CHAPTER 7

Intraclass Correlation: A Measure of Raters' Agreement

OBJECTIVE

This chapter presents a general overview of the use of Intraclass Correlation Coefficients for quantifying the extent of agreement among raters when the ratings are in the form of quantitative measurements. A high-level description of the underlying statistical models is provided as well as a discussion on the limitations associated with their use. After reading this chapter the practitioner will be able to decide which model is appropriate for the study that was conducted, and will know the related challenges that must be overcome. This chapter also describes the Bland-Altman plot, a popular graphical method for analyzing agreement between two raters. The reader will find an introduction to sample size calculations in this chapter, and a more detailed treatment of the sample size problem in subsequent chapters. Figure 7.4.1 represents a flowchart showing how to find the correct intraclass correlation coefficients based on the way the ratings were gathered and the type of analysis to be done.

CONTENTS

7.1 Introduction **186**

7.2 Statistical Models **186**

7.3 The Bland-Altman Plot **189**

7.4 Sample Size Calculations **192**

7.1 Introduction

In the past few chapters of parts I and II, I presented many techniques for quantifying the extent of agreement among raters. Although some of these techniques were extended to interval and ratio data, the primary focus has been on nominal and ordinal data. This chapter as well as the other chapters of part III, are devoted to the study of inter-rater reliability for quantitative outcomes whose possible values are defined by a continuum, as opposed to being a predetermined set of specific values.

Why do we need to care about intraclass correlation when weighted versions of the chance-corrected measures can be used to handle quantitative outcomes ? It is because the notion of "perfect" agreement associated with two raters assigning the exact same score to the same subject, does not translate well to quantitative measurements. Consider for example two electronic devices used to measure the knee joint laxity on 15 human subjects. Even if both devices are equally reliable, you would not expect them to produce the exact same quantitative measurement on the same subjects, since these values belong to a continuum. Likewise, two very competent raters that measure the height or the weight of the same human subject will likely produce slightly different numbers regardless of their proficiency level in the use of the measuring instrument. With agreement no longer referring to an exact match, the notions of chance agreement and percent agreement evaporate.

The solution to this problem is to use the portion of variation in the data that is due to subjects, and to compare it to the other portion of that variation due to raters. If the rater-induced variation exceeds that of the subject by a wide margin then the raters are said to have low inter-rater reliability. Otherwise, the raters are said to have high inter-rater reliability. But this approach will work only if the reliability experiment is designed in such a way that the different variation components can be separated. You will see in the next few sections how this task can be accomplished. Several approaches can be used to design an inter-rater reliability study, depending on the goal aimed at for the study. In the next section, I will describe a few designs commonly used in the context of inter-rater reliability analysis.

7.2 Statistical Models

Consider the reliability data shown in Table 7.1. That data represents scores that 4 raters assigned to 6 subjects, and could be interpreted in various ways depending on how it was collected. Here are 4 possible study designs (or data models) that could have produced Table 1 data:

▶ **Model 1A**: *Each subject is rated by a different group of raters*

According to this model, each row of Table 7.1 is not necessarily associated with the same set of 4 raters. Although the 4 raters are consistently labeled as 1, 2, 3, and 4, they could represent different individuals, or different measuring instruments. One may average this data row-wise to study the subject effect, but will not be able to average column-wise to obtain the rater effect. It is why this is often known as a one-factor (or one-way) model, the single factor here being the subject.

The main implication of this model is that one rater may not have the opportunity to score more than one subject. Consequently, this model makes it impossible to evaluate *Intra-rater Reliability*, which is a measure of the rater's self-consistency. However, the raters under this model still score the same subjects, making it possible to compute *Inter-rater reliability.*

The main advantage for using this model is that the raters could be located in different geographic areas, and rate local subjects. There is no need to move subjects around to allow different groups of raters to rate the same subjects. This model may also be suitable in situations where subjects are hard to recruit and the availability of the same group of raters cannot be guaranteed when a subject is able to participate in the experiment.

Table 7.1: Scores assigned by 4 raters to 6 subjects[a]

Subject	Rater				Average
	1	2	3	4	
1	9	2	5	8	6
2	6	1	3	2	3
3	8	4	6	8	6.5
4	7	1	2	6	4
5	10	5	6	9	7.5
6	6	2	4	7	4.75
Average	7.67	2.5	4.33	6.67	5.29

[a]This data is taken from Shrout & Fleiss (1979), although I replaced the terms Target and Judge with Subject and Rater respectively, and added row and column marginal averages.

▶ **Model 1B**: *Each rater rates a different group of subjects*

If Table 7.1 data were collected according to this design, then the 6 subjects may differ from rater to rater. That is, each rater scored his own set of subjects, even though I may have decided to consistently labeled them as 1, 2, 3, 4, 5, and 6. One may evaluate the rater effect by averaging Table 7.1's columns. Any row-wise averaging would be meaningless as such an operation would involve different subjects as well as different raters. Therefore, the only factor that can be studied is the rater factor, and this model will later be referred to as a one-factor or one-way model.

The main implication of this model is that it allows for the evaluation of intra-rater reliability, and not that of inter-rater reliability. Evaluating inter-rater reliability always requires different raters to score the same subjects.

▶ **Model 2**: *The Random Factorial Design*

According to this model, each subject is scored by the same group of raters. Both the subjects and the raters are random samples selected from the respective populations they represent, hence the naming "random" design. Moreover, the column and row marginal averages are meaningful, and the effects of subject and rater factors can be evaluated. It is because both factors (rater and subject) can be studied that this design is known as a "factorial design". The experimental design that produces Table 7.1 data is called a two-way factorial design.

▶ **Model 3**: *The Mixed Factorial Design*

According to this design, each subject is scored by the same group of raters, and is also in this regard a factorial design. Unlike Model 2, here only the group of subjects represents a random sample selected from a larger subject population, while the group of raters does not represent a random sample. Because the group of raters that participate in the reliability experiment is not randomly selected from a larger rater population, these raters only represent themselves. The resulting inter-rater reliability coefficient can therefore not be applied to raters beyond those in the experiment. Therefore, the subject effect is random, while the rater effect is fixed. This combination of random and fixed effects gave this design the name "Mixed Factorial Design." When the number of factors considered is limited to two as is the case for Table 7.1, it is renamed the "Two-Way Mixed Factorial Design."

Each of these models requires a different method for calculating the intraclass

correlation coefficient. Shrout & Fleiss (1979) discussed models 1A (although it was referred to as model 1), 2, and 3. The same models were also discussed by McGraw and Wong (1996), who presented methods for computing the intraclass correlation for each of them. However, these authors did not deal with the important problem of missing ratings, which is very common in inter-rater reliability experiments. The next few chapters of Part III of this book discuss the missing-rating issue extensively.

7.3 The Bland-Altman Plot

An mainly graphical method often used as an alternative to the intraclass correlation for analyzing inter-rater reliability data was proposed by Bland and Altman (1986). It combines a graphical approach and a quantitative analysis of the magnitude of the rating differences. This method can only analyze two raters at a time, and has become popular over time among researchers, although many of its users are often unaware of its limitations. In this section, I will present an overview of this method, and will discuss its merits as well as its limitations.

Suppose that we want to study the extent of agreement between the two raters labeled as 3 and 4 using Table 7.1's ratings. The Bland-Altman method is implemented as follows:

- The first step consists of creating a scatterplot that depicts the differences in ratings between raters 4 and 3 as a function of their averages. Table 7.2 shows the ratings being analyzed as well as the two series of averages and differences used to create the scatterplot of Figure 7.3.1.
- The next step is to display on the scatterplot created in the previous step, the two "limits of agreement". The dotted line at the bottom is the lower limit of agreement and the one at the top represents the upper limit of agreement. The lower limit of agreement is -1.169 while the upper limit of agreement is 5.836. This indicates that you can expect the difference between raters 4 and 3 to be as high as 3.763 and as low as 0.904. Depending on the application at hand, such a gap may be acceptable or may be too wide. Ultimately, this gap will help the researcher decide whether the extent of agreement between the two raters 4 and 3 is acceptable or not. If $\overline{d}$ is the average difference and s the standard deviation of the differences, then the lower limit of agreement is $\overline{d} - 2s/n$ and the upper limit of agreement $\overline{d} + 2s/n$.

The two steps described above summarize what is known as the Bland-Altman method. It is intuitive and fairly straightforward to apply. Bland and Altman (1986) indicated that their plot can help study the relationship between the rating pairwise differences and the associated pairwise means, which by the way are used as surrogates for the true rating associated with the subject. The study of this relationship

is one way of verifying whether the differences are independent or not. These differences must be approximately independent for the interpretation of the lower and upper limits of agreement to be valid. If these differences have for example a tendency to decrease as the averages increase, or if this relationship shows any other specific trend, this may an indication of a lack of independence. Transforming the initial ratings using the logarithm function for example may be the remedy for obtaining the independence needed.

Some researchers believe that the Bland-Altman method is the only realistic way of dealing with inter-rater agreement. That is no true. We will see in the next few chapters why the intraclass correlation is not only appropriate, but is often the better approach.

Table 7.2: Scores assigned by Raters 3 & 4 to 6 subjects

Subject	Rater #3	Rater #4	Mean Rating	Difference[a]
1	5	8	6.5	3
2	3	2	2.5	-1
3	6	8	7	2
4	2	6	4	4
5	6	9	7.5	3
6	4	7	5.5	3

[a]Difference = (Rater 4) - (Rater 3)

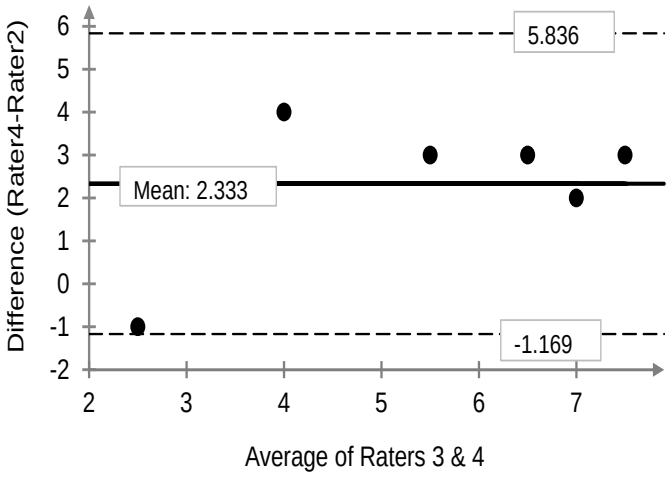

Figure 7.3.1: Rating Differences as a function of rating means

ISSUES WITH THE BLAND-ALTMAN METHOD

Part of the popularity of the Bland-Altman method stems from its graphical nature. You can look at the graph and see right in front of you the differences between the ratings obtained from the two raters you are analyzing. A simple visual exploration may even allow you to form an opinion about the extent to which they agree. Using the two limits of agreement helps you figure out how large the difference should be before it can be considered too large. Here are a few assumptions the Bland-Altman method is based upon, and which are often not satisfied:

- Bland and Altman (1986, p. 4) indicates that the "... *differences are likely to follow a Normal distribution because we have removed a lot of the variation between subjects and are left with the measurement error.*" The real problem with this assumption is that it is untrue if there is a subject-rater interaction. This is often the case when the rating is affected by the magnitude of the "true" score associated with the subjects. The subject-rater interaction does not preclude the differences from following the Normal distribution. However, the differences will be correlated and their actual standard deviation would be higher than the estimate s recommended by Bland and Altman(1986).

  If the standard deviation of the differences is underestimated then the Bland-Altman method may produce a false sense of agreement. When subjects and raters interact, inter-rater reliability is better analyzed with the intraclass correlation that relies on a formal modeling of the interaction effect.

- Another benefit of the Bland-Altman plot lies in the analysis of the the relationship between the differences and the average ratings. This relationship is important primarily because it allows you to see whether raters and subjects interact provided the average is a good surrogate for the subject's true score. The problem here is that the average is known to be close to the true value only if there is little variation in the ratings. That is if the raters are known to be in agreement, an assumption we cannot make since that very agreement is precisely what we are studying.

- The Bland-Altman method is meant for pairwise analyses only. It may not allow you to obtain a global picture of the extent of agreement among multiple raters. When the number of raters is moderately large such as 8, the number of pairwise analyses becomes as large as 28, which can be problematic.

I would recommend using the Bland-Altman plot mainly as an exploratory technique. It allows the researcher to have a first glimpse into the inter-rater reliability results. Ultimately, an intraclass correlation based on the appropriate statistical model should be calculated.

7.4 Sample Size Calculations

When designing an inter-rater (or intra-rater) reliability study, the researcher first needs to determine how many subjects and how many raters must be part of the experiment. Sometimes, there is also a need to determine the number of trials also known as the number of replicates if replication is desired. Note that replication is about a rater taking more than one measurement from the same subject. Each of the next three chapters has a section on sample size calculation. These sections provide detailed procedures showing how the number of raters, the number of subjects and the number of trials can be determined depending on which data model is chosen.

Traditionally power calculation as done in statistics is based on the test of hypothesis involving population means, and consists of finding the optimal sample size that yields the desired power[1] for the statistical test. This procedure generally requires the researcher to specify the effect size (or the detectable difference)[2], the statistical significance (also known as α or alpha), and the desired power. The approach proposed here for the ICC is slightly different. It requires the researcher to specify the desired confidence interval length (this is equivalent to specifying the effect size), the confidence level associated with the confidence interval (this often takes the values 90%, 95%, or 99%), and the anticipated ICC value. The anticipated ICC value may be known from prior studies or from a pilot experiment. If such a value is unknown then I will recommend a conservative approach based on the anticipated ICC value that will yield the largest confidence interval length.

Our investigation has revealed that you need about 5 raters to optimize your inter-rater reliability coefficient for a given total number of ratings. The total number of ratings is the product of the number of raters by the number of subjects (assuming one trial per rater and per subject). Therefore, if your experiment is going to generate 140 ratings for example, then it would be more efficient to have 5 raters and 28 subjects instead of having 10 raters and 14 subjects. A design is said to be more efficient in this context when it yields the smaller confidence interval length. Consequently increasing the number of subjects is more rewarding than increasing the number of raters beyond 5. However, if recruiting raters is cheaper than recruiting subjects then you may have to increase the number of raters beyond 5 and reduce the number of subjects.

In practice, it often happens that the researcher has to use a specific model, due

[1]The power of a statistical test represents the probability for that test to reject the "null" hypothesis when it is false. This "null" hypothesis could be the equality of two population means, or the equality of a population mean to a hypothetical value.

[2]The detectable difference is the smallest difference between the two population means under comparison, which will cause the null hypothesis to be rejected.

to various practical constraints. If you have the opportunity to chose the model you want, the question becomes which one to chose and how to chose it. The answer depends on whether you want to optimize the inter-rater reliability calculation, the intra-rater reliability calculation or both. Let us start with the inter-rater reliability optimization first.

For the purpose of optimizing the inter-rater reliability assessment, I recommend the use of models 2 or 3 if possible. Model 3, if appropriate, is expected to yield more accurate intra-rater reliability coefficients than model 2 for the same number of raters, and subjects. However, the discussions in subsequent chapters may give you the impression that using models 1A or 2 will produce similar results. This is not accurate. Model 1A allows you to use a different group of raters for each subject. While this may be convenient when the same group of raters cannot be present to rate the same subjects, the inter-rater reliability calculation comes with a price tag. The use of different groups of raters is expected to increase the variation in ratings due to the rater effect, which in turn will reduce the magnitude of the inter-rater reliability coefficient. Model 1A does not allow for an in-depth analysis of the impact of having different groups of raters, since it does not specify the mechanism underlying the selection of these raters for each subject. The choice between models 2 and 3 depends on whether the raters used in the experiment are the only ones you are interested in (model 3), or whether they are part of a larger universe of raters you like to infer to (model 2). Under models 2 and 3, each rater is expected to rate all subjects, making it easier for you to decide how many raters, subjects, and possibly trials will produce the ratings you need. This is one of the key advantages of these two models.

For the purpose of optimizing the intra-rater reliability, I still recommend the use of models 2 or 3 if possible. You may nevertheless use the simple model 1B with a single subject being rated multiple times by each of the participating raters. However, using model 1B makes the intra-rater reliability very dependent upon the one subject being rated. An alternative approach would be to use model 1A with one rater rating each subject multiple times. Again this approach will make the intra-rater reliability very dependent on the specific rater used in the experiment. This may or may not be what you want. The approach I recommend is the use of models 2 or 3, with model 2 (if appropriate) expected to produce more accurate intra-rater reliability coefficient for the same number of raters than model 3. For a fixed number of ratings per rater, you need no more than 4, 5, or 6 trials to obtain the most accurate intra-rater reliability coefficient under models 2 and 3. That is if a rater must produce 40 ratings, it would be more effective to use 8 subjects and 5 trials rather than 20 subjects and 2 trials. All these issues and many more are discussed in-depth in the next three chapters.

Figure 7.4.1 represents a decision tree showing which equations or subsections in the subsequent chapters should be used to compute the correct agreement coefficient and associated p-values and confidence intervals, depending on the model dictated by your study design. The numbering of these equations (or subsections) is descriptive, and the first digit refers to the chapter number, the second digit to the section within the chapter, and the third number to a specific equation or subsection.

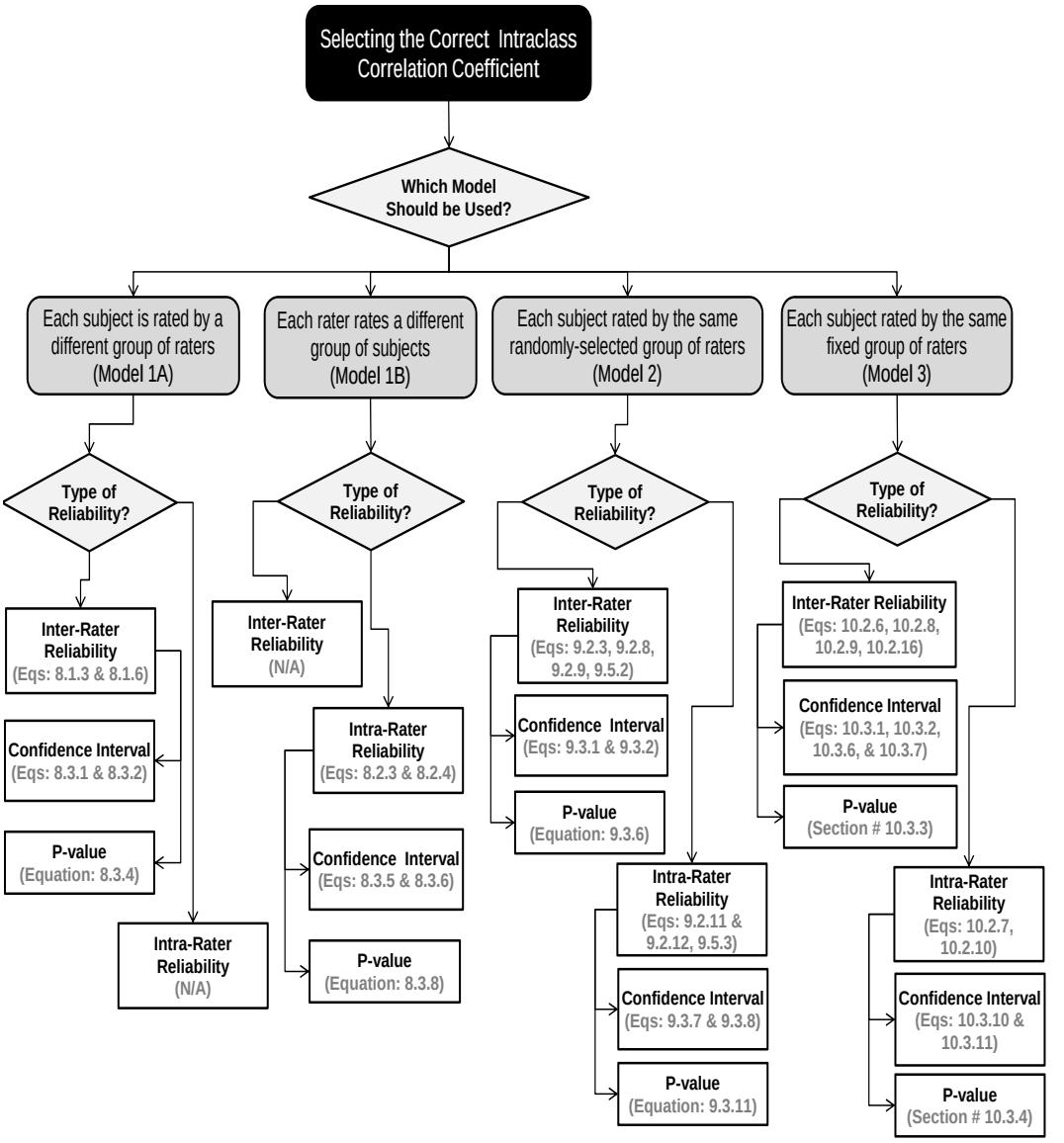

Figure 7.4.1: Choosing the Correct Intraclass Correlation