10 Statistical Analysis

Overview

Once you have recorded ERP waveforms from a sample of subjects and measured the amplitude and latencies of the components of interest, it will be time to perform statistical analyses to see whether your effects are significant. In the large majority of cognitive and affective ERP experiments, the investigators are looking for a main effect or an interaction in a completely crossed factorial design, and ANOVA-based statistical analyses are therefore the dominant approach. Consequently, this is the only approach I will describe, although other approaches can be useful in some cases.

Before I begin describing how statistical analyses are applied to ERP data, I would like to make it clear that I consider statistics to be a necessary evil. We often treat the 0.05 alpha level as being somehow magical, with experimental effects that fall below $p < 0.05$ as being "real" and effects that fall above $p < 0.05$ as being nonexistent. This is, of course, quite ridiculous. The 0.05 cutoff is purely arbitrary, and if the field had chosen a standard criterion of $p = 0.06$, we would have only a slightly higher rate of false positives (accompanied by a slightly lower rate of false negatives). Moreover, the assumptions of ANOVA are violated by almost every ERP experiment, so the p values that we get are only approximations of the actual probability of a Type I error. However, it is difficult to imagine how the publication process would work if we didn't have a commonly accepted criterion for deciding which effects to treat as real (although that may simply be a lack of imagination on my part). Unless Bayesian statistics completely take over, we are stuck with the need to evaluate statistical significance. This chapter therefore describes the common practices for dealing with this necessary evil in the context of ERPs. I'm assuming that you have already had an introductory statistics course, so this chapter focuses on the specific issues that arise when analyzing ERP data.

Before you read any more, take a look at box 10.1. It describes the most important principle for assessing statistical significance, and it supersedes everything else I will say in this chapter.

This chapter begins by describing the conventional approach to analyzing ERP data, in which ERP amplitudes and latencies are treated just like behavioral variables such as reaction time and accuracy. This approach was initially developed when the technology for ERP research was

Box 10.1
The Best Statistic

I first met Steve Hillyard when I visited UCSD during my senior year of college. When I met with Steve, I proudly told him about all the fancy multivariate statistics I had been using to analyze ERP data in my undergraduate senior thesis. He looked at me and said, "Around here, we think that replication is the best statistic." My initial thought, of course, was that this guy was a technically unsophisticated Luddite. By the time I finished my first year of grad school, however, I realized that he was absolutely correct (and wasn't a Luddite at all). Replication does not depend on assumptions about normality, sphericity, or independence. Replication is not distorted by outliers. Replication is a cornerstone of science. Replication is the best statistic.

A corollary principle—which I also learned from Steve—is that the more important a result is, the more important it is for you to replicate it before you publish it. An obvious reason for this is that you don't want to make a fool of yourself by making a bold new claim and being wrong. A less obvious reason is that if you want people to give this important new result the attention it deserves, you should make sure that they have no reason to doubt it. Of course, it's rarely worthwhile to run exactly the same experiment twice. But it's often a good idea to run a follow-up experiment that replicates the result of the first experiment and also extends it (e.g., by assessing its generality or ruling out an alternative explanation).

There are some areas of science in which the cost of running an experiment—in money or time—is so great that it is unrealistic to replicate a result before publishing it. If you are doing research of this nature, you will need to work extra hard to make sure that your results are real and not the result of a biased analysis approach. In these areas, replication is still important but usually occurs via meta-analyses across studies rather than via within-study replications.

primitive, and the data consisted of peak amplitudes or latencies measured at a few electrode sites. This approach evolved gradually as researchers began measuring other features of the waveform (e.g., mean amplitude) and were able to record from a couple dozen electrode sites. This approach is still valuable in many situations, especially when the researcher is testing a very specific hypothesis about the amplitude or latency of a component that is measured from a well-justified latency range at a reasonably small number of electrode sites (or if the data are averaged over clusters of nearby sites). A new variant of this approach—the *jackknife* approach—can dramatically improve statistical power under some conditions.

This chapter will also describe a newer and very different approach that is needed when large numbers of time points and/or electrode sites are analyzed, which leads to the *problem of multiple comparisons*. This newer approach is based on methods that were originally developed for the analysis of neuroimaging data, in which thousands of voxels must be tested and the problem of multiple comparisons is very obvious. These methods are just starting to hit the mainstream of ERP research, but I suspect they will become very common over the coming years.

Across the fields of psychology and neuroscience, the past few years have seen growing sensitivity to a variety of data analysis practices that dramatically increase the likelihood of Type I errors (i.e., significant effects that are actually bogus) (Vul, Harris, Winkielman, & Pashler,

2009; Simmons, Nelson, & Simonsohn, 2011; John, Loewenstein, & Prelec, 2012; Pashler & Wagenmakers, 2012; Button et al., 2013). This leads to a proliferation of incorrect conclusions in the literature, which is a very bad thing for scientific progress. The pressure to publish is partially responsible for leading people toward questionable data analysis practices. But there are many common practices people use that unintentionally inflate the Type I error rate. One goal of this chapter is to explain how these seemingly innocuous practices are problematic and to provide you with simple strategies for avoiding bogus significant results when analyzing ERP data.

This chapter has four main sections. First, I will review a little bit of statistical terminology that is used throughout the chapter. Second, I will describe the conventional approach to statistics. Third, I will describe the jackknife approach, which is a slight variation on the conventional approach that can dramatically improve your statistical power under certain conditions. Fourth, I will describe how the richness of an ERP data set often leads to a large number of (implicit or explicit) comparisons, which in turn complicate the analyses. In particular, the need to choose specific time windows and electrode sites can inflate the Type I error rate. This section provides several suggestions for avoiding this problem. One of them—called the *mass univariate approach*—is described in detail in online chapter 13. This chapter also describes a completely different general approach to statistics—called the *permutation* approach—which is becoming increasing popular in ERP research.

Terminology

This chapter uses a variety of basic statistical terms that should be familiar to most readers, but you may want to review some of them in the glossary before reading further. Here are the key terms: *null hypothesis; alternative hypothesis; alpha; p value; Type I error; Type II error; statistical power.*

In addition, this chapter uses the terms *experimentwise error rate* and *familywise error rate*, which may be less familiar. Imagine that you are analyzing the data from an oddball experiment with a three-way ANOVA for P3 amplitude and another three-way ANOVA for P3 latency. Each of these two ANOVAs will produce seven different p values (three main effects, three two-way interactions, and one three-way interaction), yielding 14 total p values across the two ANOVAs. If these were the only analyses in your experiment, the *experimentwise error rate* would be the probability that even one of these 14 p values was a false positive (a Type I error). With 14 p values and a standard alpha of 0.05 for each individual p value, the experimentwise error rate would be substantially higher than 5%. In other words, if the null hypothesis is actually true for all 14 of these effects, the chance of getting one or more significant p values ($p < 0.05$, uncorrected) would be greater than 5%. More generally, the experimentwise error rate is the probability that at least one p value among all the p values for a given experiment will be a false positive. The *familywise error rate* is the same concept, but refers to a subset of related p values from a given experiment (a "family" of related statistical tests, such as the seven p values from the P3

amplitude ANOVA). All else being equal, the more *p* values that are calculated in a given experiment or in a given family of analyses, the higher the experimentwise or familywise error rate will be.

The Conventional Approach

The conventional approach to ERP statistical analysis treats amplitude and latency measurements just like any other dependent variable. You obtain these values from each subject and enter them into a *t* test or ANOVA, just like you would for each subject's mean reaction time. The main difference from a behavioral analysis is that an ERP analysis will typically involve measurements from multiple electrode sites in each subject. You may also have both amplitude and latency measurements in an ERP study, and you may be measuring multiple components. However, amplitude and latency measurements are virtually always analyzed separately, just as you would analyze accuracy and RT separately in a behavioral experiment, and measurements of different components are almost always analyzed separately. Thus, a conventional ERP analysis is usually just like a behavioral analysis, except with measurements from each of several electrode sites. The inclusion of measurements from multiple electrode sites leads to two complications that you need to know about, and I will describe them in the next two subsections. First, however, I will give you a simple example of the conventional approach.

An Example of the Conventional Approach

This example is based on an unpublished oddball experiment described briefly near the beginning of chapter 1 (see figure 1.1). In this experiment, recordings were obtained from nine electrode sites (F3, Fz, and F4; C3, Cz, and C4; P3, Pz, and P4). Subjects saw a sequence of Xs and Os, and they were required to press one button for the Xs and another for the Os. In some trial blocks, X occurred frequently ($p = 0.75$) and O occurred infrequently ($p = 0.25$), and in other blocks this was reversed. We also manipulated the difficulty of the X/O discrimination by varying the brightness of the stimuli.[1]

In most cases, I recommend focusing your analyses on a single component (see strategy 1 in chapter 4). However, you may sometimes have a good reason to analyze multiple components. In addition, you may see effects for other components that are not the main focus, and you may want to analyze the data from these other components for the sake of completeness. In general, your experimentwise error rate will be lower if you analyze fewer components. If, for example, you conduct three-way ANOVAs on the amplitude and latency of five different components, you will have 70 total *p* values, and you are very likely to have several spurious significant effects. Practically speaking, the best approach is often to analyze multiple components but take the results seriously only for a few *a priori* comparisons (and rely on replication for assessing the reliability of the other significant effects).

In our example experiment, we focused on P3 amplitude, but we also measured the amplitudes and latencies of the P2 and N2 components. We first combined the data from the Xs and Os so

that we had one waveform for the improbable stimuli and one waveform for the probable stimuli. We did this for the simple reason that we didn't care if there were any differences between Xs and Os per se, and collapsing across them reduced the number of factors in the ANOVAs. The more factors are used in an ANOVA, the more individual *p* values will be calculated, and the greater is the chance that one of them will be less than 0.05 due to chance (i.e., this increases the familywise error).[2] By collapsing the data across irrelevant factors, you can avoid this problem (and avoid having to come up with an explanation for a weird five-way interaction that is probably spurious). Of course, it's a good idea to look at the waveforms separately first, just to make sure that they aren't radically different. But if you see an interaction with one of your counterbalancing factors, there is a good chance that it is spurious, so don't take it seriously until you see whether it is replicable.

The results of this experiment are illustrated in figure 10.1, which shows the ERP waveforms recorded at Fz, Cz, and Pz. From this figure, it is clear that the P2, N2, and P3 waves were larger

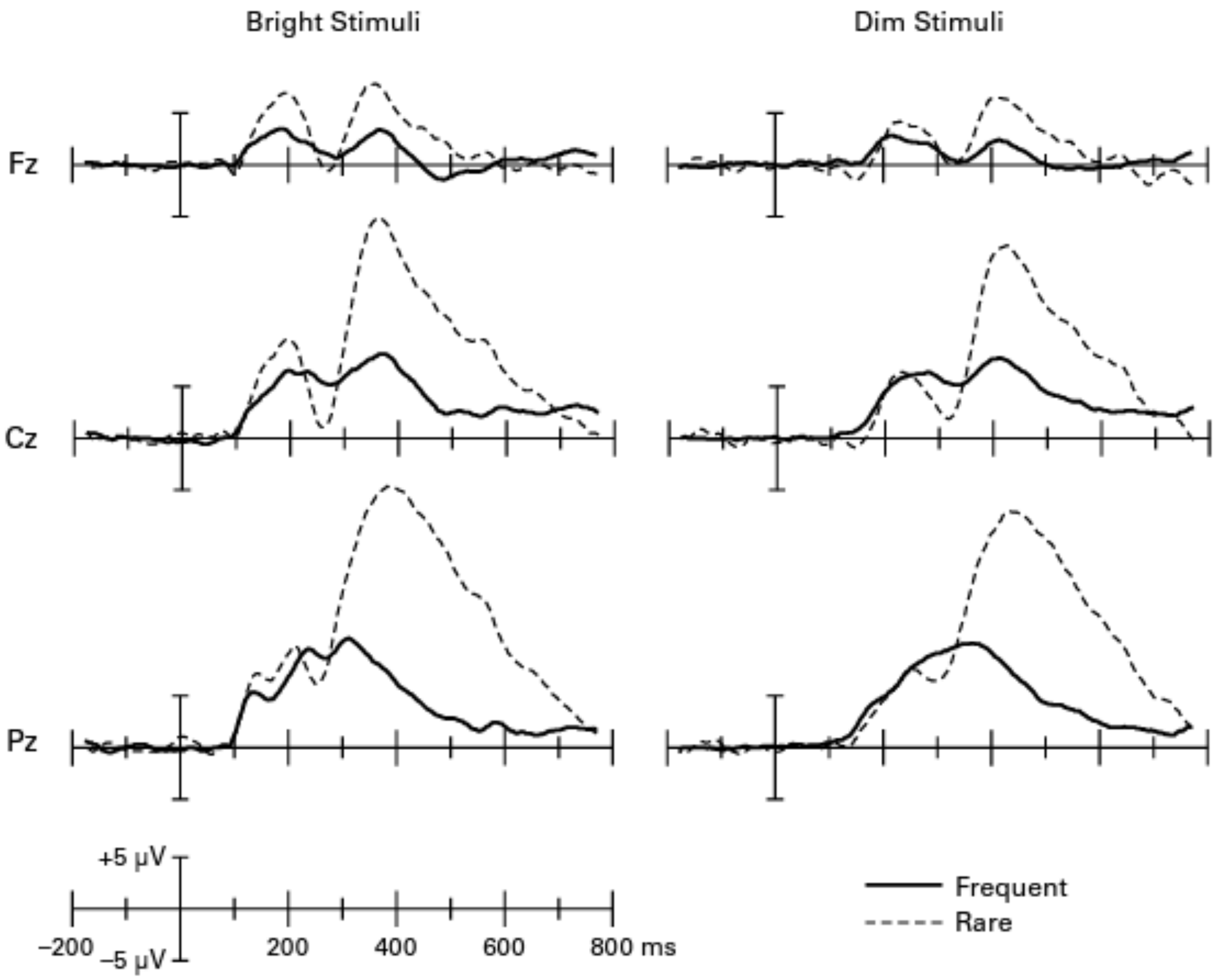

Figure 10.1
Grand average ERP waveforms from an unpublished oddball experiment in which the stimuli were either bright or dim. The data are referenced to the average of the mastoids and have been low-pass filtered (half-amplitude cutoff = 30 Hz, slope = 12 dB/octave).

for the rare stimuli than for the frequent stimuli, especially when the stimuli were bright. Thus, for the amplitude of each of these components, we would expect to see a significant main effect of stimulus probability and a significant probability × brightness interaction.

To quantify P3 amplitude, I measured the mean amplitude between 300 and 800 ms at each of the nine electrode sites in each subject, and I entered these data into a within-subjects ANOVA with four factors: stimulus probability (frequent vs. rare), stimulus brightness (bright vs. dim), anterior-to-posterior electrode position (frontal, central, or parietal), and left-to-right electrode position (left hemisphere, midline, or right hemisphere). Consistent with the waveforms shown in figure 10.1, this ANOVA yielded a highly significant main effect of stimulus probability, $F(1, 9) = 95.48$, $p < 0.001$. It also yielded a significant interaction between probability and brightness, $F(1, 9) = 11.66$, $p < 0.01$, because the difference between rare and frequent stimuli was larger for the bright stimuli than for the dim stimuli.

Using Electrode Site as an ANOVA Factor

I could have used a single factor for the electrode sites, with nine levels, but it is usually more informative to divide the electrodes into separate factors representing different spatial dimensions. That way, you can more readily determine if an electrode effect reflects a difference across hemispheres or a difference across regions within hemispheres. If you have large numbers of electrodes, you might want to average across clusters of nearby electrodes, yielding a $3 \times N$ set of values (left/middle/right × N anterior-to-posterior clusters). It is typically simplest to average across the waveforms in a cluster and then measure the amplitudes rather than measuring the amplitudes from each electrode and then averaging across the cluster (see the appendix of this book for a discussion of when the order of operations will impact your results). The downside of using separate anterior-to-posterior and left-to-right factors is that it increases the number of factors in the ANOVA and therefore increases the familywise error rate. However, this isn't a big problem if you are mainly using electrode site in your ANOVA to increase power (by including multiple sites at which the effect of interest is present) and to assist in your description of the scalp distribution of your effects.

You could, in principle, perform a separate ANOVA for each electrode site (or each left-midline-right set) rather than performing a single ANOVA with electrode site as a factor. Although this approach is occasionally appropriate, it is likely to increase the probability of both Type I and Type II errors. Type I errors will be increased because more p values must be computed when a separate ANOVA is performed for each electrode, leading to a greater probability of a spurious effect with a p value of less than .05. Type II errors may be increased because a small effect may fail to reach significance at any individual site even though the same effect would be significant in an analysis that includes multiple sites.

Even when multiple electrode sites are included in a single ANOVA, you may want to include only the sites where the component is actually present rather than including electrodes from the entire scalp. A component cannot be measured very precisely when it is small, so including these

sites may add noise to the analysis, decreasing your statistical power. In addition, it is sometimes useful to analyze only the sites at which the component of interest is large *and* other components are relatively small so that the measurements of the component of interest are not distorted as much by the other components. In the current study, for example, I used all nine sites for analyzing the P3 wave, which was much larger than the other components, but I restricted the P2 analyses to the frontal sites, where the P2 effects were large but the N2 and P3 waves were relatively small. However, when you are trying to draw conclusions about the scalp distribution of a component, it may be necessary to include measurements from all electrodes. The issue of how to select the electrode sites for a given analysis is discussed in detail later in this chapter in the section on "Choosing Time Windows and Electrode Sites: The Problem of Multiple Implicit Comparisons."

An increasingly common approach is to average across all electrode sites within the relevant region of the scalp, measure the component of interest from this average, and then conduct the statistical analysis on this single value. This approach has several virtues. First, nonlinear measures such as fractional peak latency are more robust when measured from cleaner waveforms, and averaging across multiple sites will tend to reduce the noise level. Second, it completely avoids the use of electrode factors in the ANOVA, reducing the number of *p* values calculated in the ANOVA and thereby reducing the familywise error rate. Third, it makes the analysis easier to explain, which is especially important if you are trying to reach a broad audience. I would encourage you to use this approach whenever appropriate. (If reviewers hassle you about it, you can explain to them that it reduces the familywise error rate, and then you will have educated them about a very important issue! You can also cite this chapter, thereby increasing my citation count.)

Analyzing Difference Scores

In the example experiment shown in figure 10.1, it is problematic to compare the P3 elicited by bright and dim stimuli because of the sensory differences between these stimuli. This issue was discussed in the context of a very similar Gedankenexperiment in chapter 4. One solution described in that chapter is to make rare-minus-frequent difference waves for the bright stimuli and for the dim stimuli and then compare these difference waves (see figure 4.2). Any pure effects of brightness on the ERP waveform should be the same for rare and frequent stimuli, so the rare-minus-frequent difference wave eliminates any pure effects of brightness on the waveform.[3] This is one of many ways in which difference waves can be used to isolate a specific effect.

It is often useful to measure amplitudes and latencies from difference waves and use these measures as the dependent variables in your statistical analyses. First, difference waves can help you isolate a process of interest, as described in chapters 2 and 4. Second, using difference waves will reduce the number of factors in the ANOVA, decreasing the number of *p* values being calculated and thereby decreasing the familywise error rate. Again, I would encourage you to use this approach whenever it seems appropriate.

Interactions with Electrode Site

A well-known complication in interpreting ANOVAs in ERP experiments occurs when you find an interaction between condition and electrode site. For example, it is clear from figure 10.1 that the difference in P3 amplitude between the rare and frequent stimuli was larger at posterior sites than at anterior sites. This led to a significant interaction between stimulus probability and anterior-to-posterior electrode position, $F(2, 18) = 63.92$, $p < 0.001$. In addition, the probability effect for the bright stimuli was somewhat larger than the probability effect for the dim stimuli at the parietal electrodes, but there wasn't much difference at the frontal electrodes. This led to a significant three-way interaction between probability, brightness, and anterior-to-posterior electrode position, $F(2, 18) = 35.17$, $p < 0.001$.

From this interaction, you might be tempted to conclude that different neural generators were involved in the neural responses to the bright and dim stimuli. In other words, if the scalp distribution changes, this seems like it implies a change in the underlying generators. However, as McCarthy and Wood (1985) pointed out, ANOVA interactions involving an electrode position factor are ambiguous when two conditions have different overall amplitudes. This is illustrated in figure 10.2A, which shows the ERP amplitudes that would be expected at the Fz, Cz, and Pz electrode sites from a single generator source in two different conditions, A and B. If the magnitude of the generator's activation is 50% larger in condition B than in condition A, the amplitude at each electrode will be 50% larger in condition B than in condition A. This is a multiplicative effect, and not an additive effect. That is, the voltage increased by 50% at each site, leading to an increase from 1 μV to 1.5 μV at Fz (a 0.5-μV increase) and an increase from 2 μV to 3 μV at Pz (a 1-μV increase). This shows up as an interaction in an ANOVA, even though it arises from a change in the magnitude of a single generator source. An additive effect is shown in condition C in figure 10.2A. In this condition, the absolute voltage increases by 1 μV at each site relative to condition A, which is not the pattern that would result from a change in the amplitude of a single generator source. Thus, when a single generator source has a larger magnitude in one condition than in another condition, an interaction between condition and electrode site will be obtained (as in condition A vs. condition B). In contrast, a change involving multiple generator sites may sometimes produce a purely additive effect (as in condition A vs. condition C).

To determine whether an interaction between an experimental condition and electrode site really reflects a difference in the internal generator sources, McCarthy and Wood (1985) proposed *normalizing* the data to remove any differences in the overall amplitudes of the conditions. An example of this is shown in figure 10.2B (for details of the normalization procedure, see the online supplement to chapter 10). Once the data have been normalized, the scalp distribution is the same for conditions A and B and different for condition C, which tells us that the generator has simply changed in magnitude between conditions A and B whereas the generator has changed in some way for condition C.

However, Urbach and Kutas (2002) convincingly demonstrated that this normalization procedure doesn't actually work under most realistic conditions. In most cases, I would therefore

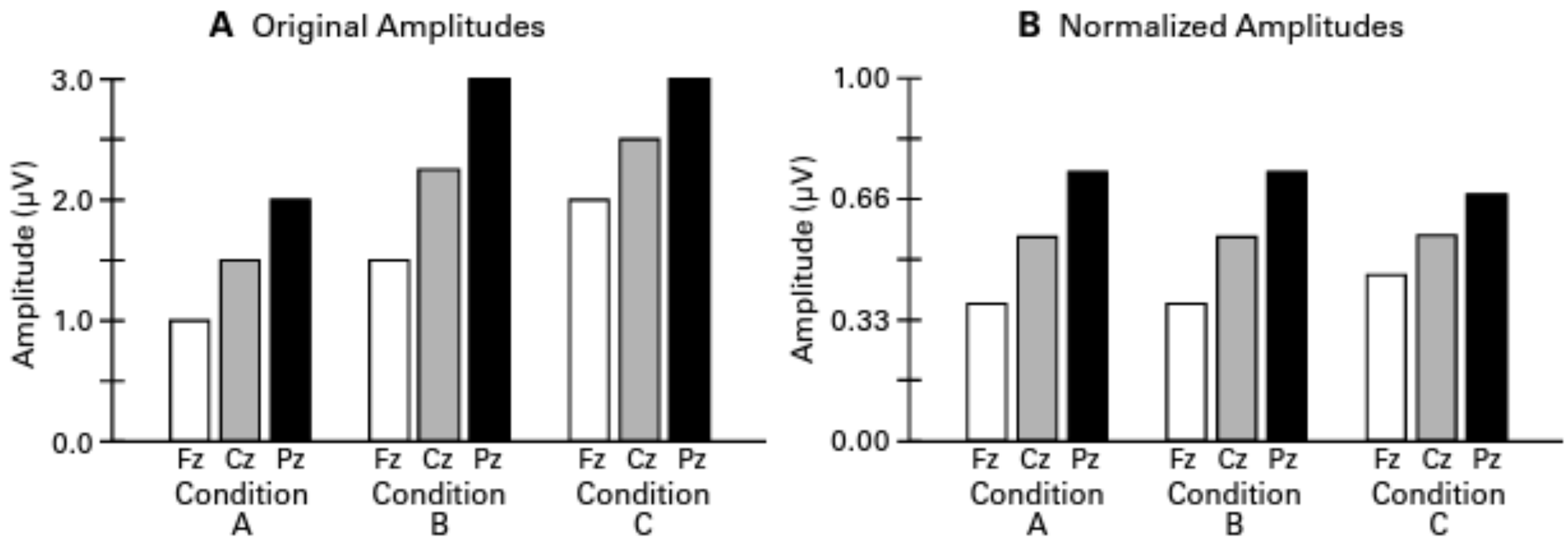

Figure 10.2
(A) Examples of additive and multiplicative effects on ERP scalp distributions. Condition B is the same as condition A, except that the magnitude of the neural generator site has increased by 50%, thus increasing the voltage at each site by 50%. This is a multiplicative change. Condition C is the same as condition A, except that the voltage at each site has been increased by 1 μV. This is an additive change, and it is not what would typically be obtained by increasing the magnitude of a single neural generator source. (B) Same as panel A, except that the amplitudes from each condition have been normalized by dividing by the vector length from that condition (see the online supplement to chapter 10 for a detailed description of this normalization procedure). Now we can see that condition A and condition B have the same scalp distribution, but condition C has a different distribution.

recommend simply reporting that you found an interaction with electrode site and then saying very little about it. If the difference between conditions is largest at the scalp sites where the component is largest (e.g., if the difference between the difference waves in the bright and dim conditions was largest at the Pz electrode site), you can simply state that the pattern of results is approximately what would be expected if a single component varied in amplitude across conditions (without even performing a formal analysis on the normalized data). This issue is discussed in more detail in the chapter 10 supplement.

Heterogeneity of Covariance and the Epsilon Adjustment

A second complication with including electrode site as an ANOVA factor is that this often leads to a violation of the assumption of *homogeneity of covariance*. You probably already know that ANOVA assumes *normality* (Gaussian distributions) and *homogeneity of variance* (equal variances across the different conditions). These assumptions are often violated, but ANOVA is fairly robust when the violations are mild to moderate, with very little change in the actual probability of a Type I error (Keppel, 1982). Unless the violations of these assumptions are fairly extreme (e.g., greater than a factor of 2), you don't need to worry about them.

However, when an ANOVA includes within-subject factors, such as electrode site, we must also assume homogeneity of covariance. The basic idea is illustrated in figure 10.3, which shows data from three subjects who were each tested in three conditions (A, B, and C). For the sake of this example, let's assume that weight is being measured. Although there is quite a bit of variance among subjects in each condition, all three subjects show exactly the same pattern of

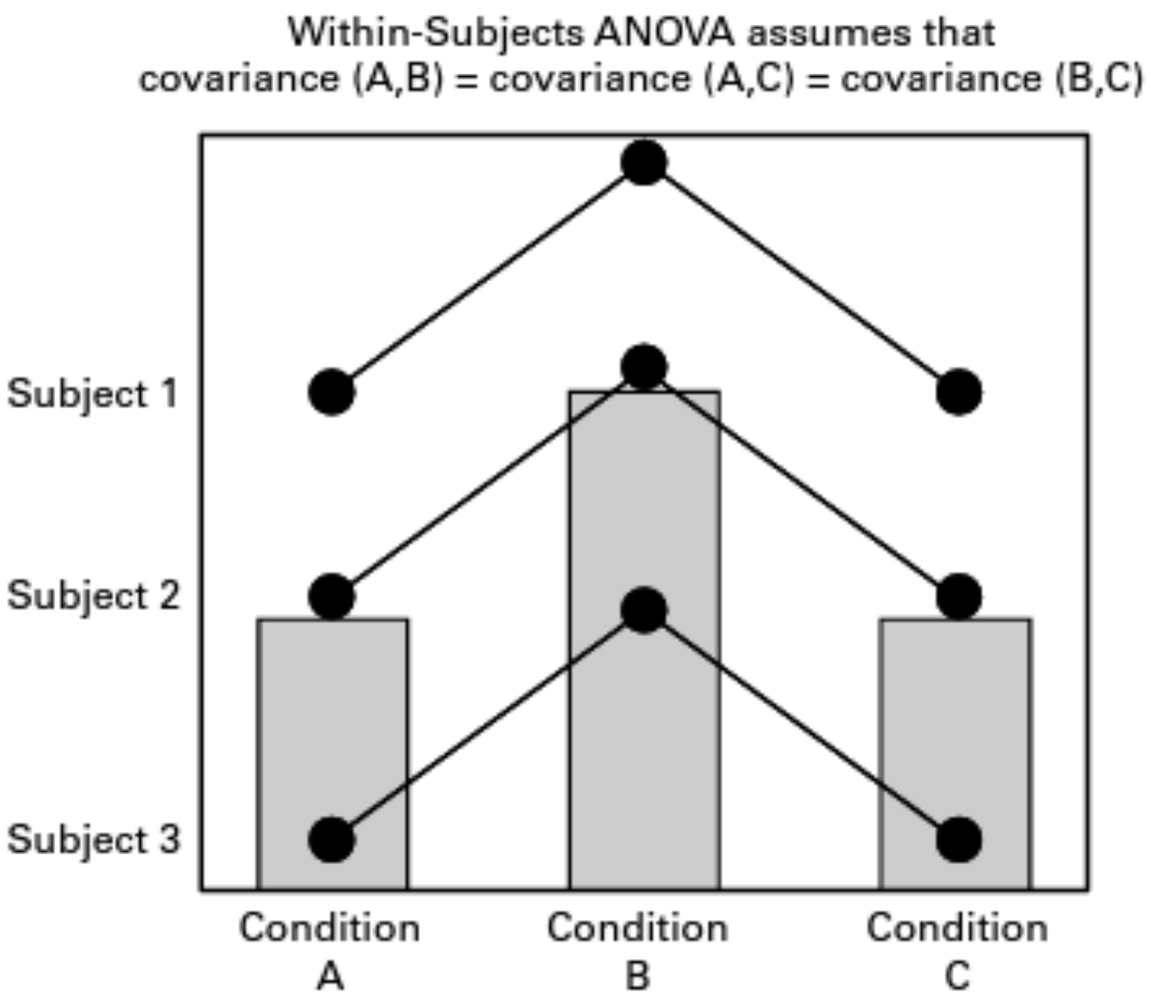

Figure 10.3
Example of a within-subjects design and the concept of homogeneity of covariance. Data are shown from three subjects in three conditions. It doesn't matter what is being measured, but you could think of the *Y* axis as representing weight.

differences among conditions. A within-subjects ANOVA factors out the overall differences among subjects, focusing on the consistency of the effect across conditions (i.e., the fact that all three subjects go up by the same amount in condition B compared to condition A and then go down by the same amount in condition C). This can dramatically increase statistical power. However, power is increased only to the extent that subjects who have high values in one condition tend to have high values across all conditions, and subjects who have low values in one condition tend to have low values across all conditions. This is equivalent to saying that a subject's score in one condition covaries with that subject's scores in the other conditions. Sometimes this covariance[4] is very strong (as in the example shown in figure 10.3). Sometimes it is weak. The assumption of homogeneity of covariance is simply the assumption that the degree of covariance between conditions A and B is equal to the degree of covariance between conditions A and C and between conditions B and C. This assumption does not apply if there are only two levels of a factor, because there is only one covariance in this case.

To see how this assumption might be violated, imagine that each subject's weight was measured three times, once at age 3, once at age 21, and once at age 22. A person's weight at age 21 will be much more strongly related to his or her weight at age 22 than to his or her weight at age 3. Thus, the covariance between ages 21 and 22 would be higher than the covariance between ages 3 and 21. This would violate the assumption of homogeneity of covariance.

Violations of the assumption of homogeneity of covariance are very common in ERP experiments that include multiple electrode sites as a factor in the analysis, because data from nearby

electrodes tend to covary more than data from distant electrodes. For example, random EEG noise at the Fz electrode will spread to Cz more than to Pz, and the correlation between the data at Fz and the data at Cz will be greater than the correlation between Fz and Pz. In addition, a real ERP signal will tend to impact nearby sites to similar degrees but will not impact distant sites to different degrees, which also creates more covariance between nearby sites.

Unfortunately, ANOVA results become very inaccurate when the covariance is heterogeneous. Violating the assumption of homogeneity of covariance leads to artificially low *p* values, such that you might get a *p* value of less than 0.05 even when the actual probability of a Type I error is 0.25. This was brought to the attention of ERP researchers very forcefully in a paper published in the journal *Psychophysiology* by Jennings and Wood (1976). The journal subsequently developed an explicit policy stating that all papers published in the journal must address this problem.

The most common solution is to use the Greenhouse–Geisser epsilon adjustment, which counteracts the inflation of Type I errors produced by heterogeneity of covariance. For each factor or interaction that has more than two within-subjects levels, a value called *epsilon* is computed along with the *F* value. The epsilon value for a given *F* value is then multiplied by the degrees of freedom for that *F* value, and the adjusted degrees of freedom are used to compute the *p* value. Epsilon varies between 0 and 1, with small values corresponding to a large heterogeneity of covariance. If the covariances are homogeneous, epsilon is near 1, and multiplying the degrees of freedom by a value near 1 doesn't change them much. Thus, little or no change in the degrees of freedom occurs if the assumption of homogeneity of covariance is met, but the degrees of freedom move downward—and the *p* value therefore gets worse—as the heterogeneity of covariance increases. This adjustment is provided by most major statistics packages and is therefore easy to use. For example, the SPSS ANOVA output contains the adjusted *p* values along with the unadjusted *p* values.

I used the Greenhouse–Geisser adjustment in the statistical analysis of the P3 amplitude data shown in figure 10.1. It influenced only the main effects and interactions involving the electrode factors, because the other factors had only two levels (i.e., frequent vs. rare and bright vs. dim). For most of these *F* tests, the adjustment didn't matter very much because the unadjusted effects were either not significant to begin with or so highly significant that a moderate adjustment wasn't a problem (e.g., an unadjusted *p* value of 0.00005 turned into an adjusted *p* value of 0.0003). However, there were a few cases in which a previously significant *p* value was no longer significant. For example, when I normalized the data before conducting the ANOVA, the main effect of anterior-to-posterior electrode site was significant before the adjustment was applied ($F[2,18] = 4.37$, $p = 0.0284$) but was no longer significant after the adjustment ($p = 0.0586$). This may seem like a bad thing, because a significant effect was made non-significant by the adjustment. However, the original *p* value was not accurate, and the adjusted *p* value is closer to the actual probability of a Type I error. In addition, when very large numbers of electrodes are used, the adjustments are usually much larger, and spurious results are quite likely to yield significant *p* values without the Greenhouse–Geisser adjustment.

It is absolutely necessary to use the Greenhouse–Geisser adjustment—or something comparable[5]—whenever there are more than two levels of a factor in an ANOVA, especially when one of the factors is electrode site. Of course, you should use this adjustment for other within-subjects factors that include more than two levels, and not just for the electrode site factor. And you should use it for analyses of behavioral data as well. If you don't, your *p* values will not be correct, and you will be likely to draw conclusions on the basis of false positives.

The Jackknife Approach

Error Variance

In traditional statistical approaches, the *p* value depends on both the size of the difference between conditions and the amount of variance within each condition. As the variance within a condition increases, it becomes less likely that the difference between conditions is real (all else being equal). The variance within a condition is called the *error variance*. As described in the supplement to chapter 8, the error variance contains both *measurement error* and *true score variance*. Measurement error occurs when we do not obtain an accurate measurement of the subject's true value. True score variance reflects the real differences between subjects. The most obvious cause of measurement error in ERP experiments is noise in the single-subject averaged ERP waveforms. If we could average together an infinite number of trials for each subject, we would dramatically reduce measurement error. And if we reduced the measurement error, we would reduce the overall error variance, which would in turn give us more statistical power.

Another common problem is that outliers can have a huge effect on the error variance. If we had a principled way to reduce the impact of outliers, this would also give us greater power.

The jackknife technique helps us reduce measurement error and the effects of outliers by measuring amplitudes or latencies from grand averages rather than from single-subject averages. Several times I have had the experience of looking at grand average waveforms, seeing a beautiful effect that was far larger than the noise in the prestimulus period, and then being very surprised to find a *p* value that was far from significant when I conducted the statistical analyses. When I then took a closer look at the individual subjects, I realized that this was because the measured values were "crazy" for some of the subjects. These crazy values dramatically increased the error variance, resulting in a *p* value that was not significant. This is one reason why I recommended—at the beginning of chapter 9—that you always compare the measurements for each subject with that subject's ERP waveforms. This might make you realize that something is wrong with your measurement procedure. Sometimes, however, there is nothing wrong with the measurement procedure, and the problem is that the waveforms from some of the subjects are noisy or unusual.

As an example, the left column of figure 10.4 shows the lateralized readiness potential (LRP) from three individual subjects, isolated with a contralateral-minus-ipsilateral difference wave (see chapter 3 for details about the LRP). The data from each subject are pretty noisy, and the

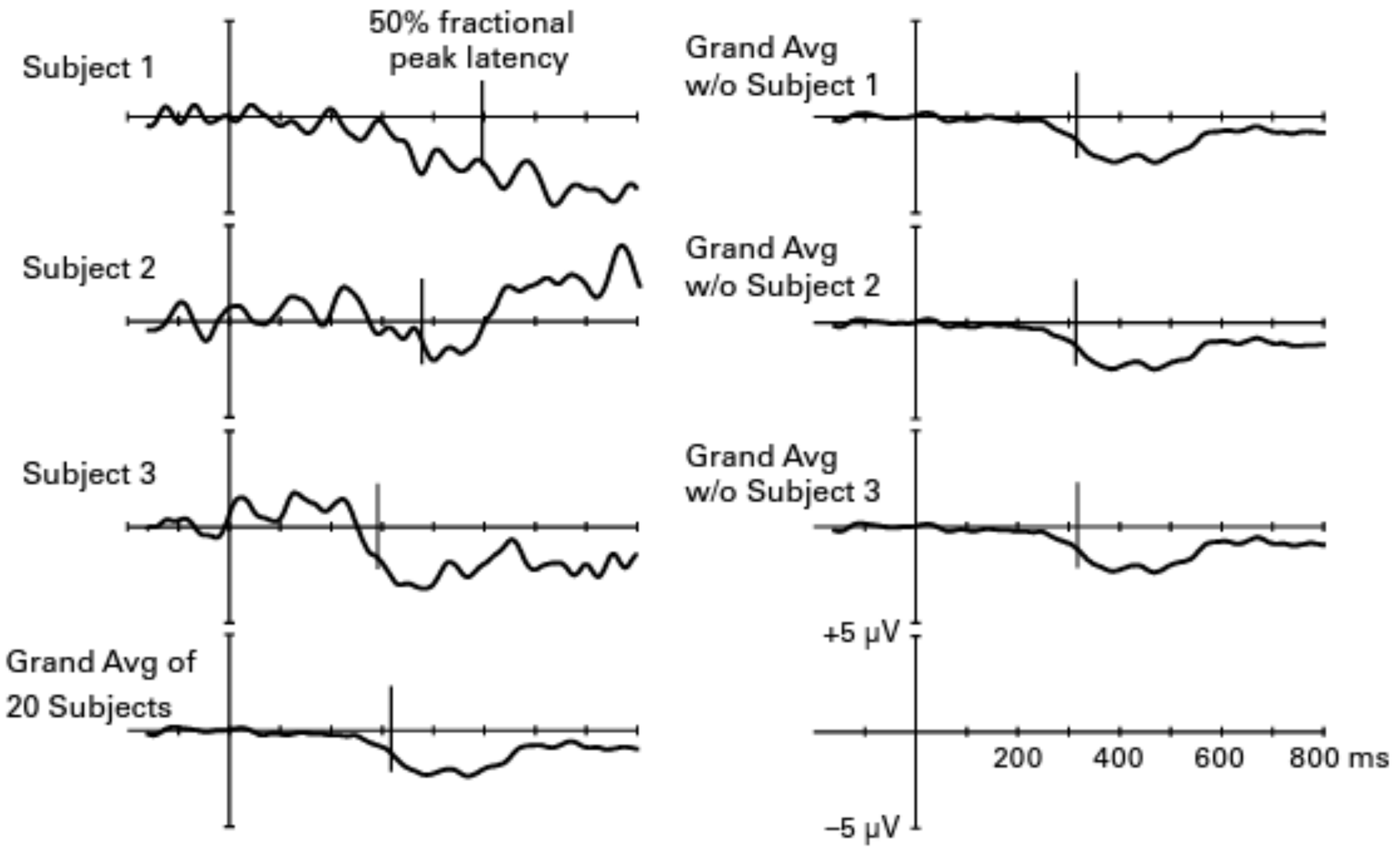

Figure 10.4
Lateralized readiness potential data collapsed across the C3 and C4 electrode sites, formed by subtracting the ERP over the hemisphere ipsilateral to the response hand from the ERP over the hemisphere contralateral to the response hand (i.e., contralateral minus ipsilateral). The left column shows averaged ERP waveforms from three individual subjects and the grand average of all 20 subjects. The right column shows leave-one-out grand averages, each of which was formed by averaging together 19 of the 20 subjects and leaving out one subject.

LRP develops much more gradually in subject 1 than in subjects 2 and 3. Consequently, the measured onset latency (measured as the 50% peak latency) of the LRP varies a great deal among these subjects. This variance gives us low statistical power.

The grand average across 20 subjects is shown in the bottom left of figure 10.4. You can see that it is very clean, and we could quantify the onset latency from this grand average with very little measurement error. However, measuring the onset latency from the grand average doesn't do us much good, because we can't do statistical analyses if we just have this one measure. This is where the jackknife technique comes in.

The jackknife technique is basically a method that makes it possible for you to measure amplitudes and latencies from your beautifully clean grand averages and still perform conventional statistical analyses.[6] This technique has been used for decades in some areas of statistics, and it was imported into ERP research by Jeff Miller and his colleagues in the late 1990s (Miller, Patterson, & Ulrich, 1998). It's easy to use, and it can dramatically improve your statistical power in some conditions, decreasing the probability of a Type II error without increasing the probability of a Type I error. When I explain how it works, you may not believe that it could possibly be a legitimate technique. However, it has been demonstrated to work both by

mathematical proofs and by rigorous simulation studies (see, e.g., Miller et al., 1998; Ulrich & Miller, 2001; Kiesel, Miller, Jolicoeur, & Brisson, 2008). It does have some limitations, however, so make sure you read this entire section before using it.

Essence of the Jackknife Approach

Imagine that you have 10 subjects who have been tested in two conditions, A and B, and you want to know if the mean onset latency of the LRP is earlier in condition A than in condition B. If you measure the onset latency from the individual subjects, as in the left column of figure 10.4, you might find that you have so much error variance that the difference between groups is far from significant. If you measure the onset latency from the grand average of each condition, you could avoid much of this measurement error. However, you would have only one measurement from each condition, so you wouldn't have a measure of the error variance, and this would make it impossible to determine if the difference between conditions was larger than would be expected by chance. In other words, you couldn't get a *p* value if you just had one measure from the condition A grand average and one measure from the condition B grand average.

To be able to assess the variance within each condition without giving up the advantages of measuring from a grand average, you can create a series of grand averages, each of which is missing one of the subjects. In our example, you would make 10 grand averages for each of the two conditions, each created by averaging together the waveforms from nine of the 10 subjects for that condition. The first grand average would include everyone except subject 1; the second would include everyone except subject 2; and so forth. Examples are shown in the right column of figure 10.4. These are called *leave-one-out* grand averages. All of the leave-one-out grand averages for a given condition will be highly similar to the others, because most of the single-subject waveforms in one leave-one-out grand average are also in the other leave-one-out grand averages. However, the individual leave-one-out grand averages will be slightly different from each other, reflecting the subject who was left out of each average.

We can measure the onset latency from each of these leave-one-out grand averages, which will work very well because these waveforms are extremely clean. We can also assess the variance across the leave-one-out grand averages for a given condition. This will not be the same as the variance across individual subjects, but there is a mathematically principled way of using the variance of the leave-one-out grand averages to estimate the variance of the single subjects. An extension of this allows us to conduct a *t* test with the values measured from the leave-one-out grand averages. Specifically, we can take the 20 onset latencies that we've measured (from 10 leave-one-out grand averages for condition A and 10 leave-one-out grand averages from condition B) and enter these values into a *t* test (just like you would ordinarily enter the 10 values from each subject in each condition into a *t* test). The resulting *t* value will be unnaturally large, because the error variance has been artificially reduced by measuring from the leave-one-out grand averages. However, we can adjust for this artificial reduction in error variance by simply dividing the *t* value by $N - 1$ (which would be 9 in this example), giving us an adjusted

t value. We can then look up the *p* value for this adjusted *t* value, and the difference between conditions will be considered significant if the *p* value is less than 0.05 (or whatever alpha level you want to use). This is the essence of the jackknife approach.

This is just as simple as it sounds: You simply make the leave-one-out grand averages and treat them as if they were single-subject ERP waveforms. The only trick is to divide the *t* value by $N - 1$ before looking up the *p* value.

You can do the same thing with a more complex experimental design, using an ANOVA instead of a *t* test. You just make the leave-one-out grand averages for each condition, measure the latencies from these leave-one-out grand averages, enter the latencies into an ANOVA just as if they were measured from single-subject averages, and compute the *F* ratios. You then need to divide the *F* ratios by $(N - 1)^2$. This works for the interactions as well as the main effects. It also works for between-subject factors. If you have *N* subjects in each group, you will again divide each *F* ratio by $(N - 1)^2$. If you have different numbers of subjects in each group, then it's a little more complicated (for the details, see Ulrich & Miller, 2001). But keep in mind that you divide by $N - 1$ for a *t* test and $(N - 1)^2$ for an *F* test.

In many cases, you might find that a *p* value of 0.20 in conventional analysis (in which you measured the latencies from the single-subject ERPs) becomes a *p* value of 0.005 when you do the same analysis with the jackknife approach, even after you've adjusted the *t* value by dividing by $N - 1$. This may seem too good to be true. However, many simulation studies have shown that the jackknife approach does not lead to an increase in the Type I error rate. If you have a real effect, the jackknife technique often helps you get a much better *p* value. However, if the null hypothesis is true, you will get a significant *p* value only 5% of the time (assuming an alpha of 0.05). I have used the jackknife approach many times. I often find a radically better *p* value with the jackknife analysis than with the conventional analysis. But not always, because sometimes the null hypothesis is true. You are no more likely to falsely reject the null hypothesis with the jackknife technique than with conventional statistics. It mainly operates by reducing the error variance, which increases your statistical power. And increased statistical power means that you can find real effects with less effort, publish more papers in better journals, and become rich and famous. What could be better than that?

The jackknife technique can also be used for the Pearson *r* correlation coefficient (Stahl & Gibbons, 2004). Imagine that you want to look at the correlation between P3 onset latency and RT. You would create pairs of latency and RT values, where the latency came from a leave-one-out grand average that left out a given subject and the RT came from that subject. If P3 onset latency tends to be later in subjects with long RTs, but we pair the latency from a grand average that is missing a subject with the RT from that subject, this will reverse the direction of correlation. Imagine, for example, that subject 4 has a very late P3 onset latency and a very long RT. The leave-one-out grand average that excludes subject 4 will have a somewhat earlier-than-usual P3 onset latency (because a subject with a late onset latency has been left out), and this will be paired with the long RT from this subject. This will turn a positive correlation into a negative correlation (or vice versa). It turns out that the adjustment for the Pearson *r* correlation coefficient

is simply to multiply the jackknife r value by –1. That is, you compute the *r* value from the pairs of values and then multiply this value by –1. I've used it, and it works.

Limitations of the Jackknife Approach

The jackknife approach can be truly amazing, but it does have some limitations. Most of these limitations are minor, but the last one I will describe is significant.

Linear Measures One limitation of the jackknife technique is that it gives you the same result as the conventional analysis if you apply it to a linear measure, such as mean amplitude. This is because the jackknife technique is really just a way of changing the order of operations in your analysis pipeline. Rather than measuring the amplitudes or latencies from the single-subject averages and then doing your statistical analysis (which involves averaging across the single-subject values), you measure from waveforms that have already been averaged across the single subjects. As described in detail in the appendix, the order of operations does not matter if all the operations are linear. Thus, the jackknife technique neither helps nor hurts if you are analyzing a linear measure such as mean amplitude.

You can actually use this to your advantage, because it gives you a way of determining whether you are using the jackknife technique correctly. You can simply measure mean amplitude in an experiment, do both the conventional analysis and the jackknife analysis, and then compare the results. The *p* values should be the same in both analyses (assuming that you divided by the appropriate adjustment factor before computing the *p* value in the jackknife analysis). The values might be slightly different due to rounding errors, but they should be very close. If you verify that this is true, then you will trust the results you get by jackknifing a nonlinear measure, such as onset latency.

Note that my examples so far have focused on onset latency. This is because the jackknife technique was originally developed to analyze onset latency, and most of the simulations have focused on onset latency. However, the same principles apply to any nonlinear measure (although you would need to be very careful when using it with peak amplitude, for reasons that will be discussed at the end of this section). Simulations examining the effectiveness of the jackknife technique using a variety of measures from several components can be found in Kiesel et al. (2008), which is on my list of the Top Ten Papers Every New ERP Researcher Should Read (see chapter 1).

Equal Sample Sizes If you have any between-groups factors, the simple adjustment procedures that I have described require that you have the same number of subjects in every group. However, it is still possible to do the adjustment if you have unequal *N*s. It's just more complicated (for a description of the adjustment procedure, see Ulrich & Miller, 2001). This is a very minor limitation.

The Jackknife *p* Value Is Sometimes Not Significant This is not really a limitation, because the null hypothesis may actually be true. The jackknife technique is no more likely than a conven-

tional analysis to give you a significant *p* value when the null hypothesis is true. In addition, even if your effect is real, you may not have enough statistical power to get a significant *p* value with any statistical technique. But you will usually have more power with the jackknife technique than with conventional statistics.

The Jackknife *p* Value Is Sometimes Worse The *p* value from the jackknife technique is sometimes worse than the *p* value from the conventional analysis. This typically happens when you have one outlier subject who has a really big effect on the leave-one-out grand averages (for a description of other conditions that may lead to poor performance by the jackknife technique, see Miller, Ulrich, & Schwarz, 2009). What should you do in this situation?

I asked Jeff Miller about this, and he told me two things. First, if the null hypothesis is true, random noise will determine whether the jackknife *p* value is better or worse than the conventional *p* value, so you can't just pick whichever one is significant. That would allow you to capitalize on random variation, which would inflate your Type I error rate. Second, it would be valid to use the jackknife technique only when the standard error of the mean is much better for the jackknifed data than for the conventional data (because this means that jackknifing is helping reduce error). In other words, although you can't just see which one gives you the better *p* value, you can decide which one to use by determining whether the jackknife technique produces a substantial improvement in the standard error. Miller et al. (1998) explain how to compute the jackknifed standard error (which is quite simple).

The One Major Limitation: Testing a Different Null Hypothesis You are testing a slightly different null hypothesis with the jackknife technique than with conventional statistics, and this is the one significant limitation of the jackknife technique. For several years, I couldn't imagine a situation in which this small difference would matter. But then my imagination improved, and I realized that this different null hypothesis could lead to misinterpretations of the results under certain conditions (see box 10.2 for the story of how this transpired). I will first explain the different null hypotheses, and then I will explain why it sometimes matters which null hypothesis you are testing.

In informal terms, the conventional and jackknife null hypotheses are as follows:

Conventional null hypothesis If we could measure an amplitude (or latency) value from the single-subject ERP waveform in every individual in the infinitely large population, the average of these measures would not differ across conditions.

Jackknife null hypothesis If we could make grand averages that included every individual in the infinitely large population and then measure an amplitude (or latency) value from these grand averages, the values from these grand averages would not differ across conditions.

These null hypotheses are equivalent except for the order of operations. When you are using a linear measure, such as mean amplitude, these two null hypotheses end up being exactly the same. When you are using a nonlinear measure, such as peak amplitude or fractional peak

Box 10.2
Imagining How the Jackknife Might Fail

Jeff Miller—who introduced the jackknife technique to ERP research—was on the faculty at UCSD when I was a graduate student there. I took his graduate ANOVA course, and he served on my dissertation committee. Jeff is an incredibly thoughtful and careful scientist, and he's particularly adept at developing quantitative techniques that can be applied to answering important questions about cognition.

I first encountered the jackknife technique when I saw the initial paper that Jeff and his colleagues wrote about it (Miller et al., 1998). I was intrigued, and I mentioned it briefly in the first edition of this book. A few years later, I did my first LRP study, and some of the latency effects that looked real were not significant in the conventional statistical analyses. I decided to try the jackknife technique, and the results were amazing! Several key effects that were far from significant in the conventional analysis were highly significant in the jackknife analysis. I was hooked.

Shortly after that, I gave a mini ERP Boot Camp at Merck Pharmaceuticals. The "audience" consisted of four extremely smart researchers who had each earned multiple advanced degrees in fields like biostatistics, mathematics, and biomedical engineering. I started talking about the jackknife technique, and the leader of the group told me that they couldn't use it, because the FDA would never allow it in a clinical trial. The reason, he explained, was that the jackknife technique tested a different null hypothesis (as explained in the main text).

I talked about this with Jeff Miller (via e-mail), and neither of us could figure out a situation in which it would matter which null hypothesis was being tested. But I am always concerned about "proof by lack of imagination" (see box 4.3 in chapter 4). So I kept thinking about it. A few years later, I was looking through the slides that I use to discuss the problem of latency variability in ERP averages, and it suddenly occurred to me that this problem also had implications for the jackknife technique. That is, the same problems that arise in creating a single-subject averaged ERP waveform from single-trial EEG epochs can also be a problem when you average together the ERPs from multiple subjects to create a grand average ERP waveform. For example, increased latency variability across subjects will lead to decreased peak amplitude in the grand average. This does not invalidate the jackknife technique, but it places some important limits on how it is applied and interpreted.

latency, they are not the same. But does it really matter which null hypothesis you're testing? After all, if two infinitely large populations are truly equivalent, then they should be the same for both null hypotheses.

It's not quite this simple. As described in box 10.2, the same problems that arise when you average multiple single trials together to create an averaged ERP waveform for a single subject can also arise when you average multiple single-subject ERPs together to create a grand average ERP waveform (or a leave-one-out grand average). Recall from figure 8.7 in chapter 8 that the peak amplitude in an averaged ERP waveform is smaller if the latency of the component varies from trial to trial. The more trial-to-trial latency variability is present, the smaller the peak amplitude will be. The same principle applies to grand averages. If you have a lot of subject-to-subject variability in the latency of a component, the peak amplitude of the grand average will

be reduced. Moreover, you may recall that the onset latency of a single-subject averaged ERP waveform reflects the trials with the earliest onset latencies, not the average of the single-trial onset latencies. Similarly, the onset of a difference between conditions in a grand average will reflect the subjects with the earliest onset of the effect, not the average of the single-subject onset times.

Imagine that you were comparing the peak amplitude of the P3 wave in a patient group and a control group using the jackknife technique. Imagine also that the latency of the P3 wave was more variable across subjects in the patient group than in the control group (which is very likely), but the amplitude of the P3 was the same in the individual patients as in the individual control subjects. In this scenario, the amplitude of the P3 peak in the patient grand average would be smaller than the amplitude of the P3 peak in the control grand average, even though there was no amplitude difference between the individual subjects in these two groups. In a conventional analysis with peak amplitude, you would see no difference between groups. This would be the correct result (although it might be confusing if you were looking at the grand averages, which do have different peak amplitudes). In a jackknife analysis, you would likely see a significant difference between the two groups. If you interpreted this result in the same way that we would interpret the result of a conventional analysis, you would conclude that the subjects in the patient group had smaller P3 amplitudes than the subjects in the control group. This would be an incorrect interpretation. The two groups do differ, but the difference is in latency variability and not in peak amplitude per se. Thus, the jackknife technique will not lead you to conclude that two groups or conditions are different if the entire waveforms are exactly the same in the two groups or conditions. However, a difference that appears to be in one aspect of the waveforms can be a result of a difference in another aspect of the waveforms (e.g., apparent differences in peak amplitude may actually reflect differences in latency variability).

This is a significant issue, but it is not insurmountable. If the jackknife effect is real, then you should see the same basic pattern of means in the conventional analysis that you see in the jackknife analysis. That is, if you look at the table of means produced by your statistics software for the conventional analysis and for the jackknife analysis, you should see the same basic pattern of differences across groups or conditions in both tables. The differences may be somewhat larger in one table than the other, and the effects might not be significant in the conventional analysis, but the patterns of means should be similar. If they are not similar, you may be seeing an unintended side effect of the process of making grand averages.

It is also worth considering the fact that the onset time of the grand average is not the same as the average of the single-subject onset times. In theory, the average of the single-subject onset times could be the same for two conditions, but the onset time of one grand average could be earlier than the onset time of the other grand average. For this to happen, however, the condition with the earlier onset time in the grand average must have some subjects with unusually early onset times and other subjects with unusually late onset times. This would lead to the same average of single-subject onset times across groups but greater variance in the group that had some subjects with very early onset times and other subjects with very late onset times. This

would be a very unusual situation. If one group has greater variance in onset times, this almost always arises from having more subjects who have longer-than-average onset times, without being balanced out by other subjects who have shorter-than-average onset times. In other words, the group with greater variability almost always has a greater mean onset latency. In addition, if you use the 50% peak latency measure, you are not attempting to measure the absolute onset time (the time at which the waveform deviates just slightly from zero), and it's likely that the value measured from the average will be a good approximation of the average onset time. This was discussed in chapter 9 (see figure 9.7), and several simulation studies have shown that the jackknife technique works very well at accurately estimating onset times, increasing statistical power without increasing the Type I error rate (Miller et al., 1998; Ulrich & Miller, 2001; Kiesel et al., 2008; Miller et al., 2009).

The bottom line is that the jackknife technique can be extremely helpful, but you need to be thoughtful and careful when using it (just like anything else in statistics). Using it to analyze peak amplitude can easily lead you astray (especially if there are differences in latency variability between groups or conditions), but you can assess this by making sure that the same pattern is present in the means from the conventional analysis. Using it to measure onset latencies requires some thought, but it is likely to work very well (both in terms of statistical power and the accuracy of the results). It's definitely a great tool to have in your data analysis "toolbox," but like all tools, it needs to be used properly.

Choosing Time Windows and Electrode Sites: The Problem of Multiple Implicit Comparisons

We're now going to switch to a very different issue that arises because of the richness of an ERP data set. If you computed a *t* test comparing two conditions at every time point and every electrode site in a typical ERP experiment, random fluctuations would be expected to lead to many large *t* values that would exceed the usual criterion for statistical significance. This is the widely known *problem of multiple comparisons*, and you would never be allowed to publish results from this approach without some kind of *correction for multiple comparisons*. However, an implicit variation on this problem arises all the time in ERP research when researchers first look at their waveforms, then find the combination of time window and electrode site where a big difference between conditions is present, then measure the amplitude or latency at that combination of time and electrode, and finally enter the resulting values into a statistical analysis. This will often lead to significant *p* values that are a result of noise rather than real effects. If you do this, you are implicitly performing multiple comparisons (by visually comparing the waveforms from multiple time points and multiple electrode sites), so you are biased to obtain a significant *p* value even if no real effect is present. I call this the *problem of multiple implicit comparisons*. It is particularly important when *a priori* information is not used to choose the time windows and electrode sites used in the statistical analyses.

Performing multiple comparisons leads to inflation of the Type I error rate, which means that the true rate of Type I errors (rejecting the null hypothesis when it is true) is greater than 5%

(or whatever alpha you are using). This can be an enormous problem in ERP research. Indeed, it is a problem in many areas of psychology and neuroscience, and it is receiving increasing attention (Simmons et al., 2011; Pashler & Wagenmakers, 2012). Over the coming years, I expect that journals, editors, and reviewers will become increasingly strict about factors that inflate the Type I error rate. For example, *Nature Neuroscience* and *Psychophysiology* have both recently instituted a methods checklist, which is designed in part to help editors and reviewers determine whether the statistical methods of a paper may have led to an inflation of the Type I error rate (Editorial, 2013; Keil et al., in press).

To make this more concrete, the following discussion will refer to an actual set of data from a published oddball experiment. The actual experiment, which was described in some detail in chapter 1 (see figure 1.4), involved rare and frequent categories of visual stimuli that were presented to schizophrenia patients and healthy control subjects (Luck et al., 2009). To keep things simple, I will discuss only the data from the control subjects, focusing on the comparison between the rare and frequent trials. In addition, I will include the data from only 12 of the subjects. Note that there wasn't anything wrong with the way we analyzed the data in the published paper, but it would be possible for someone to analyze these data in a way that capitalized on noise and inflated the Type I error rate.

Figure 10.5A shows the grand average waveforms for the rare (target) and frequent (standard) stimuli at a set of left- and right-hemisphere electrode sites (midline sites were recorded in the original experiment, but they will be ignored here). The data were recorded with a sampling rate of 500 Hz and an online band-pass of 0.05–100 Hz (half-amplitude cutoffs) using a right earlobe reference. For the analyses presented here, the data were re-referenced offline to the average of the left and right earlobes and filtered again with a low-pass filter (half-amplitude cutoff = 30 Hz, slope = 12 dB/octave). You can see a much larger P3 for the rare stimuli than for the frequent stimuli, especially at central and parietal electrode sites. You can also see a more negative voltage for the rare stimuli than for the frequent stimuli in the N2 latency range (250–350 ms) at the central and parietal electrode sites.

An Example of the Problem of Multiple Implicit Comparisons

To make the problem of multiple implicit comparisons clear in the context of ERPs, I performed a simple simulation using the data from this example experiment. The goal of the simulation was to create realistic data from two conditions in which the null hypothesis was clearly true (i.e., no real differences between conditions). To do this, I started by extracting the EEG epochs from the standard (frequent) trials in the oddball experiment. The target (rare) stimuli were left out of this simulation. For each subject, I randomly divided the 640 standard trials into two sets, 512 that that were averaged together to form condition A and 128 that were averaged together to form condition B (80% and 20%, respectively). All 640 trials were actually from the standard stimuli, and the division into conditions A and B was purely random. Thus, this simulates an oddball experiment in which there is true difference between the standards (condition A) and the targets (condition B). Any differences between them in this simulation are purely due to noise. If each

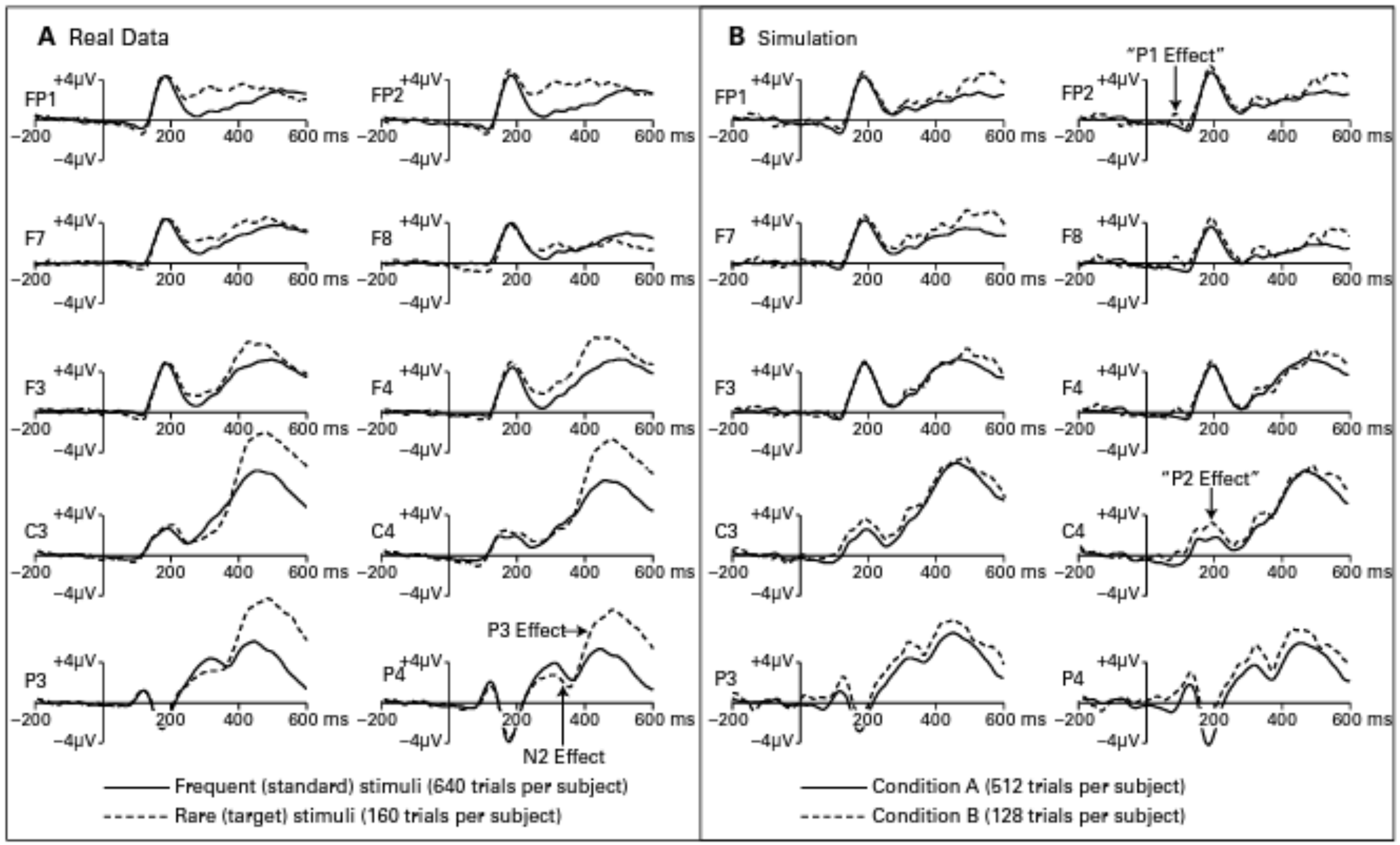

Figure 10.5
(A) Grand average ERP waveforms from an example oddball experiment in which 20% of the stimuli were letters and 80% were digits (or vice versa, counterbalanced across trial blocks). The data are from a subset of 12 of the control subjects who participated in a published study comparing schizophrenia patients with control subjects (Luck et al., 2009). (B) Grand average ERP waveforms from a simulated experiment in which the null hypothesis is guaranteed to be true. This simulation was based on the data collected on the standard (frequent) trials in panel A. The event codes from the 640 standard trials for each subject were randomly sorted into a set of 512 trials for condition A and a set of 128 trials for condition B. Averaged ERP waveforms were then created from these sets of trials. Note that the two conditions differ in P1 amplitude over the entire right hemisphere and in P2 amplitude over the central and parietal regions, but these differences are the result of false positives. The data shown in this figure are referenced to the average of the left and right earlobes and were low-pass filtered offline (half-amplitude cutoff = 30 Hz, slope = 12 dB/octave).

subject had 1 million trials instead of 640 trials, the averaged waveforms for the 800,000 trials in condition A would have been virtually identical to the 200,000 trials in condition B. However, with only 512 trials in condition A and 128 trials in condition B, some random noise was present in the averaged ERPs from these conditions. Thus, the null hypothesis was true (i.e., the data were sampled at random from a single population), but the actual waveforms were not exactly identical (as will always be the case when you are averaging together a finite number of trials). It's important that you understand this, because the idea of randomly combining the observed data in this manner to simulate the null hypothesis will come up again later in the chapter.

Figure 10.5B shows the grand average waveforms from conditions A and B in this simulated experiment. The waveforms from these conditions are fairly similar, and in this simulation we know that the differences must reflect random noise. However, if you ran an experiment with two different conditions and you saw these waveforms, you wouldn't know if the small differences between the waveforms were a result of random noise or true differences in brain activity produced by the two different conditions. You might notice that there are two interesting "effects" in the data. First, the waveforms exhibit a more positive potential in condition A than in condition B in the P1 latency range (50–150 ms), especially in the right hemisphere electrode sites. Second, the waveforms exhibit a more positive potential in condition A than in condition B in the P2 latency range (150–250 ms) at the central and parietal electrode sites.

To quantify the "P1 effect," you might measure the mean amplitude between 50 and 150 ms at all of the electrode sites and look for a condition × hemisphere interaction. I did this with the simulation data, and I found a marginally significant main effect of condition ($p = 0.051$) and a significant condition × hemisphere interaction ($p = 0.011$). I did a follow-up comparison, and I found a significant difference between conditions A and B at the right hemisphere sites ($p = 0.031$). To quantify the "P2 effect," you might measure the mean amplitude between 150 and 250 ms at the C3, C4, P3, and P4 electrode sites. I did this with the simulation data, and I found a significant difference between conditions A and B ($p = 0.026$). This seems like a real difference, but in this simulation we know with 100% certainty that this "effect" is completely bogus. If you saw these data and analyses in a journal article, they might seem like convincing evidence that these two conditions differed in P1 amplitude over the right hemisphere and in P2 amplitude over the central and parietal regions. But in fact these differences were false positives, because the two conditions were the same except for random noise.

Why are we seeing these effects that look real and are statistically significant even though the two conditions are actually just random samples from a single condition? The answer is that we have enough data points and channels that random fluctuations in the data are likely to create substantial differences at some channels and some sites. And these effects are likely to form realistic-looking clusters across nearby time points and nearby electrode sites, because EEG noise changes gradually over periods of tens of milliseconds and is blurred by the high resistance of the skull so that it is present at multiple nearby electrode sites. These data were also low-pass filtered (half-amplitude cutoff at 30 Hz, 12 dB/octave), which also causes a temporal spreading of noise in the data (see chapter 7).

Box 10.3
I Plead Guilty!

Many decades ago, when ERP researchers recorded from only a few electrodes and used peak measures of amplitude and latency based on relatively wide measurement windows, the problem of multiple implicit comparisons wasn't a big deal. There just weren't many choices that researchers could make when analyzing their data. Consequently, no one worried much about this problem. As the number of electrode sites gradually increased, and the use of measures that are more dependent on the time window became more prevalent, the problem of multiple implicit comparisons slowly increased. Because this problem increased so gradually—and the practice of looking at the waveforms to choose the analysis parameters was ingrained in the culture—it took a while before people started taking this problem seriously. By comparison, the problem of multiple comparisons was obvious from the very first days of functional neuroimaging, because researchers were measuring activity from thousands of voxels even in the first studies. Consequently, this problem has been addressed by neuroimaging analysis methods from the beginning, but most ERP researchers didn't take it very seriously until about 10 years ago.

When I think back to my own published ERP studies from the 1990s, I realize that I often used the data to guide my choices of time windows and electrode sites, and I'm certain that this inflated my Type I error rate. That is, it's very likely that more than 5% of the results that I reported as being significant were bogus. This is not an easy thing to admit! However, I have always taken seriously the idea that replication is the best statistic (see box 10.1), and most of the key effects that I reported in those studies were replicated (either in the same paper or in a subsequent paper). The ones that are likely to be bogus are the small, minor, unanticipated effects that I never bothered to replicate. But some of the major effects have not yet been replicated and might also be false positives.

If you look at my more recent papers, you will see that I am now making a concerted effort to address this problem. For example, one recent paper used the negative area measure described in chapter 9, which is less sensitive to the choice of measurement window, and this paper also used permutation statistics to control the Type I error rate (Sawaki et al., 2012). Another paper included two experiments that were nearly identical, but differed in the brightness of the stimuli (Zhang & Luck, 2009). The time windows used in the first experiment were guided by the observed waveforms, whereas the time windows used in the second experiment were based on the windows used in the first experiment (but shifted in time to reflect the fact that latencies are earlier for brighter stimuli). A third paper simply measured the amplitude in a series of consecutive 100-ms latency windows and then used time as a factor in the ANOVA (Gamble & Luck, 2011; see figure 10.6).

I hope it is now clear to you that random noise in the data can easily look like a real effect and can be statistically significant, especially if you choose to measure from a particular time window and set of electrode sites on the basis of the effects you see in the data. This is the problem of multiple implicit comparisons. And note that it was exacerbated by the inclusion of an electrode hemisphere factor in the ANOVAs, giving us more opportunities to see a bogus interaction. See box 10.3 for some further thoughts about this problem (and a shocking admission of guilt).

The waveforms shown in figure 10.5B contain some clues that the P1 and P2 "effects" are false positives. First, the P1 effect is very early for a cognitive manipulation, and it does not have the scalp distribution that one typically sees for early visual ERP components. Second, the P2 effect

begins near time zero and extends for hundreds of milliseconds, which is a common pattern for bogus effects. These issues were discussed in the section on "Baseline Correction" in chapter 8 (see figures 8.2 and 8.3). As I noted in that chapter, "it is important to look at the baseline and be very suspicious about any effects (differences between groups or conditions) that begin near time zero. . . . In addition, you should be highly suspicious about any effects that begin within 100 ms of stimulus onset unless they reflect differences in the stimuli (e.g., larger sensory responses for brighter stimuli)." It is possible for real effects to show these patterns, but these patterns indicate that the effects are probably false positives (especially if they were unexpected or if you used the waveforms to guide your selection of time windows and electrode sites).

The problem of multiple implicit comparisons leads to a somewhat counterintuitive principle:

The more-is-less principle The more conditions, time points, and electrodes are in your data, the less true statistical power you will have.

For example, if you have data from two conditions and two electrode sites, and you limit your analyses to a time window of 200–300 ms, you will have very few opportunities for noise to impact your data, and you will have very few implicit or explicit choices to make about how to analyze your data. Consequently, it will be unlikely that a bogus effect will be statistically significant. However, if you have data from 42 conditions and 128 electrode sites, and you look at the entire time period from 50 to 1500 ms poststimulus, there are thousands of opportunities for noise to produce statistically significant effects. Anything you do post hoc to avoid this inflation of the Type I error rate will reduce your statistical power. Thus, having more conditions and more electrode sites may make it more difficult for you to find the truth. This is one of several reasons why having a large number of electrodes can be problematic (see the chapter 5 supplement). See box 10.4 for additional discussion of the more-is-less principle.

Solving the Problem of Implicit Comparisons

There are several ways that you can solve the problem of multiple implicit comparisons and avoid inflating your Type I error rate. There is no one best solution for all studies, so you will need to choose the best approach given the nature of your own research. Whatever you choose, you should provide an explicit justification in your methods or results section. That way, reviewers and readers will know that you are being careful.

The best time to start thinking about this is when you are designing your experiment. You may want to limit the number of conditions and electrode sites to minimize the number of possible comparisons you could make (because of the more-is-less principle). And you may want to design the experiment so that the time windows and electrode sites that you've used previously can be used again in the current study.

A Priori Hypotheses In many cases, you can use prior research rather than the current waveforms to guide your selection of time windows and electrode sites. For example, when my

Box 10.4
When More Is Less

I first encountered the more-is-less principle when I was in graduate school and was conducting spatial cueing experiments. In these experiments, a cue (e.g., an arrow) indicates that a subsequent target is likely to appear in a particular location. The target appears in the cued location on most trials (called *valid trials*) but sometimes appears in an uncued location (called *invalid trials*). The idea is that subjects will focus attention onto the cued location, yielding improved processing (in terms of reaction time, accuracy, and/or ERP amplitudes) when the target appears at the cued location. Many experiments also include *neutral trials*, in which no information is provided about the location of the subsequent target. On these trials, one would expect that attention is broadly or randomly distributed, leading to performance that is somewhere between valid and invalid trials. After conducting several such experiments and reading many papers from other labs, I realized that performance from the neutral trials was sometimes very close to performance on the valid trials, and sometimes closer to performance on the invalid trials. But it did not seem very systematic. Eventually I realized that a little bit of random variation in performance on any of the three trial types could have a fairly substantial effect on exactly where the neutral trials fell relative to the valid and invalid trials. I and other researchers were constantly coming up with post hoc explanations for the patterns we were seeing in specific experiments, but most of these patterns were likely a result of noise (although a few have been shown to be consistent).

This led me to the following realization: The more conditions I included in an experiment, the more likely it was that I found a "weird" result in one of the conditions that was caused by random noise. Consequently, I started narrowing down my experiments to include only the essential conditions. This gave me fewer opportunities to observe random variations, and fewer weird findings to explain. And in retrospect, it gave me fewer opportunities to conduct multiple implicit comparisons, so it reduced my Type I error rate.

Of course, this approach comes at a cost: by including fewer conditions, I am putting on "theoretical blinders," and I may miss interesting results that could be seen by including more conditions. Consequently, when I am trying a completely new type of experiment, I go ahead and include lots of conditions, and I assume I will need to conduct follow-up experiments to replicate the findings of this experiment. But when I am testing a specific hypothesis in a well-developed paradigm, I focus on the smallest number of conditions that can test the hypothesis.

students conduct an N2pc experiment with highly salient target stimuli, I know from previous experience that they should measure N2pc amplitude from the lateral posterior electrode sites from approximately 175 to 275 ms. I don't have much more to say about this approach, but don't take this to mean that it isn't a useful approach. In a large proportion of experiments, this is by far the best approach.

This approach obviously can't work when you're trying a completely new experimental paradigm. When you find yourself in this situation, the best option is often to conduct a follow-up experiment so that you can use the results from the first experiment to guide the choice of time windows and electrode sites (and so that you can demonstrate the replicability of your findings, as recommended in box 10.1).

Functional Localizers This is a variant of the idea of an *a priori* hypothesis, but it is based on data that you collect at the same time as your main experiment. The idea—which is very popular in functional neuroimaging—is that you can use one very simple and well-understood manipulation to determine the time course and electrode sites of a given effect, and then you can apply this to the comparison that is the main focus of your experiment.

Imagine, for example, that you want to determine whether the N170 component is larger for smiling faces than for frowning faces. This might be a very small effect, and to maximize statistical power it would be useful to know the optimal time window and electrode sites for measuring the voltage. In fact, you might want to determine the optimal time window and electrodes separately for each individual subject. You could do this by including a condition in which you record the ERPs elicited by faces and cars, which is known to elicit a very robust N170 effect (see, e.g., figure 1.2 in chapter 1). You could then make a face-minus-car difference wave and determine the onset and offset time of the N170 effect and the electrode sites at which this effect is present. This would define the measurement window and electrode sites that you would use for measuring the amplitude of the smiling-minus-frowning difference wave. You could do this individually for each subject, or you could measure the face-minus-car difference in the grand average and apply the results uniformly to every subject.

This approach is used very rarely in ERP studies. It has a lot of potential, although it could fail under some circumstances. For example, if the time course of the smiling-versus-frowning difference is not the same as the face-versus-car difference, you won't be selecting the appropriate time window (for a discussion of limitations in neuroimaging studies, see Friston, Rotshtein, Geng, Sterzer, & Henson, 2006). Despite the limitations, it should probably be used in more ERP studies.

Collapsed Localizers This is a variant of a functional localizer, except that it does not require additional conditions in your experiment. Instead of using a contrast between one set of conditions to determine the window and electrodes that will be used to assess a contrast between a different set of conditions, the data are collapsed across the conditions of interest to determine the window and electrodes.

In our N170 example, we could simply average across the smiling and frowning faces and use the overall N170 to determine the time window and electrode sites that are best for measuring the N170. This window and set of electrodes would then be applied to measure the N170 for the separate smiling and frowning ERPs. An obvious shortcoming is that the N170, when not measured from a difference wave, contains both face-specific and face-nonspecific activity. Consequently, the timing and scalp distribution of the N170 from the collapsed waveform may be quite different from the timing and scalp distribution of the face-specific activity. Nonetheless, there are situations where this can be a reasonable approach.

Now let's consider a different example, in which an oddball paradigm is used with both a patient group and a control group. Imagine that we wanted to ask whether the peak latency of the P3 wave in a rare-minus-frequent difference wave was later in the patients than in the

controls. First, we would construct a grand average rare-minus-frequent difference wave that combines all patients and controls. We would then find the latency range and scalp sites that showed the largest P3 wave in this combined difference wave. We could then use this information to define the time window and electrode sites to be used for measuring P3 peak latency in the rare-minus-frequent difference waves from the individual subjects in each group. This approach would minimize bias in a statistical comparison between patients and controls because the differences between patients and controls were not used to select the measurement window and scalp sites used in the comparisons.

This example raises a potential problem with using a combined waveform to determine the window and electrodes. Imagine that the P3 was much larger in the control subjects than in the patients (which is a common finding for many types of patient groups). And imagine that the scalp distribution differed for the patients and controls. The larger amplitude of the P3 in the control subjects would mean that the grand average across groups would be dominated by the control subjects, and the scalp distribution that was selected on the basis of the collapsed data might therefore be more appropriate for the controls than for the patients. This could bias the results in favor of the control subjects. This kind of scenario is relatively rare, so it does not usually preclude the use of this approach. However, if the amplitudes differ across the conditions or groups that you are collapsing, you will want to think carefully about whether this approach might somehow bias your results.

Looking for Condition × Electrode Interactions Imagine that you have *a priori* knowledge that 200–300 ms is a good measurement window, but you don't know which of your 128 electrode sites will contain the effect of interest. You could simply measure the mean amplitude from 200 to 300 ms at all electrode sites and then conduct an ANOVA in which electrode site is a factor. Because the difference between conditions is likely to be large at a subset of the sites and small or even opposite at others, you probably won't see a significant main effect of condition in this ANOVA. Instead, you will be looking for a condition × electrode site interaction. Earlier in the chapter, I discussed some problems that arise with condition × electrode interactions, including the possibility of spurious results owing to the larger number of p values and the difficulty of distinguishing between multiplicative and non-multiplicative interactions. However, the former problem arises mainly when the electrode interactions were not predicted, and the latter problem arises mainly when you are trying to draw conclusions about changes in the locations of the underlying generators of the effects. If you are simply trying to conclude that two conditions are different, which will inevitably lead to larger differences at some sites than at others, you can use a condition × electrode site interaction to demonstrate your conditions are significantly different. I have used this approach successfully in several papers. In addition, if this interaction is significant, you are justified in conducting a follow-up analysis to see which electrode sites exhibit a significant main effect of condition.

The main shortcoming of this approach is that the statistical power for seeing an interaction of this nature is usually much lower than the power you would have for seeing a main effect of

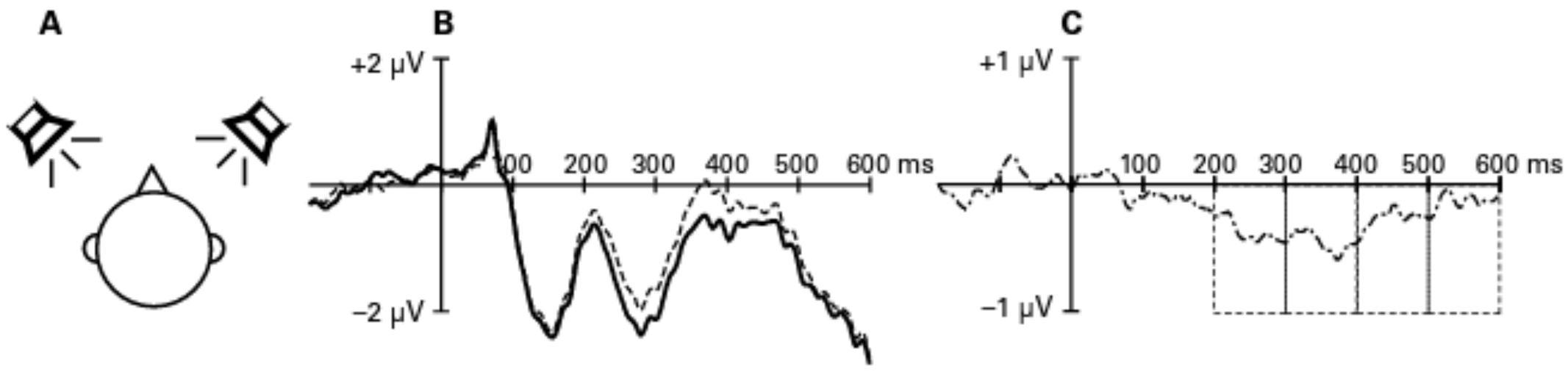

Figure 10.6
Stimuli and data from an experiment designed to find an auditory analog of the N2pc component (Gamble & Luck, 2011). (A) Stimuli. Subjects listened to pairs of auditory stimuli, presented simultaneously in separate speakers. They were asked to detect a particular target sound, which could occur unpredictably in either speaker on a given trial. (B) Grand average ERP waveforms recorded at a cluster of anterior electrode sites (F3, F7, C3, T7, F4, F8, C4, T8), contralateral or ipsilateral to the location of the target. The waveforms are referenced to the average of the mastoids and were filtered offline with a low-pass Gaussian filter with a half-amplitude cutoff at 50 Hz. (C) Difference wave (contralateral minus ipsilateral) from the data shown in panel B. The dashed rectangles show the time windows that were used to measure mean amplitude.

condition if you limited the ANOVA to the electrode sites where the effect is large. The advantage, however, is that it does not require prior knowledge of which sites will have a large effect, allowing you to avoid biasing the results by using the data to help you decide which electrodes to use. But keep in mind that looking for condition × electrode interactions in a post hoc manner can lead to an increase in the Type I error rate.

Looking for Time × Electrode Interactions An analogous time-based approach can also be used to avoid choosing a specific time window. That is, you can measure the mean voltage from multiple time windows across a broad range, and then include time as a factor in the ANOVA. For example, Marissa Gamble ran a study in my lab in which she was searching for an auditory analog of the N2pc component (figure 10.6). She found a nice contralateral effect for auditory stimuli over anterior electrode sites (which we called *N2ac* for N2-anterior-contralateral) (Gamble & Luck, 2011). We had no strong *a priori* hypothesis about the time window for this effect, because we couldn't assume that it would be the same as the time window of the visual N2pc component. We looked at the data and tried many different measurement windows, but we realized that this was inflating the probability of a Type I error.

To avoid this problem, we simply measured the mean amplitude in consecutive 100-ms windows between 200 and 600 ms, and we conducted an ANOVA in which time window was a factor. We measured N2ac amplitude as the difference in voltage between the contralateral and ipsilateral electrodes (relative to the location of the target) in each of these time windows (figure 10.6C). This difference was significantly greater than zero when we collapsed across the measurement windows (analogous to a significant main effect of contralateral-versus-ipsilateral). In addition, we found an effect of time period on this difference (analogous to an interaction between time window and contralateral-versus-ipsilateral). This justified follow-up analyses of

the individual time windows, which indicated that the difference was significantly greater than zero in the 200–300 ms, 300–400 ms, and 400–500 ms windows, but not in the 500–600 ms window.

This approach worked well given that we had no good *a priori* information about what time window to use for measuring the N2ac. That is, despite the fact that we did not inflate the Type I error rate by choosing measurement windows on the basis of the data, the very clear effects were statistically significant. However, part of the reason that it worked in this case was that the effect was present over most of our overall time range. If it had been present only in one of the four 100-ms windows, we would have had relatively low power for detecting the effect (as either a main effect or an interaction with time). If I were to analyze these data today, I would use the *mass univariate approach* described in online chapter 13. In addition, we used the multiple-consecutive-windows approach because these were our first N2ac experiments, so we had no *a priori* information to guide our selection of time windows. I would not take this approach with future N2ac experiments, because I could use our existing N2ac results to guide the choice of the time window for future experiments.

Window-Independent Measures The problem of defining the measurement window can be very severe in some experiments. An example of this, which was discussed in chapter 9 (see figure 9.2), arises when you try to measure the difference in N2 amplitude between rare and frequent trials in an oddball paradigm. Even if you measure from a difference wave, the N2 may be preceded by a P2 effect and followed by a P3 effect, which may partially cancel the N2 effect (see, e.g., the single subject shown in figure 10.7). Moreover, the latencies of these effects may vary somewhat from subject to subject. If you are measuring mean amplitude, you will need to use a very narrow window to avoid having your N2 measure canceled by P2 or P3 activity in some of the subjects, but the use of a narrow window increases the impact of noise on your measurements. In addition, the results that you get could be influenced by exactly what window you choose, and if you choose the window by looking for the period with the largest effect, this

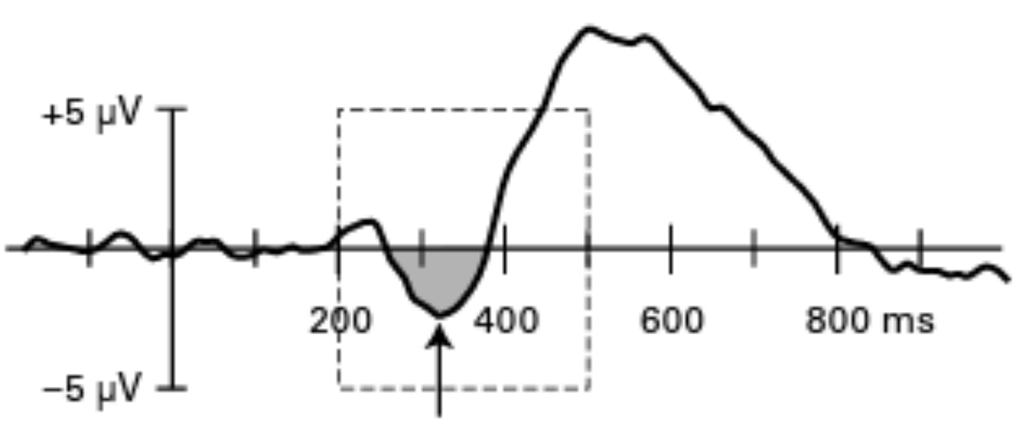

Figure 10.7
Example of how signed area can be used with a broad measurement window. The waveform here is a single-subject difference wave (rare minus frequent), and negative area was used to quantify N2 amplitude. Because only the area below the 0-μV line contributes to the measurement, the measured area would be the same with the broad window shown here (200–500 ms) as with a narrower window (e.g., 250–400 ms). This makes the measurement relatively independent of the choice of time window. The arrow points to the peak voltage.

will inflate your Type I error rate. Although I generally encourage the use of mean amplitude as a measure, this approach can be problematic when the effect you are trying to measure is surrounded by effects on other components, especially when the other components are large. Measuring from difference waves can help, but not if the other components are still present in the difference waves. The best solution, when possible, is to design the experiment so that only one component is present in the difference waves. However, that is not always possible, so you may need a different solution.

One good solution was described in detail in chapter 9; namely, the use of signed area measures. In our oddball experiment, for example, you could quantify the N2 as the area of the region falling below the zero line over a very broad time window (e.g., 200–500 ms; see figure 10.7, along with figures 9.2 and 9.3). This would ordinarily be done from a difference wave, because the zero line is much more meaningful in a difference wave.

A second solution would be to use peak amplitude. Again, this works best in a difference wave that isolates a small number of components. A wide window is possible in this situation (e.g., 200–500 ms), because the positive-going P2 and P3 waves on either side of the N2 make it unlikely that some other negative peak will be measured instead of N2 (see figure 10.7). Chapter 9 described many shortcomings of peak amplitude, but this is one situation in which it may sometimes be superior to mean amplitude.

For both of these solutions, however, you should be aware that the surrounding P2 and P3 components may still have a large impact on your N2 amplitude measurement. That is, these components are likely to overlap with the N2 and will impact the measured amplitude of the N2, even if the overall voltage in this window is negative. The use of signed area or peak amplitude may reduce the cancellation of N2 produced by the P2 and N2 waves, but it will not eliminate it. For example, if you were comparing rare-minus-frequent difference waves between two groups, a larger P3 wave in one group could make it appear that the N2 was smaller (less negative) in that group (see, e.g., figure 2.5F in chapter 2). In addition, both peak amplitude and signed area amplitude are biased: they tend to increase when the signal-to-noise ratio decreases (for an example of how we dealt with this problem when using signed area measures, see Sawaki, Geng, & Luck, 2012).

Correcting for Multiple Comparisons Instead of treating your analysis as a problem of multiple *implicit* comparisons, you could treat it as a problem of multiple *explicit* comparisons. In the oddball experiment shown in figure 10.7, for example, you could measure the voltage at every time point between 200 and 500 ms and do a *t* test comparing each voltage to zero (or comparing the voltage in the rare and frequent conditions at each time point). You could then perform a correction for multiple comparisons. This is called the *mass univariate approach*, because a massive number of individual statistical tests are used. When I wrote the first edition of this book, I didn't seriously consider raising this possibility, because the standard approach to correcting for multiple comparisons (the Bonferroni correction) has ridiculously low statistical power in this situation. However, there have been some important advances in the statistical

analysis of ERP data, and it is now possible to explicitly correct for multiple comparisons without a ridiculous loss of power. This approach is very powerful and becoming increasingly common, so I have provided an additional chapter that describes it in some detail (see online chapter 13).

Suggestions for Further Reading

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology, 48*, 1711–1725.

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields II: Simulation studies. *Psychophysiology, 48*, 1726–1737.

Jennings, J. R., & Wood, C. C. (1976). The ε-adjustment procedure for repeated-measures analyses of variance. *Psychophysiology, 13*, 277–278.

Kiesel, A., Miller, J., Jolicoeur, P., & Brisson, B. (2008). Measurement of ERP latency differences: A comparison of single-participant and jackknife-based scoring methods. *Psychophysiology, 45*, 250–274.

Maris, E. (2012). Statistical testing in electrophysiological studies. *Psychophysiology, 49*, 549–565.

McCarthy, G., & Wood, C. C. (1985). Scalp distributions of event-related potentials: An ambiguity associated with analysis of variance models. *Electroencephalography and Clinical Neurophysiology, 62*, 203–208.

Miller, J., Patterson, T., & Ulrich, R. (1998). Jackknife-based method for measuring LRP onset latency differences. *Psychophysiology, 35*, 99–115.

Miller, J., Ulrich, R., & Schwarz, W. (2009). Why jackknifing yields good latency estimates. *Psychophysiology, 46*, 300–312.

Urbach, T. P., & Kutas, M. (2002). The intractability of scaling scalp distributions to infer neuroelectric sources. *Psychophysiology, 39*, 791–808.

Urbach, T. P., & Kutas, M. (2006). Interpreting event-related brain potential (ERP) distributions: Implications of baseline potentials and variability with application to amplitude normalization by vector scaling. *Biological Psychology, 72*, 333–343.

Ulrich, R., & Miller, J. (2001). Using the jackknife-based scoring method for measuring LRP onset effects in factorial designs. *Psychophysiology, 38*, 816–827.