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 (Figures 5.11-5.14) to find the combination of number of raters and subjects that gives you the desired interval length.

Just looking at Figure 5.11 for example, you can see that to be able to determine how many subjects you need requires you to have two things:

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

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

Let us review these two requirements and see how convenient it is to satisfy them. A predicted inter-rater reliability coefficient is often obtained from one of two possible sources. These sources are the pilot and past studies. The pilot study, generally based on an arbitrarily small number of subjects and raters, will yield a rough inter-rater reliability estimate, 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.

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 the 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 0.88 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 5.11 to resolve this problem as follows:

(*a*) Compute 0.4 × ICC for both variables to obtain the desired 95% confidence interval length. This leads to 0.352 and 0.32 for SV and HR respectively.

(*b*) Using Figure 5.11 (for 2 raters), you can see that the second curve from the bottom is associated with an ICC value of 0.90, which the closest you can get to 0.88 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=0.90” curve, you can see that approximately 11 subjects will be sufficient to yield a 95% confidence interval with length that is close to 0.352. Likewise, using the “ICC=0.80” curve, it appears that approximately 32 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 32 patients. While all of them will provide HR measurements, only 11 (ideally randomly chosen) will provide SV measurements.

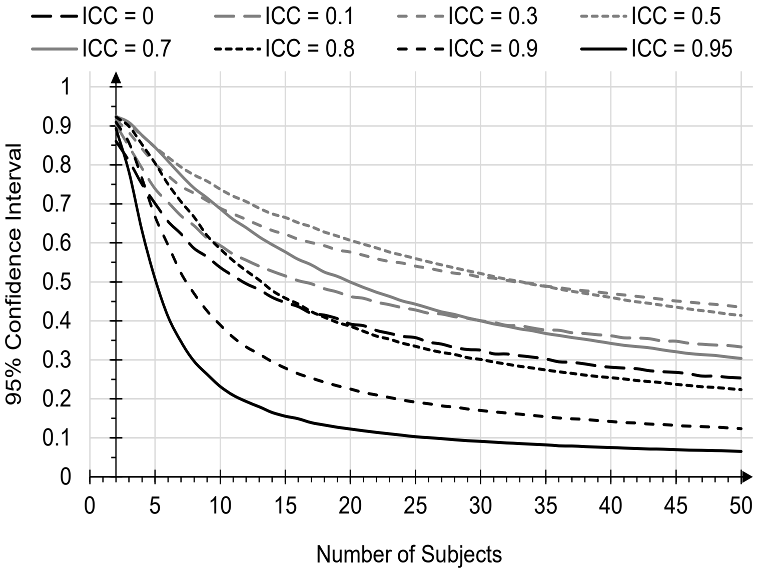

Figure 5.11: Expected length of the 95% confidence interval as a function of the number of subjects based on 2 raters.