CHAPTER 8

Intraclass Correlations in One-Factor Studies

OBJECTIVE

The objective of this chapter is to present methods and techniques for calculating the intraclass correlation coefficient and associated precision measures in single-factor reliability studies based on model 1A or model 1B. I consider situations where the quantitative measurement is studied as a function of either the rater effect or the subject effect, but not both. Intraclass correlation is first defined as an abstract construct before the computation procedures are described. Methods for obtaining confidence intervals and p-values will be presented as well. I also discuss some approaches for calculating the optimal number of subjects and raters needed while planning an inter-rater reliability experiment.

CONTENTS

8.1 Intraclass Correlation under Model 1A

Let us consider a reliability experiment where r raters must each take m measurements (or replicate measurements or trial measurements) from n subjects. However, each subject could be rated by a different group of r raters. One could say with respect to Table 8.1 that $n = 6$, $r = 4$, and $m = 1$, since there are 6 subjects, 4 raters, and 1 replicate (i.e. there is a single measurement taken by the raters on each subject). Let y_{ijk} be the abstract representation of the quantitative score assigned to subject i by rater j on the k^{th} trial. The rater may change from subject to subject as stipulated in model 1A. The mathematical translation of this model is as follows:

$$y_{ijk} = \mu + s_i + e_{ijk}, \tag{8.1.1}$$

where μ is the expected score, s_i is subject i's effect, and e_{ijk} the error effect. Both effects are assumed to be random, independent[1] and to follow the Normal distribution with mean 0, and variances σ_s^2, and σ_e^2 respectively.

8.1.1 *Defining Inter-Rater Reliability*

The *Intraclass Correlation Coefficient* (ICC) needed to measure inter-rater reliability is by definition the correlation coefficient between the two quantitative scores y_{ijk} and $y_{ij'k}$ associated with the same subject i, and the same replicate number k, but with two raters j and j'. It follows from equation 8.1.1 that this particular correlation coefficient (denoted by ρ) is given by,

$$\boxed{\rho = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2}.} \tag{8.1.2}$$

Equation 8.1.2 provides the theoretical definition of ICC as the ratio of the subject variance to the total variance (i.e. the sum of subject and error variances) based on model 8.1.1. This ratio shows that the ICC will be high when the subject variance exceeds the error variance by a wide margin. This quantity indeed represents the extent of agreement among the r raters. To see this, you must first realize that the error variance σ_e^2 is actually the variance of two factors blended together, which are the rater factor and the error factor. However, the design represented by model 1A makes it impossible to separate them[2]. Therefore, a small error variance under model

[1] Independence is taken here in a statistical sense. That is the knowledge of the magnitude of one effect tells nothing about the magnitude of the other effect

[2] You would separate the rater and error variances only if each rater scores a whole set of subjects, which is not the case under model 1A

1A, actually means that both the error and rater variances have to be small. It is that small (and unknown) rater variance that ensures a small variation in the rater's scores and a high inter-rater reliability.

8.1.2 *Calculating Inter-Rater Reliability*

Shrout and Fleiss (1979) as well as McGraw and Wong (1996) presented ways to actually compute the intraclass correlation from raw data. Their methods are based on the use on various means of squares, and assume that your dataset is complete (i.e. does not contain missing values). This could be problematic in practice as missing values are common in many applications. However, the use of the means of squares is particularly useful for planning purposes, and will help determine the required sample sizes, and number of replicates prior to conducting the actual study (see section 8.3.3). This section focuses on computation methods needed to analyze data already collected, and that may contain missing values.

The approach that I present here is a simplified version of the methods described by Searle (1997, page 474). Let m_{ij} be the number of measurements (or replicates) associated with subject i and rater j. In the case of Table 8.1, $m_{ij} = 1$ for all subjects and all raters. If rater j does not score subject i then $m_{ij} = 0$, indicating that these ratings are missing. Let M be the total number measurements collected for the whole study (i.e. M is the summation of all m_{ij} values). For Table 8.1, $M = 6 \times 4 = 24$. Here are a few quantities that we are going to need:

- $m_{i\cdot}$ = number of measurements associated with subject i. In Table 8.1 there are 4 values associated with each subject since none is missing. That is $m_{1\cdot} = m_{2\cdot} = \cdots = m_{6\cdot} = 4$.
- $m_{\cdot j}$ = number of measurements associated with rater j. In Table 8.1, there are 6 values associated with each rater since none is missing. That is, $m_{\cdot 1} = m_{\cdot 2} = m_{\cdot 3} = m_{\cdot 4} = 6$.
- $y_{i\cdot\cdot}^2$ is the squared value of subject i's total score, and T_{2s} the sum of squares of all subject total scores (i.e. the sum of $y_{i\cdot\cdot}^2$ values).
- Let T_y be the total score (i.e. the summation of all y_{ijk} values), and T_{2y} the total sum of squares (i.e. the summation of all squared scores y_{ijk}^2).

In practice, the ICC of equation 8.1.2 can only be approximated using experimental data. This is done by calculating the two variance components from the raw experimental data. While the theoretical subject variance is σ_s^2, its calculated value is denoted by $\widehat{\sigma}_s^2$ (read sigma hat s square). Likewise, the calculated error variance is denoted by $\widehat{\sigma}_e^2$. The calculated intraclass correlation coefficient associated with model

1A is denoted by ICC(1A,1)[3] and given by,

$$\boxed{\text{ICC}(1\text{A},1) = \frac{\widehat{\sigma}_s^2}{\widehat{\sigma}_s^2 + \widehat{\sigma}_e^2},} \tag{8.1.3}$$

where,

$$\widehat{\sigma}_e^2 = (T_{2y} - T_{2s})/(M - n), \tag{8.1.4}$$

$$\widehat{\sigma}_s^2 = \frac{T_{2s} - T_y^2/M - (n-1)\widehat{\sigma}_e^2}{M - k_0}, \tag{8.1.5}$$

and k_0 is the expected number of measurements per subject and is calculated by summing all factors $m_{ij}^2/m_{\cdot j}$ over all n subjects and all raters. If there is no missing rating then $k_0 = rm$.

Example 8.1

To illustrate the calculation of ICC(1A,1), let us consider the data of Table 8.1 and assume that it was collected following the 1A design. Tables 8.1 and 8.2 show the different steps for calculating the intraclass correlation coefficients. Table 8.2 aims at showing the calculation of $k_0 = 4$, obtained by summing the last column. Columns 6, 7, 9, and 10 of Table 8.1 show the numbers $T_y = 127$, $M = 24$, $T_{2s} = 728.25$, and $T_{2y} = 841$. It follows from equations 8.1.4 and 8.1.5 that,

$$\widehat{\sigma}_e^2 = (841 - 728.25)/(24 - 6) = 6.264,$$
$$\widehat{\sigma}_s^2 = (728.25 - 127^2/24 - (6-1) \times 6.264)/(24 - 4) = 1.2444.$$

After plugging these variance components into equation 8.1.3, one obtains the intraclass correlation, $\text{ICC}(1A,1) = 1.2444/(1.2444 + 6.264) = 0.1657$.
Interested readers may download the Excel spreadsheet,

`www.agreestat.com/book4/chapter8examples.xlsx`

which shows the step-by-step calculation of this ICC.

The approach used in Example 8.1 for computing $\text{ICC}(1A,1)$ yields the exact same answer as the standard approach based on means of squares advocated by Shrout and Fleiss (1979), or McGraw and Wong (1996). Their standard approach however, can only work when there is no missing rating, and is given by,

$$\text{ICC}(1A,1) = \frac{\text{MSS} - \text{MSE}}{\text{MSS} + (rm-1)\text{MSE}}, \tag{8.1.6}$$

[3] "1A" in this notation indicates that ICC is based on model 1A, and the number 1 on the right side of the comma sign indicates that each rating used in the analysis represents 1 measurement, as opposed to being an average of several measurements. Some authors (e.g. Shrout & Fleiss, 1979) discussed the situation where the rating being analyzed is an average of k measurements. However, it is unclear how useful this scenario is in practice.

8.3.3 *Sample Size Calculations under Model 1A*

When designing an inter-rater reliability study, the researcher must decide how many raters and how many subjects to recruit. The general approach I recommend here applies to model 1A only and is similar to the method proposed by Giraudeau & Mary (2011). This technique consists of finding the optimal number of observations (i.e. the total number of measurements taken on all subjects) that minimizes the length of the ICC's confidence interval. Additional information on previous efforts in this area is provided by Doros and Lew (2010).

As previously mentioned, under model 1A the rater and replicate (or error) effects are confounded. That is, they are indistinguishable and only their combined effect can be evaluated. Consequently, when designing a study under model 1A, you only need to predict two numbers:

- The optimal number of measurements per subject. Let m designate that number.
- The optimal number of subjects. Let n be that number.

Once the optimal number of measurements per subject m is calculated, then the researcher will have to decide whether that number will represent the number of raters (for inter-rater reliability assessment) or the number of replicates (for intra-rater reliability assessment) or any combination of raters and replicates. However, under model 1A there is no benefit in using raters and replicates simultaneously. These effects are confounded and only their combined impact can be evaluated. Instead, one will use raters only to evaluate inter-rater reliability, and replicates only for intra-rater reliability. For example, if $m = 10$ then one would use 10 raters and 1 measurement per rater, or 1 rater and 10 replicate measurements. Using 5 raters and 2 replicates per rater for example will weaken both the inter-rater and the intra-rater reliability.

Figure 8.3.1 depicts the expected length of the 95% confidence interval as a function of the "true" value of the ICC, under the assumption that the total number of measurements from all subjects is 20 (i.e. $n \times m = 20$). The two continuous curves represent cases where the number of subjects is 4 for the black curve and 5 for the gray one. The two dotted lines on the other hand, represent experiments based on 10 and 2 subjects. The actual value of the ICC is generally unknown, and will have to be hypothesized by the researcher at the design stage. To be more concrete, you need to know whether the raters you are investigating have high agreement or low agreement in order to develop an effective design. The purpose of the study is to measure agreement ; yet we need some information about that same agreement in order to devise an effective design. What, if you do not have any preliminary information regarding the extent of agreement among the raters under investigation ?

The solution could be to conduct a pilot study with a small number of observations (i.e. approximately 10 observations), and use the crude estimate of the ICC obtained to finalize a more refined sample size calculation.

It appears from Figure 8.3.1 that the expected confidence interval length decreases as the ICC increases. That is, if the anticipated ICC is high, then the number of observations required to obtain a narrow confidence interval (i.e. a more accurate estimation) will generally be small. In other words, the study of raters known to have high agreement is cheaper than the study of raters known to have low agreement. This is logical. If raters have high agreement, ratings collected by a few raters on a few subjects will tell the whole story. Other raters who did not participate in the study, but who agree with the participants will not add any more information to the extent of agreement. But if a group of raters are known to have low agreement, then ratings collected by a few of them will reveal very little about the extent of agreement among all raters in that heterogeneous group. A sizeable number of raters and subjects must be recruited in order to obtain an effective study.

Figures 8.3.2 through 8.3.6 depict the relationship between the 95% confidence interval length and the "True" ICC for various values of the total number of measurements. Figure 8.3.2 for example is based 40 measurements, and the different curves on that figure represent different distributions of the 40 measurements between the number of subjects and the number raters. All 6 figures 8.3.1 to 8.3.6 show that for any given ICC value, having 4 or 5 raters (or 4 or 5 measurements per subject for intr-rater reliability studies) appears to minimize the confidence interval length. Readers interested to know how these graphs are created, are invited to read the last few paragraphs at the end of this section.

How to determine the sample size ?

One possible approach that I recommend for determining the optimal sample size is the following three-step process:

- Start with an anticipated value for the intraclass correlation. The situation where you have no prior information at all about the possible magnitude of the intraclass correlation is addressed in the next subsection.
- Set your desired confidence interval length to be approximately $0.8 \times$ ICC, where ICC is the predicted extent of agreement among raters of the previous step[12].
- Review the different graphs and the expected 95% confidence interval length that is associated with the anticipated ICC value. One curve should lead to

[12]This rule of thumb is derived from the classical confidence intervals for means. The length of a 95% confidence interval length is 4 times the standard deviation (STD). If the coefficient of variation (i.e. the ratio of the STD to the actual mean) is required to be 0.2 or less then the confidence interval length will have to be 0.8 times the true mean or less ($0.8 = 0.2 \times 4$)

a confidence interval length that is sufficiently close to our desired value of $0.8 \times$ ICC. The number of observations associated with that curve will be the optimal sample size. If no curve is found matching our desired confidence interval length, one should use the closest value available.

Suppose that you anticipate the ICC to be around 0.6, a value on which the sample size determination will be based. Since $0.8 \times 0.6 = 0.48$, you are then looking for a sample size that will yield a 95% confidence interval length that does not exceed 0.48. Figure 8.3.1 (based on 20 observations) indicates that for ICC=0.6, both curves show a confidence interval length that exceeds 0.7. Consequently, you will need more than 20 observations. In Figure 8.3.2, all curves yield an interval length that exceeds 0.5 for ICC=0.6. In Figure 8.3.3 however, it appears that for ICC=0.6 a few curves yield a confidence interval length close to 0.45. Now we know that we will need approximately 60 measurements overall.

Now that you will need 60 measurements, the question becomes how should you distribute them across subjects and raters? Figure 8.3.3 also shows that using 20 subjects and 3 raters will produce the smallest confidence interval length for $ICC = 0.6$. However, a second viable option would be to use 15 subjects and 4 raters. Figures 8.3.4, 8.3.5, and 8.3.6 indicate that one may obtain much shorter intervals by increasing the number of measurements. In most cases, using approximately 3 or 4 raters is sufficient to minimize the confidence interval length.

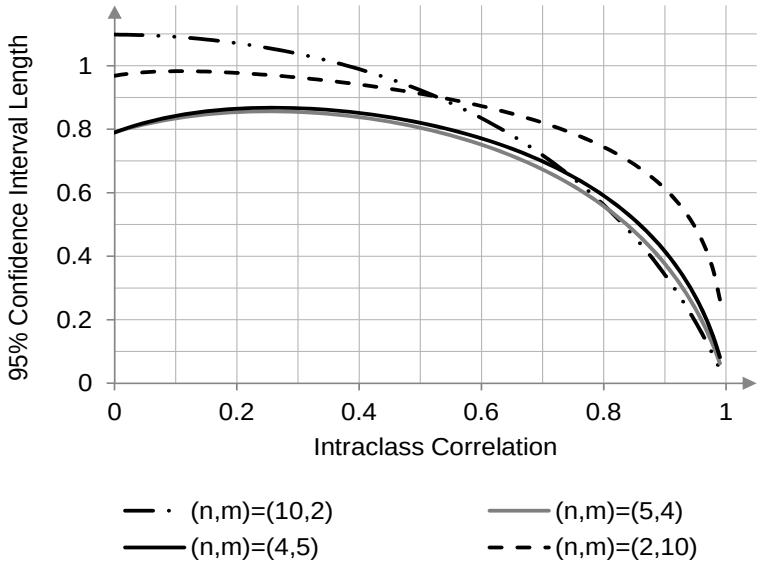

Figure 8.3.1: Expected width of the 95% confidence interval as a function of ICC for $n \times m = 20$ measurements.

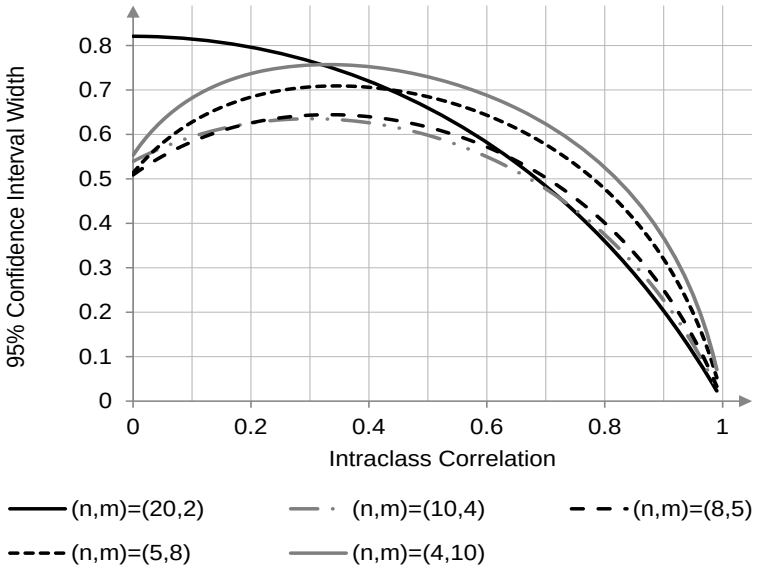

Figure 8.3.2: Expected width of the 95% confidence interval as a function of ICC for $n \times m = 40$ measurements.

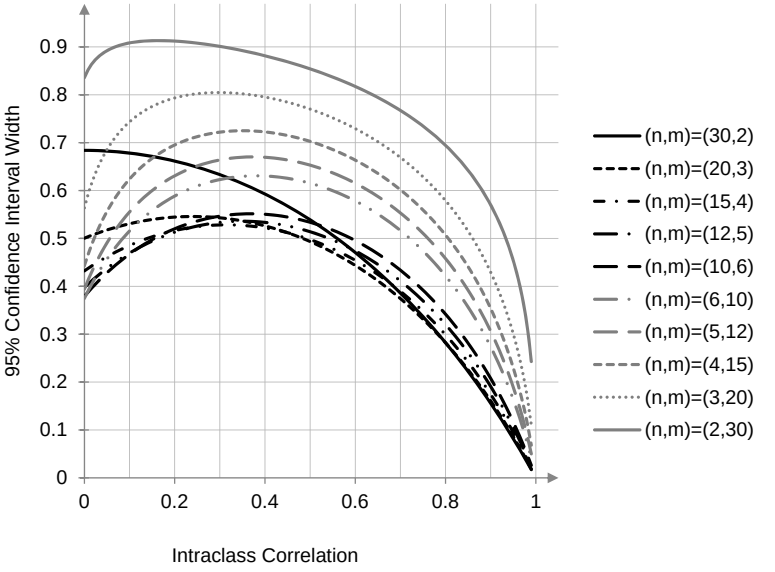

Figure 8.3.3: Expected width of the 95% confidence interval as a function of ICC for $n \times m = 60$ measurements.

How are figures 8.3.1 to 8.3.6 obtained ?

All the graphs displayed in figures 8.3.1 to 8.3.6 are obtained through a Monte-Carlo simulation. If ρ is the true value of the ICC, equation 8.3.3 shows that the ratio $F_s = \text{MSS/MSE}$ can be expressed as $F_s = F \times (1 + rm\rho/(1-\rho))$ where F is a random variable that follows the F-distribution with $n-1$ and $n(rm-1)$ degrees of freedom. To generate each curve of figure 8.3.1 for example, I proceeded as follows:

(1) I started by setting a value to n (the number of subjects) and another value to rm (the product of the number of raters by the number of replicates, which I labeled on the graph as m for simplicity) representing the total number of measurements per subject.

(2) Next, I assigned an ICC value to ρ.

(3) I then generate 100,000 random variates from the F-distribution with $n-1$ and $n(rm-1)$ degrees of freedom. These random variates allowed me to compute 100,000 F_s values as defined above.

(4) The F_s values were used as suggested in equations 8.3.1 and 8.3.2 to derive 100,000 confidence intervals and associated lengths. The average of these 100,000 interval lengths and the hypothesized value ρ produce one point on the curve.

(5) This process is repeated for each of 100 ρ values 0, 0.1, 0.2, until 0.99. For the sake of efficiency, I would generally create the confidence intervals associated with all 100 ρ values for each random variate generated. This allows me to obtain all 100 averages at once, generating thereby the 100 points needed to create the curve.

8.4 Concluding Remarks

In this chapter, I have presented two simple models the researcher can use to design an inter-rater or an intra-rater reliability study. These two models were designated as model 1A and model 1B. Model 1A describes the situation where each subject is rated by a different group of raters[14], while model 1B describes a situation when each rater rates a different group of subjects[15]. We learned that model 1A could be used to investigate both the inter-rater reliability and the intra-rater reliability, unlike model 1B that may be considered only for investigating intra-rater reliability.

Because each subject can be rated by a different group of raters under model 1A, its use presents the following advantages:

[14]There may be some overlap between the different groups of raters
[15]There may be some overlap between the different groups of subjects

- The subjects can be located in different places, and be rated by local raters.
- Even if the subjects are all located in the same place, the rating process itself can be spread over the course of a relatively long period of time without having to worry wether the same group of raters will always be available in its entirety. The subjects can always be rated by the group of raters that happens to be available.

However, the use of model 1A has some disadvantages that the researcher should be aware of. Although the measure of inter-rater reliability under model 1A is valid, its estimation using actual ratings will have a potentially high variance if each rater produces very few ratings. That is, under this model you will want to have some overlap between the different groups of raters to give each rater the opportunity to rate multiple subjects. The calculation of the intra-rater reliability coefficient under model 1A requires the use of a single rater. This will make it more sensitive to the particular rater used to generate the ratings.

Because each rater can rate a different group of subjects under model 1B, its use presents the following advantages:

- The raters can be located in different places and will have the opportunity to rate subjects recruited locally.
- Even if the raters are all located in the same place, the rating process itself can be spread over the course of a relatively long period of time without having to worry about the same group of subjects being available to each rater. Some subjects that may be unavailable to one rater can well be replaced with new ones, without the validity of model 1B being affected.

The use of model 1B has some disadvantages that should be mentioned. Evaluating inter-rater reliability under model 1B is an impossible task. It is because, different raters are not required to rate the same subjects making it impossible to assess the extent of agreement among them. As for the intra-rater reliability coefficient, it can be calculated under model 1B although the researcher will need to use a single subject, which will be rated multiple times by each rater. However the single subject being rated could be different for each rater. This will make the intra-rater reliability coefficient very sensitive to the particular subject the ratings are based upon.

If the disadvantages associated with each of the two models presented here are unacceptable to the researcher, then more elaborate models discussed in subsequent chapters must be considered. The new models explored in these chapters will describe more restrictive experimental designs, but will produce more accurate reliability coefficients.