SPORTSCIENCE · sportsci.org Latest issue
Perspectives / Research Resources This issue

Spreadsheets for Analysis of Validity and Reliability

Will G Hopkins

Sportscience 19, 36-44, 2015 (sportsci.org/2015/ValidRely.htm)
Institute of Sport Exercise and Active Living, Victoria University, Melbourne, Australia, and High Performance Sport NZ, Auckland, New Zealand. Email. Reviewer: Alan M Batterham, Health and Social Care Institute, Teesside University, Middlesbrough, UK.

This article consists of explanations and links to updated validity and reliability spreadsheets that were previously available at this site as non-reviewed draft versions. The validity spreadsheet is based on simple linear regression to derive a calibration equation, standard error of the estimate and Pearson correlation linking one-off assessments of a practical measure to a criterion measure. For analysis of consistency of repeated measurements, three reliability spreadsheets are included in one workbook: *consecutive pairwise*, for performance tests or other measurements where habituation is an issue; *one-way*, where variable numbers of repeated measurements on subjects are all equivalent; and *two-way*, where the repeated measurements on subjects come from identified but randomly selected trials (games, raters, or similar sources) with no missing data. All three spreadsheets produce an estimate of within-subject error and an intraclass (effectively test-retest) correlation. The one- and two-way spreadsheets also produce estimates of observed and pure between-subject standard deviations, the two-way spreadsheet produces estimates of observed and pure between-trial standard deviations, and both produce estimates of error and correlations (including Cronbach's alpha) for means of any chosen number of trials. All spreadsheets include log transformation for analysis when the standard deviations expressed as factors or percents (coefficients of variation) apply more accurately to the full range of subjects. Instructions are also provided for use of SPSS to perform two-way mixed-model analyses that allow missing data and inclusion of fixed or random game, rater or other effects. KEYWORDS: intraclass correlation, typical error, standard error of the estimate, standard error of measurement, alpha reliability.

Reprint pdf · Reprint docx · Slideshow
Spreadsheets: Validity · Reliability

Update Nov 2015. Reviewers of reliability studies may want you to name the **type of intraclass correlation coefficient** (ICC) produced by the spreadsheets. In the terminology of Shrout and Fleiss (1979), the consecutive pairwise spreadsheet and the two-way spreadsheet produce the ICC(3,1), where the "3" refers to the type of ICC in which the subjects is a random effect and the trials is a fixed effect, while the "1" refers to the reliability of single repeated measurements (not the mean of several measurements). This ICC is the correlation expected between the pairs of measurements in any two trials, where all subjects have the same two trials. The one-way spreadsheet produces the ICC(1,1), where the first "1" designates a model in which subjects are random and trials are not included in the model at all. This ICC is the correlation expected between any two trials randomly selected for each subject. The one- and two-way spreadsheets also produce ICC(1,n) and ICC(3,n), which refer to the reliability of the mean of n trials. None of the spreadsheets produces the ICC(2,1) or ICC(2,n): these are correlations expected when the trials are considered to be random effects, and the pure between-trial variance is added to the pure between-subject variance to give an estimate of the between-subject variance for the calculation of the ICC. This kind of correlation has no immediate practical application; the ICC(3,1) is preferable, because it is the observed correlation between measurements in two real-life trials. In the calculation of the ICC(3,1) it does not matter whether trials are treated as a fixed or a random effect.

The terms ***intra-rater*** and ***inter-rater reliability*** also need explaining. When the trials are measurements taken by the same rater, referring to *intra-rater reliability* is sensible enough, but be aware of the possible sources of error. If the rater assessed the subjects' values without the subject repeating the movement or whatever (e.g., repeated assessment of videos of a movement), the typical error represents the error contributed only by the rater, and changes in the mean represent habituation of the rater, depending on the ordering of the subjects and trials. If the subjects repeated the movement for each trial (the usual scenario), then the typical error represents a combination of variability contributed by the subjects and the rater. You can't partition the error into the two sources, but that doesn't normally matter, because subjects always need a rater. Changes in the mean between the trials represent habituation of the subjects with possibly some habituation of the rater.

The term *inter*-rater reliability can be applied when the different trials represent assessments by different raters. If the measurements are taken simultaneously on a given subject by the different raters in real time or from a single movement on a video, the typical error represents the noise contributed to the measurement by raters only, averaged over the raters, and the differences in the means represent the different bias each rater brings to the party. *Inter-rater* then seems a reasonable term. The term seems less reasonable when each subject repeats the movement or whatever for each rater, because the typical error in the analysis is a combination of within-subject variability and the variability contributed by the raters, and differences in the means represent a mixture of habituation of the subjects and bias of the raters. If you randomize or balance the order in which the raters assess each subject, you can use a mixed model to partition out the habituation and bias effects. With mixed modeling and enough subjects, you can also partition the typical error into variability contributed by the subjects and by the raters (and even by each rater, with even more subjects). In these analyses you can treat the raters either as a fixed effect (in which case you get each rater's mean and comparisons of the means) or as a random effect (in which case you get the differences in the means expressed as a standard deviation).

Update Oct 2015. I have improved the flow of information in the slides on reliability. There is also a slide on a new use for reliability: explaining how error of measurement needs to be taken into account when estimating a smallest important difference or change defined by standardization.

The spreadsheets for analysis of validity and reliability were amongst the first published at the Sportscience site. Partly for this reason they were not accompanied by dedicated peer-reviewed articles that could be cited easily by researchers. The present article corrects that omission. The article is based on a slideshow previously published only as an in-brief item. I have updated the slideshow and included it in the PDF version of this article. I have also added two new reliability spreadsheets for analysis of straightforward repeated assessments when the consecutive pairwise approach of the existing spreadsheet is not appropriate. All three reliability spreadsheets are included in a single Excel workbook.

All spreadsheets include analysis of log transformation to properly estimate errors that are more likely to be similar across the range of values of the measurements when expressed in percent units (as a coefficient of variation) or as a factor standard deviation. Between-subject standard deviations are also estimated as percents or factors when log transformation is used.

Validity Spreadsheet

The spreadsheet is intended for analysis of concurrent validity, where the researcher wants to quantify the relationship between a practical and a criterion measure. The analysis is simple linear regression, in which the criterion is the dependent variable and the practical is the predictor. The analysis therefore results in a calibration equation that can be used to predict the criterion, given a value of the practical. The standard error of the estimate is the prediction error. The spreadsheet can be used for any simple linear regression

My colleagues and I used the regression approach for reviews of tests of cycling performance (Paton and Hopkins, 2001) and rowing performance (Smith and Hopkins, 2012). I have long eschewed the method-comparison approach promoted by Bland and Altman, as explained in other peer-reviewed articles at this site on bias in Bland-Altman analyses

(Hopkins, 2004) and a Socratic dialogue on what we're trying to achieve with a validity study, an estimate of the true value of something we've measured with a less-than-perfect instrument (Hopkins, 2010).

Reliability Spreadsheets

The original spreadsheet was designed primarily for analyzing the reproducibility of measurements in the kinds of setting common in sport and exercise science, where subjects are tested either on a regular basis for purposes of monitoring, or where a few repeated tests are performed for a controlled trial or crossover (Hopkins, 2000). In such settings "performance" in the test (the measured value) is likely to change between tests, owing to the effects of habituation (such as familiarization, practice, motivation, fatigue, or even the training effect of a single test). Habituation manifests itself in two ways: a change in the mean between tests and a change in the random error that contaminates every measurement. Analysis of the tests in a consecutive pairwise manner is therefore appropriate to allow you to follow the changes in the mean and the changes in the random error.

More rarely, you have at your disposal a number of repeated measurements on a sample of subjects, and the repeated measurements are all equal, in the sense that the error of measurement is expected to be the same for every measurement. Two new spreadsheets are provided to analyze such data. Both spreadsheets are shown with simulated data that change every time you open them or modify any cell. The spreadsheets were developed from one of those in the workbook with the article on understanding statistics with simulation (Hopkins, 2007). You replace the simulated data with your own.

In the one-way spreadsheet, there are no anticipated habituation effects. With such data all that's needed to estimate the error of measurement is a statistically sound way to average each subject's standard deviation. One-way analysis of variance provides an approach, and it also yields two between-subject standard deviations: the observed subject SD (what you would expect if you calculated the SD of a single measurement on each subject), and the true subject SD (the smaller SD you would expect if you could measure each subject without the random measurement error). The between- and within-subject SD are combined into an intraclass correlation coefficient, the correlation expected between a test and retest of the subjects. All these statistics are provided by the one-way spreadsheet, along with the smaller error of measurement and higher correlation you would expect if you used the mean of a given number of repeats as each subject's value.

In the two-way spreadsheet each test is assumed to have a different mean, as might occur when some performance indicator is measured in a sample of players in a series of games. The spreadsheet summarizes the different game means as an observed SD (the typical variation in the mean of the same sample of players from game to game) and a true SD (the typical variation from game to game, excluding the within-player SD [sic], or the SD you would expect to see if you had a very large sample of players). The intraclass correlation is again the correlation expected for subjects' values between any two tests. The changes in the mean between the tests have no effect on such a correlation.

Instructions for use of SPSS to do the one-way and two-way analyses are available in Zip-compressed file. See the In-brief item in this issue. You'll need a stats package to do the two-way analysis, if there are any missing data. See below.

Computational Issues

Unfortunately I have been unable to source formulae for computing the reliability statistics in the two-way spreadsheet when there are missing data. The ANOVA routine in Excel (available via File/Options/Add-Ins) also does not allow missing values, so you will have to use either a two-way analysis of variance or mixed modeling in a statistics package. (Warning: the mixed model in the package R does not currently estimate standard errors for random effects.) If you have lots of data for players without missing data, you could use the spreadsheet by first deleting those players with missing data.

Estimates for the correlation coefficient and its confidence limits in the one- and two-way spreadsheets come from a formula using the F statistic for subjects provided by Bartko (1966). The ICC shown in the pairwise spreadsheet is a close approximation based on deriving the observed between-subject SD by averaging the between-subject variances in the two tests; its confidence limits were estimated by converting it to an F ratio. For an exact ICC, use the two-

way spreadsheet. The estimates and confidence limits for the correlation with the mean do not work in the rare situation of negative values for the ICC of single measurements, so the correlation for the mean is shown as ~0.0 and the confidence limits are not computed.

The confidence limits for the pure between-subject SD are computed from an estimate of the standard error of the variance (derived from statistical first principles and checked against the estimates provided by a mixed model in SPSS). The pure between-subject variance or its confidence limits can be negative in some samples, owing to sampling variation, but in any case it is appropriate to assume that the sampling distribution of the variance is normal rather than chi-squared. Negative variance is then converted to a negative standard deviation (by changing the sign and taking the square root), as explained above for estimation of individual responses as a standard deviation (Hopkins, 2015). For more on this issue, see the current In-brief item and follow the link there for the full editorial.

I have as yet been unable to find a way to derive the confidence limits for the errors and correlations and correlations with different raters in the two-way analysis spreadsheets. I will update the spreadsheets and this article when I find a method that can be implemented readily in the spreadsheet.

References

Bartko JJ (1966). The intraclass correlation coefficient as a measure of reliability. Psychological Reports 19, 3-11

Hopkins WG (2000). Measures of reliability in sports medicine and science. Sports Medicine 30, 1-15

Hopkins WG (2004). Bias in Bland-Altman but not regression validity analyses. Sportscience 8, 42-46

Hopkins WG (2007). Understanding statistics by using spreadsheets to generate and analyze samples. Sportscience 11, 23-36

Hopkins WG (2010). A Socratic dialogue on comparison of measures. Sportscience 14, 15-21

Hopkins WG (2015). Individual responses made easy. Journal of Applied Physiology 118, 1444-1446

Paton CD, Hopkins WG (2001). Tests of cycling performance. Sports Medicine 31, 489-496

Shrout PE, Fleiss JL (1979). Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 86, 420-428

Smith TB, Hopkins WG (2012). Measures of rowing performance. Sports Medicine 42, 343-358

Published June 2015

Validity and Reliability

Will G Hopkins (will@clear.net.nz)
Institute of Sport Exercise and Active Living, Victoria University, Melbourne

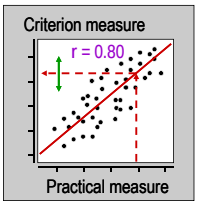
- Validity
  - Calibration equation, standard or typical error of the estimate, correlation
  - Bland and Altman's Limits of Agreement
  - Magnitude thresholds for the typical error and correlation
  - Uniformity of error and log transformation
  - Uses: calibration, correction for attenuation
- Reliability
  - Standard or typical error of measurement, (intraclass) correlation
  - Pairwise analyses; uniformity of error and log transformation
  - Magnitude thresholds for the typical error and correlation
  - Time between trials; 1- and 2-way analyses; mixed models
  - Uses: sample-size estimation, smallest effects, individual responses, monitoring
- Relationships Between Validity and Reliability
- Sample Sizes for Validity and Reliability Studies

Definitions

- Validity of a (practical) measure is some measure of its one-off association with another measure.
  - "How well does the measure measure what it's supposed to measure?"
  - Concurrent validity: the other measure is a criterion (gold-standard).
    - Example: performance test vs competition performance.
  - Convergent validity: the other measure ought to have some relationship.
    - Example: performance test vs competitive level.
  - Important for distinguishing between individuals.
- Reliability of a measure is some measure of its association with itself in repeated trials.
  - "How reproducible is the practical measure?"
  - Important for tracking changes within individuals.
- A measure with high validity must have high reliability.
- But a measure with high reliability can have low validity.

Validity

- We can often assume a measure is valid in itself…
  - …especially when there is no obvious criterion measure.
  - Examples from sport: tests of agility, repeated sprints, flexibility.
- If relationship with a criterion *is* an issue, the usual approach is to assay practical and criterion measures in 100 or so subjects.
  - Fitting a line or curve provides a calibration equation, a standard error of the estimate, and a correlation coefficient.
    - These apply only to subjects similar to those in the validity study.



- The standard (or typical) error of the estimate is a standard deviation representing the "noise" in a given predicted value of the criterion.
  - If the practical is being used to assess individuals, we should determine whether the noise (error) is negligible, small, moderate, and so on.

- To interpret the magnitude of a standard deviation, the usual magnitude thresholds for differences in means have to be halved (or you can double the SD before assessing it) (Smith & Hopkins, 2011).
- If the magnitude thresholds are provided by standardization, the smallest important difference in means is 0.2 × the between-subject SD.
- Therefore error <0.1×SD is negligible,
- This amount of error can be expressed as a correlation, using the relationship r^2 = "variance explained" = $(SD^2 - error^2)/SD^2$, where SD and error are those of the criterion.
  - Substituting error = 0.1×SD, gives r = 0.995, which can be defined as a very high validity correlation.
- The thresholds for small, moderate, large, very large and extremely large errors are half of 0.2, 0.6, 1.2, 2.0 and 4.0 × SD.
- Unfortunately the typical error of the estimate can never be greater than the observed SD, so this approach to interpretation of error yields only thresholds for moderate and large error (0.3 and 0.6 SD).
  - The corresponding correlations are 0.95 and 0.80.

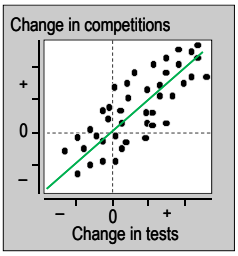
- So we need a new approach to interpret large magnitudes of the typical error and of corresponding low correlations. I'm working on it.
- If the criterion itself has random error, the criterion itself has a validity correlation with its error-free value. In this case it is possible to define very large and extremely large errors and corresponding low correlations.
  - The error thresholds are 0.1, 0.3, 0.6, 1.0 and 2.0 of the error-free SD, and the correlations are 0.995, 0.96, 0.86, 0.71, and 0.45.
- The usual thresholds for correlations representing effects in populations (0.90, 0.70, 0.50, 0.30, and 0.10) *are* appropriate to assess validity of a practical measure used to quantify mean effects in a population study.
- Uniformity of error is important. You want the estimate of error to apply to all subjects, regardless of their predicted value.
  - Check for non-uniformity in a plot of residuals vs predicteds, or just examine the scatter of points about the line.
  - Log transformation gives uniformity for many measures. Back-transform the error into a coefficient of variation (percent of predicted value).

- Some units of measurement can give spuriously high correlations.
  - Example: a practical measure of body fat in kg might have a high correlation with the criterion, but…
    Express fat as % of body mass and the correlation might be 0.00.
    So the practical measure effectively measures body mass, not body fat!
- Instead of a regression analysis you will often see a Bland-Altman plot of difference vs mean of the pairs of scores and limits of agreement.
  - This kind of analysis is limited to practical measures that are in the same units as the criterion.
  - The plot usually shows a downward trend suggesting proportional bias, but it's an artefact similar to regression to the mean (Hopkins, 2004).
  - The limits of agreement are the mean difference ±1.96× the SD of the difference scores (or more exactly, ± a value of the t statistic).
    - The limits are intended to define a range within which the measures agree. Example: if the limits of agreement for B - A were -9 to 21, a score of 109 for A would agree with scores of 100 to 130 for B.
    - Smallest important differences are not included. Avoid this approach!

- Regression analysis are intended for studies where every subject has a different value of the criterion and practical, and the aim is to produce an unbiased estimate of the criterion from the practical.
  - However, an analysis of difference scores is appropriate in validity studies where there are only a few fixed values of the criterion.
  - Example: a study of the ability of GPS to track distance run around a running track and around an agility course.
    - The mean difference between GPS and the criterion (measured with a tape or wheel) is the mean bias in the GPS measure, and the SD of the the difference is the random error from run to run, equivalent to the typical error of the estimate.
- Uses of validity: "calibration" for single assessments.
  - The regression equation between the criterion and practical measures converts the practical into an unbiased estimate of the criterion.
  - The standard (typical) error of the estimate is the random error in the calibrated value.

- Uses of validity: adjustment of effects in studies involving the practical measure ("correction for attenuation").
  - If the effect is a correlation, it is attenuated by a factor equal to the validity correlation.
  - If the effect is slope or a difference or change in the mean, it is attenuated by a factor equal to the square of the validity correlation.
- BEWARE: these two uses apply only to subjects drawn from the population used for the validity study.
  - Otherwise the validity statistics themselves need adjustment.
  - I have developed as yet unpublished spreadsheets for this purpose, useful for a meta-analysis of validity of a given measure.
- Uses of validity: calibration for change scores.
  - Sport scientists are not usually interested in "one-off" assessments.
  - Instead, they want to know how changes in a fitness test predict or track changes in competitive performance.
  - Very little research has been done on this question…

- If the athletes are tested twice, it's a simple matter of the relationship between change scores in the test and change scores in competitions.
- With multiple tests, the relationship between changes in tests and changes in competitions is best investigated with mixed modeling.
  - The modeling produces an average within-athlete slope for converting changes in tests into changes in competitions.



Reliability

- Reliability is reproducibility of a measurement when you repeat the measurement.
- It's important for practitioners…
  - because you need good reproducibility to monitor small but practically important changes in an individual subject.
- It's crucial for researchers…
  - because you need good reproducibility to quantify such changes in controlled trials with samples of reasonable size.

- How do we quantify reliability?
  Easy to understand for one subject tested many times:

Subject	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6	Mean ± SD
Chris	72	76	74	79	79	77	76.2 ± 2.8

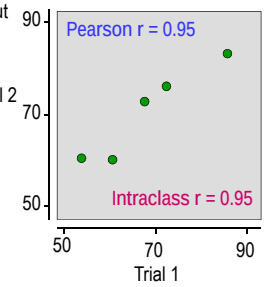
  - The 2.8 is the standard error of measurement.
  - I call it the typical error, because it's the typical difference between the subject's true value (the mean) and the observed values.
  - It's the random error or "noise" in our assessment of clients and in our experimental studies.
  - Strictly, this standard deviation of a subject's values is the total error of measurement rather than the standard or typical error.
    - It's inflated by any "systematic" changes, for example a learning effect between Trial 1 and Trial 2.
    - Avoid this way of calculating the typical error.

- We usually measure reliability with many subjects tested a few times:

Subject	Trial 1	Trial 2	Trial 2-1
Chris	72	76	4
Jo	53	58	5
Kelly	60	60	0
Pat	84	82	-2
Sam	67	73	6
	Mean ± SD:		2.6 ± 3.4

  - The 3.4 divided by √2 is the typical error (= 2.4).
  - The 2.6 is the change in the mean.
  - This way of calculating the typical error keeps it separate from the change in the mean between trials.
  - With more than two trials, analyze consecutive pairs of trials to determine if reliability stabilizes

- And we can define retest correlations: Pearson (for two trials) and intraclass (two or more trials).
  - These are calculated differently but have practically the same values.
  - The Pearson is biased slightly low with small sample sizes.
  - The ICC has slightly more bias.
- The typical error is more useful than the correlation coefficient for assessing changes in a subject.



- Important: reliability studies consist of more than five subjects!
  - And you need more than two trials to determine if there is substantial habituation in the form of changes in the mean and error between trials.
  - Analyze consecutive pairs of trials to address this issue.

- The reliability spreadsheet at Sportscience provides pairwise analyses of consecutive trials to properly assess familiarization (habituation).
  - Familiarization is common with performance tests.
  - Its effects are evident as substantial changes (improvements) in the mean and reductions in error between consecutive pairs of trials.
  - Two or more consecutive typical errors (and ICC) showing trivial changes can be averaged in the spreadsheet.
  - Typical error and changes in the mean are shown raw and standardized.
  - Scatterplots derived from pairs of trials allow assessment of uniformity of error in raw and log-transformed data.
    - Most measures have more uniform error with log transformation.
    - The spreadsheet also indicates which measures to log transform.
    - For such measures, use log transformation, even when it's not obvious in the plots.
  - Analysis of log-transformed data provides changes in the mean and typical error in percent, factor and standardized units.
    - Retest correlations are also performed with log-transformed data.

- As with validity, the standard (or typical) error of measurement is a standard deviation representing the "noise" in the measurement.
  - Interpret the magnitude of the typical error for assessing individuals by halving the usual magnitude thresholds for differences in means.
  - If the magnitude thresholds are provided by standardization, the thresholds are half of 0.20. 0.60, 1.2, 2.0 and 4.0.
  - These error thresholds can be expressed as correlations, using the relationship ICC = $SD_P^2/SD_O^2 = SD_P^2/(SD_P^2 + e^2)$, where SD_P is the pure or true (error-free) SD, SD_O is the observed SD, and e is the typical error.
  - Substituting $e = 0.1\times SD_P$, $0.3\times SD_P$, $0.6\times SD_P$, etc., the thresholds for extremely high, very high, high, moderate, and low reliability correlations are 0.99, 0.90, 0.75, 0.50, and 0.20.
  - These are less than the corresponding validity correlations but still much higher than the usual thresholds for population correlations.

- If the measure is competitive performance of solo athletes (e.g., time for 100-m run), can we assess its reliability?
  - For such athletes, magnitude thresholds for changes in the mean are given by 0.3, 0.9, 1.6, 2.5, and 4.0 × the within-athlete race-to-race SD.
  - So the thresholds for assessing the within-athlete SD itself as a measure of reliability are half these, or 0.15, 0.45, 0.8, 1.25 and 2.0.
  - The within-athlete SD is 1.0 on this scale, so competitive solo performance has a "large" error, regardless of the sport.
  - I have yet to develop a meaningful scale for interpreting the ICCs representing reproducibility of competition performance.
    - Smith & Hopkins (2011) produced a scale, but it is only for prediction of mean performance in one race by performance in another.
    - The thresholds are similar to the usual 0.90, 0.70, 0.50, 0.30, and 0.10 for population correlations.
    - A scale is needed that reflects the reproducibility of the ranking of athletes from one race to the next.

- Importance of time between trials…
  - In general, reliability is lower for longer time between trials.
  - When testing individuals, you need to know the noise of the test determined in a reliability study with a short time between trials, short enough for the subjects not to have changed substantially.
    - Exception: to assess an individual's change due specifically to, say, a 4-week intervention, you will need to know the 4-week noise.
  - For estimating sample sizes for research, you need to know the noise of the test with a similar time between trials as in your intended study.
  - A good reliability study investigates several times between trials.
    - Use a time sufficiently short for real changes in the subjects to be negligible but sufficiently long for dissipation of any transient fatigue or potentiation effects of the first trial.
    - A gap of up to one week between consecutive trials is desirable for physically demanding tests.
    - Include another cluster of trials weeks-months later for training studies.
  - But time between trials may not be an issue…

- Sometimes all trials are expected to have the same error. Examples:
  - Measurements of a performance indicator in the same player in different games. (The error may differ between playing positions.)
  - The individual Likert-scale items making up a dimension of the psyche in a questionnaire.
- For such measures, analysis of variance or mixed modeling provide better estimates of error and correlation.
- In a one-way analysis, the means of a sample of subjects are not expected to change on retesting–an unusual scenario.
- In a two-way analysis, the means of each trial are estimated, and their differences can be expressed as a standard deviation.
  - An analysis of two trials in this way is the same as a pairwise analysis.
- The Sportscience spreadsheet provides 1-way and 2-way analyses.
- Use mixed models for several sources of error arising from clustering of trials within different time points, equipment, raters, and/or items.
  - Judicious combinations of dummy variables, fixed effects and random effects provide a complete analysis of error structure.

- Uses of reliability: sample-size estimation for crossovers and controlled trials, when the dependent variable is continuous.
  - In a crossover, you calculate the change score for each subject between the intervention of interest and a control or reference treatment.
    - The effect of the intervention is the mean of the change scores.
    - Sample size is given by an acceptably narrow confidence interval (CI) for the mean change.
    - But CI = (t statistic)×(standard error of mean of change scores).
    - And standard error = (SD of change scores)/√(sample size).
    - And SD of change scores = $\sqrt{2}$ × typical error, assuming the intervention does not increase the error via individual responses.
    - Hence CI = t×$\sqrt{2}$×(typical error)/√(sample size).
    - So sample size is proportional to (typical error)^2.
    - (If there *are* individual responses, sample size may need to be bigger.)
  - In a controlled trial, the effect is the difference in the mean change in the intervention and control groups.
    - Sample size is still proportional to (typical error)^2, but ~4x as large.

- Uses of reliability: improved estimation of the smallest important difference or change defined by standardization.
  - 0.2 of the reference-group, control or baseline between-subject standard deviation provides a default value for a difference in a mean between groups or a change in a mean in a crossover or controlled trial.
  - But the typical error (e) makes the *observed* standard deviation (SD_O) greater than the *pure* standard deviation (SD_P): $SD_O^2 = SD_P^2 + e^2$.
  - And the smallest effect should obviously be defined by the most precise measurement, so $0.2 \times SD_P$ should be used, not $0.2 \times SD_O$.
    - To estimate SD_P, use $SD_P = \sqrt{(SD_O^2 - e^2)}$ or $SD_P = SD_O\sqrt{ICC}$, where ICC is the intraclass or retest correlation = SD_P^2/SD_O^2.
  - For right-now comparisons, the error or correlation should represent only technical error in the measurement (often negligible).
    - Example: performance indicators of team-sport athletes.
  - Include within-subject variability in estimation of SD_P for a given time between trials, if "stable" differences (differences between subject means) over the given time are important.
    - Example: health indicators in population studies.

- Uses of reliability: quantifying individual responses in controlled trials.
  - This "use" is really more about understanding the role of measurement error in individual responses.
  - The control group in a controlled trial is nothing more than a reliability study with two trials: one before and one after a control treatment.
  - You could analyze the two trials to get the change in the mean (expected to be trivial) and the typical error.
  - You could also analyze the intervention group to get the change in the mean (expected to show an effect) and the typical error.
  - If there are individual responses to the treatment, there is more error in the second trial, which shows up as a larger typical error.
  - This extra error represents the individual responses.
  - It can be estimated as an SD by taking the square root of the difference in the squares of the SD of the change scores.
  - To get individual responses in crossovers, you need an extra trial for the control treatment, or a separate comparable reliability study to give a standard deviation of change scores in the control condition.

- Uses of reliability: monitoring change in an individual…
  - Think about ± twice the typical error as the noise or uncertainty in the change you have just measured, and take into account the smallest important change.
  - Example: observed change = 1.0%, smallest important change = 0.5%.
    - The observed change is beneficial, but if the typical error is 2.0%, the uncertainty in the change is 1 ± 4%, or -3% to 5%.
    - So the *real* change could be quite harmful through quite beneficial.
    - So you can't be confident about the true change.
    - But if the typical error is only 0.5%, your uncertainty in the change is 1.0 ± 1.0%, or 0.0% to 2.0%.
    - So you can be reasonably confident that the change is important.
  - Conclusion: ideally, you want typical error << smallest change.
    - If typical error > smallest change, try to find a better test.
    - Or repeat the test several times and average the scores to reduce the noise. (Four tests halves the noise.)
  - The spreadsheet Assessing an individual gives chances of real change.

Relationships Between Validity and Reliability

- Short-term reliability sets an upper limit on validity. Examples:
  - If reliability error = 1%, validity error ≥ 1%.
  - If reliability correlation = 0.90, validity correlation ≤ √0.90 (= 0.95).
- Reliability of Likert-scale items in questionnaires
  - Psychologists average similar items in questionnaires to get a factor: a dimension of attitude or behavior.
  - The items making up a factor can be analyzed like a reliability study.
  - But psychologists also report alpha reliability (Cronbach's α).
    - The alpha is the reliability correlation you would expect to see for the mean of the items, if you could somehow sample another set of similar items.
    - As such, alpha is a measure of consistency of the mean of the items, not the test-retest reliability of the factor.
    - But √(alpha) is still the upper limit for the validity of the factor.

Sample Sizes for Validity and Reliability Studies

- As with all studies, the larger the expected effect, the smaller the sample size needs to be.
- Validity studies
  - n = 10-20 of given type of subject for very high validity;
  - n = 50-100 or more for more modest validity.
- Reliability studies
  - n is similar to that for validity studies, but how many trials are needed?
  - For laboratory or field tests, plan for at least four trials to properly assess familiarization (habituation) effects.
    - Such effects usually result in changes in the mean and error of measurement between consecutive trials.
    - Estimation of error requires analysis of a pair of trials.
    - Therefore error for Trials 2 & 3, if smaller than for 1 & 2, needs comparison with 3 & 4 to check for any further reduction.

This slideshow is available via the Validity and Reliability link at sportsci.org.

References

- My spreadsheets for analysis of validity and reliability. See links at sportsci.org.
- Hopkins WG (2000). Measures of reliability in sports medicine and science. Sports Medicine 30, 1-15.
- Paton CD, Hopkins WG (2001). Tests of cycling performance. Sports Medicine 31, 489-496.
- Hopkins WG (2004). How to interpret changes in an athletic performance test. Sportscience 8, 1-7. See link at sportsci.org.
- Hopkins WG (2004). Bias in Bland-Altman but not regression validity analyses. Sportscience 8, 42-46.
- Hopkins WG (2008). Research designs: choosing and fine-tuning a design for your study. Sportscience 12, 12-21, 2008. See link at sportsci.org.
- Hopkins WG (2010). A Socratic dialogue on comparison of measures. Sportscience 14, 15-21. See link at sportsci.org.
- Smith TB, Hopkins WG (2011). Variability and predictability of finals times of elite rowers. Medicine and Science in Sports and Exercise 43, 2155-2160.
- Hinckson, EA, Hopkins, WG, Aminian S, Ross K. (2013). Week-to-week differences of children's habitual activity and postural allocation as measured by the ActivPAL monitor. Gait and Posture 38, 663-667.