Enhancement of Reliability Analysis: Application of Intraclass Correlations with SPSS/Windows v.8

by Robert A. Yaffee
Statistics and Social Science Group
Academic Computing Facility
New York University
251 Mercer Street
New York, New York 10012
March 11, 1998

March 11, 1998

SPSS version 8 breaks new ground in providing for the computation of a collection of intraclass correlation coefficients. Intraclass correlations are correlations often used as reliability coefficients among evaluations of items that are deemed to be in the same category or class. They are ratios of between rating variance to total variance. They compare the covariance of the ratings with the total variance.

In the 1979 issue of Psychological Bulletin, Patrick E. Shrout, a professor in the graduate psychology department at NYU and Professor Joseph Fliess, of the Department of Biostatistics at Columbia University, wrote an article, entitled 'Intraclass Correlations: Uses in Assessing Rater Reliability." In that article, they presented guidelines according to which researchers can select the proper coefficient with which to assess rater reliability. When judges subjectively evaluate phenomena, measurement error is often found in their assessment. The careful and responsible researcher will assess this error before appling their ratings to the study of any targeted phenomena. To evaluate this measurement error, the researcher needs to be aware of the numerous available kinds of intra-class correlation coefficient, their relative advantages and disadvantages, and how they may be properly applied. As Shrout and Fleiss point out, the textbooks generally only present one or two forms of this coefficient and often do not discuss the appropriate applications for it.

Shrout and Fleiss present six kinds of intraclass correlation reliability coefficient, each of which is available in SPSS /Windows version 8. These coefficients are based on three types of models.

In Case 1, the targets are deemed a random sample of objects to be evaluated by the judges(raters). In this case, the difference of the individual judge's rating from the average rating of the judges for the ith target is the focus of interest. Therefore, the random targets happen to be the grouping variable. The judges provide the with-cell error variance for the ratings. Case 1 is therefore a one-way random effects model.

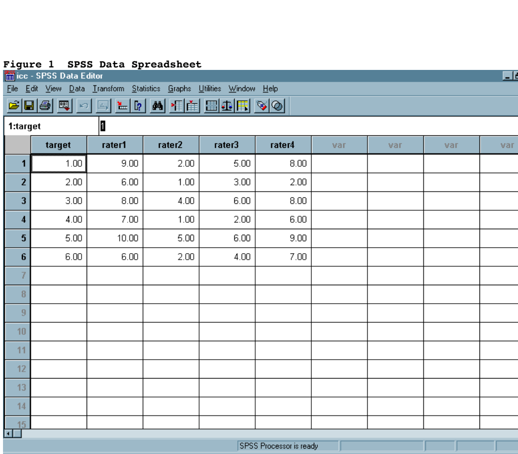
In Cases 2 and 3, the raters are deemed an important factor in the ICC computation. In Case 2, we have a two-way random effects model. Not only are the targets deemed random, but the judges are deemed a random effect as well. This derives from the fact that the judges are randomly selected from a larger population of judges who rate n targeted phenomena (which themselves are randomly selected from a larger pool of targets). There are three mean squares(variances) according to K. O. McGraw and S. P. Wong's 1996 "Forming inferences about some intraclass correlation coefficients" article in Psychological Methods, Vol.1,1, 30-46. There is the random effect of the targets, the random effect of the judges, and the residual effect as well. Case 2 is a two-way randomized block design.

In Case 3, the each target is evaluated by k raters, who are the only judges of interest. In this case, the judges are a fixed effect while the target ratings are a random effect. This is known as the two-way mixed model. For each case, there are two versions of reliability: One where the unit of analysis is the individual rating (which SPSS called single measure reliability) and the other where the unit of analysis is the mean of all the ratings (which SPSS calls average measure reliability).

Unlike earlier versions, SPSS version 8 now computes these six different types of Shrout and Fleiss intraclass reliability, but gives them different names. In addition, SPSS version 8 adds some new versions of intraclass correlation, culled from K.O. McGraw and S.P. Wong's 1996 "Forming inferences about some intraclass correlation coefficients" article in Psychological Methods, Vol.1,1, 30-46. to provide a total of ten of these correlation coefficients (Nichols, D, 1998). In their article, Shrout and Fleiss designate the intraclass reliability coefficients as ICC(case, expected unit of measurement version of reliability). The ICC for case 2 and for reliability of a single rating is called ICC(2,1), whereas the ICC for case 3 for the expected reliability of the mean of the k judges ratings is called ICC(3,k). If there are four judges, then this is ICC(3,4). Therefore, SPSS has its own names for the one-way random reliability measures. It calls ICC(1,1) one-way (random targets are the grouping variable) single measure reliability and ICC(1,k) one-way model single and average measure reliability, respectively. Where judges are a random sample, ICC(2,1) is called a two-way random effects model single measure reliability and ICC(2,k) is called a two-way random effects model average measure reliability. Where the judges used are the only one of interest, ICC(3,1) is designated two-way mixed effects model single measure reliability and ICC(3,k) is the two-way mixed effects model average measure reliability.

Shrout and Fleiss stipulate that the data set be analyzed in the following form. The variable names in the second row of the table are bolded and entered as variables in the data set. The variables are target, rater1, rater2, rater3, and rater4.

Figure 1 SPSS Data Spreadsheet



	target	rater1	rater2	rater3	rater4
1	1.00	9.00	2.00	5.00	8.00
2	2.00	6.00	1.00	3.00	2.00
3	3.00	8.00	4.00	6.00	8.00
4	4.00	7.00	1.00	2.00	6.00
5	5.00	10.00	5.00	6.00	9.00
6	6.00	6.00	2.00	4.00	7.00

From this data structure, Shrout and Fleiss show how a one or two-way anova can be constructed yielding the following variance decomposition:

Source of Variation	df	MS
Between targets	5	11.24
Within targets	18	6.26
Between judges	3	32.49
Error	15	1.02

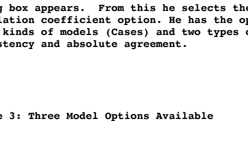
Dr. Robert M. Hamer of Virginia Commonwealth University has supplied the code for computing the Shrout and Fleiss intraclass coefficients.

```
bms= between target mean square
wms= within target mean square
jms= mean square for judges(raters)
ems=error mean square
k=number of judges

bms = ss/df for targets
msw=((ems*edf)+(jms*jdf))/(edf+jdf)
wms=msw
jms = ss/df for judges (raters)


sfsingle=(bms-wms)/(bms+(k-1)*wms)              * ICC(1,1)
sfrandom=(bms-ems)/
  ((bms)+((k-1)*ems)+((k*(jms-ems))/n))         * ICC(2,1)
sffixed=(bms-ems)/(bms+((k-1)*ems))             * ICC(3,1)
sfk=(bms-wms)/bms                               * ICC(1,k)
sfrandk=(bms-ems)/(bms+((jms-ems)/n))           * ICC(2,k)
sffixedk=(bms-ems)/bms                          * ICC(3,k)
                                With no interaction assumption
```

McGraw and Wong note that for each of these types of way models, the ICC consists of a ratio of mean squares. Although there is only one type for Case 1, there are at least two types of Cases 2 and 3, depending on the nature of the denominator of the ratio. The denominator is one of "Consistency," when the column (rater) variance is excluded from the denominator mean square, and it is one of "Absolute Agreement" when the judges (rater) variance is not excluded from the denominator. The rule of thumb, according to David Nichols, is that when the systematic variability due to raters is irrelevant, then the type of ICC used, is that of "Consistency", whereas if that variability is relevant, then "Absolute Agreement" is the type of ICC employed.

Invocation of the appropriate coefficients in SPSS is not difficult if the user is familiar with the nomenclature. In the reliability analysis, the only the rater variables are entered as variables to be analyzed. The analyst then selects in the statistics menu. A drop down menu appears and the user selects scale. Another menu appears and the user selects the the reliability analysis option. He then moves the rater1 through rater4 variables over into the items to be analyzed box.

Figure 2: Variables Loaded into Item analysis list



Then the user clicks on the statistics button and dialog box appears. From this he selects the intraclass correlation coefficient option. He has the options of the three kinds of models (Cases) and two types of reliability: consistency and absolute agreement.

Figure 3: Three Model Options Available

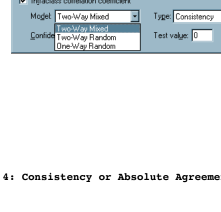
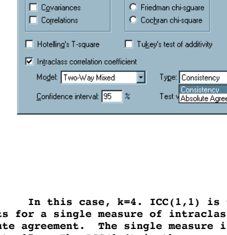


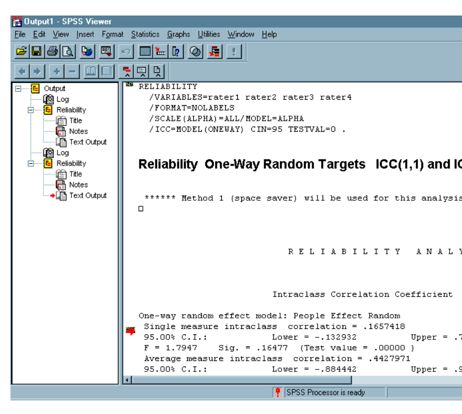
Figure 4: Consistency or Absolute Agreement Options



In this case, k=4. ICC(1,1) is the one-way random targets for a single measure of intraclass reliability with absolute agreement. The single measure is found to be equal to.17. The ICC(1,4) is the one-way random targets for an average measure with absolute agreement and it is equal to 0.44. For ICC(1,1) or ICC(1,k) absolute agreement is used.

Single measures are used for single measurements of the raters while average measurements apply when one is interested in the average rating for the k judges (raters).

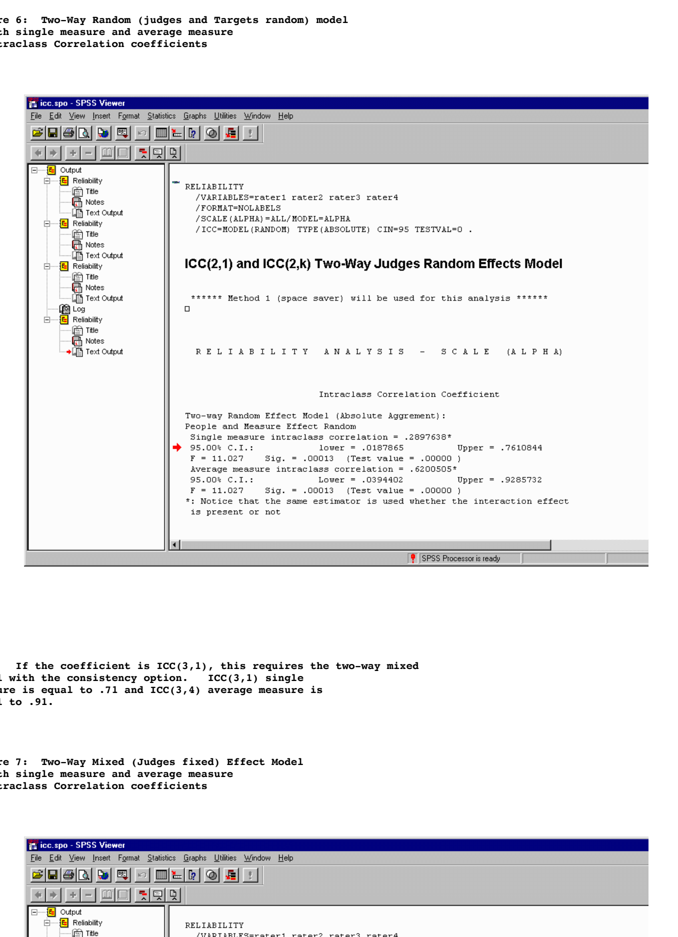
Figure 5: One-Way Random Targets models with single measure and average measure Intraclass Correlation coefficients



For Case 2 or Case 3 models, the user has the choice of either absolute agreement or consistency. He also has the choice of considering the raters as a fixed or random effect.

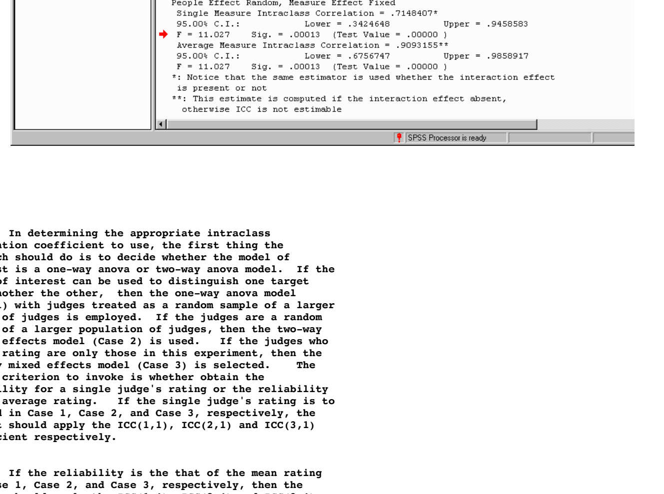
If the ICC to be analyzed is ICC(2,1), this is equal to .29 and ICC(2,4) is equal to .62. These coefficients require selection of the absolute agreement option.

Figure 6: Two-Way Random (judges and Targets random) model with single measure and average measure Intraclass Correlation coefficients



If the coefficient is ICC(3,1), this requires the two-way mixed model with the consistency option. ICC(3,1) single measure is equal to .71 and ICC(3,4) average measure is equal to .91.

Figure 7: Two-Way Mixed (Judges fixed) Effect Model with single measure and average measure Intraclass Correlation coefficients



In determining the appropriate intraclass correlation coefficient to use, the first thing the research should do is to decide whether the model of interest is a one-way anova or two-way anova model. If the model of interest can be used to distinguish one target from another the other, then the one-way anova model (Case 1) with judges treated as a random sample of a larger number of judges is employed. If the judges are a random sample of a larger population of judges, then the two-way random effects model (Case 2) is used. If the judges who do the rating are only those in this experiment, then the two-way mixed effects model (Case 3) is selected. The second criterion to invoke is whether obtain the reliability for a single judge's rating or the reliability of the average rating. If the single judge's rating is to be used in Case 1, Case 2, and Case 3, respectively, the analyst should apply the ICC(1,1), ICC(2,1) and ICC(3,1) coefficient respectively.

If the reliability is the that of the mean rating for Case 1, Case 2, and Case 3, respectively, then the analyst should apply the ICC(1,4), ICC(2,4) and ICC(3,4) respectively to this problem. In sum, if the analyst opts for a two-way random model, he may obtain either ICC(2,1) or ICC(2,4) by selecting the absolute agreement option. If the analyst opts for a two-way mixed effect model, he may obtain either ICC(3,1) or ICC(3,4) by selecting the absolute agreement option.

Number of Judges Required for Mean ICC ratings:

When a researcher decides that there is too much uncertainty to use an individual rating, he may decide to use a mean rating. When the mean rating of a number of judges is used, it is possible to ascertain the number of judges to be used.
The number of judges required should be determined from pilot study research. If NJ = number of judges required needed, and RL = lower bound from the (1-a)*100% confidence interval around the ICC, discovered in the pilot study findings, and ICC* is the minimum level of ICC acceptable --say, .75 or .80-- then

NJ = ICC*(1 - RL)/RL(1 - ICC*)

The four options not covered by Shrout and Fleiss can be found in McGraw and Wong article. Among the other forms of reliability coefficients under consideration for inclusion by SPSS now are the Cohen's non-symmetric Kappa and Cohen's multi-rater Kappa.

References:

Hamer, R.M. A SAS macro for computing intraclass correlation coeffients, Virginia Commonwealth University.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30-46.

Nichols, D. (1998). SPSS, Inc. (Personal communication), March 10, 1998.

Nichols, D. (1998). SPSS, Inc. "Choosing an Intraclass Correlation Coefficient", at http://www.utexas.edu/cc/faqs/stat/spss/spss4.html

Shrout, P.E. & Fleiss, J.L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability, *Psychological Bulletin*, Vol. 86, 2, 420-428.

Dr. Yaffee was a Senior Research/Statistical Consultant at the Information Technology Services (formerly the Academic Computing Facility) at NY.

This document has been accessed 134997 times.

last updated by RAY 6/05/03.