Sports Med 2000 Jul; 30 (1): 1-15
0112-1642/00/0007-0001/$20.00/0

Measures of Reliability in Sports Medicine and Science

Will G. Hopkins

Department of Physiology, School of Medical Sciences and School of Physical Education,
University of Otago, Dunedin, New Zealand

Abstract

Reliability refers to the reproducibility of values of a test, assay or other measurement in repeated trials on the same individuals. Better reliability implies better precision of single measurements and better tracking of changes in measurements in research or practical settings. The main measures of reliability are within-subject random variation, systematic change in the mean, and retest correlation. A simple, adaptable form of within-subject variation is the typical (standard) error of measurement: the standard deviation of an individual's repeated measurements. For many measurements in sports medicine and science, the typical error is best expressed as a coefficient of variation (percentage of the mean). A biased, more limited form of within-subject variation is the limits of agreement: the 95% likely range of change of an individual's measurements between 2 trials. Systematic changes in the mean of a measure between consecutive trials represent such effects as learning, motivation or fatigue; these changes need to be eliminated from estimates of within-subject variation. Retest correlation is difficult to interpret, mainly because its value is sensitive to the heterogeneity of the sample of participants. Uses of reliability include decision-making when monitoring individuals, comparison of tests or equipment, estimation of sample size in experiments and estimation of the magnitude of individual differences in the response to a treatment. Reasonable precision for estimates of reliability requires approximately 50 study participants and at least 3 trials. Studies aimed at assessing variation in reliability between tests or equipment require complex designs and analyses that researchers seldom perform correctly. A wider understanding of reliability and adoption of the typical error as the standard measure of reliability would improve the assessment of tests and equipment in our disciplines.

Measurement error makes the observed value of a measure differ from the true value. Anyone who takes or uses measurements should therefore have some understanding of measurement error. In my experience, the 2 most important aspects of measurement error are concurrent validity and retest reliability. Concurrent validity concerns the agreement between the observed value and the true or criterion value of a measure. Retest reliability concerns the reproducibility of the observed value when the measurement is repeated. Analysis of validity is complex, owing to the inevitable presence of error in the criterion value. I have therefore limited this article to the measurement errors that are accessible in reliability studies. These errors have a major impact on our attempts to measure changes between repeated measurements; they are also a concern for anyone interested in a single measurement.

Studying the reliability of a measure is a straightforward matter of repeating the measurement a reasonable number of times on a reasonable number of individuals. The most important measurement error to come out of such a study is the random error or 'noise' in the measure: the smaller the error, the better the measure. How best to represent this error and several other measures of reliability is a matter of debate. Atkinson and Nevill[1] contributed a useful point of view in their review of reliability in this journal recently, but I have a different perspective on the relative merits of the various measures of reliability. In the present article I justify my choice of the most appropriate measures. I also explore the uses of reliability and deal with the design and analysis of reliability studies. My approach to reliability is appropriate for most variables that have numbers as values (e.g. 71.3kg for body mass). Reliability of measures that have labels as values (e.g. female for sex) is beyond the scope of the present article.

1. Measures of Reliability

When we speak of reliability, we refer to the repeatability or reproducibility of a measure or variable. I will sometimes follow the popular but inaccurate convention of referring not to the reliability of a measure but to the reliability of the test, assay or instrument that provided the measure. I will also use the word 'trials' to mean repeated administrations of a test or assay.

Researchers quantify reliability in a variety of ways. I deal here with what I believe are the only 3 important types of measure: within-subject variation, change in the mean, and retest correlation.[2]

1.1 Within-Subject Variation

Within-subject variation is the most important type of reliability measure for researchers, because it affects the precision of estimates of change in the variable of an experimental study. It is also the most important type of reliability measure for coaches, physicians, scientists and other professionals using tests to monitor the performance or health of their clients. In these situations, the smaller the within-subject variation, the easier it will be to notice or measure a change in performance or health.

An easy way to understand the meaning of within-subject variation is to regard it as the random variation in a measure when one individual is tested many times. For example, if the values for many trials of one individual are 71, 76, 74, 79, 79 and 76, there is a random variation of a few units between trials. A statistic that captures this notion of random variability of a single individual's values on repeated testing is the standard deviation of the individual's values. This within-subject standard deviation is also known as the standard error of measurement. In plain language, it represents the typical error in a measurement, and that is how I will refer to it hereafter.

The variation represented by typical error comes from several sources. The main source is usually biological. For example, an individual's maximum power output changes between trials because of changes in mental or physical state. Equipment may also contribute noise to the measurements, although in simple reliability studies this technological source of error is often unavoidably lumped in with the biological error. When the same individual is retested on different equipment or by different operators, additional error due to differences in the calibration or functioning of the equipment or in the ability of the operators can surface. An analogous situation occurs when different judges rate the same athlete in different locations. I will deal with these and other complex examples of reliability in section 3.3.

In most situations where reliability is an issue, we are interested in the simple question of reproducibility of an individual's values obtained on the same piece of equipment by the same operator. To estimate typical error in these situations, we usually use many participants and a few trials rather than 1 participant and many trials. For example, for 5 participants in 2 trials, with the values shown in table I, the typical error is 2.9. We can still interpret the typical error of 2.9 as the variation we would expect to see from trial to trial if any one of these participants performed multiple trials.

When a group of volunteers performs 2 or more trials, there is always a change in the mean value between trials. In the above example, the means in the first and second trial are 68.4 and 69.6, respectively, so there is a change in the mean of 1.2. Change in the mean is itself a measure of reliability that I discuss in more detail in the next section. I introduce the concept here to point out that, for almost all applications of reliability, it is important to have an estimate of typical error that is unaffected by a change in the mean. The values of the change score or difference score for each volunteer yield such an estimate: simply divide the standard deviation of the difference score by √2. In the above example, the difference scores are 5, –2, 6, 0 and –3; the standard deviation of these scores is 4.1, so the typical error is $4.1/\sqrt{2} = 2.9$. This method for calculating the typical error follows from the fact that the variance of the difference score ($s_{diff}{}^2$) is equal to the sum of the variances representing the typical error in each trial: $s_{diff}{}^2 = s^2 + s^2$, so $s = s_{diff}/\sqrt{2}$.

For many measurements in sports medicine and science, the typical error gets bigger as the value of the measure gets bigger.[3] For example, several trials on an ergometer for one athlete might yield power output with a mean and typical error of 378.6 ± 4.4W, whereas a stronger athlete performing the same trials might produce 453.1 ± 6.1W. Although the absolute values of the typical errors are somewhat different, the values expressed as a percentage of their respective means are similar: 1.2 and 1.3%. This form of the typical error is a coefficient of variation. It is sometimes more applicable to every participant than the raw typical error. As a dimensionless measure, it also allows direct comparison of reliability of measures irrespective of calibration or scaling. Thus it facilitates comparison of reliability between ergometers, analysers, tests or populations of volunteers. I will refer to it in plain language as the typical percentage error.

Another measure of within-subject variation, limits of agreement, has begun to appear in reliability studies. Bland and Altman,[4] the researchers who devised this measure, realised that the difference scores between trials give a good indication of the reliability of the test. Instead of using the standard deviation of the difference scores directly, they calculated the range within which an individual's difference scores would fall most (95%) of the time. In the above example of 5 individuals tested twice, the 95% limits of agreement are –10.1 and 12.5. The interpretation of these limits is as follows: on the basis of our 2 trials with 5 participants, when we test and then retest another participant, the score in the second trial has 1 chance in 20 of being more than 12.5 higher or less than 10.1 lower than the score in the first trial. Note that the limits in this example are not quite symmetrical, because the participants showed an average improvement of 1.2 in the second trial. It is preferable to take this improvement out of each limit and express the limits as 1.2 ± 11.3.

Table I. Data from a reliability study for a variable measured twice in 5 participants

Participant	Trial 1	Trial 2
Kim	62	67
Lou	78	76
Pat	81	87
Sam	55	55
Vic	66	63

The relationship between the typical error and the limits of agreement is straightforward. Let the limits of agreement be L. As before, let the within-subject standard deviation (typical error) be s, and the standard deviation of the difference score be s_{diff}. For simplicity, we will ignore any change in the mean between the trials. It follows from basic statistical theory that $L = \pm t_{0.975,\nu} \cdot s_{diff}$, where $t_{0.975,\nu}$ is the value of the t statistic with cumulative probability 0.975 and ν degrees of freedom. But $s_{diff} = s \cdot \sqrt{2}$, so:

$$L = \pm t_{0.975,\nu} \cdot s \cdot \sqrt{2} \qquad \text{(Eq. 1)}$$

In our example of 5 participants, $s = 2.9$, $\nu = 4$ and $t_{0.975,4} = 2.8$, so the limits of agreement are $\pm(2.8)(\sqrt{2})s = \pm 3.9s = \pm 11.3$. When a reliability study has a large sample size, $t_{0.975,\nu} = 1.96$, so $L = \pm 1.96s \cdot \sqrt{2} = \pm 2.77s$, or approximately ± 3 times the typical error. This formula is still valid when the typical error is expressed as a coefficient of variation; the corresponding limits of agreement are then percentage limits.

Should researchers use the typical error or the limits of agreement as a measure of within-subject variation? Atkinson and Nevill[1] favoured limits of agreement. I believe typical error is better. Here are my reasons.

- As I have just shown, the values of the limits of agreement depend on the sample size of the reliability study from which they are estimated. In statistical terms, the limits are biased. The bias is < 5% when there are more than 25 degrees of freedom (e.g. > 25 participants and 2 trials, or > 13 participants and 3 trials), but it rises to 21% for 7 degrees of freedom (8 participants and 2 trials). In most studies of reliability, between 8 and 30 volunteers perform only 2 trials. The resulting bias ranges from 21 to < 5%, so anyone comparing the magnitude of limits of agreement between studies must account for the number of degrees of freedom between the studies. This problem does not occur with the typical error, which has an expected value independent of sample size. Defenders of limits of agreement might argue that we should compute limits of agreement in all studies by multiplying the typical error by 2.77 rather than by the exact value derived from the t statistic with the right number of degrees of freedom. In that case, though, the level of confidence of the limits would not be well defined.
- Limits of agreement apply to the special case of variability of an individual's values between pairs of trials, but they do not apply to the simplest situation of only one trial (e.g. a urine test for a banned substance). With a single trial, the user is interested in the error in the value of that trial, not in the error in the difference between the trial and some hypothetical previous or future trial. Characterising the variability of a single measurement with confidence limits for a difference score is therefore fatuous. Confidence limits for a single measurement would be more appropriate, but as a generic measure of within-subject variation this statistic would have the same bias problem as limits of agreement.
- The widespread use of 95% confidence limits to represent precision of the estimate of population parameters is not a basis for using 95% to define agreement limits for an individual participant's difference scores. Even the use of 95% for confidence intervals is debatable, but I will not go into that issue here. Instead, I will show that 95% is too stringent for a decision limit, at least when the participant is an athlete. Let us assume we are monitoring the performance of a runner with a reasonably good running test, one that has 95% limits of agreement of ± 7.0%. Proponents of limits of agreement would argue that an athlete or coach should be satisfied that something beneficial has happened between 2 trials only when there is an increase in performance of 7.0% or more. But with an observed change of + 7.0%, there is a 97.5% probability (odds of 39 to 1) that performance is indeed better, or a 2.5% probability (odds of 1 to 39) that it is worse. In my view, this degree of certainty about a true change in performance is unrealistic: an individual would or should act on less. For example, half the limits of agreement seems a more reasonable threshold for action; with an observed enhancement of 3.5%, the probability that a true enhancement has occurred is still 84%, or odds of about 5 to 1 that performance is really better. Even smaller changes in performance are worthwhile for top runners,[2] but you would need a test with better reliability to be confident that such changes were more than just chance occurrences in this simple test-retest situation with a single athlete.
- There is an extensive theoretical base for reliability, the most developed form of which is known as generalisability theory.[5,6] Variances are the common coin for all computations in this literature. Anyone wishing to perform computations using a published typical error has only to square the published value to convert it to a vari-

ance. Procedures for calculating confidence limits of the variance (and therefore of the typical error) are also available. On the other hand, limits of agreement have to be converted to a variance by factoring in the appropriate number of degrees of freedom. The conversion is straightforward for simple reliability studies, but for more complex measures of reliability involving several variance components, counting the degrees of freedom may be a challenge. I am also uncertain whether the factor that converts typical error to limits of agreement is the appropriate factor to convert the confidence limits of the typical error to confidence limits of the limits of agreement, at least for < 25 degrees of freedom.

- Which measure is better for the purpose of teaching or learning about measurement error? Although the numerical difference between them is only a factor of approximately 3, conceptually they are quite different. In my opinion the concept of typical error is self-explanatory, and it conveys what measurement error is all about: variation in the values of repeated measurements. The concept of 95% confidence limits for the difference between 2 measurements narrows the focus of measurement error to one application: decision-making in a test-retest situation. This appears to be the only situation where limits of agreement would have an advantage over the typical error, if 95% confidence limits were appropriate for decisions affecting an individual.

Researchers and editors now have to consider which of these 2 measures they will publish in reliability studies. Publishing both is probably inappropriate, because they are too closely related.

1.2 Change in the Mean

This measure of reliability is simply the change in the mean value between 2 trials of a test. The change consists of 2 components: a random change and a systematic change (also known as systematic bias).

Random change in the mean is due to so-called sampling error. This kind of change arises purely from the random error of measurement, which inevitably makes the mean for each trial different. The random change is smaller with larger sample sizes, because the random errors from each measurement tend to cancel out when more measurements are added together for calculation of the mean.

Systematic change in the mean is a non-random change in the value between 2 trials that applies to all study participants. The simplest example of a systematic change is a learning effect or training effect: the participants perform the second trial better than the first, because they benefit from the experience of the first trial. In tests of human performance that depend on effort or motivation, volunteers might also perform the second trial better because they want to improve. Performance can be worse in a second trial if fatigue from the first trial is present at the time of the second trial. Performance can also decline in a series of trials, owing to loss of motivation.

Systematic change in the mean is an important issue when volunteers perform a series of trials as part of a monitoring programme. The volunteers are usually monitored to determine the effects of an intervention (e.g. a change in diet or training), so it is important to perform enough trials to make learning effects or other systematic changes negligible before applying the intervention.

Systematic changes are seemingly less important for researchers performing a controlled study, because it is the relative change in means for both groups that provides evidence of an effect. However, the magnitude of the systematic change is likely to differ between individuals, and these individual differences make the test less reliable by increasing the typical error (see section 2.3). Researchers should therefore choose or design tests or equipment with small learning effects, or they should get volunteers to perform practice (or familiarisation) trials to reduce learning effects.

1.3 Retest Correlation

This type of measure represents how closely the values of one trial track the values of another as we move our attention from individual to individual. If each participant has an identical value in both trials, the correlation coefficient has a value of 1, and in a plot of the values of the 2 trials all points fall on a straight line. When the random error in the measurement swamps the real measurement, a plot of the values for 2 trials shows a random scatter of points, and the correlation coefficient approaches zero. The correlation also represents how well the rank order of participants in one trial is replicated in the second trial: the closer the correlation gets to 1, the better the replication.

The retest correlation is clearly a good measure of reliability, and it shares with typical percentage error the advantages of being dimensionless. However, the within-subject error is the better measure.[1,2] The main problem with retest correlation is that the value of the correlation is sensitive to the heterogeneity (spread) of values between participants. You can see this effect in a plot of points that have a strong correlation. If you focus on a small subsample of the participants in one part of the plot, the points for those individuals seem to be scattered randomly. As you expand the range of the subsample, the linearity in the scatter gradually emerges. This effect is also obvious from a formula that can be derived from the definition of reliability correlation:[7]

$$r = \text{(pure subject variance)}/\text{(pure subject variance + typical error variance)}$$

$$= (S^2 - s^2)/S^2$$

$$= 1 - (s/S)^2 \qquad \text{(Eq. 2)}$$

where S is the usual between-subject standard deviation and s is the typical error.

If the sample takes in a wide range of participants, S is much greater than s, so $(s/S)^2$ approaches zero and the correlation approaches 1. As we focus in on a homogeneous subgroup, S gets smaller until it equals s in magnitude (i.e. any apparent difference between individuals is due entirely to the random error of measurement); therefore $(s/S)^2$ approaches 1, so the correlation approaches zero. Notice that the value of the retest correlation changes as we change the sample of participants, but at no time does the test itself change, and at no time does the typical error change. The typical error therefore captures the essence of the reliability of the test, but the retest correlation does not.

An important corollary is that the typical error can often be estimated from a sample of individuals that is not particularly representative of a population, or it can be estimated from multiple retests on just a few volunteers. Either way, the resulting typical error often applies to most individuals in the population, whereas the retest correlation applies only to individuals similar to those sampled to estimate the correlation. A further important corollary is that you cannot compare the reliability of 2 measures on the basis of their retest correlations alone: the worse measure (the one with the larger typical error) could have a higher retest correlation if its reliability was determined with a more heterogeneous sample.

Suppose you are satisfied that your participants are similar to those in the published reliability study. How do you decide whether the magnitude of the published correlation is acceptable for your purposes? Authors of reliability studies sometimes give what they consider to be acceptable values. For example, Kovaleski and co-workers[8] cited the classic Shrout and Fleiss paper on reliability[9] to support their claim that a clinically acceptable correlation was 0.75[8] or 0.80.[10] It turns out that Shrout and Fleiss[9] did not assess the utility of magnitudes of retest correlations. Atkinson and Nevill[1] were of the opinion that no-one had defined acceptable magnitudes of the retest correlation for practical use, although they did cite my statistics website[11] for the relationship between retest correlation and sample size in experimental studies (see section 2.2). In fact, there is another study,[12] on acceptable values of the validity correlation, that applies to reliability. In that study, Manly and I found that a test used to assign pass-fail grades needs to have a validity correlation of at least 0.90 to keep the error rate acceptable. Assigning 3 or more grades needs a test with even higher validity. If the only source of error in a test is random error of measurement (the typical error), it is easy to show that the validity correlation is the square root of the retest reliability. Thus tests need to have reliabilities of at least $0.90^2 = 0.81$ to be trustworthy for yes-no decisions about further treatment of an individual, about selection of a team member, or for similar criterion-referenced assessments. I emphasise that this rule applies only when the between-subject standard deviation of your participants is similar to that in the reliability study.

2. Uses of Reliability

I have already mentioned how reliability affects the precision of single measurements and change scores. Anyone making decisions based on such measurements should take this precision into account. In particular, I give advice here on monitoring an individual for a real change. Another practical application of reliability is in the assessment of competing brands of equipment (section 3.3).

In research settings, an important use of reliability is to estimate sample size for experimental studies. Reliability can also be used to estimate the magnitude of individual differences in the response to the treatments in such studies. I outline procedures for these 2 uses below.

2.1 Monitoring an Individual

In section 1.1, I argued that an observed change equal in magnitude to the limits of agreement was probably too large to use as a threshold for deciding that a real change has occurred. A more realistic threshold appears to be about 1.5 to 2.0 times the typical error (or a little more than half the limits of agreement), because the corresponding odds of a real change are between 6 and 12 to 1. For example, if an anthropometrist's typical error of measurement for the sum of 7 skinfolds is 1.6mm, an observed change of at least 2 to 3mm in an athlete's skinfolds would indicate that a real change was likely.

The value of the typical error to use in such situations needs to come from a short term or concurrent reliability study, in which there is no true change in the individuals' measurements between trials. For example, the typical error of measurement between skinfold assessments taken within 1 day would be appropriate for making decisions about changes in an individual over any time frame. In contrast, the typical error for use in estimation of sample size and individual differences in experiments needs to come from a reliability study of the same duration as the experiment.

2.2 Estimation of Sample Size

Most experiments consist of a pretest, a treatment and a post-test. The aim in these studies is to measure the change in the mean of a dependent variable between the pre- and post-tests. The typical error of the dependent variable represents noise that tends to obscure any change in the mean, so the magnitude of the typical error has a direct effect on the sample size needed to give a clear indication of the change in the mean.

In this section I develop formulae for estimating sample sizes from the typical error or retest correlation. The resulting sample sizes are often beyond the resources or inclination of researchers, but studies with smaller sample sizes nevertheless produce confidence limits that are more useful than nothing at all. These studies should therefore be published, perhaps designated as pilot studies, so they can be included in meta-analyses.

I advocate a new approach to sample size estimation, in which sample size is chosen to give adequate precision for an outcome.[2] Precision is defined by confidence limits: the range within which the true value of the outcome is 95% likely to occur. Adequate precision means that the outcome has no substantial change in impact on an individual volunteer over the range of values represented by the confidence limits. Let us apply this approach to an experiment.

For a crossover or simple test-retest experiment without a control group, basic statistical theory predicts confidence limits of $\pm t_{0.975,n-1} \cdot s \cdot \sqrt{2}/\sqrt{n}$ for a change in the mean, where n is the sample size, s is the typical error and t is the t statistic. Equating this expression to the value of the confidence limits representing adequate precision, ±d say, and rearranging:

$$n = 2(t \cdot s/d)^2 \approx 8s^2/d^2 \qquad \text{(Eq. 3)}$$

The fact that sample size is proportional to the square of the typical error in this formula underscores the importance of high reliability in experimental research. For example, when the typical error of the test has the same magnitude as the smallest worthwhile effect ($s = d$), a sample of about 8 volunteers (more precisely 10) gives adequate precision in a simple experiment; a test with twice the typical error entails a study with about 4 times as many participants. This formula is easily adapted to more complex designs. For example, sample size for a study with participants equally divided between an experimental group and a control group is 4n, or $32s^2/d^2$.

Choosing the value for d depends on the nature of the outcome variable and the participants. In research on factors affecting athletic performance, d is about half the typical error of an athlete's performance between races.[2] The resulting sample sizes can be very large. For example, if race performance has half the typical error as performance in a laboratory test, a study with a control group needs a sample size of $n = 32s^2/((s/2)/2)^2 = 512$ to delimit the smallest worthwhile effect on performance.

When interest centres on experiments involving the average person in a population, Cohen[13] argued that clinical judgement should be guided by the spread of raw scores (not change scores) in the population, and suggested that the smallest worthwhile value of d is 0.2 of the between-subject standard deviation. Thus, $0.2S = d = t_{0.975,n-1} \cdot s \cdot \sqrt{2}/\sqrt{n}$, so $n = 50(t \cdot s/S)^2$. But $(s/S)^2 = 1 - r$, where r is the retest correlation, so:

$$n = 50t^2(1 - r) \approx 200(1 - r) \qquad \text{(Eq. 4)}$$

Total sample size for a study with a control group is again 4n, or $800(1 - r)$. The profound effect of reliability on sample size is again apparent: the sample size dwindles to a few individuals for a retest correlation that is nearly perfect, whereas the sample size is about 200 (800 with a control group) when the retest correlation is zero.

In the above estimate of sample size, the between-subject standard deviation, S, is made up of true between-subject variation (S_T) and an independent concurrent error of measurement (e), such that $S^2 = S_T{}^2 + e^2$. Ideally, we should consider the smallest worthwhile effect as a fraction of S_T rather than of S, so the smallest worthwhile effect should be written as $0.2S_T = 0.2\sqrt{(S^2 - e^2)}$. If e is the same as the typical error, s, it is easy to show from this equation that the sample size needs to be increased by a factor of 1/r. This factor has little effect on sample size for high retest correlations, but sample size tends to infinity as r tends to zero.

The concurrent error, e, may be different from the within-subject standard deviation, s. For example, in a 1-month study of skinfold thickness, s is the error variation between an individual's measurements separated by 1 month, but e is the error variation between an individual's skinfolds measured within a short period (e.g. the same day). Thus, s includes variation due to real changes in skinfolds between individuals, but e is simply the error in the technique of measurement. In this situation, sample size needs to be increased by a factor of $1/r_c$, where r_c is the concurrent retest correlation, $(S^2 - e^2)/S^2$.

These formulae for sample size in studies of the average person in a population appear to show a primacy for retest correlation, but I must caution researchers that use of retest correlation is justified only if the sample in the reliability study is representative of the population in the experiment. In particular, it is wrong to use a retest correlation based on one population to estimate sample size in a study of a population with a different between-subject standard deviation. Most often there will be doubt about the applicability of the correlation from a published reliability study, so you should calculate sample size using, for example, $n = 50(t \cdot s/S)^2 \approx 200s^2/S^2$. Or, if you take concurrent reliability into account, $n \approx 200s^2/(S^2 - e^2)$. Reliability studies provide estimates of s and e; S comes either from a descriptive study of the population of interest or from a reliability study of a representative sample of the population.

Reliability has the same marked effect on sample size in the traditional approach to sample size estimation, which is usually based on 80% certainty of observing statistical significance ($p < 0.05$) for the smallest worthwhile effect. The resulting sample sizes are about twice as big as those estimated using my approach. For an example related to human performance tests, see Eliasziw et al.[14]

The foregoing formulae for estimating sample size are based on the value of the typical error in the experiment itself. Of course, we do not know that value until we have performed the experiment, so we use the value from a reliability study instead. If the typical error in the experiment differs from that in the reliability study, the estimate of sample size will be misleading. For example, the time between trials may differ between the reliability study and the experiment, and this difference may have a substantial effect on the typical error. Other reasons for differences in the typical error between the experiment and reliability study include differences in equipment, researchers, environment and characteristics of the volunteers. The researcher who wants to perform a reliability study to estimate sample size for a subsequent experiment has some control over these factors, but 2 more factors that can affect the typical error are beyond his or her control. First, the treatment in the experiment may produce responses that differ between study participants. These individual differences in the response show up as an increased error in the post-test, thereby increasing the overall typical error in the experiment. Secondly, evidence from a recent study suggests that blinding participants to the treatment may increase the variability of responses between participants, again resulting in an increase in the typical error.[15] Any estimate of sample size based on typical error in a reliability study must therefore be regarded as a minimum.

2.3 Estimation of Individual Differences

When the response to an experimental treatment differs between participants, we say that there are individual differences in the response. For example, a treatment might increase the power output of athletes by a mean of 3%, but the variation in the true enhancement between individual athletes might be a standard deviation of 2.5%. In this example, most athletes would show positive responses to the treatment, some athletes would show little or no response and some would even respond negatively. Note that this figure of 2.5% is not simply the standard deviation of the difference scores, which would include variation due to typical error. When I refer to individual differences, I mean variation in the true effect free of typical error. Although the primary aim in an experiment is to estimate the mean enhancement, it is obviously important to know whether the individual differences are substantial. Analysis of reliability offers one approach to this problem.

When individual differences are present, study participants show a greater variability in the post-pre difference score. Analysis of the experimental group as a reliability study therefore yields an estimate of the typical error inflated by individual differences. Comparison of this inflated typical error with the typical error of the control group or with the typical error from a reliability study permits estimation of the magnitude of the individual differences as a standard deviation, s_{ind} (2.5% in the above example). If the experiment consists of a pre-test, an intervention and a post-test, the estimate is readily derived from basic statistical principles as:

$$s_{ind} = \sqrt{}(2s^2{}_{expt} - 2s^2) \qquad \text{(Eq. 5)}$$

where s_{expt} is the inflated typical error in the experimental group, and s is the typical error in the control group or in a reliability study. For example, if the typical error in the experimental group is 2% and the typical error in the control group or in a reliability study is 1%, the standard deviation of the individual differences (s_{ind}) is $\sqrt{}6 = 2.5\%$. Estimation of individual differences is also possible with mixed modelling,[16] which can also generate confidence limits for the estimate.

When individual differences are present, the obvious next step is to identify the participant characteristics that predict the individual differences. The appropriate analysis is repeated-measures analysis of covariance, with the likely participant characteristics (e.g. age, gender, fitness, genotype) as covariates.[16]

3. Design and Analysis of Reliability Studies

A typical published reliability study consists of several trials performed on a sample of volunteers with 1 item of equipment and 1 operator of the equipment. The results of this simple kind of study meet the needs of most users of the test or equipment, provided the study has a sufficient number of participants and trials, and provided the analysis is appropriate. I will deal with design and analysis of such studies first, then discuss more complex studies.

3.1 Design of Simple Studies

The paramount concern in the design of any study is adequate precision for the estimates of the outcome measures. In a reliability study, the most important outcome measures are the typical error and the change in the mean between trials. The rationale for choosing a sample size that gives adequate precision for the estimate of systematic change in the mean presents a conundrum: the sample size must be the same as you would use in a simple experiment to delimit the smallest worthwhile effect of a treatment, but you cannot estimate that sample size without knowing the typical error. The researcher therefore has to base sample size for a reliability study solely on consideration of precision for the typical error.

Precision is defined, as usual, by the likely range (confidence limits) for the true value. Table II shows factors for computing the likely range of the typical error in reliability studies consisting of various numbers of participants and trials. Researchers can use this table to opt for a combination of trials and participants that gives an acceptable likely range for the typical error. The definition of ‘acceptable’ depends on the intended use of the typical error. Let us consider 2 common uses: estimation of sample size in an experiment and comparison of a new test with a published test.

Table II. Factors for generating the 95% likely range of the true value of a typical error from the value observed in a reliability study consisting of different numbers of participants and trials[a]

Participants	Trials			
	2	3	4	5
7	1.94	1.55	1.42	1.35
10	1.68	1.42	1.32	1.26
15	1.49	1.32	1.24	1.21
20	1.40	1.26	1.20	1.17
30	1.30	1.20	1.16	1.14
50	1.22	1.15	1.12	1.10

a Multiply and divide an observed typical error by the factor to generate the upper and lower Tate and Klett[17] 95% confidence limits for the true value. Data were generated with a spreadsheet.[18]

Suppose we opt for 15 participants and 4 trials, and the observed typical error is 1.0%. From table II, the resulting likely range for the true typical error is 1.0 × 1.24 to 1 ÷ 1.24, or 1.24 to 0.81. The likely range for the sample size in the experiment could therefore be overestimated by a factor of 1.54 (= 1.24^2) or underestimated by a factor of 0.65 (= 0.81^2). These limits represent a large difference in the resources needed for the study, so we must conclude that 15 participants with 4 trials is hardly adequate for estimating reliability. Fifty participants and 3 trials reduce the factors to 1.32 and 0.76, which represent a more acceptable risk of wasting or underestimating resources for the experiment.

To compare the typical error of a new test with a published typical error for another test, we need the precision of the published typical error, or preferably the sample size and number of trials in that study. We then calculate confidence limits for the comparison of the typical errors, using the F ratio. For simplicity, let us assume that we perform our study with the same sample size and number of trials as in the published study, and that we obtain the same typical error. For 15 participants and 4 trials, the confidence limits for the ratio of the typical errors is 0.74 to 1.36. In other words, the typical error for our test could be as low as 0.74 of the typical error for the published test (which would make ours a far better test), or it could be as high as 1.4 of the published test (which would make ours far worse). Once again, 15 participants and 4 trials are clearly inadequate. For 50 participants in 3 trials, the confidence limits for the ratio of the typical errors are 0.82 to 1.22, from which we could conclude tentatively that there is no substantial difference between the 2 tests. Of course, if our test gave a substantially lower or higher typical error than that of the published test, we could make a firmer conclusion about the relative reliabilities, possibly with fewer participants or trials.

A further important design consideration is the number of practice trials needed before the typical error settles into its lowest value. Addressing this problem requires a reasonably accurate estimate of changes in the typical error between consecutive pairs of trials. In unpublished simulations, I have found that a sample size of at least 50 gives adequate precision for the estimate of the change in typical error. Reliability studies in which 50 or more volunteers perform 3 or more trials are rare in the literature. It seems we must accept most published reliability studies as pilot studies.

3.2 Analysis of Simple Studies

Analysis of reliability studies is straightforward when there are only 2 trials. The typical error can be derived from the standard deviation of the difference scores for each participant, and the change in the mean is simply the mean of the difference scores. For 3 or more trials, I urge researchers to check for learning effects on the typical error by performing separate analyses on consecutive pairs of trials (trials 1+2, trials 2+3, etc.). You can download a spreadsheet for this purpose.[19]

Consecutive trials with similar typical errors can be analysed together to produce a single more precise estimate of typical error for those trials. Estimates of changes in the mean between these trials will also be a little more precise when derived from a single analysis of 3 or more such trials than when derived from consecutive pairs of trials. The appropriate analysis is a linear model with participants and trials as effects and with estimation by analysis of variance or by restricted maximum likelihood. The typical error is the residual error term in such analyses, regardless of whether participants and trials are fixed or random effects, but trials has to be a fixed effect for estimation of changes in the mean.

A one-way analysis of variance with participants as the effect produces an unsuitable estimate of typical error: in such an analysis the identity of the trial is ignored, so changes in the mean between trials add to the typical error. The resulting statistic is biased high and is hard to interpret, because the relative contributions of random error and changes in the mean are unknown. For example, with 2 trials and a change in the mean equal in magnitude to the typical error, I have found in simulations that this method yields a typical error inflated by a factor of 1.23. One-way analysis of variance is equivalent to calculating a separate variance for each participant from 2 or more trials, then averaging the variances and taking the square root. Authors who have used this equivalent method have usually committed a further mistake by averaging the participants' standard deviations instead of variances. In my simulations, averaging the standard deviations underestimates the typical error by a factor of 0.82 for 2 trials and 0.90 for 3 trials; the factor tends to 1.00 for a large number of trials. If the change in the mean between 2 tests is equal in magnitude to the typical error, the 2 mistakes virtually cancel each other out.

Having opted for an appropriate method of analysis, researchers should check their data for the presence of so-called heteroscedasticity. In the context of reliability or repeated-measures analyses, this term refers to a typical error that differs in some systematic way between participants. For example, participants with larger values of a variable often have larger typical errors, and typical errors for subgroups of participants (male *vs* female, competitive *vs* recreational, etc.) may also differ. Analysing the raw values of these measures with the usual statistical procedures is problematic, because the procedures are based on the assumption that the typical error is the same for every participant. If this assumption is violated, participants with the larger typical errors have a greater influence on the value of any derived statistic, and the value of the statistic may also be biased.

The generic method to check for heteroscedasticity is to examine plots of residual values versus predicted values provided by the analysis of variance or other statistical procedure used to estimate the reliability statistics. The residuals are the individual values of the random error for each participant for each trial; indeed, the standard deviation of the residuals is the typical error. With pairwise analysis of trials, a simple but equivalent method is to plot each participant's difference score against the mean for the 2 trials.[4] If the residuals for one group of participants are clearly different from another, or if the residuals or difference scores show a trend towards larger values for participants at one end of the plot, heteroscedasticity is present. The appropriate action in the case of groups with different residuals is to analyse the reliability of the groups separately. Variation in the magnitude of residuals with magnitude of the variable can be removed or reduced by an appropriate transformation of the variable.

As noted earlier, for many variables the typical error increases for volunteers with larger values of the variable, whereas the typical percentage error tends to be similar between volunteers. For these variables, analysis after logarithmic transformation addresses the problem of heteroscedasticity and provides an estimate of the typical percentage error. To see how, imagine that the typical percentage error is 5%, which means that the observed value for every volunteer is typically (1 ± 0.05) times the mean value for the volunteer. Therefore, $\log(\text{observed value}) = \log[(\text{mean value})(1 \pm 0.05)] = \log(\text{mean value}) + \log(1 \pm 0.05) \approx \log(\text{mean value}) \pm 0.05$, because $\log(1 \pm 0.05) \approx \pm 0.05$ for natural (base e) logarithms. The typical error in the log of every individual's value is therefore the same (0.05). You obtain the estimate of the typical percentage error of the original variable by multiplying the typical error of the log-transformed measure by 100. Alternatively, if you use 100log(observed value) as the transformation, the errors in the analyses are automatically approximate percentages, as are the magnitudes of changes in the mean in the analyses. The approximation is accurate for errors or changes less than 5%, but for larger errors or changes the typical percentage error or change is $100(e^{\text{err}/100} - 1)$, where err is the typical error or change in the mean provided by the analysis of the 100log-transformed measure.[20] There is also a special way to interpret errors > 5%. For example, if the error is 23%, the variation about the mean value is typically 1/1.23 to 1.23 times the mean value, or 0.81 to 1.23. The typical variation is not 1 ± 0.23 times the mean.

When a sample is homogeneous – that is, when all participants have similar values for the measure in question – the typical error is the same for all participants, regardless of transformation. In this situation, transformation to reduce heteroscedasticity is not an issue. Analysis of the log transformed variable is still a convenient method for obtaining the typical percentage error, although an equally accurate estimate is obtained by dividing the typical error (from an analysis of the raw variable) by the grand mean of all trials. Log transformation becomes more important as the sample becomes more heterogeneous, but I have found by simulation that estimates of typical percentage error from raw and log-transformed variables differ substantially (by a factor of 1.04 or more) only when the between-subject standard deviation is more than 35% of the mean. I doubt whether any variables in sports medicine and science show such large between-subject variation, so estimates of reliability derived from untransformed variables in previous studies are probably not substantially biased.

The estimate of the typical error for the average participant may be unbiased, but participants at either end of a heterogeneous sample who differ in the typical error before transformation may still differ in the typical percentage error after log transformation. For example, with increasing skinfold thickness the typical error increases but the typical percentage error decreases (Gore C, personal communication). A simple solution to this kind of problem is to rank-order participants, divide them into several groups, then compute the typical error or typical percentage error for each group. Alternatively, it may be possible to find a transformation that gives all participants the same typical error (absence of heteroscedasticity) for the transformed variable.

For researchers interested in retest correlation as a measure of reliability, the intraclass correlation coefficient derived from a mixed model (the

ICC(3,1)[9]) is unbiased for any sample size. Use of the intraclass correlation is also the only sensible approach to computing an average correlation between more than 2 trials. The usual Pearson correlation coefficient between a pair of trials is an adequate estimate of retest correlation, although it is biased slightly high for small samples: in simulations for 7 individuals, the bias is up to 0.04 units, depending on the value of the correlation.

Authors of many previous reliability studies have provided only a correlation coefficient as the measure of reliability. Nevertheless, it is usually possible to calculate the more useful typical error or typical percentage error from their data. By rearranging the relationship $r = (S^2 - s^2)/S^2$, we get the familiar:

$$s = S\sqrt{(1 - r)} \qquad \text{(Eq. 6)}$$

where s is the typical error, S is the average of the standard deviations for the participants in each trial and r is the intraclass correlation. The typical percentage error is obtained by dividing the resulting estimate of the typical error by the mean for the participants in all trials, then multiplying by 100. This formula is exact when r is the intraclass correlation, but even for a Pearson correlation my simulations show that the formula in surprisingly accurate: for samples of 10 or more participants the resulting typical percentage error is underestimated by a factor of 0.95 at most, but for samples of 7 the bias can be a factor of 0.90.

All estimates of reliability should be accompanied by confidence limits for the true value. Statistical programs usually provide confidence limits for the change in the mean, or you can use the formula in section 2.2. Confidence limits for the typical error are derived from the chi-squared distribution. For small degrees of freedom, the upper limit tends to be skewed out relative to the lower limit. Tate and Klett[17] provided an adjustment that reduces the skewness by minimising the width of the confidence interval, although it is then not an equal-probability interval. With only slight adjustment the Tate and Klett limits can be represented conveniently by a single factor (table II).

3.3 Complex Studies

The foregoing sections concern studies aimed at determining the reliability of 1 group of individuals with 1 type of test or equipment. In this section I deal with more complex studies: reliability of the mean of several trials; comparison of the reliability of 2 groups of individuals; comparison of 2 test protocols, items of equipment or operators of the equipment; and studies of continuously graded reliability.

Researchers sometimes improve the reliability of their measurements by using the mean of multiple trials: if there are n independent trials, the typical error of the mean is $1/\sqrt{n}$ times the error of a single trial. If the multiple trials are conducted over a short period (e.g. on the same day, without recalibration of equipment), but the researcher is interested in reliability of the mean over a longer period (e.g. on different days, with recalibration), the longer period is likely to be a source of substantial error. Therefore, beyond a certain number of multiple trials no substantial increase in reliability will be possible. To determine the number of trials, researchers need to perform a reliability study with multiple trials, estimate the magnitude of the error between trials over the shorter period (e_s) and over the longer period (e_l), then choose n such that $e_s/\sqrt{n} << e_l$. The most appropriate analysis is by repeated measures with 2 within-subject effects (same day, different day), each modelled with its own within-subject error. A statistically less challenging approach is as follows: analyse reliability of the trials on the same day to determine the trial number beyond which learning effects are negligible (e.g. trial 2); now compute between-day reliability for the mean of an increasing number of contiguous same-day trials (e.g. trials 3+4, trials 3+4+5. . .) to determine the number of same-day trials beyond which there is no further increase in between-day reliability.

Comparing the reliability of 2 groups of participants is straightforward. The participants are independent of each other, so any study amounts to 2 separate reliability studies. Confidence limits for the ratio of the typical errors between corresponding trials in the 2 groups can be derived from an F ratio. Changes in the mean between corresponding pairs of trials can be compared with unpaired t tests of the difference scores.

Comparing the reliability of 2 items (protocols, equipment or operators) is possible using the above approach for 2 groups of participants tested separately. Using the same participants has more power but requires analysis by an expert. Each participant performs at least 1 trial on 1 item of equipment and at least 2 trials on the other, preferably in a balanced, randomised fashion. The analysis needs a mixed model, in which the equipment is a fixed effect, trial number is a fixed effect, participants is a random effect, and a dummy random variable is introduced to account for the extra within-subject variance associated with measures on one of the items. Confidence limits for the extra variance address the question of the difference in typical error between the items. The model also provides an estimate of the difference in learning effects between the items.

When setting up a study to compare 2 items, keep in mind that the typical error always consists of biological variation arising from the individuals and technological variation arising from the items. Since the aim is to compare the technological variation, try to make the biological variation as small as possible, because it contributes to the uncertainty in your comparison of the items. For example, when comparing the reliability of 2 anthropometrists, you would get them to measure the same individuals on the same day, to avoid any substantial biological variation. Similarly, when comparing the reliability measures of power provided by 2 ergometers, use athletes as study participants, because they appear to be more reliable than non-athletes.

The problem of a continuous gradation of reliability arises when randomly chosen items or installations of the same kind of equipment produce consistently different values. For example, one item might always give high values, another might give low values and so on. Possible sources of these differences between items include inadequate quality control in manufacture, different environmental effects at the same or different locations, and differences in calibration or other aspects of operation by different operators. When a volunteer is retested on different items of the equipment, this variation between items adds to what would otherwise be the typical error for retests on the same apparatus, with the result that the overall typical error is higher. This typical error is the one that best represents the typical error in a one-off measurement taken on a randomly chosen item of equipment. It is also the one to use in the somewhat unusual situation of repeated trials when each trial is with a different item of equipment.

Researchers who are aware of the concept of lower reliability when retesting on different items or installations have usually computed a retest correlation rather than a typical error. The appropriate correlation is the intraclass correlation ICC(2,1) of Shrout and Fleiss.[9] It is derived from the so-called fully random model, in which the identity of the participants and trials are considered random effects. Researchers have often misapplied this model to data obtained from a single item of equipment. The resulting reliability is degraded by the learning effect, not by consistent differences in values between items of equipment. The only correct way to estimate the reliability between items of equipment is to test volunteers with a sufficient number of different items. The identity of the items is a random effect, and an extra fixed effect representing trial number is introduced in the analysis to account for learning effects. The typical error for a volunteer retested on different items is derived by adding the residual variance to the variance for the items. A similar analysis is appropriate when a number of different judges rate the performance of the same athletes at different competitions; in this case, the variance corresponding to judges needs to be divided by the number of judges before it is added to the residual variance to give the typical error variance for an athlete between competitions.

Unfortunately, even the 2-way random model with the addition of a fixed trial effect would still not account for the possibility that the magnitude of the typical error itself varies between items of equipment or between judges. As far as I know, no-one has developed a theoretical framework for quantifying such continuous variability in the typical error. It is not part of generalisability theory, which is another name for mixed modelling and which can deal only with the impact of random effects I discussed in the previous paragraph. Modelling continuous differences in reliability of subjects also seems to be impossible at present. Thus, the only way to model the better reliability that you find, for example, with faster athletes or more experienced operators, is to divide the volunteers or operators appropriately into a small number of groups, then compare the typical errors between groups.

4. Conclusion

The concept of the typical error in an individual's score should be comprehensible to most researchers and practitioners in sports medicine and science. I believe the concept is easier to grasp and to apply than limits of agreement. Change in the mean value of a measure between trials is also an important component of reliability, and it needs to be kept separate from typical error. Retest correlation is difficult to use, because its value is sensitive to the heterogeneity of the sample of participants. In my opinion, observed values and confidence limits of the typical error and changes in the mean are necessary and sufficient to characterise the reliability of a measure. Publication of these data in reliability studies would substantially enhance comparison of the reliability of tests, assays or equipment. Greater understanding of the theory of reliability by researchers would also help reduce the incidence of inappropriate analyses in the literature.

Acknowledgements

Chris Gore, John Hawley, Jenny Keating, Michael McMahon, Louis Passfield and Andy Stewart provided valuable feedback on drafts of this article.

References

1. Atkinson G, Nevill AM. Statistical methods for addressing measurement error (reliability) in variables relevant to sports medicine. Sports Med 1998; 26: 217-38
2. Hopkins WG, Hawley JA, Burke LM. Design and analysis of research on sport performance enhancement. Med Sci Sports Exerc 1999; 31: 472-85
3. Nevill AM, Atkinson G. Assessing agreement between measurements recorded on a ratio scale in sports medicine and sports science. Br J Sports Med 1997; 31: 314-8
4. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986 Feb; 8: 307-10
5. Roebroeck ME, Harlaar J, Lankhorst GJ. The application of generalizability theory to reliability assessment: an illustration using isometric force measurements. Phys Ther 1993; 73: 386-401
6. VanLeeuwen DM, Barnes MD, Pase M. Generalizability theory: a unified approach to assessing the dependability (reliability) of measurements in the health sciences. J Outcome Measures 1998; 2: 302-25
7. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psych Reports 1966; 19: 3-11
8. Kovaleski JE, Heitman RJ, Gurchiek LR, et al. Reliability and effects of leg dominance on lower extremity isokinetic force and work using the Closed Chain Rider System. J Sport Rehabil 1997; 6: 319-26
9. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psych Bull 1979; 86: 420-8
10. Kovaleski JE, Ingersoll CD, Knight KL, et al. Reliability of the BTE Dynatrac isotonic dynamometer. Isokinet Exerc Sci 1996; 6: 41-3
11. Hopkins WG. A new view of statistics. Available from: http://sportsci.org/resource/stats [Accessed 2000 Apr 18]
12. Hopkins WG, Manly BFJ. Errors in assigning grades based on tests of finite validity. Res Q Exerc Sport 1989; 60: 180-2
13. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Mahwah (NJ): Lawrence Erlbaum, 1988
14. Eliasziw M, Young SL, Woodbury MG, et al. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. Phys Ther 1994; 74: 777-88
15. Clark VR, Hopkins WG, Hawley JA, et al. Placebo effect of carbohydrate feedings during a 40-km cycling time trial. Med Sci Sports Exerc. In press
16. Hopkins WG, Wolfinger RD. Estimating 'individual differences' in the response to an experimental treatment [abstract]. Med Sci Sports Exerc 1998; 30 (5): S135
17. Tate RF, Klett GW. Optimal confidence intervals for the variance of a normal distribution. J Am Statist Assoc 1959; 54: 674-82
18. Hopkins WG. Generalizing to a population. Available from: http://sportsci.org/resource/stats/generalize.html [Accessed 2000 Apr 18]
19. Hopkins WG. Reliability: calculations and more. Available from: http://sportsci.org/resource/stats/relycalc.html [Accessed 2000 Apr 18]
20. Schabort EJ, Hopkins WG, Hawley JA, et al. High reliability of performance of well-trained rowers on a rowing ergometer. J Sports Sci 1999; 17: 627-32

Correspondence and offprints: Dr *Will G. Hopkins*, Department of Physiology, Medical School, University of Otago, Box 913, Dunedin, New Zealand.
E-mail: will.hopkins@otago.ac.nz