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 · Reliability two devices

Update Apr 2024. A new workbook of two spreadsheets is now available for analysis of reliability studies in which subjects are tested on each of two occasions with two devices (either two units of the same device, or two different devices). Simultaneous measurement with the two devices allows for separate estimation of random biological variability and technical error(s). Full instructions are included in the spreadsheets. See this In-brief item for the rationale and references.

Update Jan 2017. I have now provided a full scale for validity correlations, derived via standardized magnitudes of the standard error of the estimate, where the standardization is performed with the standard deviation of *predicted* values.

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