

Is it Legitimate to Compare Conditions with Different Numbers of Trials?

Steven J. Luck

UC-Davis Center for Mind & Brain

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 waveforms from the rare stimuli than for the waveforms from 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 my purposes here, 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.

What is bias?

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 mean will be different from the population mean. However, sometimes the mean of a sample will be greater than the population mean and sometimes it will be smaller than the population mean. If I do an infinite number of experiments in which I measure the mean height of a sample of 20 subjects (with a different random sample of 20 subjects in each experiment), the average over all of those experiments (the average of this infinite set of 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, 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 a given sample of 5 subjects will tend to be farther from the

population mean than the mean of a given sample of 20 subjects. But neither the mean of 5 individuals nor the mean of 20 individuals would be expected to be consistently greater than the population mean (or consistently less than the population mean). And the mean of a sample of 5 individuals will not be consistently greater than the mean of a sample of 20 individuals.

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 5 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 mean of a group of 5 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.

Peak amplitude as a biased measure

Now let's turn to ERP waveforms and see how a measure can be biased. Figure 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 extreme negative value will tend to be more negative, just as the most extreme positive value will tend to be more positive.

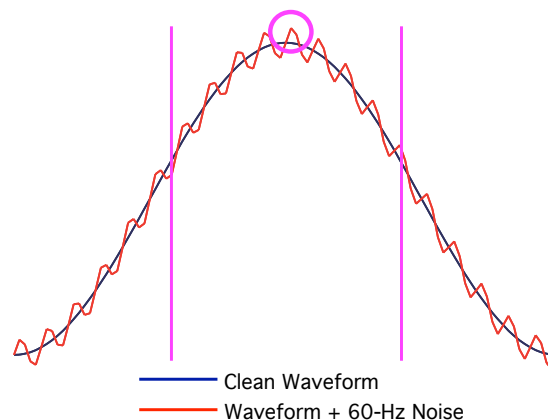


Figure 1. Example of a clean waveform and a waveform with 60-hz noise added. The pink circle shows the peak amplitude of the noisy waveform, which is greater than the peak amplitude of the clean waveform. The pink lines show the measurement window that was used to find the peak. This same window could instead be used to measure the mean amplitude (i.e., the amplitude is quantified as the mean voltage during the measurement period). The mean amplitude of the noisy waveform won't be exactly the same as the mean amplitude of the clean waveform. However, the mean amplitude of the noisy waveform is no more likely to be greater or smaller than the mean amplitude of the clean waveform. That is, mean amplitude is not biased by the noise level, whereas peak amplitude is biased to be larger for the noisier waveform.

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.

Mean amplitude as an unbiased measure

In contrast to peak amplitude, mean amplitude is an unbiased measure. In Figure 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. They won't be exactly the same, however. The important thing is 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. That is, mean amplitude is not biased (pushed in a consistent direction) by the noise level.

If you take a clean waveform and add noise to it, the mean voltage of the noisy waveform is equally likely to be greater than or less than the mean voltage of the clean waveform. If you do an infinite number of simulations in which you add random noise to a clean waveform, and measure the mean amplitude for each of these simulations, the average over this infinite set of mean amplitudes will be exactly equal to the mean amplitude of the clean waveform. In contrast, if you did the same thing with peak amplitude, the average of the infinite set of peak amplitude values would be greater for the noisy waveforms than for the clean waveforms. This is what it means to say that peak amplitude is biased by the noise level and mean amplitude is not.

What would happen in an ERP experiment 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 condition 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 .05, even though there is no real difference 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. And with mean amplitude, the probability of finding a significant difference between conditions A and B over an infinite number of experiments would be exactly .05. That is, the difference in the number of trials between conditions A and B does 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 II error.

Having more noise tends to make the average value across subjects farther away from the true value, just not in a particular direction. In other words, having fewer trials per waveform in condition A than in condition B means that the average mean value across a finite sample of subjects will be farther from the true value (the population average) in condition A than in condition B. You might think this would lead to an increase in the probability of a Type II error (rejecting the null hypothesis when it is true) compared to an experiment with equal numbers of trials per condition. But you would be wrong! Having fewer trials in a given condition increases the variance for that condition, and that will tend to decrease the p value from the statistical test (assuming that your statistical test does not assume equal variances across conditions or is robust to modest differences in variance, which is true of t-tests and F-tests). Thus, although reducing the number of trials in one condition will tend to decrease your p-value (i.e., it will decrease your statistical power), it will not increase your Type II error rate.

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 you are measuring peak amplitude (or peak latency or any other nonlinear measure), this is a reasonable approach. However, if you are measuring mean amplitude, this would be foolish because it throws away statistical power. Having 20 trials per waveform in one condition and 100 trials per waveform in another condition gives you more statistical power (better p values) than having 20 trials per waveform in each condition. Thus, if you are measuring mean amplitude, you can simply ignore the difference between condition in the number of trials per waveform. This difference is not a confound: It does not bias you to find an effect if there is no real effect present. However, throwing away a large number of trials to equate the number of trials per waveform increases the probability that you will falsely accept the null hypothesis (i.e., it increases the Type I error rate). There are many other advantages of mean amplitude over peak amplitude as well (see Chapter 6 in *An Introduction to the Event-Related Potential Technique*). 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.