CHAPTER 10

Intraclass Correlations under the Mixed Factorial Design

OBJECTIVE

This chapter aims at presenting methods for analyzing intraclass correlation coefficients for reliability studies based on a random sample of subjects and a fixed group of raters. Therefore, the rater factor is considered fixed, while the subject factor is considered random. The intraclass correlation coefficient (ICC) used for quantifying intra-rater reliability is a valid measure of reproducibility. However, the ICC used for quantifying inter-rater reliability is a valid measure of consistency, and would be a valid measure of agreement only if there is no systematic bias in the rating process from one rater. This chapter also discusses methods for obtaining confidence intervals and p-values for all types of ICCs under model 3, in addition to presenting a detailed account of the methods used to determine the optimal number of subjects during the experimental design.

CONTENTS

10.1 The Problem

In the previous chapter, I presented methods for computing the intraclass correlation coefficient as a measure of inter-rater or intra-rater reliability, under the random factorial design. The random factorial design treats both the subject and the rater effects as random, which is justified only when the subject and rater samples are randomly selected from larger subject and rater universes. However, the rater effect cannot be treated as random in certain types of reliability experiments. For example, the reliability experiment may use two measuring instruments (an existing one and a new one) to take measurements on subjects. This experiment involves two raters (i.e. the measuring instruments expected to produce comparable measurements), and no other rater is under consideration. The rater effect must be considered fixed in such a situation. A fixed rater effect combined with a random subject effect will lead to an experimental design known as the "Mixed factorial design."

There is a fundamental difference between the random and mixed factorial designs regarding the role the rater effect plays in data analysis. Unlike the random factorial model of the previous chapter, which proceeds with a direct evaluation of the rater variance, the mixed factorial model is essentially based on the analysis of the interaction between raters and subjects. The raters alone do not represent a source of uncertainty to be analyzed since their effect is fixed. The only source of uncertainty involving the raters relies on the subject-rater interaction. A large subject-rater interaction has a negative effect on inter-rater reliability ; but may have a positive effect on intra-rater reliability for a given total variation in the ratings. Note that the subject-rater interaction is high if the score difference between raters varies considerably from subject to subject, and is low otherwise. I will discuss these relationships further in subsequent sections.

If the raters strongly agree then the subject-rater interaction will be small and the mixed factorial design will rightfully yield a high intraclass correlation coefficient. However, a small subject-rater interaction or its complete absence in a well-designed experiment could well yield a high intraclass correlation without the raters being in high agreement. This typically occurs in situations where there is a large and systematic gap (across subjects) between ratings from two raters. Although these situations are unusual in practice, especially with raters having basic training, researchers may want to take a precautionary measure by testing that ratings from all raters come from distributions with a common expected value. If this hypothesis is rejected, then I will not recommend the use of a mixed factorial design for the purpose of analyzing inter-rater reliability. Nevertheless, using a mixed factorial design for the purpose of studying inter-rater reliability is generally an effective approach for quantifying the extent of agreement among raters when the rater effect is fixed. This is due to us not having to worry about the rater variance component that represents raters that did not participate in the reliability experiment.

For the purpose of analyzing intra-rater reliability, the mixed factorial design works well when the rater effect is fixed. For a given total variation in the ratings, having a high subject-rater interaction may be beneficial since it could lead to a higher coefficient of intra-rater reliability. In fact, the error variance under the mixed factorial design includes not only the variance due to random experimental errors, but also the variance due to replication, which actually measures reproducibility. Therefore, a small error variance is an indication of small variance due to replication as well, which in turn is an indication of high reproducibility. Consequently, a high subject-rater interaction for a given total variation is not problematic, and may even be a sign of a relatively small error variance and high reproducibility. The relationship between the subject-rater interaction, inter-rater and intra-rater reliability will be further discussed in subsequent sections.

10.2 Intraclass Correlation Coefficient

The mixed factorial design involves a single group of raters as well as a single group of subjects, all of which are rated by each rater. Note that the word raters in this context could designate a group of 5 individuals (for example) operating the same measuring device, or could also designate 5 trials performed by a single individual operating the same measuring device to score all subjects. The data produced in both situations can be analyzed with the same methods that are discussed in this section. The analysis results may be interpreted differently depending on the context. Therefore, developing an in-depth understanding of the nature of the reliability experiment is essential to properly interpret the data analysis.

The abstract representation of ratings under the mixed factorial design is given by,

$$y_{ijk} = \mu + s_i + r_j + (sr)_{ij} + e_{ijk}, \tag{10.2.1}$$

where y_{ijk} is the k^{th} replicate score[1] that rater j assigned to subject i. The remaining terms in model 10.1 are defined as follows[2]:

- μ is the expected value of the y-score.
- s_i is the random subject effect, assumed to follow the Normal distribution with 0 mean, and variance σ_s^2.
- r_j is the fixed rater effect, assumed to satisfy the condition,

$$\sum_{j=1}^{r} r_j = 0, \tag{10.2.2}$$

[1] Many reliability experiments only involves one trial (the first one)

[2] These are standard conditions used in all ANOVA models

10.4 Sample Size Calculations

This section addresses the problem of sample size determination when designing inter-rater and intra-rater reliability studies under the mixed factorial model. It provides guidelines and methods for calculating the number of raters and number of subjects required in order to achieve the prescribed precision level with which inter-rater or intra-rater reliability coefficients must be estimated. Before the proposed methods can be used, the researcher must specify the desired precision level. The precision of the ICC estimate can be specified by the length of the associated 95% confidence interval[10]. More information regarding the use of confidence intervals for this purpose may be found in Giraudeau and Mary (2001), and in Doros and Lew (2010). In the previous chapters, I recommended that the desired confidence interval length (L) be expressed as a fraction of the the ICC value, and suggested $L = 0.4 \times \text{ICC}$ as being a reasonable precision level for the intraclass correlation coefficient[11]. You may decide to obtain a more precise ICC estimate by setting a confidence interval length that represents an even smaller fraction of the ICC, such as $0.2 \times ICC$.

Section 10.4.1 is devoted to the inter-rater reliability coefficient, while section 10.4.2 focuses on the intra-rater reliability coefficient. Note that the inter-rater and intra-rater reliability coefficients have conflicting sample size requirements. A good inter-rater reliability study requires fewer raters and more subjects, whereas a good intra-rater reliability study requires more raters and fewer subjects. Consequently, if the same study is designed for both types of coefficients, a compromise will be necessary.

10.4.1 *Sample Size Calculations for Inter-Rater Reliability*

For a given number of ratings per rater, the most efficiency strategy for obtaining an accurate inter-rater reliability coefficient is to maximize the number of subjects by allowing the raters to produce a single rating for each subject. For example, if each rater must produce a total of 14 ratings, rather than producing 2 ratings on each of 7 subjects, the inter-rater reliability coefficient will be more precise if the design is based on 14 subjects and 1 rating per subject. If the recruitment of subjects is costly, then the researcher may want to increase the number of ratings per rater by rating some subjects more than once. This strategy will result in a loss of precision in the estimation of the inter-rater reliability coefficient.

[10] The 95% confidence interval associated with an ICC estimate is a range of values expected to contain the “true” value of the ICC 95% of the times.

[11] In the classical context of population mean estimation, this condition would ensure a coefficient of variation that does not exceed 20%.

In this section, I will first assume that there will be a single rating per rater and per subject (i.e. no replication). This will allow you to determine the number of subjects and raters that can yield your desired confidence interval length. If this optimal number of subjects is deemed unduly high, you will be able to explore some options for reducing the number of subjects while maintaining the same number of ratings per rater. This is typically achieved by increasing the number of ratings per subject for each rater. You will see how much you will lose in precision by using this alternative approach, before deciding whether it should be considered or not.

NUMBER OF RATERS AND SUBJECTS - NO REPLICATION

With the mixed factorial design, researchers generally know ahead of time how many raters will be used in the inter-rater reliability study. One retains all the raters willing to be part of the investigation. However, our investigation has revealed that, for a given number of ratings per rater, using more than 5 raters can increase the precision of the inter-rater reliability only marginally. This explains why I have investigated this problem for 2, 3, 4, and 5 raters.

Figure 10.4.1 depicts the length of the 95% confidence interval associated with the inter-rater reliability coefficient as a function of the number of subjects, and the magnitude of the ICC, when the number of raters is limited to 2. An examination of this figure reveals the following facts:

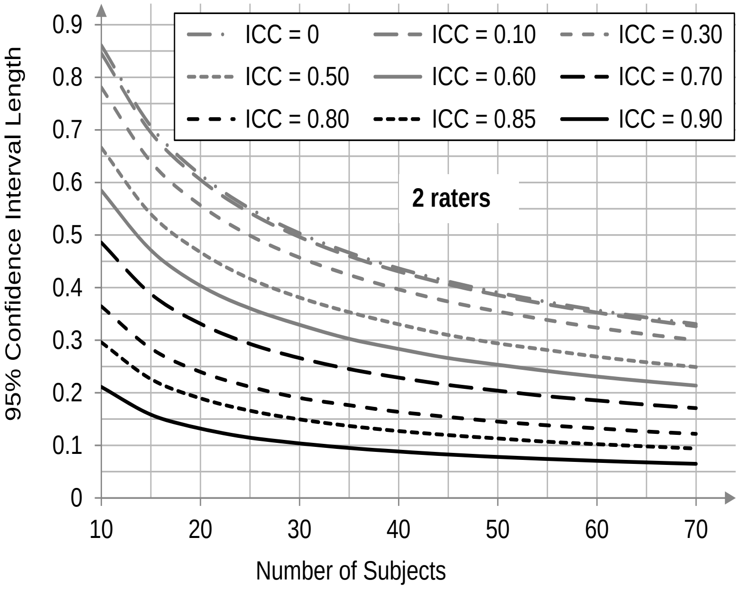


Figure 10.4.1: 95% C.I. Length by Number of Subjects for 2 Raters

- For any given ICC value, the confidence interval length decreases (i.e. precision improves) as the number of subjects increases. That is, having more subjects in the experiment can only improve the quality of the inter-rater reliability coefficient.
- For any given number of subjects, the interval length decreases as the ICC value increases. Consequently, your experiment will yield a more accurate inter-rater reliability coefficient if the extent of agreement among raters is high. This is logical because raters who agree are homogeneous with respect to the way they rate subjects. Therefore, a handful of subjects will generally be sufficient to tell an accurate story regarding the extent to which they agree.

  The lesson to be learned here is that giving some training to the raters prior to conducting the experiment is likely to pay off, and the payoff could be big. If ICC=0 indicating total absence of agreement, then achieving a desired precision level may become an impossible task.

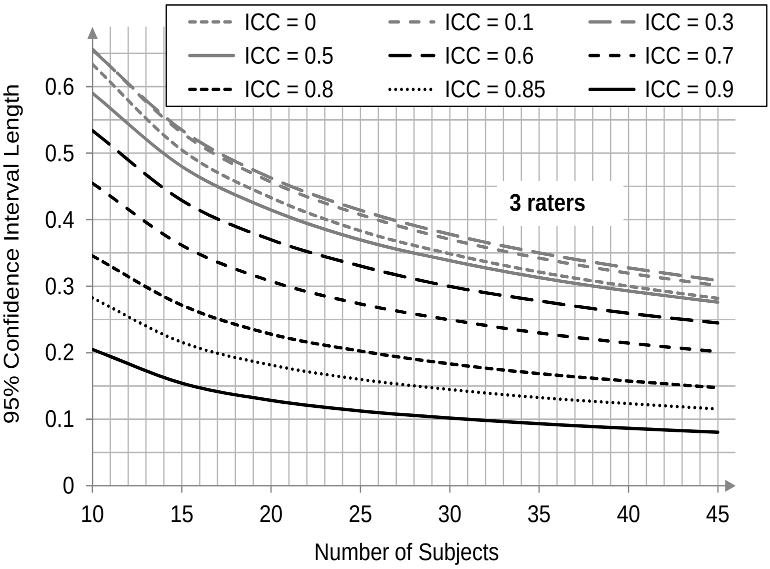
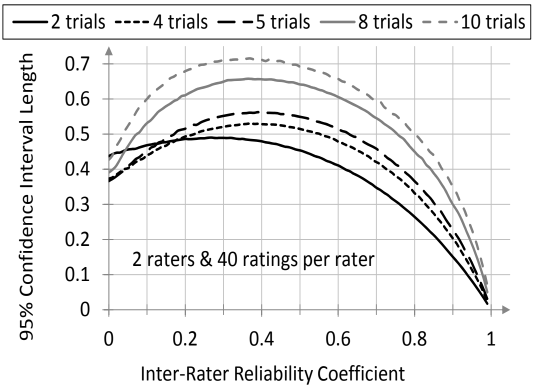


Figure 10.4.2: 95% C.I. Length by Number of Subjects for three Raters

Suppose you know that your study will involve two raters only, and want to know how many subjects to recruit. To be able to use Figure 10.4.1, you will need two things: (*i*) a predicted ICC value (i.e. a predicted inter-rater reliability coefficient), and (*ii*) the desired confidence interval length (e.g. 0.4 × ICC). The predicted ICC often comes from a pilot or a prior study. This initial ICC value is essential and obtaining it is part of the preliminary exploratory analysis necessary for an effective study design. Assume for example that this initial value is $\text{ICC}_0 = 0.70$. This leads to our desired 95% confidence interval length of $0.4 \times \text{ICC} = 0.4 \times 0.70 = 0.28$.



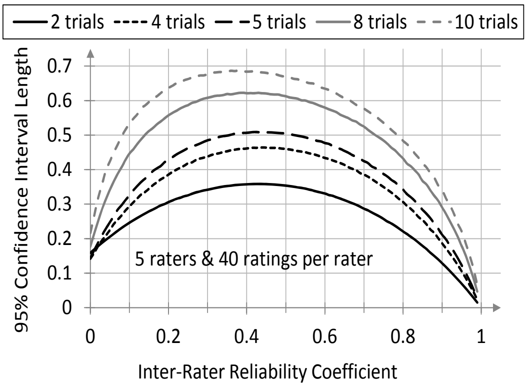


Figure 10.4.11: Interval length by IRR & # of trials (2 raters & 40 ratings per rater).

Figure 10.4.12: Interval length by IRR & # of trials (5 raters & 40 ratings per rater).

10.4.2 *Sample Size Calculations for Intra-Rater Reliability*

The purpose of an intra-rater reliability study is to evaluate the reproducibility of ratings. The researcher quantifies the extent to which the raters can replicate the rating procedures and consistently produce comparable ratings. Under the mixed factorial design, the researcher generally knows how many raters will be tested. However, because of both the subject and subject-rater interaction effects, subjects are expected to have an impact on the reproducibility of ratings. Moreover, the number trials plays a pivotal role in the quantification of reproducibility, since these trials produce a set of ratings associated with the same rater and the same subject. Therefore, the researcher needs to determine the optimal number of subjects and trials to consider for the experiment that will produce the desired confidence interval length.

Here are a few things researchers need to know upfront:

- Using only 2 trials in an intra-rater reliability is a bad thing, unless you know from prior studies or from a pilot study that reproducibility is expected to be high. Reproducibility needs to exceed 0.6 when using 2 raters, and needs to exceed 0.3 when using 3 raters or more, if you want 2 trials to produce an acceptable precision level for the intra-rater reliability coefficient. If reproducibility as measured by the ICC can be below 0.3, then using only 2 trials may lead to poor results, with estimates of intra-rater reliability that are very unreliable.
- For a given number of ratings per rater, the best results are generally obtained with 3, 4, or 5 trials. Using more than 5 trials will generally decrease the

precision of the intra-rater reliability coefficient for the same number ratings. Therefore, if you have a fixed number of ratings per rater that you cannot exceed, then I would strongly recommend using 3 or 4 trials, since a good balance must be found between the number subjects and the number of trials.

Figures 10.4.13 through 10.4.36 depict the relationship between the length of the 95% confidence interval and the magnitude of the intra-rater reliability coefficient. The confidence interval is associated with the intra-rater reliability coefficient that will be estimated using observed ratings. Its expected length is what is depicted on the graph. The magnitude of the intra-rater reliability coefficient is generally replaced at the design stage with a surrogate obtained from a pilot study or from old studies.

The figures 10.4.13 through 10.4.36 do not cover all possible scenarios, although they give you a glimpse into what you can expect in terms the precision associated with your intra-rater reliability coefficient. Let us consider figure 10.4.13 for example. It assumes an intra-rater reliability study using 2 raters and 10 ratings per subject. There are 2 possible designs that give you 10 ratings per rater. One possibility is to use 5 subjects and 2 trials, and a second possibility is to use 2 subjects and 5 trials. These 2 possibilities are represented with the continuous black curve (for 5 raters and 2 trials) and the continuous gray curve (for 2 raters and 5 trials). It follows from this chart that if you anticipate high reproducibility represented by an inter-rater reliability coefficient that exceeds 0.8, then using 2 trials is expected to yield the most accurate coefficients. However, if reproducibility is below 0.7 then using only 2 raters will yield a wide confidence interval; therefore an intra-rater reliability coefficient that is very unreliable. An examination of the remaining curves reveals that using more raters or more ratings per rater improve the precision of the intra-rater reliability coefficient.

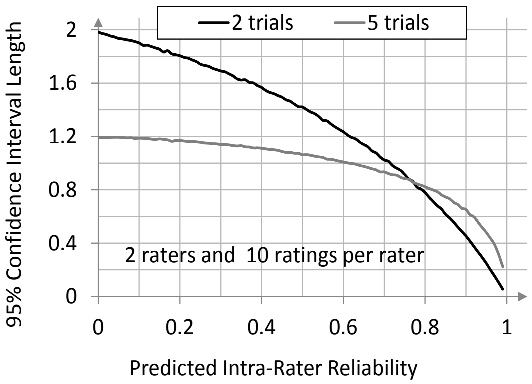


Figure 10.4.13: Interval length by IRR & # of trials (2 raters & 10 ratings per rater).

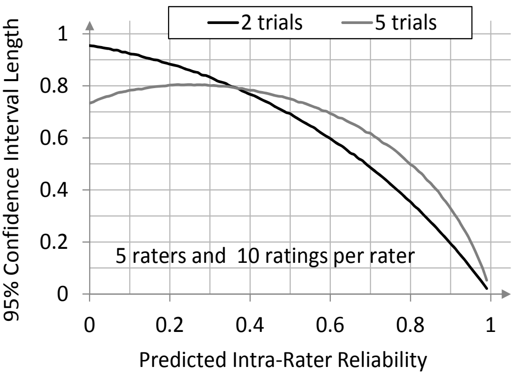


Figure 10.4.14: Interval length by IRR & # of trials (5 raters & 10 ratings per rater).

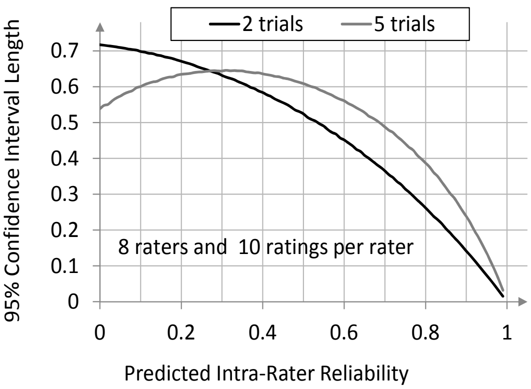


Figure 10.4.15: Interval length by IRR & # of trials (8 raters & 10 ratings per rater)

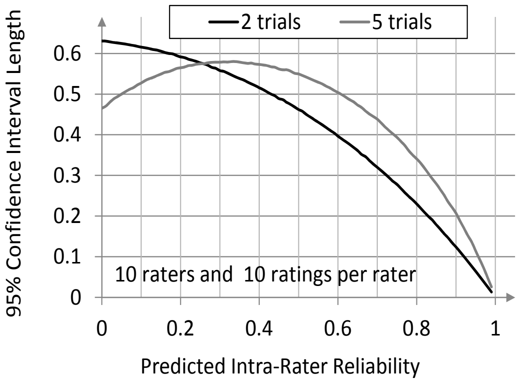


Figure 10.4.16: Interval length by IRR & # of trials (10 raters & 10 ratings per rater)

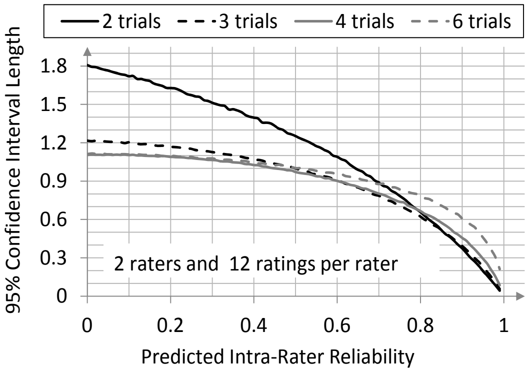


Figure 10.4.17: Interval length by IRR & # of trials (2 raters & 12 ratings per rater)

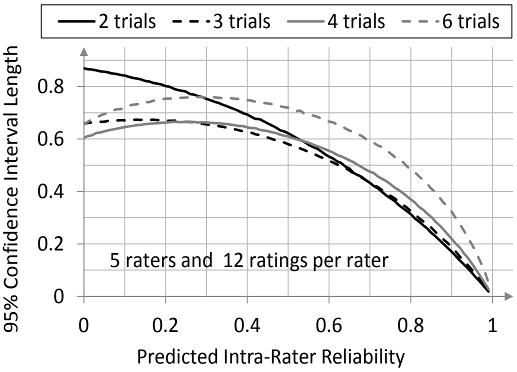


Figure 10.4.18: Interval length by IRR & # of trials (5 raters & 12 ratings per rater)

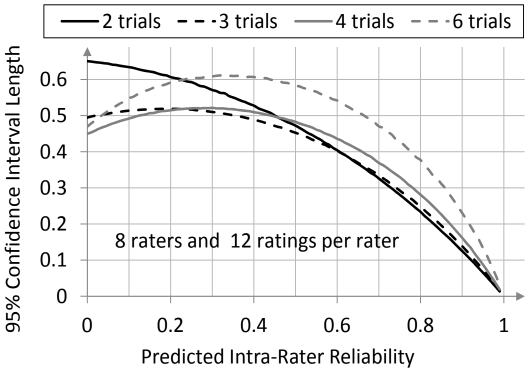


Figure 10.4.19: Interval length by IRR & # of trials (8 raters & 12 ratings per rater)

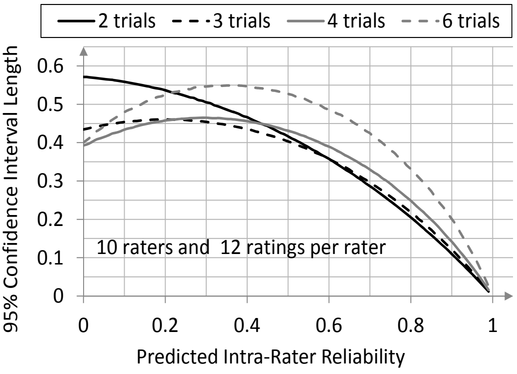


Figure 10.4.20: Interval length by IRR & # of trials (10 raters & 12 ratings per rater)