Intraclass Correlation

Andy P. Field

Volume 2, pp. 948–954

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Intraclass Correlation

Commonly used correlations such as the **Pearson product moment correlation** measure the bivariate relation between variables of different measurement classes. These are known as *interclass* correlations. By 'different measurement classes', we really just mean variables measuring different things. For example, we might look at the relation between attractiveness and career success; clearly one of these variables represents a class of measures of how good looking a person is, whereas the other represents the class of measurements of something quite different: how much someone achieves in their career. However, there are often cases in which it is interesting to look at relations between variables *within* classes of measurement. In its simplest form, we might compare only two variables. For example, we might be interested in whether anxiety runs in families, and we could look at this by measuring anxiety within pairs of twins (see [1]). In this case, the objects being measured are twins, and both twins are measured on some index of anxiety. As such, there is a pair of variables, which both measure anxiety and are, therefore, from the same class. In such cases, an intraclass correlation (ICC) is used and is commonly extended beyond just two variables to look at the consistency between judges. For example, in gymnastics, ice-skating, diving, and other Olympic sports, the contestant's performance is often assessed by a panel of judges. There might be 10 judges, all of whom rate performances out of 10; therefore, the resulting measures are from the same class (they measure the same thing). The objects being rated are the competitors. This again is a perfect scenario for an intraclass correlation.

Models of Intraclass Correlations

There are a variety of different intraclass correlations (see [4] and [5]) and the first step in calculating one is to determine a model for your sample data. All of the various forms of the intraclass correlation are based on estimates of mean variability from a one-way **repeated measures Analysis of Variance** (**ANOVA**).

All situations in which an intraclass correlation is desirable will involve multiple measures on different entities (be they twins, Olympic competitors, pictures, sea slugs etc.). The objects measured constitute a random factor (*see* **Fixed and Random Effects**) in the design (they are assumed to be random exemplars of the population of objects). The measures taken can be included as factors in the design if they have a meaningful order, or can be excluded if they are unordered as we shall now see.

One-way Random Effects Model

In the simplest case, we might have only two measures (refer to our twin study on anxiety). When the order of these variables is irrelevant (for example, with our twin study it is arbitrary whether we treat the data from the first twin as being anxiety measure 1 or anxiety measure 2), the only systematic source of variation is the random variable representing the different objects. In this case, the only systematic source of variation is the random variable representing the different objects. As such, we can use a one-way ANOVA of the form:

$$x_{ij} = \mu + r_i + e_{ij}, \tag{1}$$

in which r_i is the effect of object i (known as the *row effects*), j is the measure being considered, and e_{ij} is an error term (the residual effects). The row and residual effects are random, independent, and normally distributed. Because the effect of the measure is ignored, the resulting intraclass correlation is based on the overall effect of the objects being measured (the mean between-object variability MS_{Rows}) and the mean within-object variability (MS_{W}). Both of these will be formally defined later.

Two-way Random Effects Model

When the order of measures is important, then the effect of the measures becomes important also. The most common case of this is when measures come from different judges or raters. Hodgins and Makarchuk [3], for example, show two such uses; in their study they took multiple measures of the same class of behavior (gambling) and also measures from different sources. They measured gambling both in terms of days spent gambling and money spent gambling. Clearly these measures generate different data so it is important to which measure a datum belongs (it is not arbitrary to which measure a datum is assigned). This is one scenario in which a two-way model is used. However, they also took measures of gambling both from the gambler and a collateral (e.g., spouse). Again, it is important that we attribute data to the correct source. So, this is a second illustration of where a two-way model is useful. In such situations, the intraclass correlation can be used to check the consistency or agreement between measures or raters.

In this situation a two-way model can be used as follows:

$$x_{ij} = \mu + r_i + c_j + rc_{ij} + e_{ij}, \qquad (2)$$

where c_j is the effect of the measure (i.e., the effect of different raters, or different measures), and rc_{ij} is the interaction between the measures taken and the objects being measured. The effect of the measure (c_j) can be treated as either a fixed effect or a random effect. How it is treated does not affect the calculation of the intraclass correlation, but it does affect the interpretation (as we shall see). It is also possible to exclude the interaction term and use the model:

$$x_{ij} = \mu + r_i + c_j + e_{ij}. \qquad (3)$$

We shall now turn our attention to calculating the sources of variance needed to calculate the intraclass correlation.

Sources of Variance: An Example

Field [2] uses an example relating to student concerns about the consistency of marking between lecturers. It is common that lecturers obtain reputations for being ‘hard’ or ‘light’ markers, which can lead students to believe that their marks are not based solely on the intrinsic merit of the work, but can be influenced by who marked the work. To test this, we could calculate an intraclass correlation. First, we could submit the same eight essays to four different lecturers and record the mark they gave each essay. Table 1 shows the data, and you should note that it looks the same as a one-way repeated measures ANOVA in which the four lecturers represent four levels of an ‘independent variable’, and the outcome or dependent variable is the mark given (in fact, these data are used as an example of a one-way repeated measures ANOVA).

Three different sources of variance are needed to calculate an intraclass correlation. These sources of variance are the same as those calculated in one-way repeated measures ANOVA (see [2] for the identical set of calculations!).

The Between-object Variance (MS_{Rows})

The first source of variance is the variance between the objects being rated (in this case the between-essay variance). Essays will naturally vary in their quality for all sorts of reasons (the natural ability of the author, the time spent writing the essay, etc.). This variance is calculated by looking at the average mark for each essay and seeing how much it deviates from the average mark for all essays. These deviations are squared because some will be positive and others negative, and so would cancel out when summed. The squared errors for each essay are weighted by the number of values that contribute to the mean (in this case, the number of different markers, k). So, in general terms we write this as:

$$SS_{\text{Rows}} = \sum_{i=1}^{n} k_i(\bar{X}_{\text{Row}\,i} - \bar{X}_{\text{all rows}})^2. \qquad (4)$$

Table 1 Marks on eight essays by four lecturers

Essay	Dr Field	Dr Smith	Dr Scrote	Dr Death	Mean	S^2	$S^2(k-1)$
1	62	58	63	64	61.75	6.92	20.75
2	63	60	68	65	64.00	11.33	34.00
3	65	61	72	65	65.75	20.92	62.75
4	68	64	58	61	62.75	18.25	54.75
5	69	65	54	59	61.75	43.58	130.75
6	71	67	65	50	63.25	84.25	252.75
7	78	66	67	50	65.25	132.92	398.75
8	75	73	75	45	67.00	216.00	648.00
Mean:	**68.88**	**64.25**	**65.25**	**57.38**	**63.94**	**Total:**	**1602.50**

Or, for our example we could write it as:

$$SS_{\text{Essays}} = \sum_{i=1}^{n} k_i(\bar{X}_{\text{Essay }i} - \bar{X}_{\text{all essays}})^2. \tag{5}$$

This would give us:

$$\begin{aligned} SS_{\text{Rows}} &= 4(61.75 - 63.94)^2 + 4(64.00 - 63.94)^2 \\ &\quad + 4(65.75 - 63.94)^2 + 4(62.75 - 63.94)^2 \\ &\quad + 4(61.75 - 63.94)^2 \\ &\quad + 4(63.25 - 63.94)^2 + 4(65.25 - 63.94)^2 \\ &\quad + 4(67.00 - 63.94)^2 \\ &= 19.18 + 0.014 + 13.10 + 5.66 \\ &\quad + 19.18 + 1.90 + 6.86 + 37.45 \\ &= 103.34. \end{aligned} \tag{6}$$

This sum of squares is based on the total variability and so its size depends on how many objects (essays in this case) have been rated. Therefore, we convert this total to an average known as the *mean squared error* (*MS*) by dividing by the number of essays (or in general terms the number of rows) minus 1. This value is known as the *degrees of freedom.*

$$MS_{\text{Rows}} = \frac{SS_{\text{Rows}}}{\text{df}_{\text{Rows}}} = \frac{103.34}{n-1} = \frac{103.34}{7} = 14.76. \tag{7}$$

The mean squared error for the rows in Table 1 is our estimate of the natural variability between the objects being rated.

The Within-judge Variability (MS_{W})

The second variability in which we are interested is the variability within measures/judges. To calculate this, we look at the deviation of each judge from the average of all judges on a particular essay. We use an equation with the same structure as before, but for each essay separately:

$$SS_{\text{Essay}} = \sum_{k=1}^{p} (\bar{X}_{\text{Column }k} - \bar{X}_{\text{all columns}})^2. \tag{8}$$

For essay 1, for example, this would be:

$$\begin{aligned} SS_{\text{Essay}} &= (62 - 61.75)^2 + (58 - 61.75)^2 \\ &\quad + (63 - 61.75)^2 + (64 - 61.75)^2 = 20.75. \end{aligned} \tag{9}$$

The degrees of freedom for this calculation is again one less than the number of scores used in the calculation. In other words, it is the number of judges, k, minus 1.

We calculate this for each of the essays in turn and then add these values up to get the total variability within judges. An alternative way to do this is to use the variance within each essay. The equation mentioned above is equivalent to the variance for each essay multiplied by the number of values on which that variance is based (in this case the number of Judges, k) minus 1. As such we get:

$$\begin{aligned} SS_{\text{W}} &= s^2_{\text{essay1}}(k_1 - 1) + s^2_{\text{essay2}}(k_2 - 1) \\ &\quad + s^2_{\text{essay3}}(k_3 - 1) + \cdots + s^2_{\text{essay}n}(k_n - 1). \end{aligned} \tag{10}$$

Table 1 shows the values for each essay in the last column. When we sum these values we get 1602.50. As before, this value is a total and so depends on the number essays (and the number of judges). Therefore, we convert it to an average by dividing by the degrees of freedom. For each essay, we calculated a sum of squares that we saw was based on $k - 1$ degrees of freedom. Therefore, the degrees of freedom for the total within-judge variability are the sum of the degrees of freedom for each essay $\text{df}_{\text{W}} = n(k - 1)$, where n is the number of essays and k is the number of judges. In this case, it will be $8(4 - 1) = 24$.

The resulting mean squared error is, therefore:

$$MS_{\text{W}} = \frac{SS_{\text{W}}}{\text{df}_{\text{W}}} = \frac{1602.50}{n(k-1)} = \frac{1602.50}{24} = 66.77. \tag{11}$$

The Between-judge Variability (MS_{Columns})

The within-judge or within-measure variability is made up of two components. The first is the variability created by *differences* between judges. The second is the unexplained variability (error for want of a better word). The variability between judges is again calculated using a variant of the same equation that we have used all along, only this time we are interested in the deviation of each judge's mean from the mean of all judges:

$$SS_{\text{Columns}} = \sum_{k=1}^{p} n_i(\bar{X}_{\text{Column }i} - \bar{X}_{\text{all columns}})^2 \tag{12}$$

or

$$SS_{\text{Judges}} = \sum_{k=1}^{p} n_i(\bar{X}_{\text{Judge } i} - \bar{X}_{\text{all Judges}})^2, \tag{13}$$

where n is the number of things that each judge rated. For our data we would get:

$$\begin{aligned} SS_{\text{Columns}} &= 8(68.88 - 63.94)^2 + 8(64.25 - 63.94)^2 \\ &\quad + 8(65.25 - 63.94)^2 + 8(57.38 - 63.94)^2 \\ &= 554. \end{aligned} \tag{14}$$

The degrees of freedom for this effect are the number of judges, k, minus 1. As before, the sum of squares is converted to a mean squared error by dividing by the degrees of freedom:

$$MS_{\text{Columns}} = \frac{SS_{\text{Columns}}}{\text{df}_{\text{Columns}}} = \frac{554}{k-1} = \frac{554}{3} = 184.67. \tag{15}$$

The Error Variability (MS_{E})

The final variability is the variability that cannot be explained by known factors such as variability between essays or judges/measures. This can be easily calculated using subtraction because we know that the within-judges variability is made up of the between-judges variability and this error:

$$\begin{aligned} SS_{\text{W}} &= SS_{\text{Columns}} + SS_{\text{E}} \\ SS_{\text{E}} &= SS_{\text{W}} - SS_{\text{Columns}}. \end{aligned} \tag{16}$$

The same is true of the degrees of freedom:

$$\begin{aligned} \text{df}_{\text{W}} &= \text{df}_{\text{Columns}} + \text{df}_{\text{E}} \\ \text{df}_{\text{E}} &= \text{df}_{\text{W}} - \text{df}_{\text{Columns}}. \end{aligned} \tag{17}$$

So, for these data we obtain:

$$\begin{aligned} SS_{\text{E}} &= SS_{\text{W}} - SS_{\text{Columns}} \\ &= 1602.50 - 554 \\ &= 1048.50 \end{aligned} \tag{18}$$

and

$$\begin{aligned} \text{df}_{\text{E}} &= \text{df}_{\text{W}} - \text{df}_{\text{Columns}} \\ &= 24 - 3 \\ &= 21. \end{aligned} \tag{19}$$

The average error variance is obtained in the usual way:

$$MS_{\text{E}} = \frac{SS_{\text{E}}}{\text{df}_{\text{E}}} = \frac{1048.50}{21} = 49.93. \tag{20}$$

Calculating Intraclass Correlations

Having computed the necessary variance components, we shall now look at how the ICC is calculated. Before we do so, however, there are two important decisions to be made.

Single Measures or Average Measures?

So far we have talked about situations in which the measures we have used produce single values. However, it is possible that we might have measures that produce an average score. For example, we might get judges to rate paintings in a competition on the basis of style, content, originality, and technical skill. For each judge, their ratings are averaged. The end result is still the ratings from a set of judges, but these ratings are an average of many ratings. Intraclass correlations can be computed for such data, but the computation is somewhat different.

Consistency or Agreement?

The next decision involves whether we want a measure of overall consistency between measures/judges. The best way to explain this distinction is to return to our example of lecturers and essay marking. It is possible that particular lecturers are harsh (or lenient) in their ratings. A consistency definition views these differences as an irrelevant source of variance. As such the between-judge variability described above (MS_{Columns}) is ignored in the calculation (see Table 2). In ignoring this source of variance, we are getting a measure of whether judges agree about the relative merits of the essays without worrying about whether the judges anchor their marks around the same point. So, if all the judges agree that essay 1 is the best and essay 5 is the worst (or their rank order of essays is roughly the same), then agreement will be high: it does not matter that Dr. Field's marks are all 10% higher than Dr. Death's. This is a consistency definition of agreement.

Table 2 Intraclass correlation (ICC) equations and calculations

Model	Interpretation	Equation	ICC for example data
ICC for Single Scores			
One-way	Absolute agreement	$\frac{MS_{\mathrm{R}} - MS_{\mathrm{W}}}{MS_{\mathrm{R}} + (k-1)MS_{\mathrm{W}}}$	$\frac{14.76 - 66.77}{14.76 + (4-1)66.77} = -0.24$
Two-way	Consistency	$\frac{MS_{\mathrm{R}} - MS_{\mathrm{E}}}{MS_{\mathrm{R}} + (k-1)MS_{\mathrm{E}}}$	$\frac{14.76 - 49.93}{14.76 + (4-1)49.93} = -0.21$
	Absolute agreement	$\frac{MS_{\mathrm{R}} - MS_{\mathrm{E}}}{MS_{\mathrm{R}} + (k-1)MS_{\mathrm{E}} + \frac{k}{n}(MS_{\mathrm{C}} - MS_{\mathrm{E}})}$	$\frac{14.76 - 49.93}{14.76 + (4-1)49.93 + \frac{4}{8}(184.67 - 49.93)} = -0.15$
ICC for Average Scores			
One-way	Absolute agreement	$\frac{MS_{\mathrm{R}} - MS_{\mathrm{W}}}{MS_{\mathrm{R}}}$	$\frac{14.76 - 66.77}{14.76} = -3.52$
Two-way	Consistency	$\frac{MS_{\mathrm{R}} - MS_{\mathrm{E}}}{MS_{\mathrm{R}}}$	$\frac{14.76 - 49.93}{14.76} = -2.38$
	Absolute agreement	$\frac{MS_{\mathrm{R}} - MS_{\mathrm{E}}}{MS_{\mathrm{R}} + (MS_{\mathrm{C}} - MS_{\mathrm{E}}/n)}$	$\frac{14.76 - 49.93}{14.76 + (184.67 - 49.93/8)} = -1.11$

The alternative is to treat relative differences between judges as an important source of disagreement. That is, the between-judge variability described above (MS_{Columns}) is treated as an important source of variation and is included in the calculation (see Table 2). In this scenario, disagreements between the relative magnitude of judge's ratings matters (so, the fact that Dr Death's marks differ from Dr Field's will matter even if their rank order of marks is in agreement). This is an absolute agreement definition. By definition, the one-way model ignores the effect of the measures and so can have only this kind of interpretation.

Equations for ICCs

Table 2 shows the equations for calculating ICC on the basis of whether a one-way or two-way model is assumed and whether a consistency or absolute agreement definition is preferred. For illustrative purposes, the ICC is calculated in each case for the example used in this article. This should enable the reader to identify how to calculate the various sources of variance. In this table, MS_{Columns} is abbreviated to MS_{C}, and MS_{Rows} is abbreviated to MS_{R}.

Significance Testing

The calculated intraclass correlation can be tested against a value under the null hypothesis using a standard F test (*see* **Analysis of Variance**). McGraw and Wong [4] describe these tests for the various intraclass correlations we have discussed; Table 3 summarizes their work. In this table, ICC is the observed intraclass correlation whereas ρ_0 is the value of the intraclass correlation under the null hypothesis. That is, it is the value against which we wish to compare the observed intraclass correlation. So, replace this value with 0 to test the hypothesis that the observed ICC is greater than zero, but replace it with other values such as 0.1, 0.3, or 0.5 to test that the observed ICC is greater than known values of small, medium, and large-effect sizes respectively.

Fixed versus Random Effects

I mentioned earlier that the effect of the measure/judges can be conceptualized as a fixed or random effect. Although it makes no difference to the calculation, it does affect the interpretation. Essentially, this variable should be regarded as random when the judges or measures represent a sample of a larger population of measures or judges that could have been used. In other words, the particular judges or measures chosen are not important and do not change the research question that is being addressed. However, the effect of measures should be treated as fixed when changing one of

Table 3 Significance test for intraclass correlations (Adapted from McGraw, K.O. & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients, *Psychological Methods* **1**(1), 30–46.)

Model	Interpretation	*F-ratio*	*Df 1*	*Df 2*
ICC for Single Scores				
One-way	Absolute agreement	$\frac{MS_{\mathrm{R}}}{MS_{\mathrm{W}}} \times \frac{1-\rho_0}{1+(k-1)\rho_0}$	$n-1$	$n(k-1)$
Two-way	Consistency	$\frac{MS_{\mathrm{R}}}{MS_{\mathrm{E}}} \times \frac{1-\rho_0}{1+(k-1)\rho_0}$	$n-1$	$(n-1)(k-1)$
	Absolute agreement	$\frac{MS_{\mathrm{R}}}{aMS_{\mathrm{C}}+bMS_{\mathrm{E}}}$		
		In which;	$n-1$	$\frac{(aMS_{\mathrm{C}}+bMS_{\mathrm{E}})^2}{\frac{(aMS_{\mathrm{C}})^2}{k-1}+\frac{(bMS_{\mathrm{E}})^2}{(n-1)(k-1)}}$
		$a = \frac{k\rho_0}{n(1-\rho_0)}$		
		$b = 1 + \frac{k\rho_0(n-1)}{n(1-\rho_0)}$		
ICC for Average Scores				
One-way	Absolute agreement	$\frac{1-\rho_0}{1-ICC}$	$n-1$	$n(k-1)$
Two-way	Consistency	$\frac{1-\rho_0}{1-ICC}$	$n-1$	$(n-1)(k-1)$
	Absolute agreement	$\frac{MS_{\mathrm{R}}}{cMS_{\mathrm{C}}+dMS_{\mathrm{E}}}$		
		In which;	$n-1$	$\frac{(cMS_{\mathrm{C}}+dMS_{\mathrm{E}})^2}{\frac{(cMS_{\mathrm{C}})^2}{k-1}+\frac{(dMS_{\mathrm{E}})^2}{(n-1)(k-1)}}$
		$c = \frac{\rho_0}{n(1-\rho_0)}$		
		$b = 1 + \frac{\rho_0(n-1)}{n(1-\rho_0)}$		

the judges or measures would significantly affect the research question (*see* **Fixed and Random Effects**). For example, in the gambling study mentioned earlier it would make a difference if the ratings of the gambler were replaced: the fact the gamblers gave ratings was intrinsic to the research question being addressed (i.e., do gamblers give accurate information about their gambling?). However, in our example of lecturers' marks, it should not make any difference if we substitute one lecturer with a different one: we can still answer the same research question (i.e., do lecturers, in general, give inconsistent marks?). In terms of interpretation, when the effect of the measures is a random factor then the results can be generalized beyond the sample; however, when they are a fixed effect, any conclusions apply only to the sample on which the ICC is based (see [4]).

References

[1] Eley, T.C. & Stevenson, J. (1999). Using genetic analyses to clarify the distinction between depressive and anxious symptoms in children, *Journal of Abnormal Child Psychology* **27**(2), 105–114.

[2] Field, A.P. (2005). *Discovering Statistics Using SPSS*, 2nd Edition, Sage, London.

[3] Hodgins, D.C. & Makarchuk, K. (2003). Trusting problem gamblers: reliability and validity of self-reported gambling behavior, *Psychology of Addictive Behaviors* **17**(3), 244–248.

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

[5] Shrout, P.E.F. & Fleiss, J.L. (1979). Intraclass correlations: uses in assessing reliability, *Psychological Bulletin* **86**, 420–428.

ANDY P. FIELD