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 6.10 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:

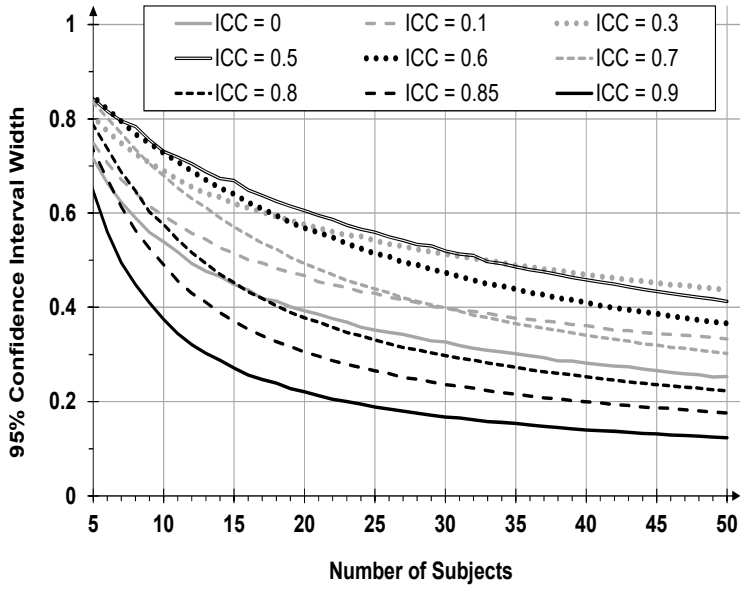

Figure 6.10: 95% C.I. Length by Number of Subjects for 2 Raters

- 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 reaches its maximum value when ICC is in the neighborhood of 0.5, and reaches its minimum value as ICC increases towards 1. Consequently, your experiment will yield a more accurate inter-rater reliability coefficient estimate if the extent of agreement among raters exceeds is high (i.e. exceeds 0.85). This is logical because raters who agree are homogeneous with respect to the way they rate subjects. There-

fore, 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. However, if the ICC is in the neighborhood of 0.5, then achieving a desired precision level may become an impossible task. In other words, if the raters agree near as much as they disagree, then obtaining an accurate measure of agreement can be difficult.

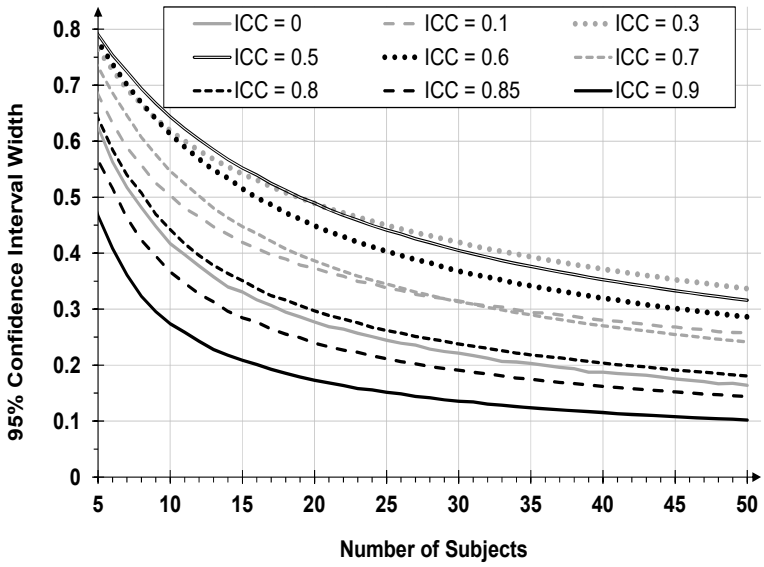

Figure 6.11: 95% C.I. Length by Number of Subjects for three Raters

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 6.10, 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. $0.4 \times$ 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 $\text{ICC}_0 = 0.80$. This leads to our desired 95% confidence interval length of $0.4 \times \text{ICC}_0 = 0.4 \times 0.80 = 0.32$. For simplicity, you may consider your target interval length to be 0.30.

Note that this initial ICC_0 value could be different (ideally not by too much) from the actual and unknown "true" ICC you are after, and which is used in Figure

6.10. Figure 6.10 reveals that if the "true" ICC = 0.85 (see the second curve from the bottom), then you will need about $n = 20$ subjects to achieve your target interval length of 0.30. If ICC = 0.80 (see the third curve from the bottom), then you will need about $n = 30$ subjects to achieve the precision goal. Note that with ICC = 0.90, only $n = 13$ subjects are needed. This discussion gives a rough idea of the number of subjects you will need to achieve the target confidence interval length.

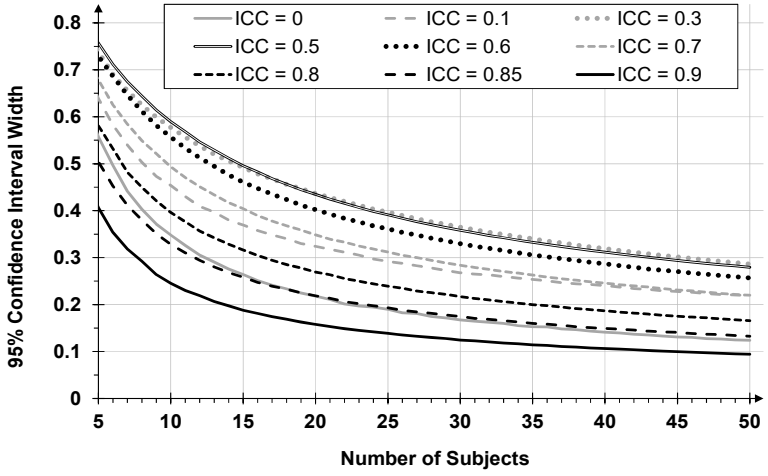

Figure 6.12: 95% C.I. Length by Number of Subjects for 4 Raters

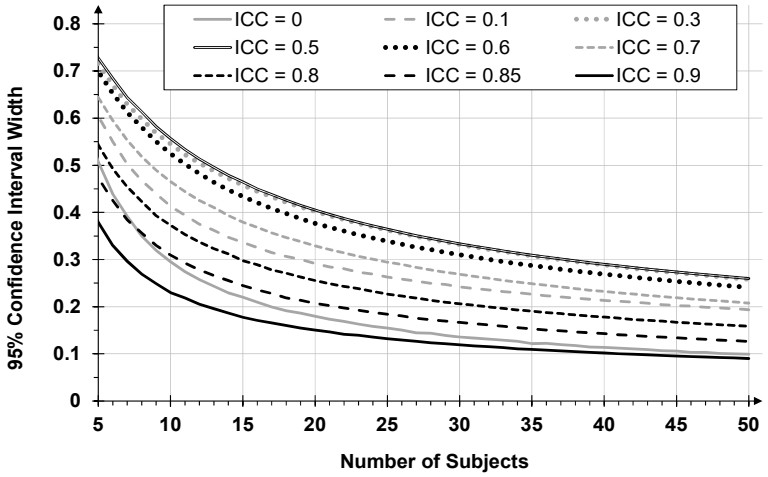

Figure 6.13: 95% C.I. Length by Number of Subjects for 5 Raters

Figure 6.11 shows that with 3 raters in your experiment and the same target confidence interval length of 0.30, you will need about 9 subjects if ICC=0.90, will need 15 subjects if ICC = 0.85, and will need 20 subjects if ICC = 0.80. Four and five raters will require even fewer subjects to achieve the precision level as seen in Figures 6.12 and 6.13, although beyond 4 raters, the gain in precision become marginal.

NUMBER OF RATERS AND SUBJECTS - WITH REPLICATION

We just saw that an inter-rater reliability experiment based on 2 raters requires 30 subjects to achieve a 95% confidence length of 0.30 if ICC = 0.80. The question I want to answer now is the following: "Can one use fewer subjects and 2 trials per subject and still be able to achieve a reasonable interval length?" Figures 6.14 through 6.21 can help answer this question.

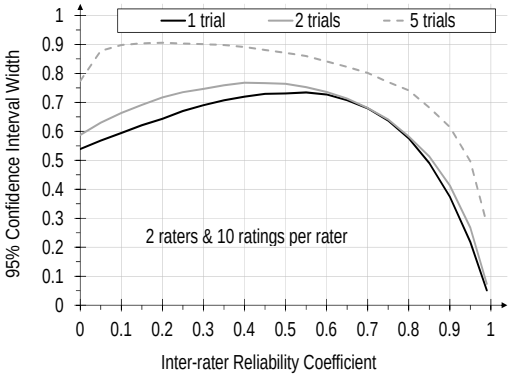

Figure 6.14: Interval length by inter-rater reliability & number of trials (2 raters & 10 ratings per rater).

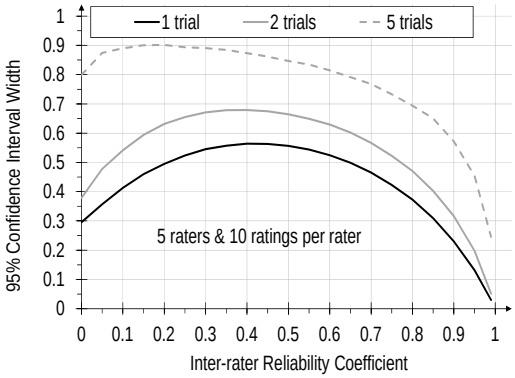

Figure 6.15: Interval length by inter-rater reliability & number of trials (5 raters & 10 ratings per rater).

Figure 6.14 shows the 95% confidence interval length as a function of the "true" inter-rater reliability coefficient and the number of trials, for a study that is based on 2 raters and 10 ratings per rater. Since the continuous black curve is associated with 1 trial, it represents an inter-rater reliability experiment with 2 raters, 10 subjects and 1 trial (i.e. $10 \times 1 = 10$ ratings per rater). Likewise the gray curve describes the relationship between the precision of the estimated inter-rater reliability and its actual value for an experiment based on 2 raters, 5 subjects and 2 trials (i.e. $5 \times 2 = 10$ ratings per rater). As for the dotted gray line it refers to an experiment with 2 raters, 2 subjects and 5 trials (i.e. $2 \times 5 = 10$ ratings per rater). Figures 6.14 through 6.21 all indicate that the smaller the number of trials, the more accurate the estimate of the inter-rater reliability coefficient. Using more than 3 trials will likely result in a substantial loss of precision for a given number of ratings per subject.

Figure 6.18 shows that an inter-rater reliability experiment that is based on 2 raters, 6 subjects and 5 trials (i.e. $6 \times 5 = 30$ ratings per rater) will yield a 95% confidence interval length of about 0.45 if ICC = 0.80. We saw earlier that 30 subjects and 1 trial will yield an interval length of 0.30 when ICC = 0.80. Consequently, reducing the number of subjects and increasing the number of trials for the same number of ratings per rater has a negative impact on the precision of estimates that can be substantial.

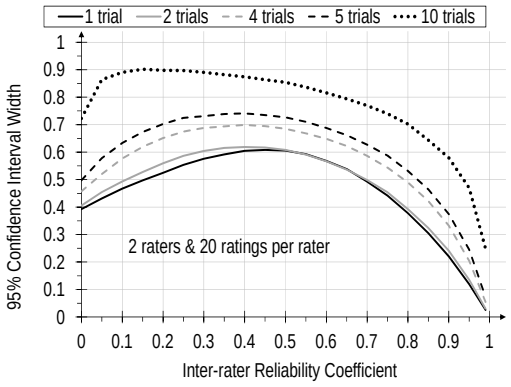

Figure 6.16: Interval length by IRR & # of trials (2 raters & 20 ratings per rater).

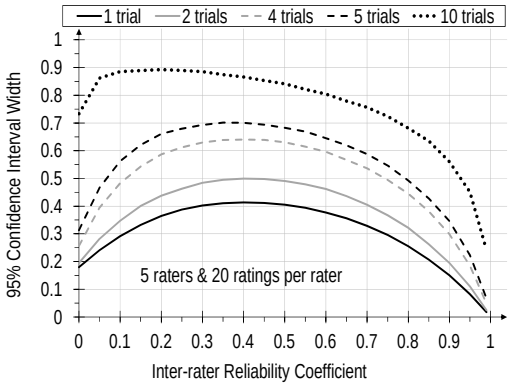

Figure 6.17: Interval length by IRR & # of trials (5 raters & 20 ratings per rater).

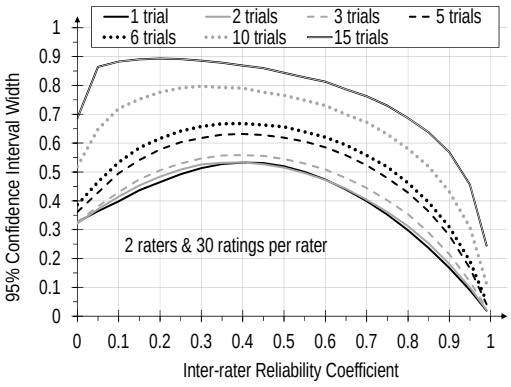

Figure 6.18: Interval length by IRR & # of trials (2 raters & 30 ratings per rater).

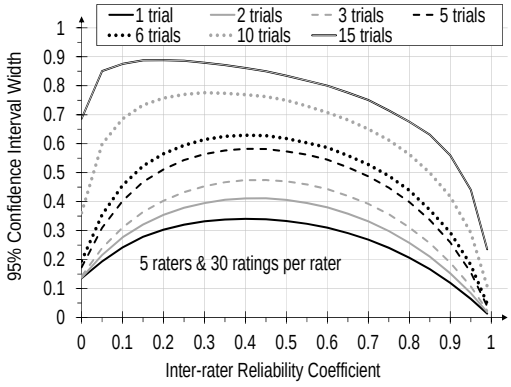

Figure 6.19: Interval length by IRR & # of trials (5 raters & 30 ratings per rater).