The Statistical Evaluation of Errors in Exposure Measurement

A Self-Instructional Guide for Postgraduate Students

Graham Dunn
Health Sciences Research Group
University of Manchester

January 2011
INTRODUCTION

Although you may not have thought of it in this way, as postgraduate students you will already have had quite a lot of experience of the techniques of assessment. You may have very little experience in exposure measurement but you will have experience of several analogous procedures. You will have had, for example, your own knowledge, skills and understanding assessed at many stages of your career. Similarly, you will be very familiar with many of the common physical measuring devices (rulers, thermometers, balances, and so on). Returning to educational assessments and considering them with respect to your own academic career, you will probably be quite familiar with the idea that examinees can have ‘good’ and ‘bad’ days. On some occasions you can write an essay apparently without any effort. On others it appears to be impossible to write the first sentence. You may have luckily chosen to revise just the right topics or you may have made a serious error of judgement in revising those that do not appear to have been in the mind of the examiner! Your performance can be influenced by the state of both your physical and mental health. You will not perform well, for example, if you are suffering or recovering from a bout of influenza, or if an important relationship has apparently ended the night before the exam. The examination room may be far too hot for your liking (but not apparently for others) or it may be cold or draughty. How many of you have confidently left an examination room to subsequently discover from conversation with your peers that you must have completely misunderstood an exam question or worse, to have misread the instructions at the top of the exam paper? If you are confident, not suffering from extraneous stress and (relatively) relaxed, however, you may perform better than either you or your teachers ever dreamt of. Your final score can also be influenced by the examiner. She, too, can have her ‘good’ or ‘bad’ days! One examiner might consider handwriting, spelling and grammar to be vitally important, whilst another is quite happy to ignore them as long as he is satisfied that he has understood the student’s argument. One examiner might like a particular style of reasoning which appears to be an anathema to another. The sources of prejudice and hidden biases are limitless.

"Anyone who regularly plays a game with objective scoring such as golf or bridge, is acutely aware of the variability in human performance. No one operates at his or her personal best on all occasions, be the domain one of physical or mental activity. This inconsistency stems from a variety of factors, depending on the nature of measurement. Among the important factors are subtle variations in physical and mental efficiency of the test taker, uncontrollable fluctuations in external conditions, variations in specific tasks required of individual examinees, and inconsistencies on part of those who evaluate examinee performance. Quantification of consistency and inconsistency in examinee performance constitutes the essence of reliability analysis." (Feldt and Brennan, 1989; p105).

Moving on to exposure measurement, consider dietary intake as a risk factor for cancer of the colon. It is clear that, like the assessment of examination performance, the assessment of fibre intake, for example, can be subject to many sources of variation. We do not consume exactly the same amount of dietary fibre day after day - even if we intend to - and even the most straightforward of measuring instruments such as weighing machines vary from clinic to clinic or from day to day within any one clinic. There may be problems with memory recall in the case of retrospective assessments, limits to a proxy respondent's knowledge and memory of a subject's exposures, under reporting of socially desirable behaviours and over-reporting of the good ones, and so on. White et al. (2008) list the following sources of ‘error’ (which are subsequently expanded
in their Table 3.1):

- Faulty design of the measuring instrument (assay, questionnaire, diary, interview etc).
- Errors or omissions in the protocol for the use of the instrument.
- Poor execution of the protocol during data collection.
- Limits due to subject characteristics (see above).
- Errors during data entry and analysis.

We must accept that all exposure assessments are fallible. It is the purpose of a reliability study to determine how fallible they are, and perhaps to yield information by which we can compare the performance of different measurement or assessment techniques. Having accepted that all measurements contain error (and we would never know when a particular measurement were, in fact, error-free) it is important to be able to estimate the likely size of this error. If we were to repeat the required assessment we would be very unlikely to get the same result but would hope that it would be 'close'. With the aid of the results of a reliability analysis we should be able to define what 'close' means. After working through this booklet we should also be able to design and analyse simple reliability studies to estimate the characteristics of tests (standard errors of measurement and reliabilities) which are needed to evaluate the fallibility of exposure measurements.

In this text, many of the ideas are presented through the use of guessing the lengths of pieces of string. Epidemiologists and clinicians might regard the use of this example as frivolous and unrealistic in the context of exposure assessment. I agree that measuring lengths of pieces of string is trivial in comparison with the measurement of environmental pollution or radiation exposure but that they do have many features in common. Having understood the common features, you can then move on to the more difficult and intractable problems of exposure measurement.

The structure of this text is the following. First, Assignment 1 tells you how to carry out simple reliability (precision) studies using pieces of string as objects being 'assessed'. There are also simple arithmetic exercise which familiarise students with the manipulation and interpretation of test characteristics. This is followed by an Interlude in which some of the statistical ideas are clarified in terms of exposure assessments. Assignment 2 concerns the graphical presentation and analysis of data arising from a method comparison study There is then a section summarising the main statistical ideas concerning exposure measurement. Finally there is a set of Appendices which can be used as a student resource. Appendices B1 to B4 provide essential technical material (proofs and derivations) which goes beyond that provided in the main text. These are provided for the enthusiast (and MSc Biostatistics student!) who is not prepared to accept mathematical details on trust and who needs a primer before tackling more technical texts. Also built into the whole package is a considerable amount of redundancy. Key ideas are discussed in several different places and at different levels. The hope being that repeated exposure to the ideas will eventually lead to an intuitive understanding of the material.
Before introducing Assignment 1, it may prove useful to provide a detailed list of aims and objectives for the text as a whole:

**GENERAL AIMS**

- To explain the concept and importance of the quantitative evaluation of measurement errors.
- To introduce the student to statistical models for measurements and, in particular, the estimation of reliability and the standard error of measurement.
DETAILED OBJECTIVES

1. ATTITUDES

The acquisition of theoretical knowledge and associated technical skills should be motivated by the following changes in orientation of the student.

- To move away from the treatment of exposure assessments as if they were infallible. To realise that all assessments are fallible and that some are more fallible than others!

- To move towards thinking that there is always a need to give a measure of fallibility of an assessment in addition to that assessment itself. For example, to move away from the idea of presenting an assessment as a point estimate to the routine use of an estimate with its standard error (or perhaps, the routine use of confidence intervals).

2. KNOWLEDGE

After working through this text the learner should:

- Be able to explain the basic concepts of the psychometricians' Classical Test Theory (truth and error, true scores, reliability, standard error of measurement, parallel test forms).

- Be able to describe the various sources of measurement error and the design of reliability studies to evaluate them (test-retest, alternative forms, split-halves, inter-rater agreement).

3. SKILLS

After the completion of working through this programmed text the student should have acquired the following skills:

- To be able to estimate a standard error of measurement from published values of the reliability coefficient the standard deviation of the observed measurements (and vice versa).

- To be able to design a simple reliability study in order to estimate reliability coefficients and associated standard errors of measurement.

- To be able to estimate the reliability of a shortened (enlarged) questionnaire scale via the use of the Spearman-Brown Prophesy Formula and to estimate the effect of changing population heterogeneity on the reliability of assessments through the use of attenuation formulae.

- To be able to design and analyse a simple method-comparison study.
ASSIGNMENT 1

How Long is a Piece of String?

Part 1

BIAS AND PRECISION

Think of a piece of string, say, 10cm (about 4in) long. Imagine holding it in your hand and, without knowing its actual length, guessing its length. Now consider the possibility of forgetting your guess and then repeating the exercise. If this procedure is repeated over and over again you will finish up with a series of guesses, the average of which (that is, their arithmetic mean) is your best estimate of the string's actual length. If you now measure the length with a rule, the difference between this 'true' length and your average is a measure of bias in your guesses (measurements). A measure of variability of your guesses provides an estimate of your precision as a measuring instrument. The higher your precision the lower the variability of your guesses. Variability is most commonly described through the use of the measurement's standard deviation (or the square of the standard deviation - the variance). One can regard the fluctuations of your guesses about their mean value as random measurement errors. The standard deviation of the guesses themselves is also the standard deviation of these measurement errors. In psychological measurement theory, the standard deviation of the measurement errors is known as the standard error of measurement. Another name is repeatability.

Exercise 1.1: Estimation of Bias and Precision

In practice, the above experiment is obviously an impossible one to carry out (unless you are prepared to wait a long time between the repeated guesses). The following experiment, however, is a feasible one:

Take ten strips of paper. Imagine a distance (length) of 10cm (or 4in). Take each strip of paper in turn and, without reference to either your previous attempts or to a rule, mark each strip with two crosses which you believe to be this distance apart.

Now, having made your ten attempts, measure the distance between each of these ten pairs of crosses with a rule. Calculate their mean and their standard error of measurement (standard deviation).

An example.

The following set of ten guesses were obtained by the author when he carried out the exercise:

\[
\begin{array}{cccccccc}
10.2 & 9.8 & 10.4 & 9.0 & 9.1 & 10.1 & 10.2 & 10.7 & 9.0 & 9.5 \\
\end{array}
\]

Their mean is 9.80 (i.e. bias is -0.20) and the standard error of measurement is about 0.62. The variance of the errors (the square of the standard error of measurement) is 0.39.
Exercise 1.2: Dependence of Bias and Precision on True Values

Now try an experiment in which the aim is to discover whether your bias and precision are dependent on (correlated with) the actual length you are trying to reproduce. Repeat Exercise 1.1 a further five times, using distances of 1cm, 2cm, 5cm, 15cm, and 20cm. (A better experimental design might involve mixing these six distances up in a single exercise, but only attempt this if you are sure that you won't get confused!)

For each set of ten reproduced lengths, calculate the mean and standard error of measurement. Investigate the relationship between these statistics and the true lengths through the use of two simple graphs: (a) a plot of the mean (or bias) against the true length (here the mean is plotted on the vertical- or y-axis; the truth on the horizontal- or x-axis), and (b) a plot of the standard error of measurement (or variance) against the nominal length. Interpret your findings.

Example.

The following is an example of some data from a similar experiment (based on a table given on page 289 of Guilford, 936). Each standard error of measurement (s.e.m.) is based on 50 reproductions of the nominal length.

<table>
<thead>
<tr>
<th>Length(mm)</th>
<th>s.e.m.(mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>2.22</td>
</tr>
<tr>
<td>40</td>
<td>4.45</td>
</tr>
<tr>
<td>60</td>
<td>7.71</td>
</tr>
<tr>
<td>80</td>
<td>7.56</td>
</tr>
<tr>
<td>100</td>
<td>10.08</td>
</tr>
<tr>
<td>120</td>
<td>11.27</td>
</tr>
<tr>
<td>140</td>
<td>11.27</td>
</tr>
<tr>
<td>160</td>
<td>13.34</td>
</tr>
<tr>
<td>180</td>
<td>10.84</td>
</tr>
<tr>
<td>200</td>
<td>17.64</td>
</tr>
</tbody>
</table>

Clearly, Guilford's subject gets less precise as the nominal length increases - a result that is not really surprising given the nature of the task.

Exercise 1.3: Dependence of Standard Error of Measurement on Measurement Scale

Using the data obtained in Exercise 1.1, convert all of your guesses to millimetres. What is their standard error of measurement? Now repeat the exercise, having converted all your guesses to inches (1 inch=2.54cm; 1cm=0.394 inches).

In the case of the author's guesses, the required standard errors are 6.2mm and 0.24 inches, respectively. If you are not already aware of the relationship, note that these two figures are obtained by multiplying the original standard error of measurement (0.62cm) by the conversion factor directly (0.62 X 10, or 0.62 X 0.394, respectively). It is not necessary to convert each individual measurement and then recalculate their standard error of measurement - although it
might be worth doing once or twice, just in case you are not convinced! The main point of this exercise, however, is not to demonstrate arithmetical tricks, but to demonstrate that the standard error of measurement is scale dependent. If two different measuring instruments are to be compared through these standard errors then it is important that they are calibrated using the same units.

**Part 2**

**A SIMPLE RELIABILITY STUDY**

Now we return to actual pieces of string. Cut ten pieces of string, one each of the following lengths (cm):

2.1  3.0  5.3  7.2  8.4  9.7  11.1  13.5  15.7  19.2

Put them into random order (don't be too concerned by what is meant by 'random' here) and label them from 1 through to 10. Taking each piece of string in turn, guess its length, say, to the nearest millimetre. Do not look at a rule during this experiment nor look at any of your previous guesses. Wait a convenient amount of time (an hour or two, or even until the next day) and then repeat the exercise. This is an example of a test-retest reliability study. The aim of the study is to ask how well do your second guesses agree with the first ones.

In an epidemiological study subjects' dietary fat intake might be assessed by questionnaire on one day and then assessed again, say, two or three days (weeks, months or even years, depending on the stability of the trait being assessed) later using the same questionnaire. This would be a test-retest reliability study. If, on the other hand, subjects could be re-assessed on the same day through the use of an alternative form of assessment, assumed to be equivalent in terms of its measurement characteristics then reliability and precision could be estimated from these pairs of assessments. In terms of the pieces of string, their length could be guessed once by one rater and again by entirely different rater with no knowledge of the first ratings. Again we would assume that the two raters had no (relative) biases and had the same precision.

Following on from Part 1 of this Assignment, you might expect that we use these replicate measurements of estimate a standard error of measurement. An alternative strategy might be to calculate the correlation between the first set of guesses and the second set. Later in the text, we will demonstrate the relationship between the two.

**Exercise 2.1: Standard Error of Measurement**

Taking each piece of string in turn, subtract your second guess from the first. This difference score is simply the difference between the errors of measurement for the two guesses (the true length cancels out). Square each of the differences and then calculate the mean of these ten squared differences. Assuming that the measurement errors in the first set of guesses are uncorrelated with those in the second, and also assuming that you have not improved or declined in terms of precision between the first and second set of guesses, it can be shown that the variance of the individual errors is simply half of the mean of the squared differences. The required standard error of measurement is, therefore, the square root of the variance of the individual errors. In other words,
the standard error of measurement is the square root of the mean of the squared differences divided by root two. Calculate this statistic from your data.

Example:

The following data arose from a similar experiment (Dunn, 2004) in which the lengths of 15 pieces of string were estimated to the nearest 1/10th inch.

<table>
<thead>
<tr>
<th>String</th>
<th>1st Guess</th>
<th>2nd Guess</th>
<th>Difference</th>
<th>Squared Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5.0</td>
<td>5.5</td>
<td>-0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>B</td>
<td>3.2</td>
<td>3.5</td>
<td>-0.3</td>
<td>0.09</td>
</tr>
<tr>
<td>C</td>
<td>3.6</td>
<td>4.4</td>
<td>-0.8</td>
<td>0.64</td>
</tr>
<tr>
<td>D</td>
<td>4.5</td>
<td>4.5</td>
<td>0.0</td>
<td>0.00</td>
</tr>
<tr>
<td>E</td>
<td>4.0</td>
<td>5.1</td>
<td>-1.1</td>
<td>1.21</td>
</tr>
<tr>
<td>F</td>
<td>2.5</td>
<td>2.6</td>
<td>-0.1</td>
<td>0.01</td>
</tr>
<tr>
<td>G</td>
<td>1.7</td>
<td>1.8</td>
<td>0.1</td>
<td>0.01</td>
</tr>
<tr>
<td>H</td>
<td>4.8</td>
<td>4.7</td>
<td>0.1</td>
<td>0.01</td>
</tr>
<tr>
<td>I</td>
<td>2.4</td>
<td>2.3</td>
<td>0.1</td>
<td>0.01</td>
</tr>
<tr>
<td>J</td>
<td>5.2</td>
<td>5.6</td>
<td>-0.4</td>
<td>0.16</td>
</tr>
<tr>
<td>K</td>
<td>1.2</td>
<td>1.1</td>
<td>0.1</td>
<td>0.01</td>
</tr>
<tr>
<td>L</td>
<td>1.8</td>
<td>1.8</td>
<td>0.0</td>
<td>0.00</td>
</tr>
<tr>
<td>M</td>
<td>3.4</td>
<td>3.6</td>
<td>-0.2</td>
<td>0.04</td>
</tr>
<tr>
<td>N</td>
<td>6.0</td>
<td>6.2</td>
<td>-0.2</td>
<td>0.04</td>
</tr>
<tr>
<td>O</td>
<td>2.2</td>
<td>2.1</td>
<td>-0.1</td>
<td>0.01</td>
</tr>
</tbody>
</table>

MEAN: 0.166

The variance of the individual errors is 0.083 (0.166/2). The standard error of measurement is, therefore, 0.288 inches (the square root of 0.083).

Note that we have assumed in this exercise that the standard error of measurement is not dependent on the actual length being estimated. This assumption will usually be made in epidemiological surveys, but note that it is not necessarily true.

Exercise 2.2: Correlation

Plot the second guess for each piece of string against first one. They would be expected to be quite close to a straight line of unit slope. A measure of how close the points lie to a straight line is the correlation coefficient (or more precisely, the product-moment correlation coefficient). Calculate this statistic for your data. This correlation can be used as an estimate of your reliability as a measuring instrument. A better estimate of reliability, however, is provided by the intra-class correlation coefficient. This can be calculated through creating a new table by entering the data for each piece of string twice - first giving the estimates in the order obtained, and second, by giving them in the reverse order. For the data above, for example, the entries would be:
<table>
<thead>
<tr>
<th>String</th>
<th>1st Estimate</th>
<th>2nd Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5.0</td>
<td>5.5</td>
</tr>
<tr>
<td>A</td>
<td>5.5</td>
<td>5.0</td>
</tr>
<tr>
<td>B</td>
<td>3.2</td>
<td>3.5</td>
</tr>
<tr>
<td>B</td>
<td>3.5</td>
<td>3.2</td>
</tr>
<tr>
<td>C</td>
<td>3.6</td>
<td>4.4</td>
</tr>
<tr>
<td>C</td>
<td>4.4</td>
<td>3.6</td>
</tr>
</tbody>
</table>

... and so on.

The correlation coefficient calculated from this doubled table is the intra-class correlation. For the above data, the product-moment correlation is 0.979 and the intra-class correlation is 0.963. For any set of data arising from a test-retest reliability study these two coefficients are likely to be fairly similar, with the intra-class correlation never being the higher of the two. Calculate the intra-class correlation for your own data.

**Exercise 2.3: Analysis of Variance**

The data from a test-retest reliability study can be analysed through the use of a one-way analysis of variance (the single explanatory factor being the identity of the piece of string). This will partition variability in the measurements into that between the pieces of string and that within them (measurement error). The above string data yield the following analysis of variance table:

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Square(M.S.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>String</td>
<td>65.729</td>
<td>14</td>
<td>4.695</td>
</tr>
<tr>
<td>Error</td>
<td>1.245</td>
<td>15</td>
<td>0.083</td>
</tr>
<tr>
<td>Total</td>
<td>66.974</td>
<td>29</td>
<td></td>
</tr>
</tbody>
</table>

The first thing to note is that the Error Mean Square is the variance of the measurement errors (0.083, as before). The standard error of measurement is therefore the square root of the Error Mean Square obtained from a one-way analysis of variance.

The second statistic that can be easily calculated from the ANOVA table is the intra-class correlation. This is given by:

\[ r_i = \frac{\text{String M.S. - Error M.S.}}{\text{String M.S.} + \text{Error M.S.}} \]

\[ = \frac{(4.695-0.083)}{(4.6+0.083)} \]

\[ = 0.965 \]
Exercise 2.4: Alternative Designs

Returning to the ten different lengths of string, cut two pieces of string corresponding to each of the lengths. Number these pieces of string from 1 through to 20, making sure that you know which pairs of numbers correspond to pieces of string of the same length. Put these twenty pieces of string into random order and then ask a colleague to estimate the length of each piece of string in turn, again making sure that no reference is made to a rule or to other pieces of string.

This experiment provides 20 observations which in which the pairs of guesses corresponding to each length of string cannot be ordered. There are simply two unordered duplicates corresponding to each length. They cannot be placed in columns labelled '1st Estimate', '2nd Estimate', and so on.

Analyses the data using a one-way analysis of variance. Estimate the standard error or measurement and the intra-class correlation.

Where might such a design arise in practise? Consider ten video-recordings of a dietary assessment interview. Each video-recording is watched and rated by two (or perhaps more) raters, but no individual rater rates more than one video. Similarly, we could replace the videos with questionnaire responses which are to be marked by a panel of raters. Two raters rate each questionnaire with no one rater seeing more than one questionnaire.

Finally, consider the design of a reliability study in which, for example, each subject (piece of string, dietary exposure interview, smoking questionnaire) is measured, rated or assessed by two (or more) particular raters. This design would enable us to estimate the relative biases of the two or more raters, and these biases should be taken into account when calculating standard errors of measurement or intra-class correlations. Here the appropriate method of analysis would involve a two-way analysis of variance.
Part 3
ARITHMETICAL EXERCISES

The aim of the following series of exercises is to provide you with practice in the handling of the arithmetical manipulations applicable to (a) calculating standard errors of measurement from published reliability data, (b) allowing for changes in population heterogeneity in use of reliability coefficients, and (c) the use of the Spearman-Brown Formula in the calculation of the reliability of shortened forms of tests (or of longer forms and the average of repeated assessments). N.B. Like a manual or research paper, the Exercises are likely to contain redundant information.

Exercise 3.1 (Adapted from Gulliksen, *Theory of Mental Tests*, 1950)

<table>
<thead>
<tr>
<th>Test</th>
<th>Mean</th>
<th>St.Dev</th>
<th>No. Items</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>18.4</td>
<td>4.2</td>
<td>30</td>
<td>0.72</td>
</tr>
<tr>
<td>B</td>
<td>28.9</td>
<td>9.8</td>
<td>60</td>
<td>0.96</td>
</tr>
<tr>
<td>C</td>
<td>37.2</td>
<td>8.1</td>
<td>50</td>
<td>0.90</td>
</tr>
<tr>
<td>D</td>
<td>63.7</td>
<td>10.4</td>
<td>100</td>
<td>0.86</td>
</tr>
<tr>
<td>E</td>
<td>39.2</td>
<td>11.5</td>
<td>75</td>
<td>0.92</td>
</tr>
</tbody>
</table>

(a) Calculate the standard errors of measurement associated with the above five tests.

(b) Another investigator repeats the administration of Test A to a new group of subjects and finds a mean of 25.3 and a standard deviation of 8.4. About what reliability would you expect the test to have for this new group?

(c) It is reported that Test B has been administered to a new group and the reliability coefficient is only 0.90. What would account satisfactorily for this lowered reliability without indicating any faults of test administration or scoring?

(d) Test C is administered to a new group with the following results: mean 31.9, standard deviation 12.7, reliability 0.96. Are these results in reasonable agreement with those reported in the table for Test C?

(e) Test D is reported as having a mean of 68.2, a standard deviation of 14.6, and a reliability of 0.98. Are these results in reasonable agreement with those reported in the table for Test D?

(f) Estimate the reliability of all of the above tests if they were to be doubled in length (or, alternatively, the reliability of the average of two replicate measurements).

(g) What would the reliability of Test C be if it were halved in length?
INTERLUDE

The Fallibility of Exposure Assessments

Truth and Error

I hope that by this stage in the text you will have been convinced that no measurement is ever error-free. Even the most accurately made and calibrated physical instrument (a chemical balance, for instance) will display some error. If we represent a measurement by the letter X and the error associated with it by the letter e, we can think of 'the truth' (symbolised by the Greek letter \( \tau \) or tau) as being the observation minus error:

\[
\text{Truth} = \text{Observation} - \text{Error}
\]

or

\[
\tau = X - e
\]

Alternatively (by rearrangement of the equation),

\[
X = \tau + e
\]

Our problem is that we can never know the value of e. But, the use of the methods described earlier does, however, allow us to estimate the statistical properties of e. We can use a reliability study to estimate its standard deviation (the standard error of measurement), for example. Accepting that in exposure measurement (and in all other disciplines, including physics and chemistry, for that matter) our scales (units) of measurement are all rather arbitrary, we could ignore bias and define the truth to be the average of all possible measurements that would be obtained in an indefinite sequence of repetitions of the measurement process (that is, an expected value). Note that this is a statistical definition of truth. It should not be confused with Platonic truth, the idea that there is something 'real' out there which needs us to discover it.

Extracting Standard Errors of Measurement From Published Studies

A well-presented test manual or research paper giving the results of reliability investigations should provide estimates of a test's or instrument's standard error of measurement. In many cases, however, it simply provides an estimate of reliability (in the guise of a correlation coefficient). A reliability coefficient is a scale-independent index of the performance of a test which can be interpreted as meaning 'the proportion of the test's variability which is not accounted for by measurement error'. If it has a value of 1 then the test is perfectly reliable. If it has a value approaching zero then it is useless. Another way of interpreting the value of the reliability coefficient is as a measure of how well the test can distinguish between the individuals being tested. If two tests are administered to a common sample of subjects, their respective reliability coefficients will enable the investigator to decide which of the two tests is the most precise.

It is common to assume that a test's standard error of measurement is a fixed characteristic of a test.
and that it is unaffected by the subjects' true scores. A reliability coefficient, on the other hand, changes with the variability of the sample of subjects being investigated. If this variability is measured by the standard deviation of the actual measurements, say, $\sigma_x$, and the standard error of measurement is represented by $\sigma_e$, then the following equation applies:

$$\sigma_e = \sigma_x \sqrt{1-R}$$

Alternatively,

$$R = 1 - \frac{\sigma_e^2}{\sigma_x^2}$$

For a fixed standard error of measurement, the reliability, $R$, will increase as the sample becomes more heterogeneous (higher $\sigma_x$) and vice versa.

The variation in $R$ with $\sigma_x$ might appear to be rather academic, but be careful in using published values of $R$. Questionnaire rating scales, for example, are typically investigated using clinical samples. It is naive to assume that the rating scale will have the same reliability when used in a general community survey, for example.

If a test has a reliability of $R_a$ when used on a sample with a standard deviation of $\sigma_{xa}$, then its reliability ($R_b$) when used on a sample with standard deviation of $\sigma_{xb}$ is given by:

$$R_b = 1 - \frac{(1-R_a)\sigma_{xa}^2}{\sigma_{xb}^2}$$

**The Precision and Reliability of Average Scores**

Suppose that you are in a lucky position of obtaining two or more independent test results (of the same type) on a particular subject. These might be, for example, results of repeated assessment, multiple-rating of a video recording of an interview or responses to a questionnaire, and so on. It makes a lot of sense to estimate his or her true score using the mean of the available test results. But what is the standard error of measurement of the resulting mean? What is its reliability?

The standard error of measurement of the mean of $k$ test results is simply the test's standard error of measurement divided by the square root of $k$. If the test's standard error of measurement is 4 points, for example, and we have scores provided by four independent examiners, then the standard error of measurement of their average mark is 2 points.

The reliability of the average of $k$ scores is provided through the use of the well-known Spearman-Brown Prophesy Formula. Here

$$R_k = \frac{kR_1}{1+(k-1)R_1}$$

where $R_k$ is the reliability of the average and $R_1$ is the reliability of the original test. Note that, strictly speaking, the term reliability applies to a whole sample or population of individual's rather than an individual test score in isolation.
If we are evaluating, say, a exposure rating scale on a sample of subjects, it is often useful to compare the sub-total scores obtained from the sum of the even items and the sum of the odd items (or preferably sub-totals from a random partition of the questionnaire into two equal-sized sub-scales). The standard error of measurement for the whole test is simply the square root of the mean of the squared differences between the sub-test totals. The reliability of the whole test is again calculated using the Spearman-Brown Formula (here k=2). So

$$R = \frac{2r}{1+r}$$

where R is the reliability of the whole test and r is the correlation (preferably intra-class) between the sub-total scores (the split-half correlation).

**Reliability versus Validity**

Be careful to distinguish the everyday use of words such as reliability from that used in the technical literature. The following extract illustrates this point.

"Regardless of how reliability is defined and quantified, it is somewhat unfortunate that this word was adopted originally for the phenomenon under consideration. In everyday conversation, reliability refers to a concept much closer to the measurement concept of validity. Weather reports, for example, are thought to be unreliable if they are frequently contradicted by prevailing conditions a day or two later. Medical tests are said to be unreliable if they often give false cues about the condition of the patient. In both of these contexts, the information might be highly reliable in the measurement sense. Meteorologists might have no reason to doubt the accuracy of the temperature, air pressure, and wind velocity measurements used in determining the forecast. Several experts, each using this information independently, might arrive at precisely the same forecast day in and day out. Similarly, the medical test might yield consistent, but often erroneous, conclusions about patients, if repeated several times on each individual. In measurement nomenclature these measures would be said to be highly reliable because they are self-consistent. But they would have questionable validity. The distinction is often lost on lay-people and many educators, however, and is a potential source of confusion when the results of testing are summarized." (Feldt and Brennan, 1989, p106)

"Validity is an integrative evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p13).

As the reader may have already realised, this text has little to say concerning the assessment of validity.
ASSIGNMENT 2

Pieces of String: Whose Guess is Best?

How Long is a Piece of String?

Measurements for 15 pieces to nearest 1/10th inch

<table>
<thead>
<tr>
<th>Measured length</th>
<th>Graham's Guess</th>
<th>Brian's Guess</th>
<th>Andrew's Guess</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>G</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>6.3</td>
<td>5.0</td>
<td>4.8</td>
<td>6.0</td>
</tr>
<tr>
<td>4.1</td>
<td>3.2</td>
<td>3.1</td>
<td>3.5</td>
</tr>
<tr>
<td>5.1</td>
<td>3.6</td>
<td>3.8</td>
<td>4.5</td>
</tr>
<tr>
<td>5.0</td>
<td>4.5</td>
<td>4.1</td>
<td>4.3</td>
</tr>
<tr>
<td>5.7</td>
<td>4.0</td>
<td>5.2</td>
<td>5.0</td>
</tr>
<tr>
<td>3.3</td>
<td>2.5</td>
<td>2.8</td>
<td>2.6</td>
</tr>
<tr>
<td>1.3</td>
<td>1.7</td>
<td>1.4</td>
<td>1.6</td>
</tr>
<tr>
<td>5.8</td>
<td>4.8</td>
<td>4.2</td>
<td>5.5</td>
</tr>
<tr>
<td>2.8</td>
<td>2.4</td>
<td>2.0</td>
<td>2.1</td>
</tr>
<tr>
<td>6.7</td>
<td>5.2</td>
<td>5.3</td>
<td>6.0</td>
</tr>
<tr>
<td>1.5</td>
<td>1.2</td>
<td>1.1</td>
<td>1.2</td>
</tr>
<tr>
<td>2.1</td>
<td>1.8</td>
<td>1.6</td>
<td>1.8</td>
</tr>
<tr>
<td>4.6</td>
<td>3.4</td>
<td>4.1</td>
<td>3.9</td>
</tr>
<tr>
<td>7.6</td>
<td>6.0</td>
<td>6.3</td>
<td>6.5</td>
</tr>
<tr>
<td>2.5</td>
<td>2.2</td>
<td>1.6</td>
<td>2.0</td>
</tr>
</tbody>
</table>

The above table gives information on the lengths of 15 pieces of string. The column headed 'R' provides measurements made with a rule. The columns labelled 'G', 'B' and 'A' are independent guesses made by Graham, Brian and Andrew, respectively. R, G, B and A are all examples of random variables. The guesses, G, B and A, are clearly examples of error-prone or fallible measures. R, on the other hand, might be regarded as the 'truth' but in reality it is also fallible. It is
merely much more accurate. All four measurements are, therefore, fallible manifest variables. Furthermore, it is obvious that these four measures will be highly correlated and this covariation or correlation arises from the fact that they are all measures of the same concept: length. Corresponding to each piece of string is a true but unknown length. Although R will be a more reliable indication of this length than G, B or A, the truth will always remain unknown. The true length, therefore, is a relatively straightforward example of a latent variable.

Summary Statistics for String Data:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ruler (R)</td>
<td>4.2933</td>
<td>1.9692</td>
</tr>
<tr>
<td>Graham (G)</td>
<td>3.4333</td>
<td>1.4563</td>
</tr>
<tr>
<td>Brian (B)</td>
<td>3.4267</td>
<td>1.6294</td>
</tr>
<tr>
<td>Andrew (A)</td>
<td>3.7667</td>
<td>1.7987</td>
</tr>
</tbody>
</table>

Correlation Matrix
Lower triangular form

<table>
<thead>
<tr>
<th></th>
<th>Ruler</th>
<th>Graham</th>
<th>Brian</th>
<th>Andrew</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ruler</td>
<td>1.0000</td>
<td>0.9802</td>
<td>0.9811</td>
<td>0.9899</td>
</tr>
<tr>
<td>Graham</td>
<td></td>
<td>1.0000</td>
<td>0.9553</td>
<td>0.9807</td>
</tr>
<tr>
<td>Brian</td>
<td></td>
<td></td>
<td>1.0000</td>
<td>0.9684</td>
</tr>
<tr>
<td>Andrew</td>
<td></td>
<td></td>
<td></td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Exercise 2.1: Simple Calibration of Guesses against the Ruler

(a) First, comment on the above summary statistics.

(b) Plot Graham's guesses against the value given by the Rule. Interpret your findings.

(c) Now consider the regression of Graham's guesses (G) against the Rule (R). This is done by fitting a simple regression model of the form:

\[ G = a + bR + e \]
where $a$ and $b$ are constant coefficients (parameters) and $e$ is a random variable representing the unsystematic or random error in Graham's guesses. The parameters $a$ and $b$ correspond to the intercept and slope coefficients measuring the systematic bias in Graham's guesses. (If there were no biases relative to $R$, $a$ would be zero and $b$ would be unity). In this model $G$ is referred to as the dependent variable and $R$ is the independent variable. We assume that $R$ and $e$ are uncorrelated. It is also assumed that the $e$'s have a mean of zero and a constant variance, $\text{Var}(e)$.

Fit this model using a simple linear regression program (or by hand), using ordinary least squares (OLS) as a fitting criterion. Interpret your results. What is the standard error of measurement of Graham's Guesses - that is, $\sqrt{\text{Var}(e)}$?

(d) The above model implies the following predictions for the variance of $G$ and for the covariance of $G$ and $R$:

$$\text{Var}(G) = \text{Var}(a + bR + e)$$
$$= b^2 \text{Var}(R) + \text{Var}(e)$$

and,

$$\text{Cov}(G, R) = \text{Cov}(a + bR + e, R)$$
$$= b \text{Var}(R)$$

(See Appendix B1 for the derivation of these relationships).

The following are the estimates of the covariances for the string data:

**Covariance Matrix\(^1\)**

**Lower triangular form**

<table>
<thead>
<tr>
<th></th>
<th>Ruler</th>
<th>Graham</th>
<th>Brian</th>
<th>David</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ruler</td>
<td>3.8778</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Graham</td>
<td>2.8110</td>
<td>2.1210</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brian</td>
<td>3.1480</td>
<td>2.2669</td>
<td>2.6550</td>
<td></td>
</tr>
<tr>
<td>David</td>
<td>3.5062</td>
<td>2.5690</td>
<td>2.8341</td>
<td>3.2352</td>
</tr>
</tbody>
</table>

\(^1\)Using N-1 as the divisor.
By equating the observed (i.e. estimated) variances and covariance with their predicted values, solve the three simultaneous equations to obtain estimates for a, b and Var(e). Comment on your results.

It also follows from our simple model for Graham's guesses that

\[ \text{Mean}(G) = a + b \text{Mean}(R) \]

Again, equate observed and expected values (after first substituting your estimates of b) to obtain an estimate of the intercept, a.

(e) Repeat (d) for Brian's and Andrew's guesses. Interpret your results for all three sets of guesses. Which is the best guess?

Exercise 2.2: Comparative calibration without a 'gold standard'

Usually, a measure such as R will not be available in an epidemiological study. So, now forget that we have ever seen R and consider ways in which we might calibrate the guesses in the absence of this 'gold standard'.

(a) Produce all possible plots for each pair of guesses (G vs. B, G vs. A, B vs. A).

(b) For each pair of guesses, plot the difference between them against their mean.
   (i.e. three plots: G-B vs. (G+B)/2, etc.)

(c) Estimate the level of agreement for each pair of guesses - using the intra-class correlation (from a one-way ANOVA model) as the appropriate measure of agreement.

Interpret your findings. Do you think you can decide which is the best set of guesses from these analyses? If so, which is the best? Why?
(d) Consider a latent (hidden) true length (T) and a simple measurement model for the three guessed lengths G, B and A, simultaneously. We will assume that we have no knowledge of the truth. The measurement model is

\[
\begin{align*}
G &= a + bT + E_G \\
B &= c + dT + E_B \\
A &= e + fT + E_A
\end{align*}
\]

and

This model implies the following set of predictions (see Appendix B4):

\[
\begin{align*}
\text{Var}(G) &= b^2\sigma_T^2 + \sigma_{E_G}^2 \\
\text{Var}(B) &= d^2\sigma_T^2 + \sigma_{E_B}^2 \\
\text{Var}(A) &= f^2\sigma_T^2 + \sigma_{E_A}^2 \\
\text{Cov}(G,B) &= bd\sigma_T^2 \\
\text{Cov}(G,A) &= bf\sigma_T^2 \\
\text{Cov}(B,A) &= df\sigma_T^2
\end{align*}
\]

Where \(\sigma_T^2\) is the variance of T, and \(\sigma_{E_G}^2\), \(\sigma_{E_B}^2\) and \(\sigma_{E_A}^2\) are the unknown variances of \(E_G\), \(E_B\) and \(E_A\), respectively.

In addition:

\[
\begin{align*}
\text{mean}(G) &= a + b\text{mean}(T) \\
\text{mean}(B) &= c + d\text{mean}(T) \\
\text{mean}(A) &= e + f\text{mean}(T)
\end{align*}
\]

Produce a set of simultaneous equations by equating the above expected values for the variance and covariance with their observed (estimated) values. Show that it is possible to solve these equations to get unique estimates for \(\sigma_{E_G}^2\), \(\sigma_{E_B}^2\) and \(\sigma_{E_A}^2\), but not for the regression coefficients or \(\sigma_T^2\). Show, however, that it is possible to calibrate one instrument against another - that is, to get a unique estimate, for example, of the ratio \(b/d\). Note that it is also impossible to estimate the intercept terms. Show that the estimates of \(\sigma_{E_G}^2\), \(\sigma_{E_B}^2\) and \(\sigma_{E_A}^2\) are:

\[
\begin{align*}
\sigma_{E_G}^2 &= \text{Var}(G) - \text{Cov}(G,B)\cdot\text{Cov}(G,A)/\text{Cov}(B,A) \\
\sigma_{E_B}^2 &= \text{Var}(B) - \text{Cov}(B,A)\cdot\text{Cov}(G,B)/\text{Cov}(G,A)
\end{align*}
\]
\[ \sigma_A^2 = \text{Var}(A) - \text{Cov}(B,A) \cdot \text{Cov}(G,A)/\text{Cov}(G,B) \]

What are their numerical values? Based on the estimates of \( \sigma_G^2 \), \( \sigma_B^2 \) and \( \sigma_A^2 \), which do you think is the best guess? Can you foresee any problems in simply using these variances as measures of precision?

(d) Now fix the value of b to be 1 and a to be 0. This implies that we are scaling the guesses in terms of those produced by Graham. Show that it is now possible to get the following estimates:

\[ d = \frac{\text{Cov}(B,A)}{\text{Cov}(G,A)} \]
\[ f = \frac{\text{Cov}(B,A)}{\text{Cov}(G,B)} \]

\[ \text{mean}(T) = \text{mean}(G) \]
\[ c = \text{mean}(B) - d \cdot \text{mean}(T) \]
\[ e = \text{mean}(A) - f \cdot \text{mean}(T) \]

\[ \sigma_G^2 = \text{Var}(G) - \text{Cov}(G,B) \cdot \text{Cov}(G,A)/\text{Cov}(B,A) \]
\[ \sigma_B^2 = \text{Var}(B) - \text{Cov}(B,A) \cdot \text{Cov}(G,B)/\text{Cov}(G,A) \]
\[ \sigma_A^2 = \text{Var}(A) - \text{Cov}(B,A) \cdot \text{Cov}(G,A)/\text{Cov}(G,B) \]

and

\[ \sigma^2 = \frac{\text{Cov}(G,B) \cdot \text{Cov}(G,A)/\text{Cov}(B,A)}{\text{Cov}(G,B) \cdot \text{Cov}(G,A)/\text{Cov}(B,A)} \]

What are their numerical values? Note that \( \sigma_G^2 \), \( \sigma_B^2 \) and \( \sigma_A^2 \) are unchanged as are the ratios of the regression coefficients (b/d, for example). Reliabilities for the three guesses can be estimated from the following (see AppendixB4):

\[ R_G = \frac{\sigma^2}{\sigma^2 + \sigma_G^2} \]
\[ R_B = \frac{d^2 \sigma^2}{d^2 \sigma^2 + \sigma_B^2} \]

and

\[ R_A = \frac{f^2 \sigma^2}{f^2 \sigma^2 + \sigma_A^2} \]
What are their numerical values?

Whose is the best guess now?
OVERVIEW OF CLASSICAL TEST THEORY

AN INTRODUCTION TO CLASSICAL TEST THEORY

(a) The Underlying Model

Here we start by assuming that an individual test score (X) is the sum of two components: the true score (τ) and a measurement error (e). In terms of simple algebra:

\[ X = \tau + e \]  

(1)

The 'true score', τ, is a hypothetical, unobservable or latent, quantity. It is defined in terms of the average score that would be obtained if the measurement process could be repeated indefinitely (in statistics this average is known as an *expected value*. Note that truth is defined in terms of an average or expected value only - it does not represent truth in a Platonic sense. As the true score can never, in practice, be known, the size of the measurement error is never actually known. We can estimate τ, however, (by the observed score, for example) and get some idea of the precision of this estimate from the variability of the errors (measured by their standard deviation, for example). In test theory the standard deviation of the errors, represented here by \( \sigma_e \), is usually referred to as the *standard error of measurement*.

Now consider two alternative testing instruments, Test 1 and Test 2. For Test 1

\[ X_1 = \tau_1 + e_1 \]  

(2)

For Test 2

\[ X_2 = \tau_2 + e_2 \]  

(3)

The true scores, \( \tau_1 \) and \( \tau_2 \), are not necessarily the same. If, however, \( \tau_1 \) and \( \tau_2 \) are perfectly correlated, then the two tests are referred to as *congeneric tests*. This might arise, for example, with two tests, thought to be measuring the same ability, which have different lengths. If \( \tau_1 = \tau_2 \) then the tests are *tau-equivalent*. Finally, if the tests are tau-equivalent and if the standard errors of measurement for the two tests are equal then the tests are *parallel*.

Returning to Equation (1), it is usual to make the following assumptions:

- The mean (expected value) of all of the possible e's is zero. This implies that the mean of the Xs for a given individual is τ.

- The correlation between τ and e is zero.

- The correlation between the true score on one measurement and the error of another is zero.

- The correlation between measurement errors on distinct measurements is zero.
Two supplementary assumptions are also usually made. These are:

- The variance of the errors, $\sigma_e^2$, is fixed (not dependent on the true score, $\tau$). In cases in which the variance of the errors does, in fact, vary with the true score, $\sigma_e^2$ is used to denote the average variance of the errors.

- The errors of measurement are distributed according to the normal or Gaussian distribution.

From Equation (1) and the first five assumptions the following fundamental relationship can be derived:

$$\sigma_X^2 = \sigma_T^2 + \sigma_e^2$$  \hspace{1cm} (4)

where $\sigma_X^2$, $\sigma_T^2$, and $\sigma_e^2$ are the variances of $X$, $T$ and $E$, respectively. The Greek letter $\sigma$ is used to denote a population or expected value of a standard deviation (or standard error), corresponding with the sample statistic which is usually denoted by the Roman equivalent, s. Note that, by definition, the standard error of measurement $\sigma_e$ is the square root of the variance of $e, \sigma_e^2$.

In summary, an observed measurement or test score is assumed to be the sum of two independent (uncorrelated components): truth and error. The implication of this assumption is that the variability of the observed score (measured by its variance) is the sum of the variabilities of the two separate components.

(b) The Definition of Reliability

The reliability of a test, $R$ (or, in more technical descriptions, the Greek equivalent, $\rho$), is defined by the ratio of the variability of the quantity being measured ($T$) to the variability of the observations ($X$). That is,

$$R = \frac{\sigma_T^2}{\sigma_X^2}$$  \hspace{1cm} (5)

An alternative way of writing Equation (5) is

$$R = 1 - \frac{\sigma_e^2}{\sigma_X^2}$$  \hspace{1cm} (6)

Starting with this definition together with the assumptions listed in Section (a) it is possible to demonstrate mathematically that this reliability is identical to the correlation between two parallel tests. This provides a common method of estimation of reliability.

(c) The Relationship Between Reliability and the Standard Error of Measurement

Re-arrangement of Equation (6) produces

$$\frac{\sigma_e^2}{\sigma_X^2} = 1 - R$$

and, therefore
\[ \sigma_e^2 = (1 - R) \sigma_x^2 \]  

(7)

Taking square roots of both sides of this equation yields

\[ \sigma_e = \sigma_x \sqrt{1 - R} \]  

(8)

where \( \sqrt{1-R} \) is the square root of 1-R and \( \sigma_x \) is the standard deviation of the observed test scores.

(e) Reliability of Tests used on Different Populations

Consider a test which has a standard error of measurement of about 3. When used on a population with a standard deviation of 15 the test's reliability can be found from Equation (6). That is,

\[ R = 1 - \frac{3^2}{15^2} \]

\[ = 1 - \frac{9}{225} \]

\[ = 0.96 \]

Remember that a fundamental assumption of classical test theory is that the standard error of measurement is a fixed characteristic of a test. The reliability of a test, on the other hand, is not fixed. Consider the use of a screening questionnaire as a selection device for individuals of high exposure. Suppose, for example, that the standard deviation of the exposure measurement in this restricted population is 10. The reliability of the test when used in these circumstances is

\[ R = 1 - \frac{9}{100} \]

\[ = 0.91 \]

It cannot be stressed too much that the reliability of a test is not a fixed characteristic of that test. It is population-dependent.

(e) The Spearman-Brown Prophesy Formula

This is used to calculate the reliability of the sum (or average) of the scores from k parallel sub-tests, each with a known reliability of, say, \( R_1 \). The reliability of the sum, \( R_k \), is given by

\[ R_k = \frac{kR_1}{1 + (k-1)R_1} \]  

(9)

Note that the reliability of the average of k tests or sub-tests is the same as the reliability of their sum. Consider, for example, the average of the scores from 10 assessments on an individual, each with a reliability of 0.70. The reliability of this average is given by

\[ R_{10} = \frac{10 \times 0.7}{1 + 9 \times 0.7} \]
That is,

\[ R_{10} = \frac{7}{1+6.3} \]

= 0.96

In the special case when k=2, then Equation (9) is equivalent to

\[ R_2 = \frac{2R}{1+R_1} \] (10)

The latter expression is used to calculate the reliability of a whole test derived from estimation of the reliability through the correlation between split-halves. Another application of the Spearman Brown formula is in the theoretical interpretation of Cronbach's coefficient of internal consistency (Cronbach's alpha coefficient). Here one takes the correlation between all possible split-halves for a multi-item test. The Spearman Brown Formula is then used to calculate the reliability of the whole test for each of the possible ways of producing the split-halves. It can be shown mathematically that, although it is usually calculated in a much simpler way, Cronbach's alpha coefficient is equal to the average value of all the possible split-half reliabilities.
FURTHER READING

Elementary Texts


Intermediate Texts


For Reference Only


Appendix B1
EXPECTED VALUES

Means

First consider measurements that can have only certain discrete values. A typical example is the number of correct answers to a series of questions designed to test a subject's ability in mathematics. Another would be the coding of a five-point severity scale for depression, for example. The possible values for these two measurements would all be integers (whole numbers), but this is not always the case. The first score could be converted to the corresponding proportion of correct answers, and it would still only take particular discrete values. An intelligence quotient will, of course, be measured by integer values but it will usually be treated as if it were (potentially, at least) measured on a continuous scale.

Let a discrete measurement be represented as a random variable X. Let the k possible values of X be labelled as $x_1, x_2, \ldots, x_k$ (with a typical values being denoted by $x_i$). Furthermore, let the probability of the observation of $x_i$ be represented by $\text{Prob}(X=x_i)$ or, more simply, by $P(x_i)$. The sequence of probabilities $P(x_1), P(x_2), \ldots, P(x_k)$ defines a probability distribution of the random variable X.

The expected value or mean of this random variable is defined by

\[
E(X) = \mu_X = \sum_{i=1}^{k} x_i P(x_i) = x_1 P(x_1) + x_2 P(x_2) + \ldots + x_k P(x_k)
\]

(1)

where the symbol $\sum$ means 'add over values of $x_i P(x_i)$ from $i=1$ to $i=k$. The expression $E(X)$ is read as 'expected value of X' or 'expectation of X'.

Consider an example. Suppose that a five-point scale for severity of depressed mood is coded from 1 (no depression) to 5 (severe depression) and suppose that the five categories have probabilities of being observed as follows:

<table>
<thead>
<tr>
<th>Code</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.4</td>
</tr>
<tr>
<td>2</td>
<td>0.3</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>0.05</td>
</tr>
<tr>
<td>5</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Then the mean, or expected value is given by

\[
(1 \times 0.4) + (2 \times 0.3) + (3 \times 0.2) + (4 \times 0.05) + (5 \times 0.05)
\]

or

0.4 + 0.6 + 0.6 + 0.2 + 0.25
Variances

If we denote the mean of a random variable $X$ by $\mu_X$, then the variance of this random variable is defined by

$$\text{Var}(X) = E(X - \mu_X)^2 \quad (2)$$

In words, the variance of $X$ is the expected value of the squared deviation of the values of $X$ from its mean. That is, it is the mean of these squared deviations. In the above numerical example the deviations and their squares are given as follows:

<table>
<thead>
<tr>
<th>Code</th>
<th>Probability</th>
<th>Deviation(D)</th>
<th>$D^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.4</td>
<td>-1.05</td>
<td>1.1025</td>
</tr>
<tr>
<td>2</td>
<td>0.3</td>
<td>-0.05</td>
<td>0.0025</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
<td>0.95</td>
<td>0.9025</td>
</tr>
<tr>
<td>4</td>
<td>0.05</td>
<td>1.95</td>
<td>3.8025</td>
</tr>
<tr>
<td>5</td>
<td>0.05</td>
<td>2.95</td>
<td>8.7025</td>
</tr>
</tbody>
</table>

The variance of this variable is therefore:

$$(1.1025 \times 0.4) + (0.0025 \times 0.3) + (0.9025 \times 0.2) + (3.8025 \times 0.05) + (8.7025 \times 0.05)$$

$$= 1.2165$$

The standard deviation of $X$ is simply the square root of its variance; that is, 1.1030.

Means and Variances for Continuous Measures

Now consider a measurement that is continuous (potentially, at least, if not in practice). Typical examples would be measurements of weight, time or length. If a continuous measurement is represented by the random variable $Y$, then the probability of observing a value of $Y$ between $y$ and $y+dy$ (where $dy$ is an infinitesimally small increase in $Y$) is denoted by $p(y)dy$. The term $p(y)$ is a probability density function (or p.d.f. for short) of the random variable $Y$. In this case the expected value of $Y$ is defined by integrating over all of the range of possible values of $Y$. That is

$$E(Y) = \int p(y)dy \quad (3)$$

The variance of $Y$ is defined as in Equation (2), but in this case integration is carried out over the range of $Y$ values rather than simple summation.

Covariation and Correlation

Quite often one is interested in the way in which two types of measurement, such as a subject's depression score and anxiety score, co-vary. If these two measurements are represented by two random variables, $D$ and $A$, with expected values $\mu_D$ and $\mu_A$, respectively, then the covariance of $D$ and $A$ is

$$\text{Cov}(D, A) = E((D - \mu_D)(A - \mu_A))$$

$$= \int \int (y - \mu_Y)(z - \mu_Z)p(y,z)dydz$$

where $p(y,z)$ is the joint probability density function of the random variables $Y$ and $Z$. The covariance measures the linear dependence between $Y$ and $Z$. A positive covariance indicates that the variables tend to move in the same direction, while a negative covariance indicates that they tend to move in opposite directions.
and A can be defined by the following expression:

\[
\text{Cov}(D,A) = E[(D-\mu_D)(A-\mu_A)] \tag{4}
\]

It should be clear from this definition that the covariance of a random variable with itself is, in fact, that variable's variance. If the standard deviations of D and A are denoted by \(\sigma_D\) and \(\sigma_A\), respectively, then the product-moment correlation (or correlation for short) between D and A is

\[
\text{Corr}(D,A) = \frac{\text{Cov}(D,A)}{\sigma_D \sigma_A} \tag{5}
\]

Note that \(\text{Cov}(D,A)=\text{Cov}(A,D)\).

**Variances of Sums and Differences of Random Variables**

Consider a pair of measurements, \(X_1\) and \(X_2\), with means \(\mu_1\) and \(\mu_2\), respectively. Let

\[
Y = X_1 + X_2
\]

Then

\[
E(Y) = E(X_1+X_2) = E(X_1)+E(X_2) = \mu_1+\mu_2 \tag{6}
\]

In words, the mean of the sum of two variables is the same as the sum of their means. Also,

\[
\text{Var}(Y) = E[(X_1+X_2-(\mu_1+\mu_2))^2]
\]

\[
= E[(X_1-\mu_1)+(X_2-\mu_2)]^2
\]

\[
= E[(X_1-\mu_1)^2+(X_2-\mu_2)^2+2(X_1-\mu_1)(X_2-\mu_2)]
\]

\[
= E[(X_1-\mu_1)^2+E(X_2-\mu_2)^2+2E[(X_1-\mu_1)(X_2-\mu_2)]]
\]

\[
= \text{Var}(X_1)+\text{Var}(X_2)+2\text{Cov}(X_1,X_2) \tag{7}
\]

In particular, if the two variables are independent (uncorrelated), then the variance of their sum is the sum of their separate variances (that is, the covariance term drops out). That is

\[
\text{Var}(X_1+X_2) = \text{Var}(X_1)+\text{Var}(X_2) \tag{8}
\]

Similarly, it can be shown that

\[
\text{Var}(X_1-X_2) = \text{Var}(X_1)+\text{Var}(X_2)-2\text{Cov}(X_1,X_2) \tag{9}
\]
And, again, when \( X_1 \) and \( X_2 \) are uncorrelated:

\[
\text{Var}(X_1 - X_2) = \text{Var}(X_1) + \text{Var}(X_2)
\]  

(10)

The variance of the difference of two uncorrelated random variables is the \textit{sum} of their separate variances (\textit{not their difference}).

The variance of \( 2X \) is obtained from Equation (7) by noting that \( \text{Cov}(X,X) = \text{Var}(X) \) and, therefore:

\[
\text{Var}(2X) = 4\text{Var}(X)
\]  

(11)

In general, for any number \( a \):

\[
\text{Var}(aX) = a^2\text{Var}(X)
\]  

(12)

If \( X_1 = \tau + e_1 \) and \( X_2 = \tau + e_2 \) and \( e_1 \) and \( e_2 \) both have variance \( \text{Var}(e) \) and are uncorrelated with each other and with \( \tau \) (as in Classical Test Theory, for example), then

\[
\text{Var}(X_1 + X_2) = \text{Var}(2\tau + e_1 + e_2)
\]

\[
= 4\text{Var}(\tau) + 2\text{Var}(e)
\]  

(13)

where \( \text{Var}(e) \) is the variance of the errors. Similarly,

\[
\text{Var}(X_1 - X_2) = \text{Var}(e_1 - e_2)
\]

\[
= 2\text{Var}(e)
\]  

(14)

Expressions (12) and (13) lead to estimators of \( \text{Var}(\tau) \) and \( \text{Var}(e) \) when one can equate these expected values with observed sample statistics (the sample variances for the sums and differences of two parallel test scores).
Appendix B2
EXPECTED MEAN SQUARES (VARIANCE COMPONENTS)

One-Way Models

Consider a simple reliability study in which each of a sample of subjects is measured k times (using parallel tests). We assume that individual test score (X) is the sum of two components: the true score (τ) and a measurement error (e). In terms of simple algebra:

\[ X = \tau + e \]  

(1)

The observed score is a random variable with variance \( \sigma_X^2 \). Similarly, the true score is a random variable with variance \( \sigma_\tau^2 \) and e is a third random variable with variance \( \sigma_e^2 \). If the \( \tau \)'s for different subjects are assumed to be uncorrelated and the \( e \)'s are also uncorrelated both with each other and with the \( \tau \)'s, then it can be shown that

\[ \text{Var}(X) = \text{Var}(\tau) + \text{Var}(e) \]  

(2)

\[ \sigma_X^2 = \sigma_\tau^2 + \sigma_e^2 \]

Now, if we carry out a one-way analysis of variance on data arising from the simple reliability study above, we will obtain two estimated mean squares, \( s_B^2 \) and \( s_W^2 \). These are the between-subjects mean square and the within-subjects mean square, respectively. With a knowledge of the algebra of expected values, it is possible to demonstrate the truth of the following to equalities:

\[ E(s_B^2) = k\sigma_\tau^2 + \sigma_e^2 \]  

(3)

and

\[ E(s_W^2) = \sigma_e^2 \]  

(4)

By equating the observed (estimated) mean squares with their expected values, we can estimate the variance components to the right of the equality in Equation 2 by

\[ \sigma_e^2 = s_W^2 \]  

(5)

and

\[ \sigma_\tau^2 = (s_B^2 - s_W^2)/k \]  

(6)

These two estimates can, in turn, be used to estimate reliability (intra-class correlation):

\[ r_i = \sigma_\tau^2/(\sigma_\tau^2 + \sigma_e^2) \]

\[ = (s_B^2 - s_W^2)/[s_B^2 + (k-1)s_W^2] \]  

(7)
In the simplest case (k=2), this estimate becomes
\[ r_i = \frac{(s^2_B - s^2_W)}{(s^2_B + s^2_W)} \] (8)

**Two-Way Models**

Now consider a reliability study in which a sample of \( n \) patients is each rated by a sample of \( k \) raters, giving a total of \( kn \) observations. We can carry out a two-way analysis of variance on data provided by a study such as this and estimate three mean squares: \( s^2_P \), \( s^2_R \) and \( s^2_E \). These correspond to the mean square between patients (P), between raters (R) and error (E), respectively. Assuming that there are no patient-rater interactions, the measurement model for this study is the following:

\[ X = m + b + e \] (9)

where \( m \) is a subject's universe score, \( b \) is the rater effect, and \( e \) is random measurement error. Making the usual analysis of variance assumptions about the statistical independence of these effects, it follows that:

\[ \sigma^2_X = \sigma^2_m + \sigma^2_b + \sigma^2_e \] (10)

where \( \sigma^2_m \), \( \sigma^2_b \) and \( \sigma^2_e \) are the variance components due to patients, raters and measurement errors, respectively.

Using the algebra of expected values, it can be shown that:

\[ E(s^2_P) = k\sigma^2_b + \sigma^2_e \] (11)
\[ E(s^2_R) = n\sigma^2_b + \sigma^2_e \] (12)
and
\[ E(s^2_E) = \sigma^2_e \] (13)

By equating observed and expected values of the three mean squares we obtain the following estimates for the variance components:

\[ \sigma^2_e = s^2_E \] (14)
\[ \sigma^2_b = (s^2_P - s^2_E)/k \] (15)
and
\[ \sigma^2_b = (s^2_R - s^2_E)/n \] (16)

These in turn can be used to estimate various reliability coefficients. The reliability of scores made by randomly selected raters is given by:

\[ R = \frac{\sigma^2_m}{\sigma^2_m + \sigma^2_b + \sigma^2_e} \] (17)

If, however, a single identified rater is to be used in a future study, then the reliability of that rater's
scores is given by:

\[
R = \frac{\sigma_m^2}{(\sigma_m^2 + \sigma_e^2)}
\]  
(18)

Finally, if, say, \(m\) raters are to provide scores on each subject and inferences are to be based on the means of the \(m\) scores, then the reliability of these means is given by:

\[
R = \frac{\sigma_m^2}{(\sigma_m^2 + (\sigma_b^2 + \sigma_e^2)/m)}
\]  
(19)

Returning to the reliability expression given in Equation (17), this can be re-expressed as

\[
R = \frac{n(s_p^2 - s_E^2)}{(ns_p^2 + ks_R^2 + (nk-n-k)s_E^2)}
\]  
(20)

The expression in Equation (20) is the estimator of the intra-class correlation coefficient for the two-way random effects model (both raters and patients being regarded as random). If the observations, rather than being made on an interval- or ratio-scale, are either binary (yes/no) or on an ordered categorical scale (0/1/2/3 etc), then it is possible to demonstrate that, for all practical purposes, this intra-class correlation is equivalent to Cohen's *kappa coefficient* or *weighted kappa* (with quadratic weights), respectively. Kappa and weighted kappa are chance-corrected measures of agreement for categorical ratings (see Streiner & Norman, 1995).
Appendix B3
RELIABILITY OF THE MEAN OF K PARALLEL TEST SCORES

Consider observed scores \( X_i \) from test i, where

\[
X_i = T + E_i
\]

and

\[
\text{Var}(X_i) = \text{Var}(T) + \text{Var}(E_i)
\]

The true score, \( T \), is identical for (common to) all of the \( k \) tests. Similarly, for these tests to be parallel, \( \text{Var}(E_i) \) is not dependent on the choice of \( i \) (that is, the standard errors of measurement for all tests are the same). The reliability of scores provided by a single test is given by:

\[
R = \frac{\text{Var}(T)}{\text{Var}(T) + \text{Var}(E)}
\]

where, for simplicity, we have dropped the subscript \( i \) from the measurement error, \( E \). In the following we will also drop the subscript for the observed score, \( X_i \), when referring to its variance, \( \text{Var}(X_i) \) (that is, we will refer to \( \text{Var}(X) \)).

Now consider the mean of the scores from two parallel tests. The mean is given by \((X_1 + X_2)/2\), the true scores corresponding to this mean is simply \( T \), and the measurement error is \((E_1 + E_2)/2\). The variance of the true score of the mean is unchanged at \( \text{Var}(T) \), but that for the error component is smaller:

\[
\text{Var}((E_1 + E_2)/2) = \frac{\text{Var}(E)}{2}
\]

The reliability of the mean is therefore:

\[
R_2 = \frac{\text{Var}(T)}{\text{Var}(T) + \text{Var}(E)/2}
\]

where the subscript 2 in \( R_2 \) is given to indicate the reliability of the mean of (or, equivalently, the sum of) the two parallel tests. If we note that \( \text{Var}(T) = R \cdot \text{Var}(X) \) and \( \text{Var}(E) = (1-R) \cdot \text{Var}(X) \), \( \text{Var}(X) \) being the common variance of the observed scores for the two tests, then Equation (5) can be written as:

\[
R_2 = \frac{R \cdot \text{Var}(X)}{R \cdot \text{Var}(X) + (1-R) \cdot \text{Var}(X)/2}
\]

\[
= \frac{R}{R + (1-R)/2}
\]

With further re-arrangement, this becomes:

\[
R_2 = \frac{2R}{1+R}
\]
This is the well-known Spearman-Brown Prophesy Formula for the reliability of the full scale in terms of the common reliabilities (correlations) of split-halves.

Moving to the mean of $k$ parallel tests, again the mean true score is $T$ and the error of measurement for the mean is $(E_1+E_2+E_3+ \ldots +E_k)/k$ or $SE/k$. The variance of the error component of the mean score is given by:

$$Var(SE/k) = Var(E)/k$$  \hspace{1cm} (7)

The reliability of the mean of $k$ parallel test scores, $R_k$, is therefore given by:

$$R_k = Var(T)/(Var(T)+Var(E)/k)$$  \hspace{1cm} (8)

Again, this can be re-arranged, after substituting $R Var(X)$ for $Var(T)$ and $(1-R) Var(X)$ for $Var(E)$, to give:

$$R_k = kR/[1+(k-1)R]$$  \hspace{1cm} (9)

This is the more general Spearman-Brown Prophesy Formula. Note that, in general, the Prophesy Formula given by Equation (9) can be used for values of $k$ which are less than one. If $R$ is the reliability of a full-length test, for example, then the reliability of each of two randomly-split halves is provided by:

$$R_{0.5} = (R/2)/(1-(R/2))$$
Appendix B4
RELATIVE CALIBRATION WITHOUT A GOLD STANDARD (COVARIANCE COMPONENTS)

The following model may be postulated to explain the relationships between three sets of measurements (Graham's, Brian's and Andrew's guesses of the lengths of pieces of string, for example):

\[ X_{ik} = \alpha_k + \beta_k \mu_i + e_{ik} \]  

or

\[ E(X_{ik}|\mu_i) = \alpha_k + \beta_k \mu_i \]

where \( X_{ik} \) is the kth measurement on the ith specimen or subject (k=1,2 or 3), and \( \mu_i \) is the unknown 'true' value for that subject, and \( e_{ik} \) (with a mean of zero) is the measurement error associated with \( X_{ik} \). [Note that in the context of Classical Test Theory the true scores, \( T_i = \alpha_k + \beta_k \mu_i \), are those corresponding to congeneric tests.] In this model, \( X_{ik}, \mu_i \) and \( e_{ik} \) are all assumed to be random variables. Two parameters of this model, that is, \( \alpha_k \) and \( \beta_k \), jointly describe the measurement bias (calibration) characteristic of instrument k; \( \alpha_k \) and \( \beta_k \), respectively, being the intercept and slope of the line relating true value to that of the observed indicator. Let the population mean of the \( \mu_s \), \( E(\mu) \), be M. Assuming that measurement errors are uncorrelated to the true values, then the variance of \( \mu_i \), denoted by \( \sigma^2 \), and the variance of the \( e_{ik} \), denoted by \( \sigma_k^2 \), jointly predict the variance of the \( X_{ik} \) as follows:

\[ \text{Var}(X_{ik}) = \beta_k^2 \sigma^2 + \sigma_k^2 \]  

The covariance of pairs of indicators, \( X_{ik} \) and \( X_{ik'} \), \( \text{Cov}(X_{ik}, X_{ik'}) \), is provided by:

\[ \text{Cov}(X_{ik}, X_{ik'}) = \beta_k \beta_{k'} \sigma^2 \]

Dropping the i's from the subscripts for simplicity, and explicitly considering k=1, 2 or 3, we have the following sets of equations:

\[ \text{Var}(X_1) = \beta_1^2 \sigma^2 + \sigma_1^2 \]  

\[ \text{Var}(X_2) = \beta_2^2 \sigma^2 + \sigma_2^2 \]  

\[ \text{Var}(X_3) = \beta_3^2 \sigma^2 + \sigma_3^2 \]  

\[ \text{Cov}(X_1, X_2) = \beta_1 \beta_2 \sigma^2 \]  

\[ \text{Cov}(X_1, X_3) = \beta_1 \beta_3 \sigma^2 \]  

\[ \text{Cov}(X_2, X_3) = \beta_2 \beta_3 \sigma^2 \]

In addition:

\[ E(X_1) = \alpha_1 + \beta_1 M \]  

\[ E(X_2) = \alpha_2 + \beta_2 M \]  

\[ E(X_3) = \alpha_3 + \beta_3 M \]
For the time being we will ignore M and the $\alpha$s and concentrate on the estimation of the parameters contributing to the pattern of variances and covariances. Given an observed variance-covariance matrix for a set of three measurements on each of a sample of subjects we can simply equate these summary statistics to their predicted values (in equations 5-10, above) and solve the resulting six simultaneous equations for the unknown parameters ($\beta_1, \beta_2, \beta_3, \sigma_1^2, \sigma_2^2, \sigma_3^2$ and $\sigma^2$). But there is a catch! Note that there are 7 parameters describing the pattern of variances and covariances, but only 6 independent summary statistics from which to estimate them. The model is what is called under-identified. The problem arises from the fact that we have not fixed the scale of measurement. As it expands and contracts (via changes in $\sigma^2$) the $\beta$s will adjust accordingly (but note that their values relative to one another will remain the same). So, we need to do something to fix the scale of measurement. One way of doing this (which we will use here) is to fix the scale by comparing the second two instruments with the first, and this is done by fixing the value for regression coefficient for the first instrument. It could be any value, but here we fix $\beta_1 = 1$. There are other alternatives (psychometricians or sociologists would be more likely to fix $\sigma_2 = 1$, on the assumption that the scale of measurement for a behavioural or social indicator is usually completely arbitrary), but we will not discuss them in detail here. Having introduced this constraint the estimation problem is now straightforward. Let the observed covariance between $X_k$ and $X_{k'}$ be represented by $S_{kk'}$, with the observed variance of $X_k$ being $S_{kk}$. Then, it is straightforward to show that the remaining 6 parameters are estimated as follows:

$$\beta_2 = \frac{S_{23}}{S_{13}}$$  
$$\beta_3 = \frac{S_{23}}{S_{12}}$$  
$$\sigma^2 = \frac{S_{12} S_{13}}{S_{23}}$$  
$$\sigma_1^2 = S_{11} - \frac{S_{12} S_{13}}{S_{23}}$$  
$$\sigma_2^2 = S_{22} - \frac{S_{12} S_{23}}{S_{13}}$$  
$$\sigma_3^2 = S_{33} - \frac{S_{13} S_{23}}{S_{12}}$$  

Similarly, we can equate observed and predicted means in equations 11-13 to solve for $\alpha_1, \alpha_2, \alpha_3,$ and M. Again there is an identifiability problem and we solve this by setting $\alpha_1=0$ (and, again, there are mathematically equivalent alternatives such as setting $M=0$). having done this, we have estimates provided by:

$$M = \text{mean}(X_1)$$  
$$\alpha_2 = \text{mean}(X_2) - \beta_2 M$$  
$$\alpha_3 = \text{mean}(X_3) - \beta_3 M$$  

where, in equations 20 and 21, the parameters $\beta_2, \beta_3$ and M are replaced by their estimates.

Finally, note that the reliability of the kth instrument, $R_k$, is provided by

$$R_k = \frac{\beta_k^2 \sigma^2}{(\beta_k^2 \sigma^2 + \sigma_k^2)}$$

and it is estimated by replacing the parameters to the right of this equation by their estimates.