Chapter 9 Supplement

Comparing Conditions with Different Numbers of Trials or Different Levels of Noise

In many ERP experiments, the averaged ERP waveforms will be based on a greater number of trials in some conditions than in others. In an oddball experiment, for example, fewer trials will be averaged together for the rare stimuli than for the frequent stimuli. All else being equal, waveforms created by averaging together a larger number of trials will have less noise than waveforms created by averaging together a smaller number of trials. Consequently, a comparison between waveforms from conditions with different numbers of trials will typically be a comparison between waveforms with different amounts of noise. The noise level can also differ across conditions for other reasons. For example, waveforms from different groups of subjects may have different noise levels if one group is more prone to artifacts (e.g., skin potentials, movement artifacts) than the other group, even if both groups have the same number of trials per waveform. For the issues described in this supplement, it doesn't matter whether the increased noise level is due to differences in the number of trials or some other factor.

Are differences in noise level a problem? The answer can be "yes" or "no" depending on whether you are measuring peak amplitude or mean amplitude. That is, differences in noise level may or may not bias the results of a study, and this depends on whether the waveforms are quantified with peak amplitude or mean amplitude measures. Generally speaking, peak amplitude is a biased measure that will tend to lead to larger values in conditions with greater noise (e.g., due to smaller numbers of trials contributing to the averaged waveforms), but mean amplitude is an unbiased measure that you can legitimately use even when noise levels differ across conditions. However, unpacking these ideas takes a bit of work (for additional discussion and detailed simulations, see Clayson, Baldwin, & Larson, 2013).

Biased versus Unbiased Measures

First, it is important to be clear what "bias" means. Bias is a consistent shift in one direction. For an unbiased measure, the average value over an infinite number of experiments would be equal to the true value. For a biased value, the average value over an infinite number of experiments would be different from the true value.

To make this clear, let's not think about ERP waveforms yet. Instead, let's think about a simple value that you could measure from a set of individuals: height. If I take the mean height of a sample of 20 individuals from an infinitely large population, this *sample mean* will be different from the mean of the infinitely large population (the *population mean*). However, if we collect many different samples of 20 individuals, sometimes the mean of a sample will be greater than the population mean and sometimes it will be smaller than the population mean. If we collect an infinite number of samples of 20 subjects (with 20 different subjects in each sample), and we calculate the sample means from each of those samples, the average of all the sample means will be equal to the population mean. This is what it means for a measure to be unbiased. It may be a noisy, inaccurate measure in any given sample, but it does not have a systematic tendency to be different from the true value in a particular direction.

The mean height is an unbiased measure irrespective of the number of individuals in each sample. The mean height of one particular sample of five subjects will usually be farther from the population mean than the mean of a particular sample of 20 subjects. But the mean of five individuals would not be expected to be consistently greater than (or consistently less than) the population mean.

Imagine that we have two populations, A and B, and the average height of each population is 200 cm. If I do an experiment in which I measure the height of five people from population A and 20 people from population B, this difference in sample size will not tend to make the mean of group A larger than the mean of group B. Half the time, such an experiment will lead to a greater average height in group A than group B, and half the time we will get a greater average height in group B than in group A. The sample mean of a group of five individuals will typically be farther away from the true population mean than the mean of a group of 20 individuals, but not in a particular direction. This is what it means to be unbiased. An unbiased measure isn't necessarily a good measure, but at least it doesn't tend to produce an effect in a particular dimension.

As an example of a biased measure, consider finding the tallest person in a sample. The more people are in a sample, the more likely it would be that someone in that sample is really tall. For example, if we found the tallest person in a sample of five subjects from group A and the tallest person in a sample of 20 subjects from group B, it's more likely that the tallest person would be in group B. Sometimes the tallest person would be in group A, but because there are more opportunities to find a really tall person in a larger sample, the tallest person will usually be in group B. This is an example of a biased measure (although in this case it is biased to have a larger value in the largest group, which is the opposite of what happens with peak amplitude).

Peak Amplitude as a Biased Measure

Now let's turn to see how peak amplitude is a biased measure in ERP experiments. Figure S9.1 shows an artificial ERP waveform (one cycle of a sine wave) with no noise and the same waveform with 60-Hz noise added. As you can see, the peak of the noisy waveform is greater than the peak of the clean waveform. The mean voltage of the noise that was added to the waveform is zero, but some points are positive and others are negative, and the peak measure finds the most positive value (assuming we are searching for a positive peak). That's what makes peak amplitude a biased measure: Because it finds the most extreme positive value, and the extremes are bigger when the waveform is noisier, the peak tends to be larger for conditions with noisier waveforms. Note that this would be true with a negative peak as well: When the waveform is noisier, the most

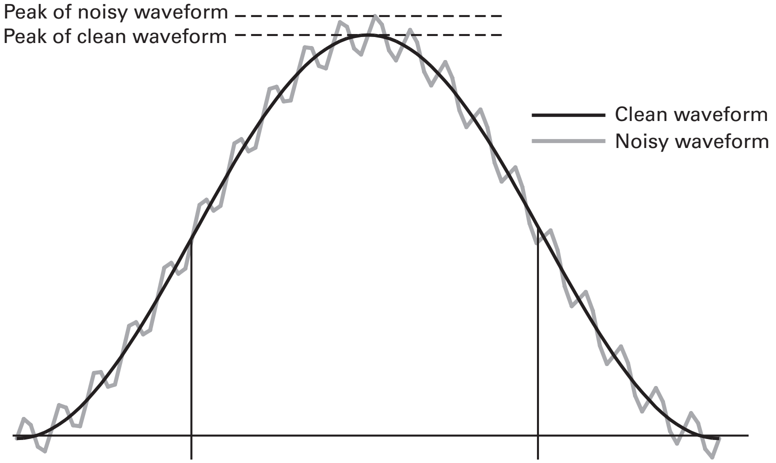

Figure S9.1
Example of a clean waveform and a noisy waveform (created by adding 60-Hz noise to the clean waveform). The broken lines show the peak of the clean and noisy waveforms. Note that the peak is higher for the noisy waveform than for the clean waveform, which will almost always be true when one of two otherwise-identical waveforms has greater high-frequency noise. In contrast, the mean amplitude over a broad time period (indicated by the shaded region) is generally less influenced by the noise and is just as likely to be smaller or larger for the noisier of two otherwise-identical waveforms. In other words, mean amplitude is not biased by the amount of noise, whereas peak amplitude is biased to be larger for noisier waveforms.

extreme negative value will tend to be more negative, just as the most extreme positive value will tend to be more positive (for simulations showing this bias, see Clayson et al., 2013).

You should keep in mind that a bias is a tendency, not an inevitability. That is, the peak of a noisier waveform will not be greater than the peak of a cleaner waveform in every case. However, it will be greater more often than not. Thus, when you compare two conditions in which the noise levels are different (usually due to differences in the number of trials), a finding of a greater peak amplitude in the condition with noisier waveforms could reflect this bias rather than a real difference between the conditions. However, if the peak amplitude is smaller in the condition with noisier waveforms, this could not be a direct consequence of the greater noise level in this condition. For example, a *smaller* P3 amplitude for rare trials than for frequent trials could not be explained by the smaller number of trials contributing to the waveform for the rare trials. However, the smaller number of trials does distort the apparent size of the effect, making it smaller than it should be. Thus, if you want an accurate assessment of the difference in amplitude between two conditions, you should make sure that you are measuring in a way that is not biased by the number of trials, even if this bias goes in the opposite direction of the observed effect.

Mean Amplitude as an Unbiased Measure

Mean amplitude is less influenced by high-frequency noise than peak amplitude. In figure S9.1, for example, the positive and negative bumps caused by the 60-Hz noise will cancel each other out, and the mean voltage over the measurement window will be quite close in the noisy and clean waveforms. Mean amplitude is also an unbiased measure. This means that the noise will sometimes cause the mean amplitude to be larger for the noisy waveform than for the clean waveform and sometimes cause it to be smaller, with an equal likelihood of larger and smaller values for the noisy waveform (for simulations, see Clayson et al., 2013).

If you measure the mean amplitude from a noisy waveform (e.g., a waveform created by averaging together a small number of trials), and you also measure the mean amplitude from a clean waveform (e.g., a waveform created by averaging together a large number of trials), you will find that the mean amplitude will tend to be farther away from the true amplitude for the noisy waveform than for the clean waveform. Consequently, if you measure the mean amplitude from many subjects, you will have greater variance for noisy waveforms than for clean waveforms, and this will reduce your statistical power. However, the mean amplitude across subjects will sometimes be larger for the noisy waveforms and sometimes be smaller for the noisy waveforms, with equal likelihood. More noise increases the Type II error rate (the probability that you accept the null hypothesis when there is a real effect), but it does not increase the Type I error rate (the probability that you reject the null hypothesis when the null hypothesis is true). In other words, more noise reduces your statistical power, but it does not cause you to incorrectly reject the null hypothesis more than 5% of the time. This is what it means to say mean amplitude is not biased by the noise level. In contrast, peak amplitude will increase the likelihood that you conclude that there is a real difference between a noisy condition and a clean condition when the true amplitude is actually equivalent across conditions.

What Happens in ERP Experiments with Different Numbers of Trials per Condition?

Imagine a within-subjects experiment ($N = 10$ subjects) in which condition A has 20 trials averaged together for each subject's waveform, and condition B has 100 trials averaged together for each subject's waveform. And imagine that there is no real difference between the single-trial waveforms in conditions A and B (i.e., the null hypothesis is true). In this imaginary experiment, the averaged ERP waveforms in conditions A and B will be equivalent, except that each will have random noise added to them, with greater noise in condition A than in condition B.

Because the noise is greater in condition A, the peak amplitude measured from the averaged ERP waveforms will tend to be greater in condition A than in condition B. The peak amplitude won't be greater for condition A in every subject, but it will typically be greater in condition A than in condition B for most subjects. And if you repeated this experiment an infinite number of times, most of the experiments would have a larger average amplitude (across subjects) in condition A than in condition B. Moreover, the probability of finding a significant difference between conditions A and B would be greater than 0.05, even though there is no real difference in amplitude between conditions. Again, this is what it means to say that peak amplitude is biased by the noise level.

If we measured mean amplitude rather than peak amplitude from the averaged ERP waveforms, we would not have these problems. With mean amplitude, we would expect that half of the subjects would have a greater amplitude in condition A and half would have a greater amplitude in condition B (assuming that the null hypothesis is true). Moreover, the probability of finding a significant difference between conditions A and B over an infinite number of experiments would be exactly 0.05. That is, the difference in the number of trials between conditions A and B would not artificially increase the probability that we conclude that the conditions are different. In other words, it does not increase the probability of a Type I error.

What Should You Do if You Have Different Numbers of Trials per Condition?

So, what should you do when analyzing the data from an actual experiment? One common approach is to randomly select a subset of the trials from the condition with more trials so that the number of trials contributing to the averaged ERP waveforms is the same across conditions. If, for some reason, you must measure peak amplitude, this is a reasonable approach.

However, if you are measuring mean amplitude, this would be foolish because it throws away statistical power. If you are measuring mean amplitude, then having 20 trials per waveform in one condition and 100 trials per waveform in another condition will give you a lower Type II error rate (lower likelihood of accepting the null hypothesis when it is false) than when you have 20 trials per waveform in each condition, without increasing the Type I error rate (no change in the likelihood of rejecting the null hypothesis when it is true). Thus, if you are measuring mean amplitude, you can simply ignore the difference in the number of trials between conditions. This difference is not a confound: It does not bias you to find an effect if there is no real effect present. Moreover, throwing away a large number of trials to equate the number of trials per waveform reduces your statistical power but does not provide any benefit. Thus, in the vast majority of cases, you should simply measure mean amplitude and ignore the fact that the number of trials per waveform differs across conditions.