Chapter 8 Supplement

A Closer Look at the Number of Trials

In this supplement, I will provide a more detailed discussion of how the number of trials and the resulting SNR will impact your results. If you are using conventional statistical approaches—such as *t* tests, analyses of variance, correlations, or multiple regression—the desired SNR is related to your *statistical power*. Statistical power is the probability that the data from a given experiment will allow you to reject the null hypothesis if the null hypothesis is false. In other words, statistical power is the likelihood that a real difference between conditions or groups will lead to a statistically significant effect in your experiment. This is, of course, what really matters to you in the long run (i.e., in terms of finding the truth, getting your experiments published in top journals, getting a good faculty position, getting grants, and living happily ever after).

The one simple fact that I can tell you about the number of trials in an average is this: The more trials you average together in your ERPs, the less random variation will occur, and the greater your statistical power will be. That is, more trials mean a greater likelihood of obtaining a significant effect (if a real effect is present). In an ideal world, you would want an infinite number of trials. In the real world, however, increasing the number of trials has its costs. You therefore need to ask yourself whether the benefits of increasing the number of trials (i.e., the increased statistical power) is worth these costs.

Statistical Power

The effect of the number of trials on statistical power is complicated by the fact that other factors also influence power. Three main factors influence power in a typical ERP experiment. The first factor is the size of the effect: All else being equal, a 5-µV difference between conditions or groups is more likely to be significant than a 1-µV difference. The second factor is the number of subjects: All else being equal, an effect is more likely to be significant with a larger sample size than with a smaller sample size. The third factor is the amount of variability across subjects: All else being equal, an effect is more likely to be significant if the variability across subjects is smaller.[1]

The variability across subjects can be subdivided into two subcomponents. The first is *measurement error*, which is the variation that is introduced during the measurement process. The second is *true score variance*, which is the variability across individuals that would be obtained if each subject could be measured without error. Imagine, for example, that we are measuring height in a sample of 100 people. If we have a stretchy rubber measuring tape, we would not get the same measurement from an individual if we measured multiple times. And if we measured only once from each individual, we would have a poor measure. This error in measurement would lead us to a larger estimate of the variability across individuals than we would get if we used a more stable measuring tape. Measurement error is closely related to *reliability*, which is the ability to obtain the same measured value when we measure the same thing multiple times. If we have low measurement error, then we will have high reliability. In the case of ERPs, the number of trials in each average will influence the measurement error and the reliability. All else being equal, the more trials we average together, the lower the measurement error will be, and the more reliable the measure will be.[2]

A More Formal Analysis

Figure S8.1 illustrates how increasing the number of trials in each average reduces the overall variance and therefore improves your statistical power, but in a way that depends on the amount of true score variance across subjects. The *Y* axis in this figure shows the total variance across subjects. When you compute a *t* or *F* value, this number (or something like it) is in the denominator. Therefore, as the total variance across subjects gets bigger, your *t* or *F* value gets smaller, and your *p* value gets worse. Thus, you want the total variance to be as small as possible (all else being equal, of course). The *X* axis in the figure represents the number of trials that are being averaged together for each subject. Three different situations are shown in the three lines in the figure. In one case (indicated by the squares), the single-trial EEG is very noisy, but the true variance across subjects is small (i.e., the variance across subjects would be small if we eliminated measurement error). With a small number of trials, most of the variability across subjects is due to the EEG noise, because this hasn't been reduced very much by averaging and because there isn't much variance due to real differences among subjects. As the number of trials increases, the measurement error decreases, and the overall variance across subjects drops a great deal (reaching an asymptote at the true score variance). This is the situation in which you gain the most by increasing the number of trials. In other words, if a large portion of the variation across subjects is due to noise in the averaged waveforms, then increasing the number of trials will lead to a substantial decrease in overall variance and an increase in statistical power (all else being equal).

In contrast, if the variance due to EEG noise is small relative to the real differences across subjects (triangles in figure S8.1), then increasing the number of trials has a relatively modest effect on the overall variance (and therefore on statistical power). For example, when the variance due to EEG noise is 10 and the true score variance is 30, the overall variance with one trial is 40 (because variances simply add; see box 8.1), and the overall variance with an infinite number of trials is 30 (because you are still left with the true variance after all measurement error has been removed). You will still get an improvement in statistical power by increasing the number of trials in this case. However, the gain in statistical power will be relatively modest compared to what you would achieve by increasing the number of trials in a case with lower true score variance. In this situation, your best bet might be to increase the number of subjects or design the experiment in a way that increases the size of the effect.

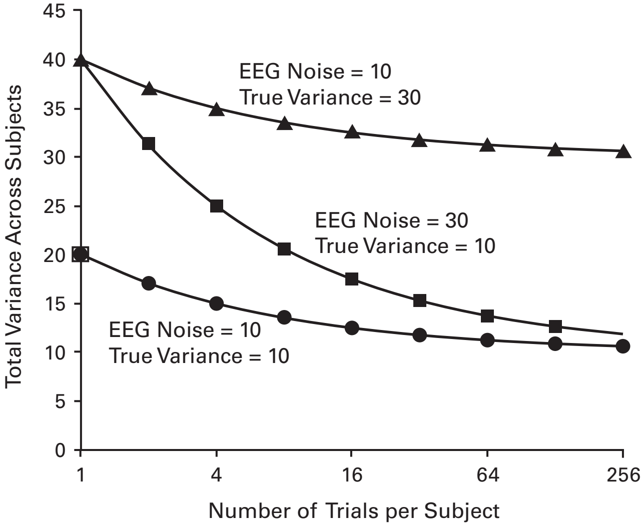

Figure S8.1
Effects of the number of trials being averaged together for each subject on the total variance across subjects. This is for a hypothetical experiment in which the amplitude of some component is measured from averages of *T* trials in each of *N* subjects and then the variance is calculated from these measures. Different curves are shown for different combinations of true score variance (the variance across subjects that would be obtained with an infinite number of trials per average) and EEG noise (the variance across subjects that would be obtained if each of the measures were obtained from a single trial in each subject if all subjects had identical ERPs except for the EEG noise).

Figure S8.1 also shows a case in which the EEG noise and true score variance are equal (indicated by the circles). In this case, increasing the number of trials from one to infinity will cut the overall variance in half, and increasing from one to 64 almost cuts it in half. I suspect that this is the typical situation for most ERP experiments.

Note that if you do everything that was discussed in chapter 5 to record clean EEG data, the variance due to EEG noise will become smaller relative to the true score variance. This means that you do not need to include as many trials to obtain a good overall variance.

A number of papers have been published addressing the number of trials needed to obtain “adequate” reliability for ERPs (e.g., Olvet & Hajcak, 2009; Pontifex et al., 2010; Marco-Pallares, Cucurell, Muente, Strien, & Rodriguez-Fornells, 2011). In my view, these studies are only narrowly useful because the impact of the number of trials on statistical power will depend on the amount of variance due to EEG noise and the amount of true score variance. The amount of variance due to EEG noise will vary across laboratories (e.g., due to variations in the use of procedures that yield clean EEG recordings), across tasks (e.g., due to noise that may vary across tasks, such as alpha EEG oscillations and muscle noise), and across subject populations (e.g., due to variations between groups of people in factors like skin potentials and movement artifacts). The amount of true score variance will also vary across subject populations (e.g., patient groups are usually more heterogeneous than control groups). Moreover, statistical power also depends on the effect size and the number of subjects. Thus, the number of trials needed to obtain a given level of statistical power or reliability in one study may be very different from the number needed in another study.

Notes

1. If this is not obvious to you, I would recommend refreshing your memory about statistics by reading about *t* tests in any good introductory statistics textbook. Also, when I speak of variability here, I am referring to unexplained variability (i.e., variability that cannot be factored out by a covariate such as age). If you are interested in examining individual differences among subjects rather than averages across subjects, increasing the number of trials is especially valuable because it brings you closer to the true value for each subject.

2. When talking about the number of trials, I keep using the phrase "all else being equal." The reason is that increasing the number of trials may have psychological effects in addition to statistical effects. If you run the same condition for 10,000 trials, brain activity during the last 100 trials is unlikely to be the same as brain activity during the first 100 trials. For example, I typically find that behavioral performance improves over the first 10 min or so, then levels off, and then begins to decline after about 45 min (for relatively repetitive and boring tasks). These changes are modest, and I don't think they have much impact on my results. However, I try not to run experiments that are so long that subjects become stressed-out zombies.