is simply the formula for predicting one score from another with r_{xx} substituted for r_{xy}.

The table gives both raw score and Z-score values.

Measures of dispersion in reliability problems		
Question	**Predicted score**	**Standard deviation**
Given an obtained score, what is the best estimate of the range within which the true score lies? (But see the Confidence Interval module also)	(a) $r_{xx}x + M_t$ (b) $Z_x\sqrt{r_{xx}}$	(a) $\sigma_x\sqrt{r_{xx}(1-r_{xx})}$ (b) $\sqrt{(1-r_{xx})}$
Given a true score, what is the best estimate of the range within which obtained scores will lie?	(a) T (b) $Z_t\sqrt{r_{xx}}$	(a) $\sigma_x\sqrt{1-r_{xx}}$ (b) $\sqrt{1-r_{xx}}$
Given an obtained score, what is the best estimate of the range within which a re-test score will lie?	(a) $r_{xx}x + M_x$ (b) $r_{xx}Z_{x_1}$	(a) $\sigma_x\sqrt{1-r_{xx}^2}$ (b) $\sqrt{1-r_{xx}^2}$

6. Tests for the reliability of a difference between test scores.

This Section is concerned with the reliability of a difference between two scores. The scores can be scores on full tests or on subtests. There are two main categories of situation here.

The first is the one where both of the scores being compared are on the same test.

The second is where the scores being compared are on two different tests

The main situations in which we might want to compare two scores derived from one test or subtest are:

1. When we want to compare two scores for one individual who has taken the same test (or subtest) twice
2. When we want to compare an individual's score on one test with their score on a different test (or subtest)
3. When we want to compare the scores of two different individuals who have taken the same test (or subtest)

Be warned though, right at the outset, that the procedure is designed to tell us only whether there is reason to suppose that there is a real difference between true scores. A reliable difference does **not** mean that there is an **abnormally** large difference between scores.

The table below provides values of $Z_{difference}$ significant at various probability levels for both one- and two-tailed tests of significance.

Conventional significance level values of $Z_{difference}$ for tests of the reliability of differences		
Level of significance	**One-tailed**	**Two-tailed**
.05	1.645	**1.96**
.01	2.326	**2.576**
.001	3.090	**3.291**

6.1 Differences between Scores on the Same Test for the Same Individual on Two Occasions

The problem here is to find whether a change in scores obtained on two separate occasions is likely to be due to chance or whether it represents a change in true scores. To solve the problem, we need the distribution of differences between two obtained scores when the true scores are in fact the same. If we had the standard deviation of the distribution of differences between obtained scores when the true scores are the same, we could work out a Z score for the difference we obtain and look this up in tables for the normal curve. We could then see what proportion of differences, when true scores do not differ, would be larger than the one we have obtained. This proportion would give us the probability that the two obtained scores in fact represent two identical true scores. It is not too difficult to work out what the distribution should be. We will use deviation scores to make the derivation simpler.

σ^2_{diff} is the variance of the difference between scores:

(1) $$\sigma^2_{diff} = \frac{\sum\left((t_1 + e_1) - (t_2 + e_2)\right)^2}{N}$$

(2) But as we are interested in the situation where $t_1 = t_2$, this becomes:

$$\frac{\sum(e_1 - e_2)^2}{N} = \frac{\sum(e_1^2 + e_2^2 - 2e_1e_2)}{N}$$
$$= \frac{\sum e_1^2}{N} + \frac{\sum e_2^2}{N} - \frac{2\sum e_1e_2}{N}$$

The first two terms are error variances whose square roots will be standard errors of measurement, and the third term is a covariance term. It equals $2r_{e_1e_2}\sigma_{e_1}\sigma_{e_2}$ As the correlation between errors scores is zero the term is equal to zero.

This leaves us with two error variances for the same test, i.e., $2\sigma^2_{meas}$ but the error variance of a test equals $\sigma_x^2(1 - r_{xx})$ so the variance of the difference scores will equal $2\sigma_x^2(1 - r_{xx})$ or $\sigma_x^2(2 - 2r_{xx})$

If we take the square root of this we get the standard deviation of differences in scores, due to errors of measurement, between two individuals on the same test.

The usual formula is therefore:

$$\sigma_x\sqrt{(2 - 2r_{xx})}$$

Using Z-scores the standard deviation becomes 1, so a test for the significance of the difference is:

$$Z_{diff} = \frac{Z_{x_1} - Z_{x_2}}{\sqrt{2 - 2r_{xx}}}$$

A possible complication here is that, if an individual takes the same test twice, there are might well be practice effects. If these are known the formula is modified to take account of them.

If we call practice effect ***p,*** and remembering that we are working in Z score***s,*** this gives the modified formula:. For this purpose Z_{x2} is the second test, i.e., the one on which a practice effect might occur.

$$Z_{diff} = \frac{Z_{x_1} - \left(Z_{x_2} - \frac{P}{\sigma_x} \right)}{\sqrt{2 - 2r_{xx}}}$$

6.2 The reliability of a difference between 2 scores for one individual on two different tests.

In this case the distribution of interest is the distribution of differences between obtained scores on **two different** tests or subtests when the scores on each test are in fact the same.

For this to be a sensible procedure the scores on each test should be in comparable units. e.g. T scores. IQs with the same means and standard deviations or Z scores, because we are not interested in differences in the Scores as such. but in differences in the individuals' relative standing on the two tests.

The derivation of the formula for σ^2_{diff} follows the same steps as those above. except that we have *x* and *y* as our deviation scores.

(1) $$\sigma^2_{diff} = \frac{\sum \left(\left(t_x + e_x \right) - \left(t_y + e_y \right) \right)^2}{N}$$

(2) But as we are interested in the situation where $t_1 = t_2$, this becomes:

$$\frac{\sum(e_x - e_y)^2}{N} = \frac{\sum(e_x^2 + e_y^2 - 2e_x e_y)}{N}$$
$$= \frac{\sum e_x^2}{N} + \frac{\sum e_y^2}{N} - \frac{2\sum e_x e_y}{N}$$

As before there are two error variances and a covariance term. Once more the covariance term involves a correlation between error scores and so becomes equal to zero.

Remembering that error variance equals $\sigma_x^2(1 - r_{xx})$, we can express the above equation as

$$\sigma^2_{diff} = \sigma_x^2(1 - r_{xx}) + \sigma_y^2(1 - r_{yy})$$

Working with Z-scores the standard deviations will equal 1, so the formula for the standard deviation of a difference, due to unreliability, between scores on tests X and Y becomes

$\sqrt{(1 - r_{xx}) + (1 - r_{yy})}$ or perhaps easier $\sqrt{2 - (r_{xx} + r_{yy})}$

The test for a difference between scores is therefore:

$$Z_{diff} = \frac{Z_x - Z_y}{\sqrt{2 - (r_{xx} + r_{yy})}}$$

Example

On a test of intelligence involving visual material an individual scores at the 75th percentile, while on a test involving verbal material the score is at the 50th percentile. Is there any reason for supposing that the difference between the test materials is affecting the individual's performance, if the reliabilities of the tests are .80 and .84 respectively?

Answer. Placing these values in the formula we get:

$$Z_{diff} = \frac{.67 - 0}{\sqrt{2 - (.80 + .84)}} = \frac{.67}{.6} = 1.12$$

Consulting tables for the normal distribution we find that the two-tail p value for this difference is.26. Over a quarter of differences would be larger than the one we have found.

6.3 Differences between Two Individuals on the Same Test

This test is included for the sake of completeness. In clinical practice we seldom meet this problem, which is likely to occur more often in selection and educational settings.

The problem is to decide how likely it is that two different obtained scores represent differences in the true scores of two individuals. Is A more intelligent than B? Is A more anxious than B? Does A have a poorer memory than B? And so on.

Using Z-scores, the test for the significance of the difference is:

$$Z_{diff} = \frac{Z_{x_1} - Z_{x_2}}{\sqrt{2 - 2r_{xx}}}$$

Problem

A obtains a score of 90 on a test with a mean of 100 and a standard deviation of 10. On the same test B obtains a score of 104. If the reliability coefficient of the test is .755, with what degree of certainty can we conclude that B's score is really higher than A's?

Answer

$$Z_{diff} = \frac{-1-0.4}{\sqrt{2-(2\times.755)}} = \frac{-1.4}{\sqrt{.49}} = \frac{-1.4}{0.7} = 2.0$$

The difference between A's score and B's score is 14 points. A difference of 14 points is, therefore, 2 standard deviations away from the mean of differences obtained on a chance basis when true scores are actually the same. Less than 5 per cent of chance differences will be as large as this. On a two-tail test $p < .05$.

Sometimes in clinical practice we do not know or are not sure about one or both of the reliabilities of the tests or measures we are using. Perhaps, for example, reliabilities reported for the tests vary considerably in studies using those tests.

But if we were to calculate the minimum reliabilities required for an observed difference to be statistically significant, we might find that even at the lowest reliabilities reported, the difference is a reliable one.

In such cases it is possible to re-arrange the formula above to tell us what the average reliability of our two tests would need to be for the difference we observe to be unlikely to be due simply to errors of measurement. For example, to calculate the reliabilities required for an observed difference (Zx - Zy) to be significant at a required level, we use the formula:

$$\bar{r}_{xx} = \frac{2-\left(\frac{Z_1 - Z_2}{z_{nd}}\right)^2}{2}$$

Z_{nd} is the z in the normal distribution corresponding to the desired level of significance, e.g., for $p < .05$, (2 tail), $Z_{nd} = 1.96$.

The formula gives the value of the mean (of the two reliabilities) required for the observed difference to be reliable.

Example

You have administered two tests of visual memory – one involving meaningful shapes or figures and the other having essentially random figures as stimuli.

The tests have not yet been commercially published, but from your reading of the literature they seem valid. Reliability data reported for one test range from .5 to .75, and for the other test from .6 to .8.

The person to whom you have administered the test obtains a Z score of 1.5 on the first test, and a Z score of 0.2 on the other. How big would the mean correlation need to be for this difference to be judged reliable at the two-tail .05 level?

Filling in the appropriate values in the formula we get the following sum:

$$\bar{r}_{xx} = \frac{2 - \left(\frac{1.5 - 0.2}{1.96}\right)^2}{2} = 0.78$$

This value is higher than the mean of the highest reliabilities reported for the two tests, so we cannot conclude that this is a reliable difference.

It is also possible to calculate, given the reliability coefficients of the tests, how big a difference (Zx - Zy) is needed for significance at any desired level. The formula is:

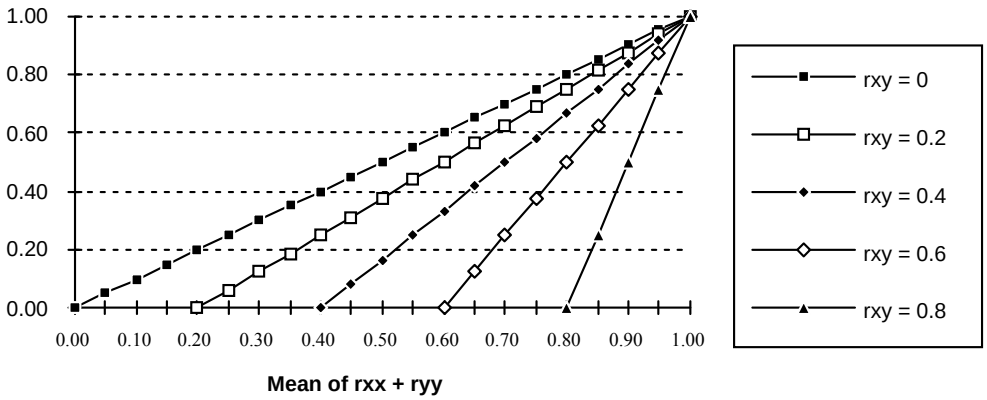
$$\left(Z_x - Z_y\right)_{p<x} = z_{nd}\sqrt{2 - (r_{xx} + r_{yy})}$$

7. The reliability of difference scores.

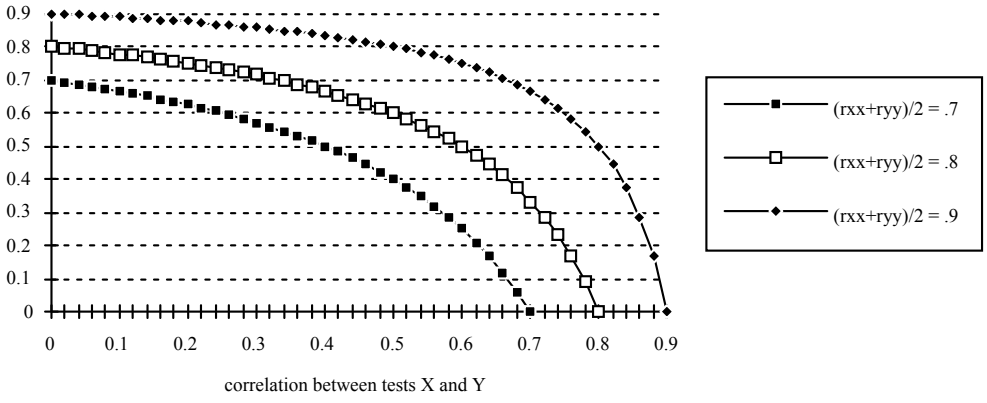
When we deal with differences observed between 2 scores we are dealing with a variable of much lower reliability than we might guess. The formula for the reliability of difference scores, i.e. (X - Y), is:

$$r_{(x-y)(x-y)} = \frac{r_{xx} + r_{yy} - 2r_{xy}}{2(1 - r_{xy})}$$

In practice clinicians often make great use of observed differences between subtest and test scores in interpreting test results. So the low reliability of difference scores should be emphasized. Notice from the formula and the graph below that the reliability of difference scores decreases as the reliability of the tests on which the difference score is based decreases:



It also decreases as the correlation between the tests increases.



It is probably also worth recalling that difference scores present another problem. This arises in attempts to find predictors of change in treatment and the like. It is quite common for investigators to see whether the pre-treatment score is correlated with change in score following treatment. The problem with this is that there is inevitably such a correlation as a result of artifact:

$$r_{Z_1(Z_2-Z_1)} = \frac{\sum Z_1(Z_2 - Z_1)}{N\sigma_{Z_1}\sigma_{(Z_2-Z_1)}} = \frac{\sum Z_1 Z_2}{N(1)\left(\sqrt{2-2r_{12}}\right)} - \frac{\sum Z_1^2}{N\left(\sqrt{2-2r_{12}}\right)} = \frac{r_{12}-1}{\sqrt{2-2r_{12}}}$$

As the correlation between the two tests will always be less than 1.0, this means that, in the absence of any "true" (non-artefactual) relationship between initial score and response to intervention, there will be a negative correlation between initial scores and amount of change recorded after treatment.

8. How to make a test more reliable (or less reliable!)

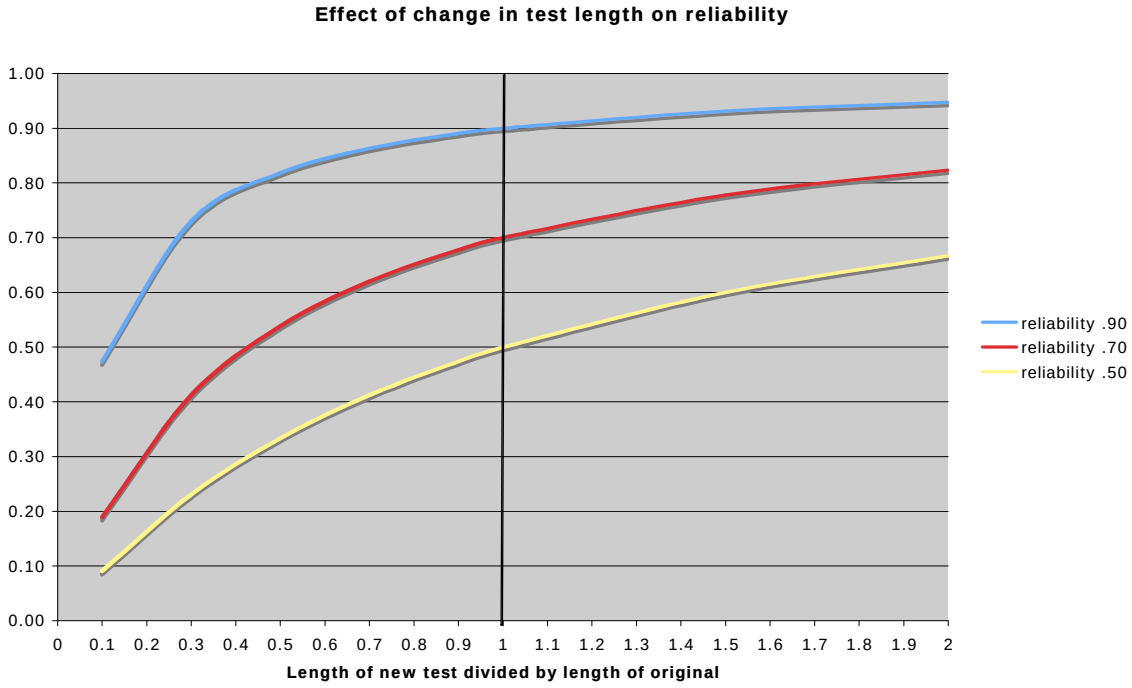
The general principle is that lengthening a test makes it more reliable. Shortening a test makes it less reliable. The test is of course lengthened by adding items which measure the same variable as the existing items. The formula to use to estimate the reliability of the lengthened or shortened test is called the Spearman-Brown Prophecy Formula. It is as follows:

$$\hat{r}_{xx} = \frac{mr_{xx}}{1+(m-1)r_{xx}}$$

Where

m = (length of modified test) divided by (length of original test)

The effects of changes in test length on reliability are shown graphically below.



Change in test length required to attain a desired level of reliability.

The formula can be manipulated to give the change in test length required to achieve a given reliability:

$$m = \frac{r_{xx(desired)}\left(1-r_{xx}\right)}{r_{xx}\left(1-r_{xx(desired)}\right)}$$

9. Effects of reliability on validity.

The highest correlation a test can possibly have with anything is equal to the square root of its reliability coefficient. This is because error scores on the first test have zero correlation with error scores on the second test. Thus the correlation will equal the correlation of the true scores on each test. The reliability coefficient tells us the percentage of variance which is true variance. As correlation coefficients are equal to the square root of the variance accounted for, the square root of the reliability coefficient sets the upper limit to the correlation a test can have with another variable. To complicate matters the criterion variable is also likely to be unreliable, so a more general formula which allows for this as well is used. Hence the formula for the maximum possible correlation between two tests is:

$$r_{xy}(\max) = \sqrt{r_{xx}}\sqrt{r_{yy}}$$

10. What would the correlation be if we had completely reliable measures?

It might sometimes be justifiable to wonder for reasons of theory what the relationship between two variables would be in the absence of any errors of measurement. To estimate this hypothetical relationship the formula for correction for attenuation is used. This is as follows:

$$\hat{r}_{XY} = \frac{r_{xy}}{\sqrt{r_{xx}}\sqrt{r_{yy}}}$$

where $\hat{r}_{XY}$ is the error free correlation.

Closely related to these formulas are those for estimating the effects of a change in reliability on the validity of measures. How does validity change if the reliability of the predictor and/or the criterion changes? If both reliabilities change the formula is:

$$\hat{r}_{xy} = r_{xy}\sqrt{\frac{r'_{xx}r'_{yy}}{r_{xx}r_{yy}}}$$

r'_{xx} and r'_{yy} are the changed reliabilities.

If only one of the reliabilities changes, the formula is:

$$\hat{r}_{xy} = r_{xy}\sqrt{\frac{r'_{xx}}{r_{xx}}}$$

11. Special Section on Estimating the limits in which the true score will lie.

As this has been a matter of some controversy, the following might be helpful?

There has been some dispute about how to measure the range of error attached to scores obtained on a test.

Traditionally this was done by using the standard error of measurement.

The logic behind this was straightforward enough. A test score was made up of the true score plus error. Error variance was equal to $\sigma_x^2(1-r_{xx})$, which means that the standard deviation of error attaching to scores is $\sigma_x\sqrt{(1-r_{xx})}$.

This was the standard deviation of the distribution of obtained scores around the true score. If we were able to administer the same test to a person lots of times in a short period and if there were no practice effects, then the mean score would equal the true score, and the standard deviation of the distribution of scores around that mean would equal $\sigma_x\sqrt{(1-r_{xx})}$

Knowing that in a normal distribution only 5 percent of cases would fall outside the range $M_x \pm 1.96\sigma_x$, we could argue as follows.

The obtained score X comes from a distribution of obtained scores about the mean obtained score for that individual.

The obtained score is from somewhere in the distribution, and there will only be 2.5 chances in 100 that it will be more than 1.96 standard deviations away from the mean in either direction.

So, given, that the mean of the distribution is also the true score for that individual, we can cater for the possibility that the obtained score is actually higher than the true score by subtracting 1.96 standard errors of measurement from it. And, of course we can cater for the possibility that it is below the mean by adding 1.96 standard errors of measurement to it.

This led to the traditional way of stating the range within which the true score was likely to lie

$$Range = X \pm z_{nd}\sigma_x\sqrt{(1 - r_{xx})}$$

Where: z_{nd} is the z value required for significance at the desired level in the normal distribution tables.

This assumption and therefore this method has been under fire for some decades now, e.g.,

Lord, F. M. and Novik, M. R. (1968) *Statistical theories of mental test scores*. Menlo Park, California: Addison Wesley

Stanley, J. C. (1971), Reliability. In Thorndike, R. L. (ed) *Educational measurement*. (Second Edition, pp. 356 – 442), Washington, DC, American Council on Education.

Nunally, J. C. (1978) *Psychometric theory*. (Second edition) New York, McGraw-Hill

Dudek, F. (1979) The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin*, **86**, 335 – 337

Glutting, J. L., McDermot, P. A., and Stanley, J. C. (1987) Resolving differences among methods of establishing confidence limits for test scores. *Educational and Psychological Measurement*, **47**, 607 – 614

Charter, R. A. and Feldt, L. S. (2001) Confidence intervals for true scores: I s there a correct approach? *Journal of Psychoeducational Assessment*, **19**, 350 - 364

The main thrust of the argument against the traditional position is that the setting of confidence limits around the true score should be treated as a prediction problem. The criterion should be the estimated true score, not the obtained score.

The confidence limits should be set around the predicted true score. The main reason being that, unless the correlation between and obtained scores is 1.0, there will be a regression effect, and the true score will be closer to the mean of the test than the obtained score.

So what is the predicted true score?

In Z-scores it will be $$\hat{Z}_t = r_{xt} Z_x$$

Where:

r_{xt} is the correlation between obtained and true scores. It will therefore equal $\sqrt{r_{xx}}$

The standard error of estimate associated with this predicted true score will be, as usual, the standard deviation of the criterion variable (in this case, true scores) multiplied by the square root of (1 minus the squared correlation coefficient between predictor and criterion). Thus we will get.

$$\hat{Z}_t = \sqrt{r_{xx}} \times Z_x$$

and true scores will be normally distributed about this predicted true score with a standard deviation (in Z-scores) of:

$$\sqrt{1-\left(\sqrt{r_{xx}}\right)^2} = \sqrt{1-r_{xx}}$$

So if the obtained score is 2 standard deviations above the mean and the reliability of the test is .81, the predicted Z_t will be 1.8, i.e., 2 x .9, and the standard error of estimate will be the square root of .19. which is about .44.

We would therefore set the 95 percent confidence interval for the true score as

$$1.8 \pm (1.96 \times .44) = .94 \; to \; 2.66$$

How does this compare with the range resulting from using the traditional approach?

The traditional approach would give us the (z-score) range:

$$2 \pm (1.96 \times .44) = 1.14 \; to \; 2.86$$

When we leave Z-scores for the world of test scores, and supposing that our test has a mean of 100 and a standard deviation of 15, the range according to the traditional method becomes 117.1 to 142.9 (note we had previously calculated this same 95% confidence interval on page 3, using a raw score calculation rather than a z-score calculation, the results are the same).

When we come to the newer method we have a decision to make

It is this. Which standard deviation should we use? The standard deviation of obtained scores, or, the standard deviation of true scores. There is a difference between the two.

To remind you. The standard deviation of true scores is going to be the square root of the true variance,

So, $$r_{xx} = \frac{\sigma_t^2}{\sigma_x^2} \quad \therefore r_{xx}\sigma_x^2 = \sigma_t^2 \quad \therefore \sigma_t = \sigma_x\sqrt{r_{xx}}$$

Thus, the standard deviation of true scores will be, not 15, but 13.5. So the predicted true score Z of 1.8 becomes, in raw scores, 124.3 and the standard error of estimate becomes .44 x 13.5 = 5.94

The range within which we can be 95 percent confident that the true score lies becomes

$$124.3 \pm 13.5 \times .44 \times 1.96 = 112.7 \; to 135.9$$

Comparison of 95% confidence intervals for the two methods and their results can be summarised thus:

Estimates	**Traditional**	**Modern**
True score	130	124
Lower limit of range	117	112
Upper limit of range	143	136

You might have noticed that not only are the obtained and the true score different but so are their standard deviations, this has worried some (e.g., Nunally (1978), and Lord and Novik (1968)). Is it perhaps cheating to use a metric with a smaller standard deviation? Of necessity, it makes the range look smaller. Might this not give the impression that the test is more accurate than it is?

Perhaps we should convert every thing to the same scale?

If we convert true scores to a scale with a mean of 100 and a standard deviation of 15, then we find that the estimated true score becomes 127 and the range becomes 115.1 to 139.9.

So we present our table again with these values added.

Estimates	**Traditional**	**Modern**	**Obtained score standard deviation for obtained and true scores**
True score	130	124	127
Lower limit of range	117	112	114
Upper limit of range	143	136	140
Range of scores	26	24	26

But, given that a test score is intended to tell us a person's relative standing on a test, there is a strong case for predicting the Z score on the true score distribution (Z_t) from the Z-score on the distribution of obtained scores (Z_x), and then converting the predicted Z_t into the same units as the obtained score.

Given that many major tests have reliabilities of .90 and above, any differences seen in the above table would be less.

If reliability was .90, the Z of 2.0 on the test the modern method would give us 1.9 as the predicted Z-score on the true score distribution. The error of estimate would be .32,

This would compare with the traditional method's Z-score estimate of 2, with a standard error of estimate of .32.

If the mean and standard deviation of the test were 100 and 15 respectively, the traditional method would predict a 95% Confidence Interval of 130 plus or minus 9.3 (120.7 – 139.3).

If we convert these Z-scores in the modern manner, we get a predicted true score of 127, with a 95% Confidence Interval of plus or minus 8.8 (118.2 – 135.8)

If we were to equalize the standard deviations of the two scales (at 15 points of IQ) we would get an estimated true score of 127 plus or minus 9.3 points (117.7 – 136.3).

But all of this is a bit controversial. For more on these dilemmas read the references cited above.

12. Coefficient Alpha

Coefficient Alpha is the leading method for assessing reliability from the internal consistency of a test

These methods assess reliability from by using the inter-correlations between the items of the test.

Essentially every item is correlated with every other item and the average correlation found. An adjustment is then made for the number of items in the test.

Probably the most famous predecessor of coefficient Alpha was the Kuder Richardson 20 formula. This formula had the disadvantage that it applied only to items scored dichotomously, whereas Cronbach's Coefficient Alpha does not have this limitation.

The formula for Coefficient Alpha is:

$$r_{kk} = \frac{k}{k-1}\left(1 - \frac{\Sigma\sigma_i^2}{\sigma_t^2}\right)$$

Where:

r_{kk} = the reliability of a k item test

k = the number of items

σ_i^2 = the variance of an item

σ_t^2 = the variance of the whole test

Why does it work?

The variance of a test with *k* items can always be split into two parts (a) the sum of the item variances and (b) twice the sum of the item covariances. To make this easier to understand let us suppose we have a very short test consisting of only two items A and B.

If we apply the formula for the variance to this combination of items, the variance of this combination will be:

$$\sigma^2_{(a+b)} = \frac{\Sigma\left((A+B) - \left(\frac{\Sigma(A+B)}{N}\right)\right)^2}{N}$$

This simplifies to:

$$\sigma^2_{(a+b)} = \frac{\Sigma(A - M_a + B - M_b)^2}{N}$$

which equals

$$\frac{\Sigma(a+b)^2}{N} = \frac{\Sigma a^2}{N} + \frac{\Sigma b^2}{N} + \frac{2\Sigma ab}{N}$$

but, because $r_{ab} = \frac{\Sigma ab}{N\sigma_a\sigma_b}$ $then$ $\frac{\Sigma ab}{N} = r_{ab}\sigma_a\sigma_b$

So

$$\sigma^2_{(a+b)} = \sigma^2_a + \sigma^2_b + 2r_{ab}\sigma_a\sigma_b$$

This of course is just an example of the rule that the variance of a composite equals the sum of the variances of its component parts plus two times the sum of the covariances.

The correlation between item A and item B will reflect the extent to which they consistently measure the true score on the test. If the correlation between them was zero then the variance of the composite would be simply the sum of the two item variances σ_a^2 and σ_b^2

In the coefficient Alpha formula this would make the right hand term of the equation = zero, i.e.,

$$\left(1-\frac{\sigma_a^2+\sigma_b^2}{\sigma_a^2+\sigma_b^2}\right)=0$$

The greater the average correlation between the component items of the test, the higher will be the reliability of the test.

Note also that the reliability of the test is dependent on the number of items used in the test.

Why does increasing the length of a test increase its reliability? Supposing that we have a 5 item test, then the crucial division we have to calculate for coefficient Alphas will be:
1 minus the sum of the 5 item variances divided by (the sum of those same 5 item variances plus the sum of the 20 covariances between those items,

If we know increase the number of test items to ten, then the number of variances in both the numerator and denominator will rise to 10, but the number of covariance terms in the denominator will rise to will rise to 90 . As the added items would not have been added to the test unless they correlated with the items already in the test, the consequence of this will be to reduce the value of the ratio of $\frac{\Sigma\sigma_i^2}{\sigma_t^2}$ and thus increase the size of Alpha.

An equivalent formula for Coefficient Alpha is:

$$r_{kk}=\frac{k\bar{r}_{ij}}{1+(k-1)\bar{r}_{ij}}$$