CHOOSING AN INTRACLASS CORRELATION COEFFICIENT

David P. Nichols
Principal Support Statistician and
Manager of Statistical Support
SPSS Inc.
From SPSS Keywords, Number 67, 1998

Beginning with Release 8.0, the SPSS RELIABILITY procedure offers an extensive set of options for estimation of intraclass correlation coefficients (ICCs). Though ICCs have applications in multiple contexts, their implementation in RELIABILITY is oriented toward the estimation of interrater reliability. The purpose of this article is to provide guidance in choosing among the various available ICCs (which are all discussed in McGraw & Wong, 1996). To request any of the available ICCs via the dialog boxes, specify Statistics->Scale->Reliability, click on the Statistics button, and check the Intraclass correlation coefficient checkbox.

In all situations to be considered, the structure of the data is as N cases or rows, which are the objects being measured, and k variables or columns, which denote the different measurements of the cases or objects. The cases or objects are assumed to be a random sample from a larger population, and the ICC estimates are based on mean squares obtained by applying analysis of variance (ANOVA) models to these data.

The first decision that must be made in order to select an appropriate ICC is whether the data are to be treated via a one way or a two way ANOVA model. In all situations, one systematic source of variance is associated with differences among objects measured. This object (or often, "person") factor is always treated as a random factor in the ANOVA model. The interpretation of the ICCs is as the proportion of relevant variance that is associated with differences among measured objects or persons. What variance is considered relevant depends on the particular model and definition of agreement used.

Suppose that the k ratings for each of the N persons have been produced by a subset of j > k raters, so that there is no way to associate each of the k variables with a particular rater. In this situation the one way random effects model is used, with each person representing a level of the random person factor. There is then no way to disentangle variability due to specific raters, interactions of raters with persons, and measurement error. All of these potential sources of variability are combined in the within person variability, which is effectively treated as error.

If there are exactly k raters who each rate all N persons, variability among the raters is generally treated as a second source of systematic variability. Raters or measures then becomes the second factor in a two way ANOVA model. If the k raters are a random sample from a larger population, the rater factor is considered random, and the two way random effects model is used. Otherwise, the rater factor is treated as a fixed factor, resulting in a two way mixed model. In the mixed model, inferences are confined to the particular set of raters used in the measurement process.

In the dialog boxes, when the Intraclass correlation coefficient checkbox is checked, a dropdown list is enabled that allows you to specify the appropriate model. If nothing further is specified, the default is the two way mixed model. If either of the two way models is selected, a second dropdown list is enabled, offering the option of defining agreement in terms of consistency or in terms of absolute agreement (if the one way model is selected, only measures of absolute agreement are available, as consistency measures are not defined). The default for two way models is to produce measures of consistency.

The difference between consistency and absolute agreement measures is defined in terms of how the systematic variability due to raters or measures is treated. If that variability is considered irrelevant, it is not included in the denominator of the estimated ICCs, and measures of consistency are produced. If systematic differences among levels of ratings are considered relevant, rater variability contributes to the denominators of the ICC estimates, and measures of absolute agreement are produced.

The dialog boxes thus offer five different combinations of options: 1) one way random model with measures of absolute agreement; 2) two way random model with measures of consistency; 3) two way random model with measures of absolute agreement; 4) two way mixed model with measures of consistency; 5) two way mixed model with measures of absolute agreement.
In addition, you can specify a coverage level for confidence intervals on the ICC estimates, and a test value for testing the null hypothesis that the population ICC is a given value.

Each of the five possible sets of output includes two different ICC estimates: one for the reliability of a single rating, and one for the reliability for the mean or sum of k ratings. The appropriate measure to use depends on whether you plan to rely on a single rating or a

combination of k ratings. Combining multiple ratings of course generally produces more reliable measurements.

Note that the numerical values produced for the two way models are identical for random and mixed models. However, the interpretations under the two models are different, as are the assumptions. Since treating the data matrix as a two way design leaves only one case per cell, there is no way to disentangle potential interactions among raters and persons from errors of measurement. The practical implications of this are that when raters are treated as fixed in the mixed model, the ICC estimates (for either consistency or absolute agreement) for the combination of k ratings require the assumption of no rater by person interactions. The estimates for the reliability of a single rating under the mixed model and all estimates under the random model are the same regardless of whether interactions are assumed. See McGraw & Wong for a discussion of the assumptions and interpretations of the estimates under the various models.

As a final note, though the ICCs are defined in terms of proportions of variance, it is possible for empirical estimates to be negative (the estimates all have upper bounds of 1, but no lower bounds). In the next issue, we will discuss the problem of negative reliability estimates.

Reference:

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. Psychological Methods, Vol. 1, No. 1, 30-46 (Correction, Vol. 1, No. 4, 390).