CHAPTER 9

Intraclass Correlations under the Random Factorial Design

OBJECTIVE

The objective of this chapter is to present methods for calculating various intraclass correlation coefficients and associated precision measures, in reliability studies where the rater and subject factors are fully crossed. Each rater is expected to rate all participating subjects, but may take more measurements on some subjects and less on others. The rater and subject samples are both assumed to have been randomly selected from larger rater and subject populations, which represent the primarily interest of the researcher. I define two types of intraclass correlation coefficients: (i) the intraclass correlation coefficient for quantifying inter-rater reliability, and (ii) the intraclass correlation coefficient for quantifying intra-rater reliability. For both types of intraclass correlation, methods for obtaining confidence intervals, p-values, and optimal sample sizes (i.e. required number of subjects and raters during the design of experiments) will be presented as well.

CONTENTS

9.1 The Issues

The Intraclass Correlation Coefficient (ICC) that is associated with Model 1A, is the ratio of the subject variance to the sum of the subject and error variances. What was termed error variance in the previous chapter is in reality the variance of a combination of three effects, which are the rater effect, a possible rater-subject interaction effect[1], and the experimental error effect. Because these three effects are blended together, they are interdependent, and their combined variance is expected to be higher than if the experiment was designed to keep them independent[2]. Therefore, the researcher can improve the magnitude of the ICC substantially by designing the experiment so as to keep all the factors at play independent from one another. This is accomplished by getting each rater to score all subjects. Such a design is known as the factorial design and is the subject of this chapter.

The ICC associated with Model 1B on the other hand, quantifies the intra-rater reliability and was defined in the previous chapter as the ratio of the rater variance to the sum of the rater and error variances. Once again, the error variance in the context of model 1B is actually the variance of the combined effect due to the subject, the rater-subject interaction and the experimental error. The experimental design that underlies model 1B (i.e. each rater scores a different group of subjects) has blended these three effects into one. Consequently, the variance of the combined effect will often be high, reducing thereby the magnitude of the ICC. If an experiment is designed so that the rater, rater-subject interaction, and error effects are independent from one another, then the variance due to their interdependency will be eliminated leading to a higher ICC for the same amount of data collected. This is the factorial design mentioned in the previous paragraph.

There are different types of factorial designs that may achieve different objectives. We will now review some of them.

Types of Factorial Designs

The factorial design, is an experimental design where each rater is expected to rate all subjects participating in the experiment. The main advantage of this design is that all the factors involved in the experiment are kept independent from one another. That is, you can fix a specific rater and study the subject effect; just as you may

[1] The rater-subject interaction can be seen as the portion of the rater effect that may be attributed to the specific subject being rated.

[2] Note that if a and b are 2 dependent effects, then their combined variance will be $var(a+b) = var(a) + var(b) + 2cov(a,b)$, where $cov(a,b)$ is the covariance between a and b. If the effects are independent, the covariance term will vanish, the joint variance will decrease (assuming a positive covariance, which is usually the case in agreement studies).

fix a specific subject so as to study the rater effect. If two measurements or more are taken from one subject by the same rater, then one may study the rater-subject interaction effect independently from the experimental error.

Rater-subject interaction is bad for both inter-rater and intra-rater reliability, but is sometimes unavoidable. It induces more variation in the data, in addition to the portion of total variation that is due to raters and subjects. This extra variation will further reduce the magnitude of the ICC. Figure 9.1.1 depicts the reliability data of Table 8.1 of chapter 8. Without interaction, all 4 curves associated with the raters would be reasonably parallel, which is the case for raters 1, 2, and 3. Rater 4 however, appears to assign scores to subjects with a gap with other raters that changes from subject to subject. This is an indication of the existence of rater-subject interaction. Rater 4 alone is likely to bring the ICC down in a significant way.

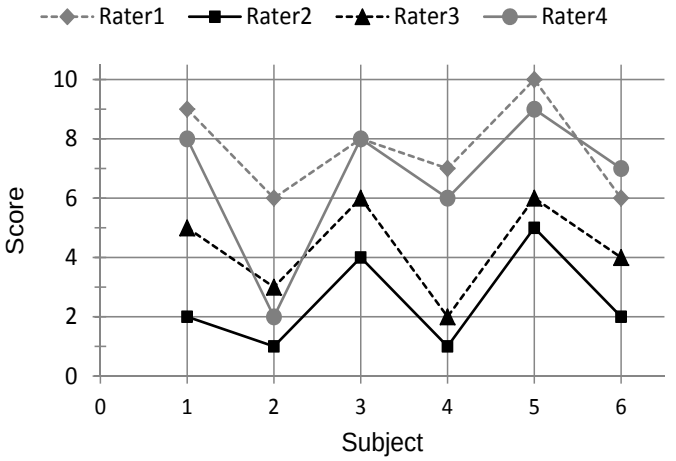

Figure 9.1.1: Ratings of 6 subjects by rater

Two types of factorial designs involving the subject and rater factors are the random and mixed factorial designs. The random factorial design is a design where the rater and the subject effects are random, while the mixed factorial design is one where the rater effect is fixed and the subject effect random.

In the random factorial design, the raters participating in the experiment are selected randomly from a larger universe of raters, and the participating subjects are selected randomly from a larger universe of subjects. The subjects and raters in their respective universes are actually those the researcher wants to investigate in the first place. The samples representing subgroups of these universes are used to minimize the costs of conducting experiments. It is the desire to draw meaningful conclusions about entire universes from their smaller representative samples that creates the need to use statistical methods.

In the mixed factorial design on the other hand, only participating subjects are selected randomly from a larger subject universe. The participating raters are not tied to any other group of raters. They represent themselves, and are the only ones being investigated by the researcher. The study findings will only apply to these raters, and cannot be generalized to raters who did not participate in the experiment. For example, consider a reliability experiment whose purpose is to evaluate the consistency level between two measuring devices used in rheumatology clinical examinations. The researcher in this case, will want the study findings to be limited to the two specific measuring devices being investigated, and not be generalized to other devices that may not be similar to those used in the experiment. Experiments based on mixed factorial designs will often yield a higher ICC than those based on the random factorial designs, because no variation is generated by the rater effect when the design is mixed.

In this chapter, I will focus on the statistical methods used for analyzing experimental data based on the random factorial design. Methods needed for analyzing mixed factorial designs will be discussed in the next chapter.

9.2 The Intraclass Correlation Coefficients

The random factorial design involves a single group of raters as well as a single group of subjects, all of which are rated by each rater. That is the rater and subject factors are fully crossed. Table 9.1 shows lung functions data of 15 children representing their peak expiratory flow rates. Four measurements were taken on each subject by 4 raters. The raters here could represent 4 individuals operating the same measuring device, or one individual using the same measuring device on 4 different occasions. The data produced in these two scenarios can be analyzed with the same methods discussed in this section, although the results may be interpreted differently depending on the context.

Table 9.1 data were generated by a single group of 4 raters, each of whom rated all members of the same group of 15 subjects. We assume that the 4 raters are representative of a larger pool of raters they were selected from. Likewise, the 15 children are assumed to represent the larger population of children of interest they were randomly selected from.

The resulting scores are described mathematically as follows:

$$y_{ijk} = \mu + s_i + r_j + (sr)_{ij} + e_{ijk}, \tag{9.2.1}$$

where y_{ijk} is the score assigned to subject i by rater j on the k^{th} trial[3]. The remaining terms of model 9.2.1 are defined as follows:

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

It follows from Table 9.4 that the p-value is smaller than 0.05 only for a γ_0 value that is equal to or smaller than 0.5. Consequently, we can claim that the observed ratings are consistent with the hypothesis that the "true" intra-rater reliability coefficient as measured by $ICC_a(2,1)$ exceeds any of these values.

9.4 Sample Size Calculations

When designing an intra-rater or an inter-rater reliability study, the researcher often needs to know how many raters and subjects must be recruited to obtain an accurate estimation of the agreement coefficient. An agreement coefficient is accurate if its confidence interval is "reasonably" narrow, or its confidence interval length (denoted by L) is "reasonably" small. 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). Now, I expect you to wonder how small is considered reasonably small - a common question asked by researchers. *I suggest to consider a confidence interval length L to be "reasonably" small if it is below* 0.4 *times the magnitude of the intraclass correlation coefficient*[13]*, whether it represents the inter-rater or the intra-rater reliability coefficient.*

This section presents methods and techniques for calculating the number of subjects, number of raters, and number of trials required to obtain a prescribed confidence interval length for the intraclass correlation coefficient. Sample size requirements for optimizing the estimation of the inter-rater reliability coefficient will be discussed first in section 9.4.1. A similar discussion devoted to the intra-rater reliability coefficient is deferred to section 9.4.2.

9.4.1 *Sample Size Calculations for Inter-Rater Reliability Studies*

For the purpose of designing an inter-rater reliability study, you must decide about the number of raters, number of subjects, and possibly the number of trials per rater and per subject if replication is an option you want to consider. One reason you may want to consider producing multiple ratings for the same subject is if intra-rater reliability will be analyzed in addition to inter-rater reliability. Another reason for using multiple trials is to increase the number of ratings when the costs associated with the recruitment of subjects are high, and should be minimized.

[13]Traditionally in statistics, the mean is considered reasonably accurate if its coefficient of variation (i.e. the variation per unit of measurement, also known as CV) does not exceed 10%, or 0.1 (some practitioners may want it smaller). For a 95% confidence interval, this coefficient of variation can be represented as the ratio of the interval length to 4 times the mean. So, $L/(4\mu) \leq 0.1$ leads to $L \leq 0.4\mu$. Here we replace μ with the intraclass correlation of interest.

However, for a given number of ratings per rater, the strategy that yields the most accurate inter-rater reliability coefficient requires the use of a single trial per subject and per rater. Therefore, I propose to first determine the optimal number of raters and subjects for a given confidence interval length, assuming one rating per rater and per subject. If you want to consider two trials or more per subject while maintaining the same number of ratings, then I will show the different options[14] and their impact on the precision of the inter-rater reliability coefficient.

Assuming that the number of raters is limited to 2, figure 9.4.1 depicts the expected length of the 95% confidence interval as a function of number of subjects, and the "true" inter-rater reliability, measured by the intraclass correlation coefficient (referred to here as ICC). Figures 9.4.2, 9.4.3, and 9.4.4 show similar graphs for 3, 4, and 5 raters respectively. All 4 figures reveal a number of important relationships that should be mentioned:

- For any ICC value, the length the of 95% confidence interval decreases as the number of subjects increases. That is, a larger number of subjects will improve the precision of your inter-rater reliability coefficient.
- For any given number of subjects, larger ICC values lead to a smaller interval length (i.e. more accurate inter-rater reliability coefficients). This suggests that raters with high extent of agreement among raters lead to inexpensive studies. This fact has an important practical implication. Giving your raters some training before the experiment is a good investment that is expected to pay off in terms of a reduced number of subjects required to achieve a certain precision level.

Our investigation has revealed that for a given number of ratings per rater, having more than 5 raters will not improve the precision of the inter-rater reliability in any meaningful way. Once you have recruited 5 raters as part of the experiment, the precision of your inter-rater reliability coefficient will improve faster with a larger number of subjects than with a larger number of raters. Unless the cost of recruiting subjects exceeds that of recruiting raters, I would recommend adding more subjects once the number of raters reaches 5. You may want to review all 4 graphs to find the combination of number of raters and subjects that gives you the desired interval length.

Just looking at figure 9.4.1 for example, you can see that to be able to determine how many raters you need requires you to have two things:

(i) a predicted (or anticipated) value for your inter-rater reliability coefficient,

[14] These options involve reducing the number of subjects while increasing the number of trials, and will necessarily result in a loss of precision for the inter-reliability coefficient. Therefore, a compromise is necessary between the number of subjects and the number of trials.

(*ii*) the desired length for your 95% confidence interval.

Let us us review these two requirements, and see how convenient it is to meet them. A predicted inter-rater reliability coefficient is often obtained from one of two sources. These sources are the pilot and previous studies. The pilot study, often based on an arbitrarily small number of subjects and raters, will yield a crude estimate of the inter-rater reliability, and give you a first look at the extent of agreement among raters. If you cannot obtain any anticipated ICC value, one possibility is to determine your sample size based on a moderate hand-picked initial value, such as 0.5. The only adverse consequence of your choice of value may be an experiment that will not produce and inter-rater reliability as accurate as the design has promised.

As an example, suppose a hospital emergency department is designing an inter-rater reliability study to evaluate the extent of agreement among physicians with respect to the measurement of stroke volume (SV) and heart rate (HR) of patients using the Impedance Cardiography (ICG) technology. We also assume that previous studies revealed an inter-rater reliability measured by the ICC to be 0.86 for SV and 0.8 for HR. We know that the emergency department wants to have only two physicians participate in the experiment, but does not know how many patients should be recruited. You can use Figure 9.4.1 to resolve this problem as follows:

(*a*) Compute $0.4 \times$ ICC for both variables to obtain the desired 95% confidence interval length. This leads to 0.344 and 0.32 for SV and HR respectively.

(*b*) Using Figure 9.4.1 (for 2 raters), you can see that the second curve from the bottom is associated with an ICC value of 0.85, which the closest you can get to 0.86 the anticipated ICC for the SV variable. Therefore, this curve will be used to determine the required sample size for the SV variable. The same procedure allows you to determine that the third curve from the bottom (associated with ICC of 0.8) is the one to use for obtaining the required number of subjects for the HR variable.

(*c*) Using the "ICC=0.85" curve, you can see that approximately 15 subjects will be sufficient to yield a 95% confidence interval with length that close to 0.344. Likewise, using the "ICC=0.80" curve, it appears that approximately 25 subjects will yield a 95% confidence interval with length that is close to 0.32.

Since SV and HR measurements must be taken during the same inter-rater reliability experiment, the right approach would be to recruit a total of 25 patients. While all of them will provide HR measurements, only 15 (ideally randomly chosen) will provide SV measurements.

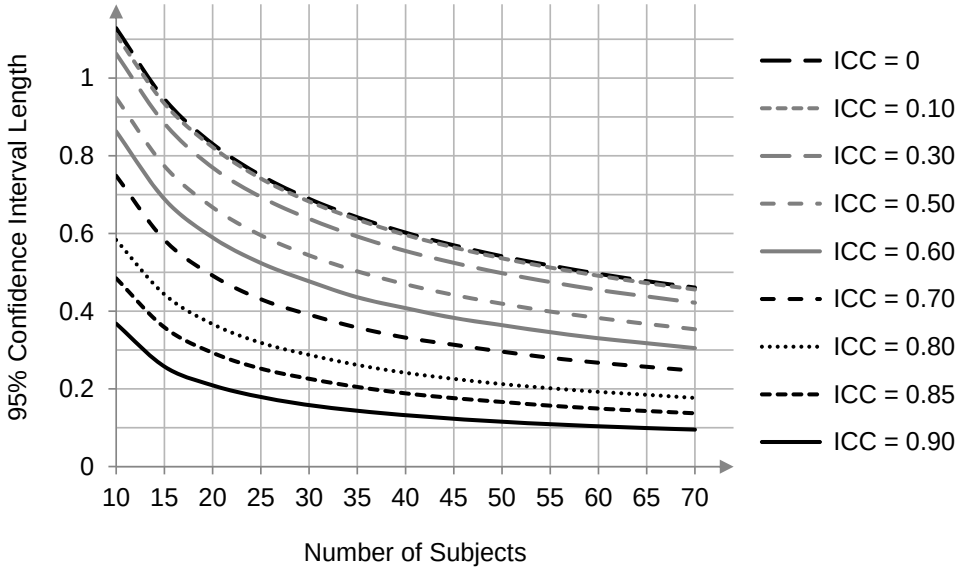

Figure 9.4.1: Expected length of the 95% confidence interval as a function of the number of subjects based on 2 raters.

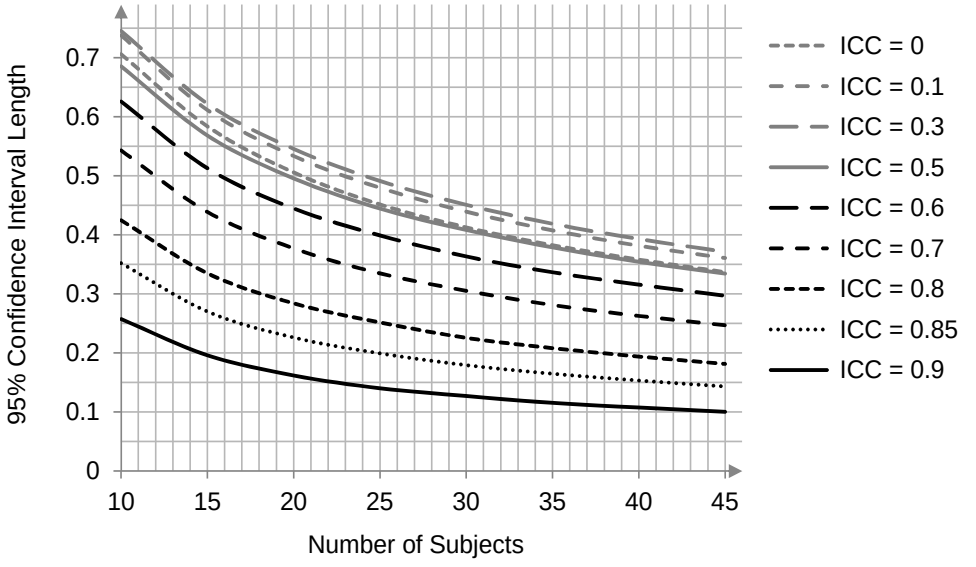

Figure 9.4.2: Expected length of the 95% confidence interval as a function of the number of subjects based on 3 raters.

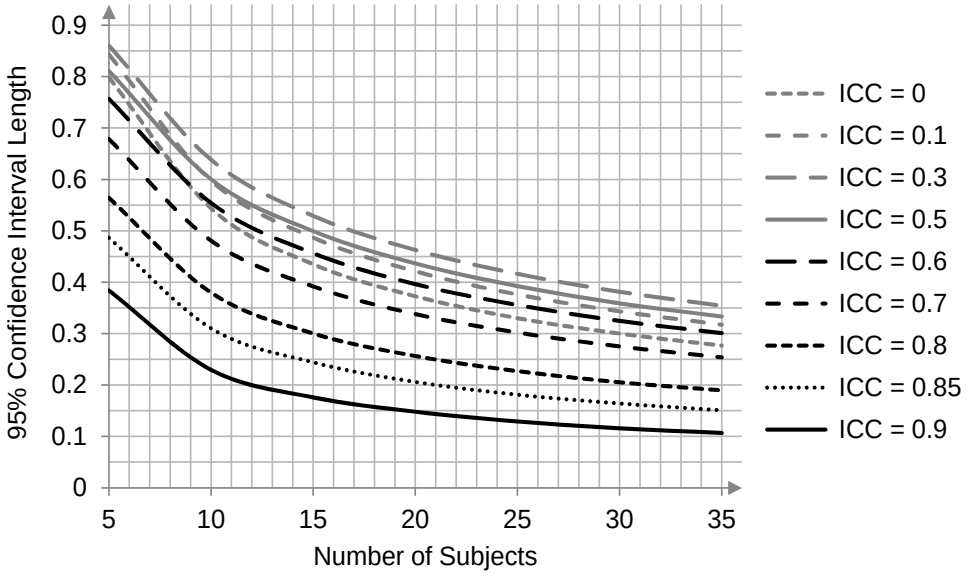

Figure 9.4.3: Expected length of the 95% confidence interval as a function of the number of subjects based on 4 raters.

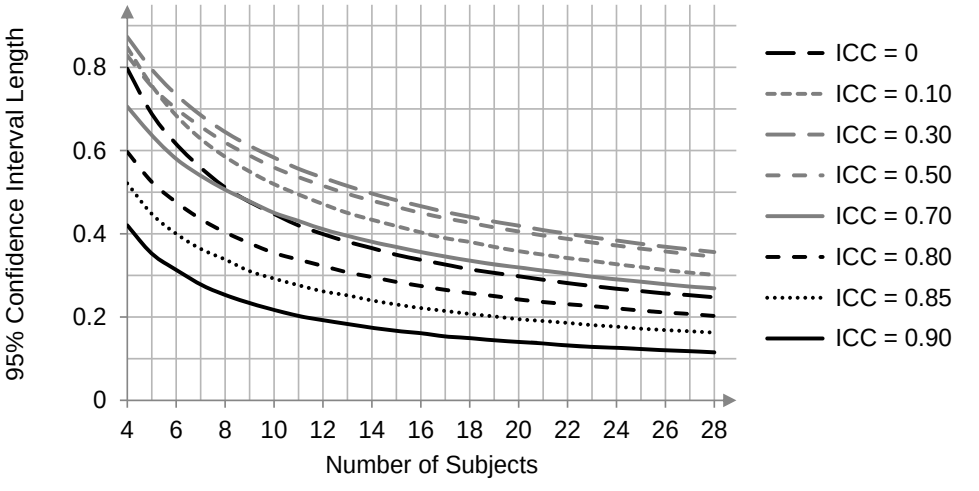

Figure 9.4.4: Expected length of the 95% confidence interval as a function of the number of subjects based on 5 raters.

9.4.2 *Sample Size Calculations for Intra-Rater Reliability Studies*

We just learned from the previous section that for a given number of ratings, using more than 5 raters will generally not improve the precision of the inter-rater reliability coefficient much. The situation is different when designing an intra-rater reliability study. More raters will generally improve the precision of the intra-rater reliability coefficient in a way that most practitioners will appreciate. The figures 9.4.8 through 9.4.9 can assist you in determining the appropriate number of raters, subjects, and trials when designing an intra-rater reliability study. Let us now see how these graphs could be used.

All 12 figures (9.4.8 through 9.4.19) reveal that in general you will need 4 or 5 trials to obtain the most accurate intra-rater reliability coefficient for a given number of raters and a given number of ratings per rater. That is for 2 raters and 20 ratings per rater for example, the optimal design consists of using 5 subjects and 4 trials or 4 subjects and 5 trials (see Figure 9.4.11). Using only 2 trials can be particularly damaging if the raters involved in the experiment have low reproducibility. However, if you know from a pilot study or from previous studies that the raters used have high reproducibility, but simply want to confirm it with a formal study, you may achieve good results using only 2 trials. If the subjects are human subjects who must endure some discomfort during the trial, then you will want to limit the number of trials per subject and increase the number of subjects instead. Since all study designs generally face specific challenges, the researcher will want to determine the best way to use these graphs.

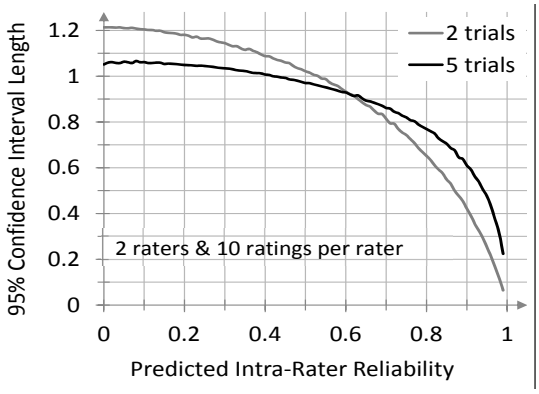

Figure 9.4.8: Interval length by intra-rater reliability (2 raters & 10 ratings per rater).

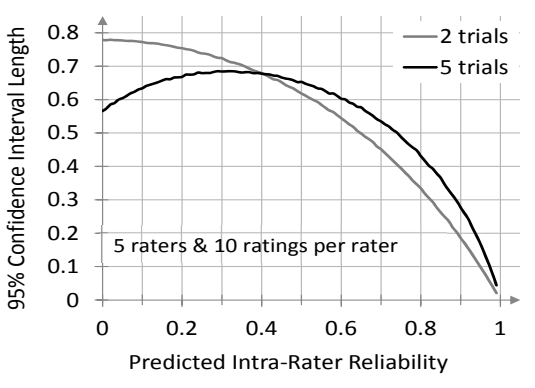

Figure 9.4.9: Interval length by intra-rater reliability (5 raters & 10 ratings per rater).

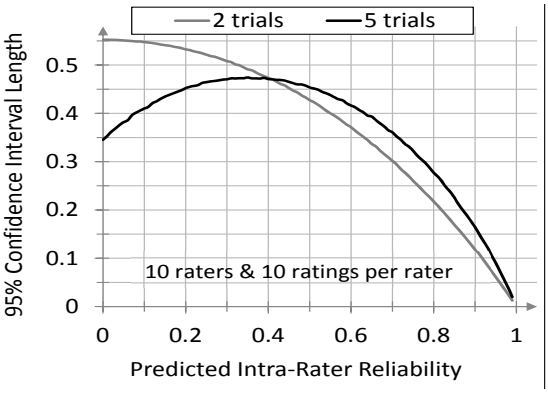

Figure 9.4.10: Interval length by intra-rater reliability (10 raters & 10 ratings per rater).

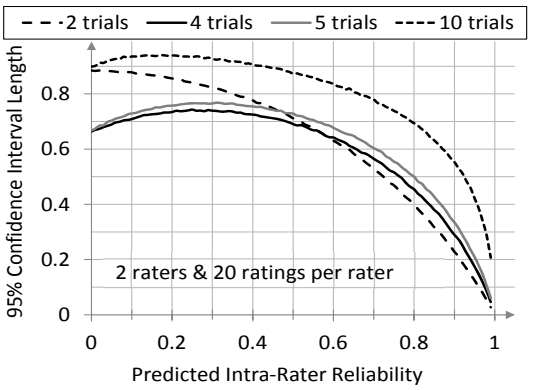

Figure 9.4.11: Interval length by intra-rater reliability (2 raters & 20 ratings per rater)

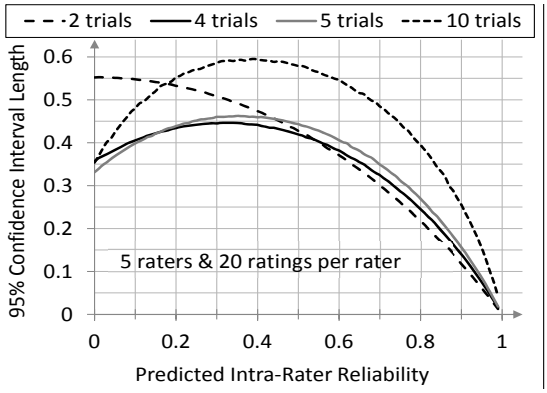

Figure 9.4.12: Interval length by intra-rater reliability (5 raters & 20 ratings per rater).

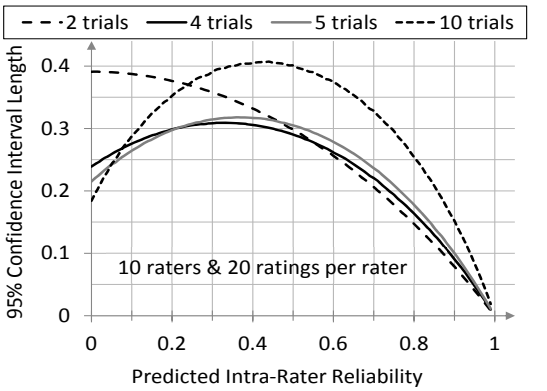

Figure 9.4.13: Interval length by intra-rater reliability (10 raters & 20 ratings per rater).

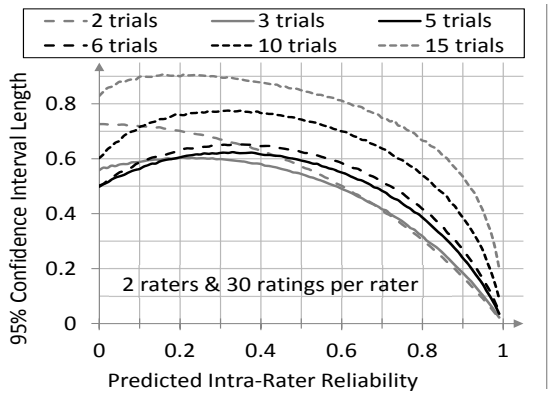

Figure 9.4.14: Interval length by intra-rater reliability (2 raters & 30 ratings per rater).

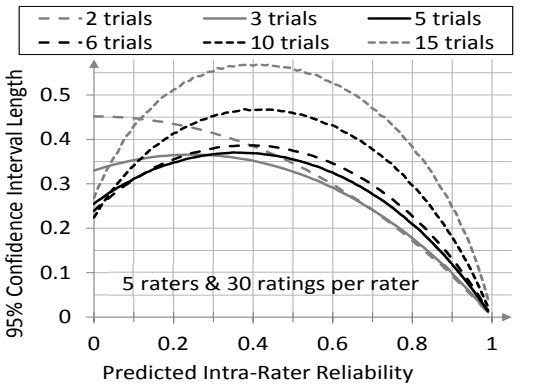

Figure 9.4.15: Interval length by intra-rater reliability (5 raters & 30 ratings per rater).