8 Baseline Correction, Averaging, and Time–Frequency Analysis

Overview

This chapter mainly focuses on baseline correction and averaging, which are very simple processes. Baseline correction is achieved by subtracting the average prestimulus voltage from the waveform, and averaging simply consists of summing together a set of EEG epochs and then dividing by the number of epochs. Simple, right? Well, these processes are simple, but their effects can be quite complex, and they can lead to important misinterpretations of the results. I will therefore spend quite a bit of time explaining how they really work and how they can lead you to an incorrect conclusion if you're not careful.

I will begin the chapter by discussing the epoch-extraction and baseline-correction processes that typically precede averaging. I will then describe how averaging decreases the noise in your data and how the number of trials will influence the *p* value you get at the end of your experiment. In this section, I will provide some general advice about how many trials you will need to average together. I will also demonstrate how different subjects can have very different-looking average ERP waveforms and discuss how these differences arise.

I will then discuss trial-to-trial variations in the amplitude and latency of an ERP component, focusing on how latency variability can lead to incorrect conclusions and providing some solutions to this problem. This issue will be described in greater detail in online chapter 11, which also introduces a useful mathematical concept called *convolution*.

The chapter will end with an introduction to time–frequency analysis, in which the time course of specific frequency bands is extracted from the data. Time–frequency analysis will be explored in more detail in online chapter 12, but it's really just averaging with an additional preprocessing step.

Extracting Epochs of EEG Data

As discussed in chapter 5, the EEG is usually recorded as a continuous signal for an entire trial block, and event codes are present to indicate the occurrence of stimuli and responses. Prior to averaging, it is necessary to extract fixed-length *epochs* or *segments* of data from the continuous

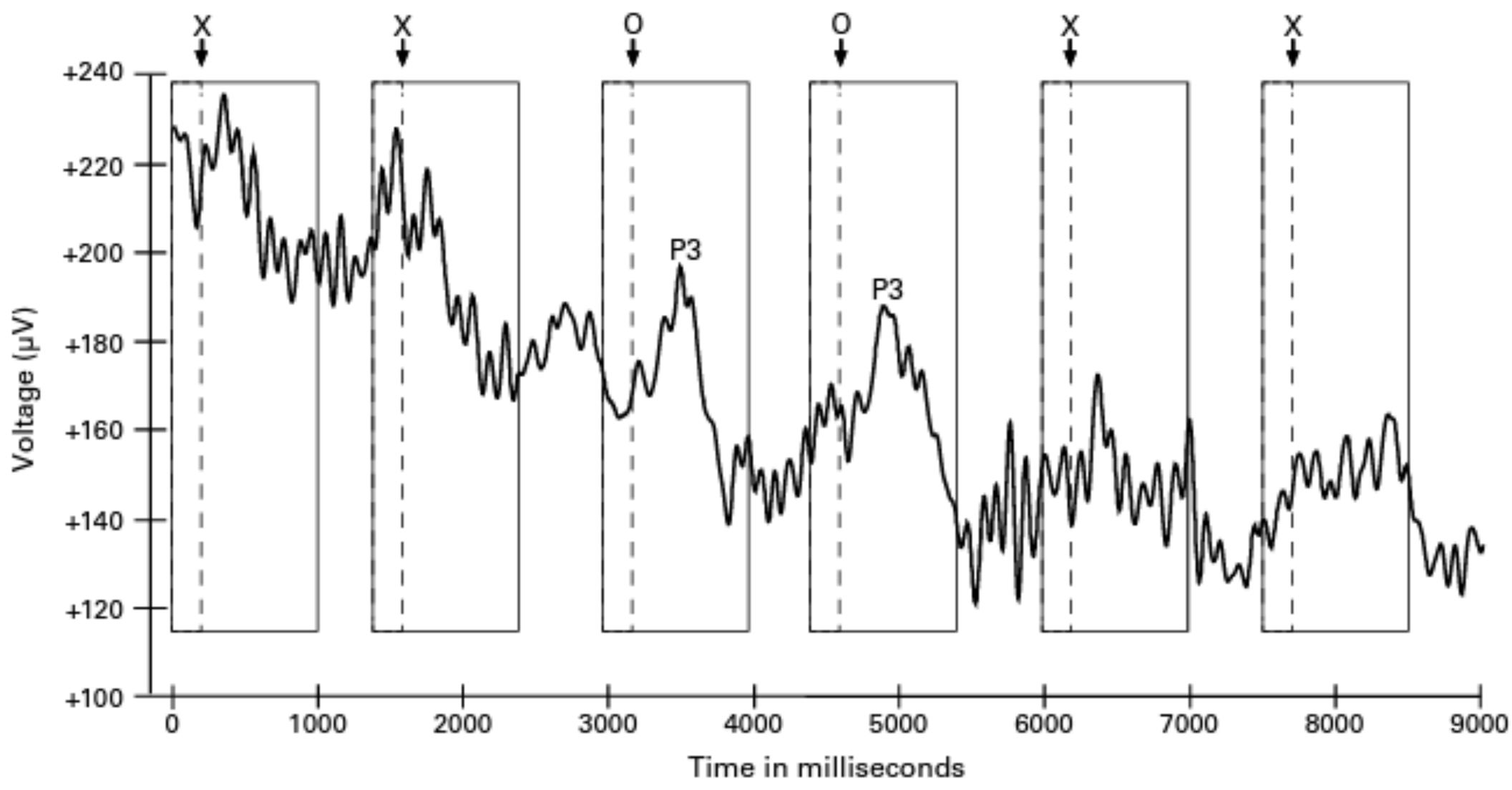

Figure 8.1
Epoch extraction. This example shows 9000 ms of EEG from the Pz electrode in an oddball experiment, with frequent X standards and rare O targets. By chance, there were two consecutive targets in this particular EEG segment. A large P3 component is visible after each target. Prior to averaging, 1000-ms segments of the EEG are extracted. Each segment starts 200 ms prior to an event code and continues until 800 ms after the event code. Note that the EEG has a substantial voltage offset (which is common when the data have not been high-pass filtered), and it also drifts downward due to a changing skin potential. Baseline correction is used to remove this drift, usually at the time of epoch extraction.

EEG, time-locked to the event codes of interest. Each epoch includes a baseline period prior to the event code (typically 100–200 ms) as well as a period after the event code (typically 500–1500 ms, depending on which components will be examined). This is illustrated in figure 8.1, which shows a 9000-ms period of continuous EEG from an oddball experiment in which the letter X was frequent (90%) and the letter O was rare (10%). This particular period of EEG happened to contain four Xs and two Os, which may not seem to fit with the 90%/10% probabilities, but that's the sort of thing that happens with a random sequence. You can see that each of the two rare O stimuli elicited a large P3 wave that is visible in the raw EEG.

Prior to averaging across trials, epochs of data were extracted from the continuous EEG, beginning 200 ms prior to each stimulus and ending 800 ms after each stimulus. Most current EEG recording systems will ordinarily save the entire continuous EEG onto the hard drive, but some systems give you the option of saving only discrete epochs around the event codes. Saving the epochs rather than the continuous EEG usually saves disk space, but I generally recommend against it during recording because this limits your options during analysis. For example, a reviewer may ask for a longer baseline period or for response-locked averages, and you will be out of luck if you did not save the continuous EEG. Also, as described in chapter 7, it is usually

better to apply high-pass filters to long periods of continuous EEG rather than to EEG epochs. Note that when the epochs are long relative to the rate of stimulation, one epoch may begin before the previous one has ended (which is not a problem).

Baseline Correction

After you extract the epochs from the continuous EEG, you will typically perform a baseline correction procedure. This is necessary because factors such as skin hydration and static charges in the electrodes may cause an overall vertical offset in the EEG (as was discussed in detail in chapter 5). Figure 8.1 illustrates this offset, showing that the EEG voltage is well over 100 µV throughout the entire 9000-ms interval shown in the figure. If we did not do some kind of correction procedure, the voltage offset would have a huge impact on our ERP amplitude measurements. For example, if we averaged the EEG epochs for the rare stimuli and then measured the P3 amplitude without performing some kind of baseline correction, the measured value would be approximately 180 µV for this subject. For another subject or another period of time, the measured value might be something like –270 µV (if there was a negative voltage offset for that subject or that time period). These differences in voltage offset across time periods and subjects are completely random and are unrelated to brain activity. If we did not remove these offsets, they would add tremendous uncontrolled variance to the data, making it nearly impossible to see significant differences between conditions or groups.

Factors like skin hydration, skin potentials, and static electrical charges often vary slowly over time, causing the offset voltage to drift gradually upward and downward. In the EEG shown in figure 8.1, for example, the voltage gradually drifts downward across the 9000-ms interval. We need to correct for this drift or else it will add substantial error variance to our amplitude measurements, reducing statistical power.

Baseline correction is a simple procedure that can minimize these offsets and drifts. In most cases, the time-locking event for the epoch is a stimulus, and we can assume that the voltage during the prestimulus period can provide a good estimate of the voltage offset for that trial (because it contains the offset but does not contain any stimulus-elicited ERP activity). If we simply subtract this estimate of the offset from the entire epoch, this will eliminate the offset. This is shown in figure 8.2A. In this figure, my data-analysis program computed the average voltage during the prestimulus period to estimate the voltage offset, and then it subtracted this value from every point in the epoch. This just shifts the waveform upward or downward (depending on whether the voltage offset is positive or negative), centering the prestimulus period around the 0-µV line (hint: if this is working correctly, the area of the waveform above the zero line should be the same as the area below the line during the prestimulus period).

In many cases, this baseline correction works beautifully to minimize voltage offsets and gradual drifts. However, it assumes that the prestimulus period contains only the voltage offset, and this assumption is not always correct. For example, the prestimulus period may contain ERP activity from the end of the preceding trial or it may contain preparatory activity that occurs

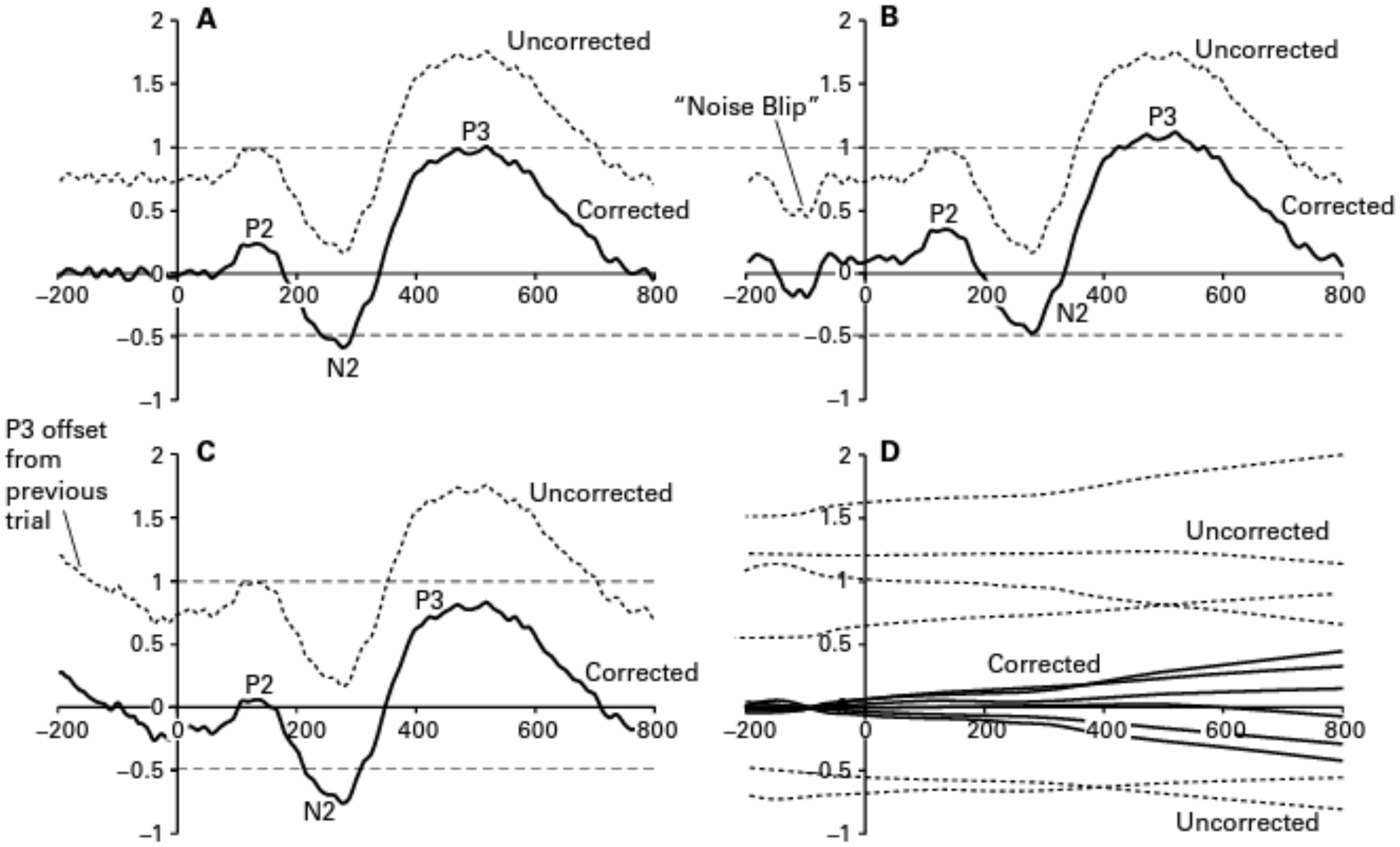

Figure 8.2
Baseline correction. (A) Without baseline correction, the averaged ERP waveform is shifted vertically (reflecting the overall voltage offset of the EEG). The vertical shift is corrected by computing the average voltage during the prestimulus interval and subtracting this voltage from each point in the waveform. This is ordinarily done on the single-trial EEG epochs, but it can also be done on the averaged ERP waveforms. (B) In this example, a negative-going voltage deflection (noise blip) is present during the prestimulus period, and the baseline correction therefore fails to "push" the waveform down as far as it should go. This affects the apparent amplitude of the P2, N2, and P3 components. Horizontal dashed lines at –0.5 and +1.0 μV are provided to facilitate comparison of the amplitudes in panels A and B. (C) In this example, overlapping P3 activity from the previous trial is present during the prestimulus interval, causing the ERP waveform to be pushed too far downward by the baseline correction procedure. (D) Illustration of how trial-to-trial variance tends to increase as time passes from the center of the baseline period. Each dashed line represents slow drift in a single trial without any baseline correction (and without any stimulus-elicited ERP activity, just to make the drift clearer). The solid lines represent the baseline-corrected single trials. The baseline correction shifts each waveform vertically so that they are all together during the baseline period. As time moves away from the baseline period, the signal tends to drift farther and farther away from the baseline voltage.

when the subject anticipates the onset of a stimulus. As we shall see, baseline correction is necessary, but it may lead to unanticipated consequences, so you should think carefully about what it is doing to your data.

Baseline correction is usually performed on the EEG immediately after the data are epoched. This is necessary to avoid problems with some types of artifact rejection (see chapter 6), and it can make the epoched data easier to view. Thus, most software systems perform baseline correction at this point. Some systems allow you to perform baseline correction again after averaging so that you can change the baseline period (assuming that the new baseline period falls within your epochs). You will typically get the same result by performing baseline correction before or after averaging, except when baseline correction affects artifact rejection (see the appendix of this book for a discussion of the factors that influence the order in which you perform operations such as baseline correction, artifact rejection, averaging, and filtering).

High-pass filters will eliminate much of the EEG offset, and you might think that baseline correction would not be needed if you've already applied a high-pass filter. However, filtering is a fairly blunt instrument for removing the offset. In contrast, baseline subtraction is based on the very precise assumption that the voltage during the prestimulus period should contain nothing except the offset and noise. When this assumption is met, baseline correction is the best way to remove the offset. When it is not met, it is still better than filtering alone, because filtering can create systematic differences across conditions if used as the only method of baseline correction. An explicit baseline correction procedure is therefore necessary in almost all conventional ERP studies.

Effects of Baseline Correction on Amplitude Measurements

Although baseline correction is usually performed at the time of epoching, it has a large (and often unappreciated) effect on the amplitude measurements that you make on your averaged ERP waveforms at the end of your data processing pipeline. Specifically, baseline correction involves subtracting the mean baseline voltage from the entire waveform, which therefore affects the amplitude at each point in the waveform. Once you've subtracted the baseline, the voltage at each time point in the waveform represents the difference between that point and the average baseline voltage, and anything that influences the baseline therefore influences your poststimulus amplitude measurements. To illustrate this, figure 8.2B shows how a "noise blip" (a small, random voltage transient) during the prestimulus baseline period influences the amplitude of every component in the ERP waveform. These types of voltage blips are very common in ERP data (and usually reflect EEG noise that remains after averaging a finite number of trials). The negative-going voltage blip in figure 8.2B causes the mean of the prestimulus voltage to be an underestimate of the actual voltage offset, and the voltage subtracted from the ERP waveform is therefore too small. As a result, the waveform isn't "pushed" far enough downward by the baseline correction procedure. This causes the measured amplitude of the N2 to be smaller (less negative) and the measured amplitude of the P3 to be larger (more positive) than they should be. Horizontal lines are drawn at –0.5 μV and +1.0 μV in the figure so that you can see how the peaks are shifted upward in the presence of the noise blip.

Thus, when you measure the amplitude of some poststimulus time period, you should always realize that you are actually measuring the difference between this period and the baseline voltage. Consequently, any noise in the baseline will create noise in your measurements, thus decreasing your statistical power.

Figure 8.2C shows how overlap from the previous trial can influence the amplitude measurements of the ERP components on the present trial. This example simulates the effect of a very short interval between stimuli (~600 ms), which causes the final portion of the ERP from one trial (in this case, the P3 wave) to overlap with the prestimulus period of the next trial. This causes the prestimulus voltage to be an overestimate of the actual offset voltage (greater positivity), and this causes the waveform to be shifted too far downward when the mean prestimulus voltage is subtracted from the waveform. This in turn causes the measured N2 to be artificially large and the measured P3 value to be artificially small. Unlike random noise, this overlap would be consistent across subjects, biasing the voltage measurements toward more negative values for everyone. The same is true of preparatory activity and the offset of blinks during the prestimulus interval (see figure 6.3 in chapter 6). Overlap confounds are on my list of *Top Ten Reasons to Reject an ERP Paper* (see online chapter 15), and overlap is discussed in more detail in online chapter 11. I also encourage you to read a paper on overlap by Marty Woldorff (1993), which is on my list of *Papers Every New ERP Researcher Should Read* (see the end of chapter 1).

As discussed in detail by Urbach and Kutas (2006), baseline correction can have large effects on your ERP waveforms, and you should not assume that it is a benign process that simply centers your waveform on some kind of true zero voltage. All of your poststimulus measures are essentially difference scores between the baseline period and your poststimulus measurement period. This is analogous to the no-Switzerland principle that was described in the context of the reference electrode in chapter 5. That is, just as the voltage at a given electrode site is really the difference between that site and the reference, the voltage at a given time is really the difference between that time and the baseline period.

Figure 8.2D shows an important principle about baseline correction; namely, that slow drifts in voltage tend to cause more and more deviation away from zero the farther you get away from the baseline period. In the first 100 ms after the baseline period, for example, there has not been much time for the voltage to drift away from the mean voltage during the baseline period. By 600 ms after the baseline period, however, there has been more time for the voltage to drift away from the baseline. The drift will sometimes be positive and sometimes be negative, leading to trial-to-trial variation in the amplitude that increases as you get farther and farther away from the baseline period. This increase in variance will make your amplitude measurements less reliable later in the waveform than they are earlier in the waveform (all else being equal). This will tend to reduce statistical power for measurements made very late in the waveform. For this reason, ERPs tend to be especially good for assessing brain activity in the first second or so after an event, but they provide a relatively poor measure of activity many seconds or minutes after an event (assuming you are using the pre-event period as the baseline). If you are interested,

Box 8.1
Accumulation of Variance over Time

The effects of drifts after the baseline period can be understood in terms of a fundamental principle of statistics. Specifically, if you add two uncorrelated random variables to create a new random variable, the variance of the new random variable is equal to the sum of the variances of the two original variables (this is called the Bienaymé formula). For example, if you measure the weight and the annual salary of each individual in a population, and you calculate the weight plus the annual salary of each individual, the variance of this sum will be equal to the variance of the weight alone plus the variance of the salary alone. To apply this principle to EEG drift, imagine that the voltage drifts up and down randomly over time, creating trial-to-trial variations in the mean amplitude measured from 0 to 100 ms after a 100-ms baseline period. If the voltage continues to drift up and down randomly from 100 to 200 ms after the baseline period, this will add even more variation from trial to trial. If we assume that the drift is truly random, then the variance that occurs from 100 to 200 ms will simply add to the variance that occurred from 0 to 100 ms. Thus, if the trial-to-trial variance is X μV in each 100 ms period, then the variance after 200 ms will be $2X$ μV, the variance after 300 ms will be $3X$ μV, and so forth. Note that this is only an approximation because the Bienaymé formula is only true for uncorrelated variables, and the EEG in one time period will be correlated with the EEG in the next time period. Even in this case, however, the variance grows as more and more time passes.

box 8.1 provides a more precise way of describing how the variance across trials increases as time passes from the baseline period.

The fact that the measurements become less reliable as you move farther away from the baseline period has important implications for how you should choose the baseline period. In many language experiments, for example, subjects see or hear sentences containing many words, and the ERP elicited by the last word of each sentence is of interest. You might think it would be best to use the voltage prior to the onset of the sentence as the prestimulus baseline, because it is not contaminated by overlapping activity from prior words (see, e.g., figure 3.13 in chapter 3). However, researchers often use the interval just prior to the final word as the baseline—even though it is contaminated by overlap from the preceding word—because this reduces the amount of slow drift between the baseline period and the ERP elicited by the final word, thus increasing the reliability of the measurements. Although this baseline is contaminated by overlap from the previous word, this overlap is not usually a problem if it is identical across conditions (see online chapter 11 for a detailed discussion). Consequently, it may be worth using a baseline that is contaminated by overlap if it allows the baseline period to be closer in time to the measurement period.

The Importance of Looking at the Baseline

Here's an important bit of practical advice: Whenever you look at a set of ERP waveforms—whether your own or someone else's—you should look at the prestimulus baseline activity before

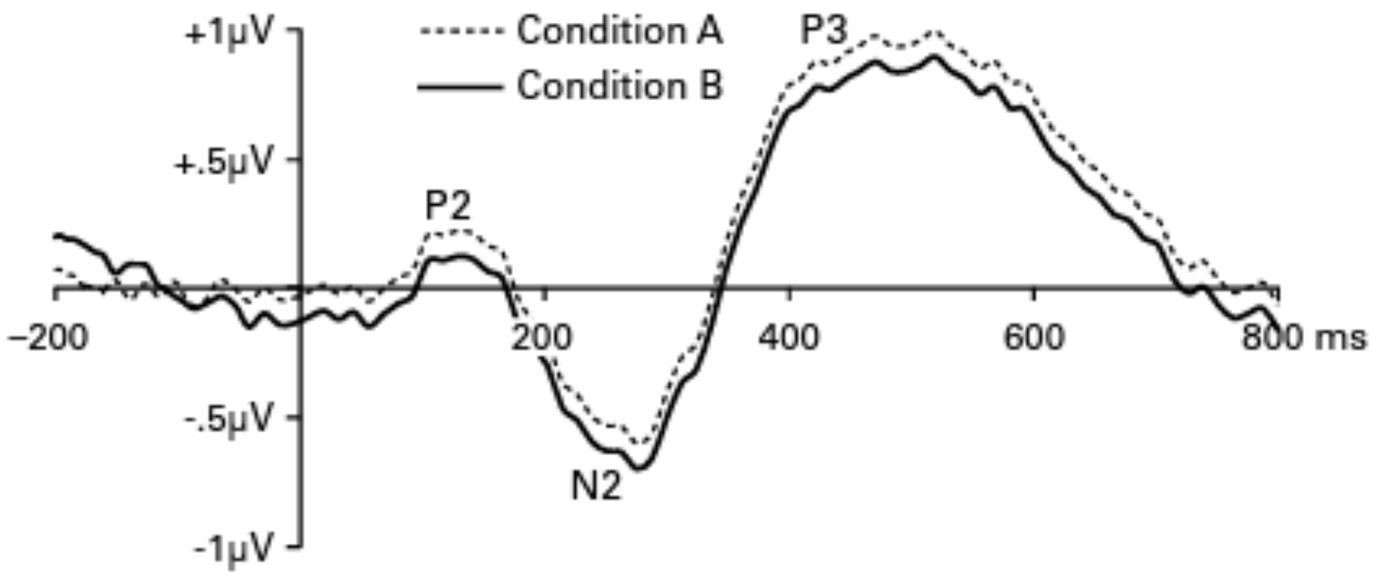

Figure 8.3
Example of how differences between conditions in prestimulus activity can lead to the artifactual appearance of differences in poststimulus activity. The only true differences between these waveforms are in the prestimulus baseline period.

looking at anything else. The baseline interval will tell you several important things. First, it will tell you how much noise remained in the data after averaging. Specifically, if the voltage fluctuations in the prestimulus baseline are as big as the experimental effects in the poststimulus period, you should be very skeptical about whether the effects are real (even if they are statistically significant, which will happen for bogus effects one out of every 20 times).

Second, the baseline will tell you if the waveforms are contaminated by overlapping activity from the previous trial or preparatory activity that occurred before the onset of the stimulus in the present trial. Such effects are not always problematic, but you should definitely think about them.

Third, if the differences between conditions or groups begin in the prestimulus interval or shortly thereafter, they are likely some kind of artifact. This is illustrated in figure 8.3, which shows data from a simulated oddball experiment in which the voltage in the prestimulus baseline period differs between the two experimental conditions. Because the differences in the prestimulus period are subtracted from the entire waveform during the baseline correction process, the prestimulus difference leads to an artifactual difference between conditions during the poststimulus period. This makes the P3 in condition B appear to be smaller than the P3 in condition A, even though the difference is actually in the prestimulus activity. If you did not look carefully at the baseline, you might reach the incorrect conclusion that the P3 differed across conditions. However, if you look at the baseline, you can see that the "experimental effect" begins at time zero, which is of course impossible. Thus, 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. You should also be a little suspicious if you're reading a paper and it doesn't show a prestimulus baseline in the ERP waveform plots (although I have to admit that I am a co-author on a paper in which this happened by accident—see Hopf, Vogel, Woodman, Heinze, & Luck, 2002).

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 very rare that a purely cognitive manipulation influences activity this early. Attention can influence ERP responses within the first 100 ms when attention was focused prior to the onset of the stimulus (reviewed by Luck & Kappenman, 2012a), but most early effects turn out to be a result of noise, artifacts, or confounds.

From the examples shown in figures 6.2 and 6.3 of chapter 6, you might think that baseline correction is a bad idea. After all, baseline correction causes noise and overlap in the prestimulus period to distort the poststimulus voltages. However, if you tried to look at your ERPs without baseline correction, you would be faced with enormous variability from one subject to the next owing to random differences in EEG offset (as illustrated in figure 8.2A). You could try to filter out the EEG offset with a high-pass filter, but this ends up having all the same problems as baseline correction and also creates additional distortions as described in chapter 7 and online chapter 12. Consequently, baseline correction using the mean of the prestimulus voltage is the best option in 99.9% of experiments, and the problems of noise blips and overlap are best solved by recording clean data and by designing experiments in which the overlap is identical across conditions (or using one of the strategies described in online chapter 11 to minimize overlap).

Specific Recommendations for Baseline Correction

The optimal length of a baseline period reflects a balance between the benefits and costs of using a long baseline period. Under ideal conditions, using a longer baseline period will give you a more accurate estimate of the true voltage offset (because little noise blips will tend to cancel out if you're averaging over more points in your baseline). This in turn gives you a more precise measure of the poststimulus amplitudes.

However, if your baseline period is too long, this can have three negative consequences. First, a longer baseline means that you will be performing artifact rejection over a longer time period, which may substantially increase the number of trials that you reject (especially if subjects blink during the intertrial interval). Second, it will move much of the baseline interval farther away from the time period of the components you will be measuring, therefore increasing the gradual drift that occurs as time passes from the baseline interval (see figure 8.2D). Third, the longer your prestimulus baseline interval, the more likely it will be that ERP activity from the previous trial will contaminate your baseline.

In most cases, I recommend a baseline period that is at least 20% of the overall epoch duration (e.g., you could use a 100-ms prestimulus baseline with a 500-ms epoch or a 400-ms prestimulus baseline with a 2000-ms epoch). I typically use a 200-ms prestimulus baseline with an 800-ms poststimulus period when I want to see late components such as P3, and I use a 100- or 200-ms prestimulus baseline with a poststimulus period of 300–500 ms if I'm focusing only on earlier components such as P1 and N2pc.

In most cases, it's a good idea for the baseline period to be a multiple of 100 ms, because this will tend to cancel out alpha-frequency (10 Hz) EEG oscillations. That is, an equal number of the negative and positive portions of an alpha cycle will be present within a given 100-ms period,

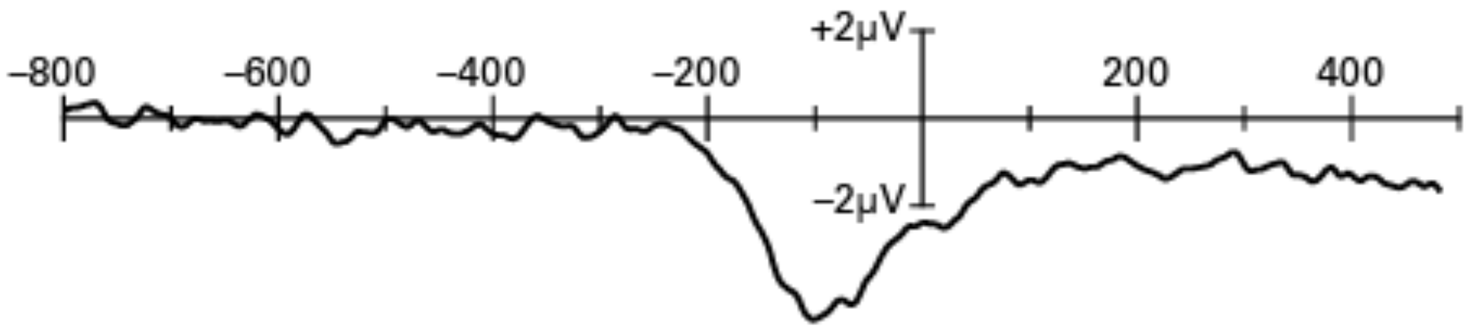

Figure 8.4
Response-locked lateralized readiness potential difference wave, formed by a contralateral-minus-ipsilateral subtraction, collapsed across the C3 and C4 electrode sites. Prior to averaging, the individual epochs were baselined from –800 to –600 ms with an epoch that extended from –800 to +500 ms (relative to the time of the behavioral response).

and these portions will therefore cancel each other out. A 200-ms baseline period includes two full cycles of the alpha oscillation, and I find this does a good job of minimizing the effects of alpha activity.

Up to this point, I have been limiting the discussion to stimulus-locked averages because baseline correction is a little more complicated for response-locked averages. Figure 8.4 shows an example of a response-locked LRP difference wave, formed by a contralateral-minus-ipsilateral subtraction (see the section on "Response-Related ERP Components" in chapter 3 for details). Time zero is the onset of the response. You can see from the waveform shown in figure 8.4 that substantial ERP activity was present in the 200-ms period immediately prior to time zero, so this period does not provide a good estimate of the EEG offset.

To determine an appropriate baseline period, the first step is usually to create an average with a long pre-response interval and then use this to determine when the waveform deviates from a flat line. For example, the data shown in figure 8.4 were initially baseline corrected from –800 to –600 ms, with an overall epoch that extended from –800 to +500 ms. It was clear that the waveform was essentially flat from –800 ms until approximately –250 ms and that the LRP was mostly complete by +200 ms. It would therefore be reasonable to re-epoch the EEG from –500 to +200 ms (relative to the response) and use a new baseline period of –500 to –300 ms.

An alternative solution would be to compute the average voltage over the 200 ms prior to each stimulus on each trial and subtract this prestimulus-defined voltage from each response-locked epoch. However, the onset time of the stimulus relative to the response varies quite a bit from trial to trial (because the reaction time varies from trial to trial), so it is not easy in practice to use the prestimulus interval as the baseline in response-locked data.

Averaging

Basics of Signal Averaging

After you have epoched and baselined the EEG, the next step is usually artifact rejection (as described in chapter 6), and then you are ready to average the data. The actual averaging process is quite simple. The EEG epochs are aligned with respect to the time-locking event, and the

voltages from all the EEG epochs at a given time point are averaged together. This type of averaging is sometimes called *signal averaging*. Note that this is equivalent to summing the single-trial EEG waveforms and then dividing by the number of trials.

Signal averaging is used to attenuate noise so that the event-related brain activity can be seen more easily. The EEG data collected on a single trial consists of event-related brain activity (i.e., the ERP) plus other activity that is unrelated to the event (e.g., other EEG activity, skin potentials, muscle artifacts, ocular artifacts, induced environmental noise, etc.). The ERP is assumed to be largely identical on each trial, whereas the other activity is assumed to be completely random with respect to the time-locking event. That is, the noise at a given time point is positive-going on some trials and negative-going on other trials, and the positives and negatives will cancel each other out when the values are averaged together across trials, leaving just the consistent ERP activity.

When I was in graduate school, Steve Hillyard taught me some "advanced" terminology for describing positive and negative voltage deflections, which I would like to share with you. Upward-going deflections are called *uppies*, and downward-going deflections are called *downies*.[1] Because of noise, some trials will have an uppie at a given time point and others will have a downie, and these uppies and downies cancel out when many trials are averaged together. Given a finite number of trials, the uppies and downies will not be perfectly equal and will not cancel out perfectly, so some noise will remain in the averaged ERP waveform. However, the uppies and downies that remain in the averaged waveform tend to become smaller and smaller as more and more trials are averaged together.

Figure 8.5 illustrates how increasing the number of trials leads to reduced noise in the averaged ERP waveforms for the targets in the oddball experiment shown in figure 8.1. The left column in figure 8.5 shows the EEG epochs for eight different target trials. The P3 wave for this subject was quite large, and it can be seen in every trial as a broad positivity between 300 and 700 ms. However, there is also quite a bit of variability from trial to trial in the exact shape of the P3 wave, and this is at least partly due to random EEG fluctuations (although the P3 itself may also vary from trial to trial). The other components are too small to be clearly visible on the single trials.

The right column in figure 8.5 shows how the effects of the random EEG fluctuations are minimized as more and more trials are averaged together. Each row shows the average of the single trials up to that row (e.g., row 3 shows the average of the first three targets and row 6 shows the average of the first six targets). If you look at the prestimulus baseline period, you can see that the uppies and downies in the averaged waveforms get progressively smaller as more trials are averaged together. The poststimulus waveshape also becomes progressively more stable as more trials are averaged together. Note that the P3 is quite nice looking in figure 8.5 when only eight trials were averaged together. This is because the single-trial EEG noise was fairly small and the P3 wave was very large. You will typically need far more trials than this in your averages. I will provide some specific advice about this later in the chapter.

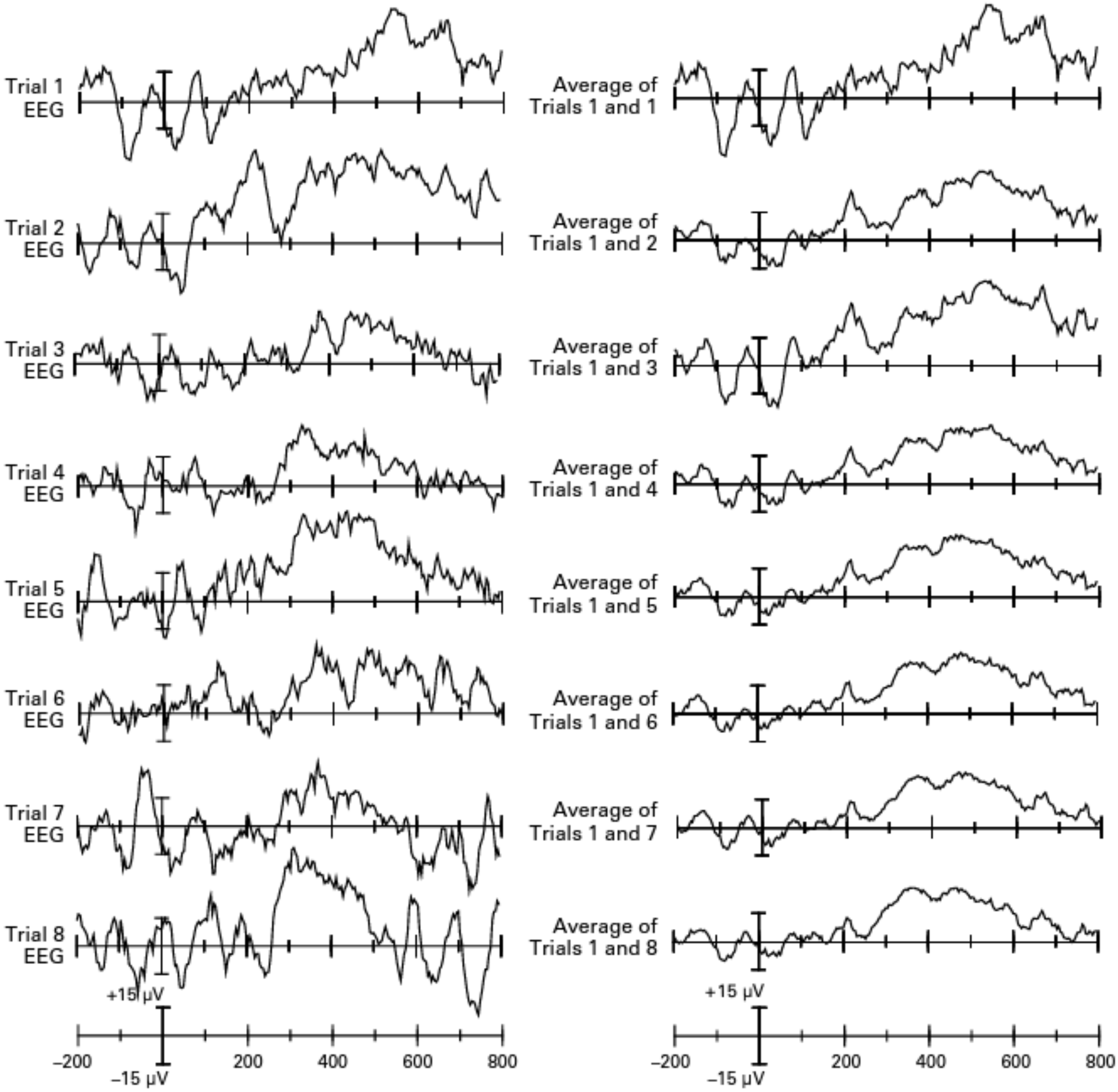

Figure 8.5
Example of signal averaging. The left column shows segments of EEG for each of several target trials in the oddball paradigm shown in figure 8.1, time-locked to stimulus onset. The right column shows the effects of averaging 1, 2, 3, 4, 5, 6, 7, or 8 of these EEG segments.

Signal-to-Noise Ratio

Now that I've provided an informal description of how averaging reduces noise, it's time to get more precise. The absolute amount of noise is not as important as the size of the noise relative to the size of the signal. Thus, researchers typically focus on the *signal-to-noise ratio* (SNR), which is simply the size of the signal divided by the size of the noise. For example, if the signal of interest is a 20-μV P3 wave, and the peak-to-peak EEG noise on a typical trial is 50 μV, we would say that the SNR is 20:50, or 0.4. If we assume that the signal is the same on every trial, then averaging will keep the size of the signal the same but will reduce the noise, thereby increasing the SNR.[2]

The noise in an average decreases progressively as the number of trials increases, and this leads to an increase in the SNR. However, this increase is not linear. For example, doubling the number of trials does not double the SNR. Instead, the SNR increases in proportion to the square root of the number of trials, so doubling the number of trials increases the SNR by the square root of 2. Let's make this more precise by using S to denote the size of the signal, N to denote the size of the noise on a typical single trial, and T to denote the number of trials. The SNR on a single trial is simply S/N (the signal divided by the noise), and the SNR in an average of T trials is equal to $(S / N)\sqrt{T}$ (the single-trial SNR multiplied by the square root of the number of trials). This is a very important fact that you should commit to memory, so I will repeat it: *The SNR increases in proportion to the square root of the number of trials*.

To make this more concrete, consider an experiment in which we are measuring a P3 wave that has an amplitude of 20 μV. If the noise in the EEG is typically 50 μV on a single trial, then the SNR on a single trial will be 20:50, or 0.4. If you average two trials together, then the SNR of the average will be the single-trial SNR multiplied by $\sqrt{2}$ (which is approximately 1.41), leading to an SNR of 0.566. To double the SNR from 0.4 to 0.8, it is necessary to average together four trials (because $\sqrt{4} = 2$). To quadruple the SNR from 0.04 to 1.6, it is necessary to average together 16 trials (because $\sqrt{16} = 4$). Thus, doubling the SNR requires quadrupling the number of trials, and quadrupling the SNR requires increasing the number of trials by a factor of 16. This relationship between the number of trials and the SNR is rather sobering because it means that achieving a substantial increase in the SNR requires a very large increase in the number of trials. You can only quadruple the number of trials so many times before the length of the recording session becomes impractical. This leads to a very important principle: It is often easier to improve the SNR of your averaged ERP waveforms by decreasing sources of noise than by increasing the number of trials.

Although the square root principle makes it difficult to achieve a good SNR, this principle has a positive side as well. Specifically, if you need to decrease the number of trials for some reason, the SNR does not decrease linearly with the decrease in the number of trials. For example, if you reject 20% of trials due to artifacts, this does not reduce your SNR by 20%. Instead, with 80% of the original number of trials, your new SNR will be $\sqrt{0.8}$ times your original SNR (or approximately 89% of your original SNR).

You might be wondering what an acceptable SNR would be. This is not an easy question to answer. In fact, the concept of SNR is not quite as simple as I've led you to believe because the

signal of interest and the noise of interest will depend on exactly how you are quantifying your components (e.g., with a mean amplitude measure, an onset latency measure, etc.). Moreover, your SNR doesn't need to be as high if you have a lot of subjects or are looking for a large effect. In general, you shouldn't worry about the specific value of the SNR. Instead, you should just do everything you can (within reason) to make your SNR as good as it can be (by increasing the signal and decreasing the noise).

How Many Trials Do You Need?

One of the most common questions I get from new ERP researchers is, "How many trials do I need to include in an average?" This is related to the question of what an acceptable SNR would be. These are simple questions, but they do not have a simple answer. A detailed discussion of the underlying theoretical issues is provided in the online supplement to this chapter. Here I will provide concrete advice based on my own experience. This reflects the particular kinds of experiments that my lab conducts as well as the details of how we record and analyze our data. Your results should be similar to my lab's results if you conduct similar experiments and follow the advice about recording and analysis provided in this book. If your experiments are quite different, you will need to look at the literature and see what other people in your area of research do. But when you read papers to get this information, take a look at the waveforms and see how noisy they are. This may make you want to increase the number of trials in your experiments. You should also keep in mind that you will want even more trials if you are attempting to look at individual differences.

In my lab's basic science experiments, the experimental manipulations are almost always within subjects rather than between groups. This allows the statistical analyses to factor out the effects of overall differences between subjects (which can be large, as discussed in the next section), and it generally increases the statistical power. You will likely need more trials and/or more subjects if your research involves between-group comparisons or explicit analyses of individual differences.

Our basic science experiments usually have an *N* of 12–16 subjects per experiment (this is the total after a few subjects have been automatically excluded because of excessive artifacts, as discussed in chapter 6). The subjects in these experiments are college students and are therefore more homogeneous in terms of cognitive abilities than the broader population. They are also generally cooperative, able to understand the instructions, and stay focused on the task. These factors minimize variance and therefore increase our statistical power. You will need more trials and/or more subjects if you test more heterogeneous or less cooperative subjects.

The number of trials we include in these experiments depends on the expected size of the effect, which is related to the size of the ERP component we are measuring. When we are looking at small components (e.g., the visual P1 wave), we typically test 300–500 trials per condition in each subject. When we are looking at somewhat larger components (e.g., N2pc), we typically test 150–200 trials per condition. When we are looking at very large components (e.g., P3 or N400), we typically test 30–40 trials per condition. This is the minimum number of trials per

subject in the conditions with the fewest trials (e.g., we might test 200 trials in an oddball experiment to get 40 targets and 160 standards for each subject). In addition, this is the number of trials in the averaged waveforms after combining across theoretically unimportant factors that were manipulated for the purpose of counterbalancing (e.g., we might combine 20 trials in which X was the target with 20 trials in which O was the target to achieve our 40 target trials).

In our experiments comparing schizophrenia patients with control subjects, we are typically trying to find a group × condition interaction (e.g., a smaller difference between two conditions in patients than in control subjects). Statistical power for these interactions is usually substantially lower than the power for detecting a main effect in a within-subjects design. In addition, we typically observe greater variance across individuals than in our basic science studies due to greater true score variance (because the patients and control subjects are more heterogeneous than the college students in our basic studies) and due to greater measurement error (because the EEG is noisier and more trials are rejected because of artifacts). We therefore test more subjects than in our basic science studies, with 20–30 subjects per group. When possible, we also double the number of trials per subject compared to our basic science studies. This means that we usually test fewer conditions per experiment in our schizophrenia studies than in our basic science studies.

All of these numbers reflect the number of trials that we initially include in the experiment, assuming that artifacts will lead to the rejection of 10% to 25% of these trials in our basic science studies and 20% to 50% of these trials in our schizophrenia studies. You can adjust upward or downward if you anticipate a different rate of artifacts or if you will be using artifact correction instead of artifact rejection.

Individual Differences in Averaged ERP Waveforms

Most ERP waveforms shown in journal articles are *grand averages*, which is the term used by ERP researchers to refer to waveforms created by averaging together the averaged waveforms of the individual subjects (see, e.g., figures 1.2 and 1.5 in chapter 1). Single-subject waveforms are shown only rarely in published papers (although they were more common in the early days of ERP research because primitive computer systems made it difficult to create and plot grand averages). The use of grand averages masks the variability across subjects, which can be both a good thing (because variability across subjects makes it difficult to see differences across groups or conditions) and a bad thing (because the grand average may not accurately reflect the pattern of individual results). When you look at your own single-subject data, you may be surprised at how different the subjects look from each other and from the grand averages you have seen in published papers.

Figure 8.6 shows the single-subject waveforms from an N2pc experiment (similar to the experiment shown in figure 3.8 of chapter 3). The left column shows waveforms from a lateral occipital electrode site in five individual subjects (from a total set of nine subjects). As you can see, there is tremendous variability in these waveforms. Everyone has a P1 peak and an N1 peak,

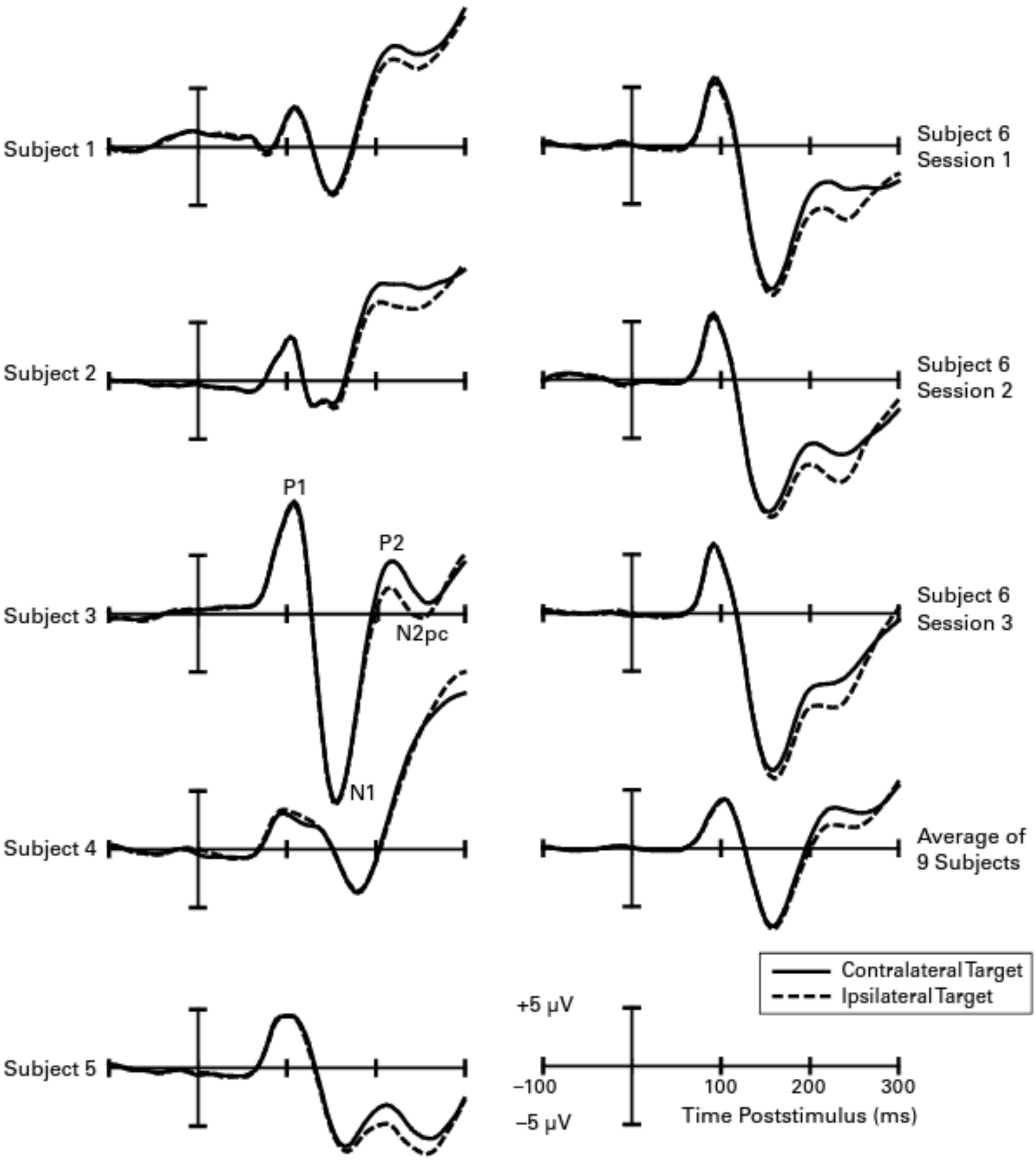

Figure 8.6
Example of individual differences in averaged ERP waveforms. Data from an N2pc experiment are shown for six individual subjects (selected at random from a total set of nine subjects). Subjects 1–5 participated in a single session, and subject 6 participated in three sessions. The grand average of all nine subjects from the experiment is shown in the lower right portion of the figure.

but the relative and absolute amplitudes of these peaks are quite different from subject to subject (compare, e.g., subjects 2 and 3). Moreover, not all of the subjects have a distinct P2 peak, and the overall voltage from 200 to 300 ms is positive for three subjects, near zero for one subject, and negative for one subject. This is quite typical of the variability that one sees in an ERP experiment. I didn't fish around for unusual examples—this is truly a random selection.

What are the causes of this variability? To answer this question, we first need to determine how much of the variability reflects stable individual differences among the subjects and how much reflects random trial-to-trial noise that remains after averaging together a finite number of trials. You can get some idea of this by looking at the first three rows of the right side of figure 8.6, which show the waveforms from subject 6, who participated in three sessions of the same experiment. You can see a little bit of variability in subject 6's waveforms from session to session, but this variability is very small compared to the variability you can see across the five subjects shown on the left side of the figure. The waveforms were quite stable across sessions for subject 6 because we averaged together hundreds of trials to create the averaged waveforms for each session, resulting in a good SNR. You can tell that the SNR was quite good because there is very little noise in the prestimulus baseline period. Some of the differences in the waveforms from session to session reflect random trial-to-trial noise that has not been eliminated by averaging, but some of the differences may reflect real changes in the subject from session to session, ranging from global state factors (e.g., number of hours of sleep the previous night) to shifts in task strategy. John Polich has published an interesting series of studies showing that the P3 wave is sensitive to a variety of global factors, such as time since the last meal, body temperature, and even the time of year (see review by Polich & Kok, 1995).

Given the session-to-session reliability shown on the right side of figure 8.6, the differences in the waveforms across subjects on the left side of the figure mainly reflect stable differences between subjects. However, you cannot assume that this will always be the case. If we had averaged together fewer trials or if we hadn't minimized other sources of noise, we would have had greater session-to-session variability in subject 6, and more of the variance between subjects would have reflected noise rather than stable differences.

When noise is minimal, as in figure 8.6, there are several potential causes of the stable differences in waveforms across subjects. One major factor is the idiosyncratic folding pattern of the cortex. As was discussed in chapter 2, the location and orientation of the cortical generator source of an ERP component has a huge influence on the size of that component at a given scalp electrode site. Every individual has a unique pattern of cortical folding, and the relationship between functional areas and specific locations on a gyrus or in a sulcus may also vary. Although I've never seen a formal study of the relationship between cortical folding patterns and individual differences in ERP waveforms, I've always assumed that this is the most significant cause of waveform variation in healthy young adults (for additional discussion, see Kappenman & Luck, 2012). There are certainly other factors that can influence the shape of the waveforms, including drugs, age, psychopathology, and even personality. But in experiments that focus on homogeneous groups of healthy young adults, these factors probably play a relatively small role.

The differences in waveforms among subjects are usually ignored in ERP studies. In many cases, this is very reasonable. However, this variability can be problematic when you are trying to measure a component using the same time window and set of electrode sites in all subjects (see the discussion of figure 9.2 in chapter 9). Large and stable differences across subjects can actually be a good thing if you are trying to study individual differences in psychological or neural processes. However, if the differences reflect nonfunctional factors such as cortical folding patterns, much of the variance across individuals may be unrelated to psychological or neural processes. Nonetheless, some of the individual differences can be related to psychological or neural factors, especially if you isolate specific processes with difference waves, as described in chapter 4. My former graduate student, Ed Vogel, has used difference waves to show beautiful and robust correlations between ERP effects and behavioral performance in studies of working memory (see, e.g., Vogel, McCollough, & Machizawa, 2005; Drew, McCollough, Horowitz, & Vogel, 2009; Anderson, Vogel, & Awh, 2011, 2013; Tsubomi, Fukuda, Watanabe, & Vogel, 2013).

The waveforms in the bottom right portion of figure 8.6 show the grand average of the nine subjects in this experiment. An important attribute of the grand average waveforms is that the peaks are smaller than those in most of the single-subject waveforms. This might seem odd, but it is perfectly understandable. The time point at which the voltage reaches its peak values for one subject are not the same as for other subjects, and the peaks in the grand averages are not at the same time as the peaks for the individual subjects. Moreover, there are many time points at which the voltage is positive for some subjects and negative for others. Thus, the grand average is smaller overall than most of the individual-subject waveforms. This is an example of a principle that will be discussed later in the chapter; namely, that latency variability leads to reduced peak amplitudes in averaged waveforms.

The (Non-)Problem of Amplitude Variability

Signal averaging is based on several assumptions, the most obvious of which is that the neural activity related to the time-locking event is the same on every trial. This assumption is clearly an oversimplification because neural and cognitive processing will obviously vary from trial to trial. If the amplitude of a given component varies from trial to trial, we are violating this assumption, but this type of violation is not usually a problem. For example, if the amplitude of the N2 wave varies from trial to trial, then the N2 wave in the averaged ERP waveform will simply reflect the average amplitude of the N2 wave across trials. This is no different from the typical use of averaging in science, in which the mean of a set of values is used as a measure of central tendency. Thus, we can just treat the voltage values in an averaged ERP waveform as a series of measures of central tendency, one at each time point in the waveform.

The mean is not always a very good measure of central tendency, and this is true for almost any measure of psychological or neural activity including ERPs. For example, if N2 amplitude is very large on half the trials and very small on the other half, the N2 in the averaged waveform

will be an intermediate value that was not present on any individual trial. Similarly, N2 amplitude could be small on most trials but very large on a few trials (i.e., a skewed distribution). ERP researchers are not usually very concerned about these possibilities, but they are occasionally important. For example, P3 amplitude is typically smaller in schizophrenia patients than in healthy control subjects when measured from averaged ERP waveforms, and this could reflect either a consistent reduction in P3 amplitude on every trial or a complete absence of the P3 on a subset of trials coupled with normal P3 amplitude on the majority of trials. Ford, White, Lim, and Pfefferbaum (1994) conducted a single-trial analysis to distinguish between these possibilities and found evidence that patients have both a reduced likelihood of having a P3 wave and a reduced amplitude on the trials that contained a P3.

In most cases, you can ignore the likely possibility that amplitudes vary across trials. This very rarely impacts the conclusions you will draw from your experiments.

The Problem of Latency Variability

Although trial-to-trial variability in amplitude is not usually problematic, trial-to-trial variability in latency (also called *latency jitter*) can be a big problem. The timing of neural processes always has some variability, and the variability tends to become greater for later, cognitive components. For example, the trial-to-trial variability of a relatively late component like N400 will tend to be greater than that of an earlier component like P1. P3 latency is especially variable because the onset of the P3 wave depends on the amount of time required for stimulus categorization (see chapter 3), and this can vary widely from trial to trial. The effects of latency variability on averaged ERPs are illustrated in figure 8.7A, which shows four individual trials in which a P3-like ERP component occurs at different latencies, along with the average of the four trials. The first thing you should notice is that the peak amplitude of the averaged waveform is much smaller than the peak amplitude of the individual trials. This is particularly problematic when the amount of latency variability differs across experimental conditions. You can see this by comparing panels A and B of figure 8.7: the amplitudes of the single trials are identical in these two panels, but panel B has less latency variability than panel A, and the averaged waveform in panel B therefore has a greater peak amplitude than the averaged waveform in panel A. Thus, if two experimental conditions or groups of subjects differ in the amount of latency variability for some ERP component, they may appear to differ in the amplitude of that component even if the single-trial amplitudes are identical. This could lead you to the incorrect conclusion that there was a difference in amplitude. For example, it is important to ask whether the reduced P3 amplitude observed in schizophrenia patients relative to healthy control subjects reflects greater variability in P3 timing rather than smaller single-trial amplitudes. A difference in latency variability seems likely because schizophrenia patients usually exhibit greater variability in reaction time than do control subjects. Ford et al. (1994) examined this possibility and found that patients did indeed have greater latency variability than controls but that this did not fully account for their reduced P3 amplitude.

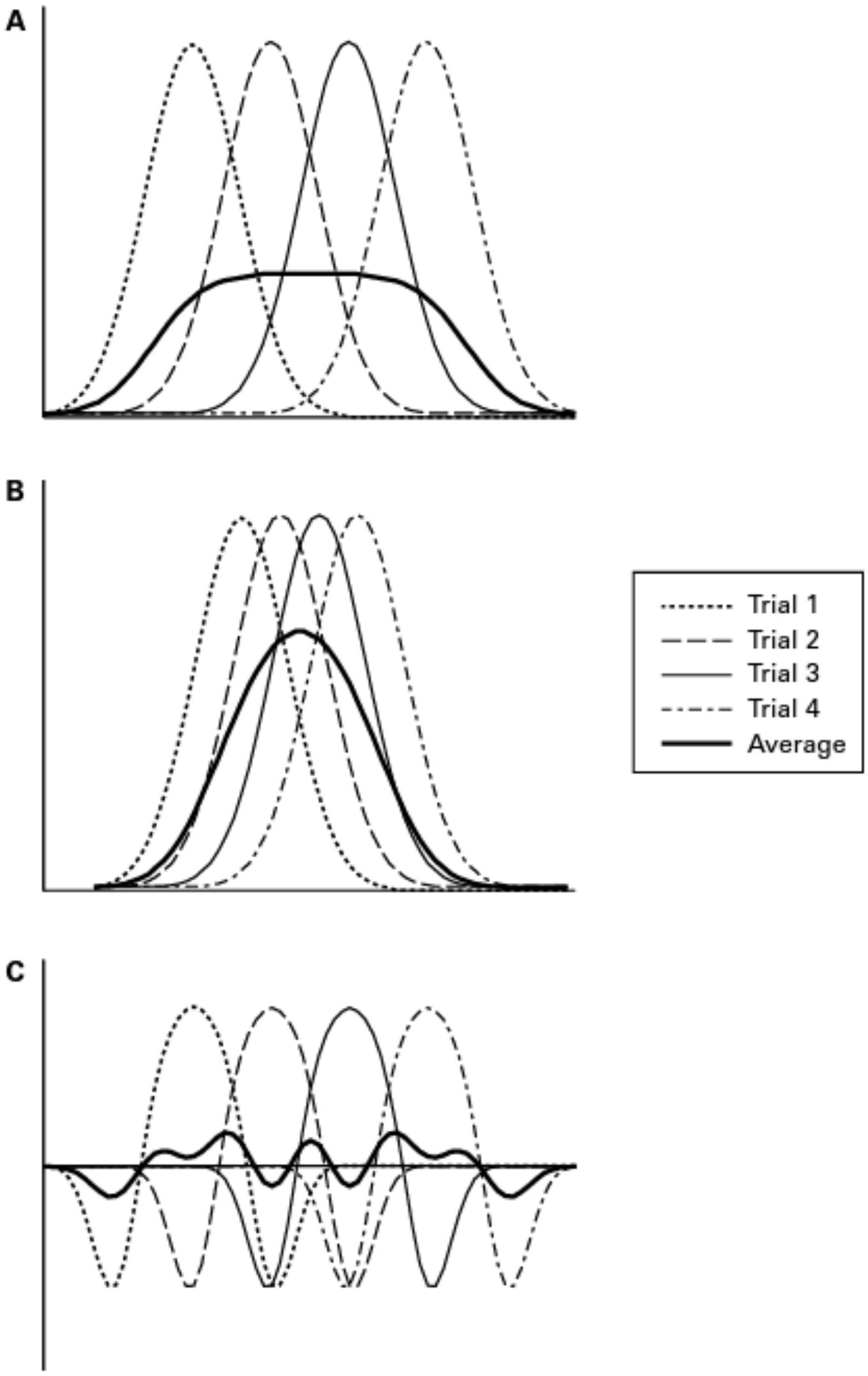

Figure 8.7
Example of the problem of latency variation. Each panel shows four single-trial waveforms, along with the average of the four waveforms. The same waveforms are present in panels A and B, but there is greater latency variability in panel A than in panel B, leading to a smaller peak amplitude and broader temporal extent in the averaged waveform for panel A than for panel B. Panel C shows that when the single-trial waveforms are not monophasic, but instead have both positive and negative subcomponents, latency variability may lead to cancellation in the averaged waveform.

When an ERP response contains both positive and negative portions, latency variability may cause the positive part of the response on one trial to be at the same time as the negative portion on another trial, leading to cancellation (see figure 8.7C). In extreme cases, the cancellation will be complete, and the ERP will be completely absent from the averaged waveform. For example, imagine that a sinusoidal oscillation is triggered by a stimulus but varies randomly in phase from trial to trial (which is not just a hypothetical problem—see Gray, König, Engel, & Singer, 1989). Such a response will average to zero and will be essentially invisible in an averaged response.

Example of Latency Jitter

A real example of latency jitter is shown in figure 8.8 (from Luck & Hillyard, 1990). In this experiment, we examined the P3 wave during two types of visual search tasks. In one condition (*parallel search*), subjects searched for a target with a distinctive visual feature that "popped out" from the display and could be detected immediately no matter how many distractor items were present in the stimulus array. In the other condition (*serial search*), the target was defined by the absence of a feature; in this condition, we expected that the subjects would search the array one item at a time until they found the target. Reaction time (RT) was expected to increase as the number of items in the array (the set size) was increased in the serial search condition, whereas no effect of set size was expected in the parallel search condition. This was the pattern of results that we obtained, replicating many previous visual search experiments (see, e.g., Treisman & Souther, 1985; Treisman & Gormican, 1988).

We also expected to see consistent differences across conditions in the trial-by-trial variability of RT. In the parallel search condition, the target pops out immediately, so the amount of time required to find the target should be relatively consistent from trial to trial. In the serial search condition, however, subjects shift attention randomly from one item to the next until they find the target, and this leads to variability in the amount of time that passes before the target is found. At set size 4, for example, the target could be the first, second, third, or fourth item searched, but at set size 12, the target might be found anywhere between the first item and the 12th item. We thus expected that RTs in the serial search condition would vary a great deal from trial to trial and that the variability would increase as the set size increased. This is exactly what we found.

Because the P3 is linked to the categorization of a stimulus as a target (see chapter 3), we expected that the timing of the P3 on each trial would be tightly linked to the timing of the behavioral response. Consequently, we expected that P3 latency would be relatively constant in the parallel search condition, because the amount of time required to find the target should not vary much from trial to trial in this condition. However, because the amount of time required to find the target varies greatly from trial to trial in the serial search condition, especially at the large set sizes, we expected that P3 latency would become progressively more variable as the set size increased in this condition. The increased latency variability would be expected to decrease the peak amplitude of the P3 wave in the averaged ERP waveforms.

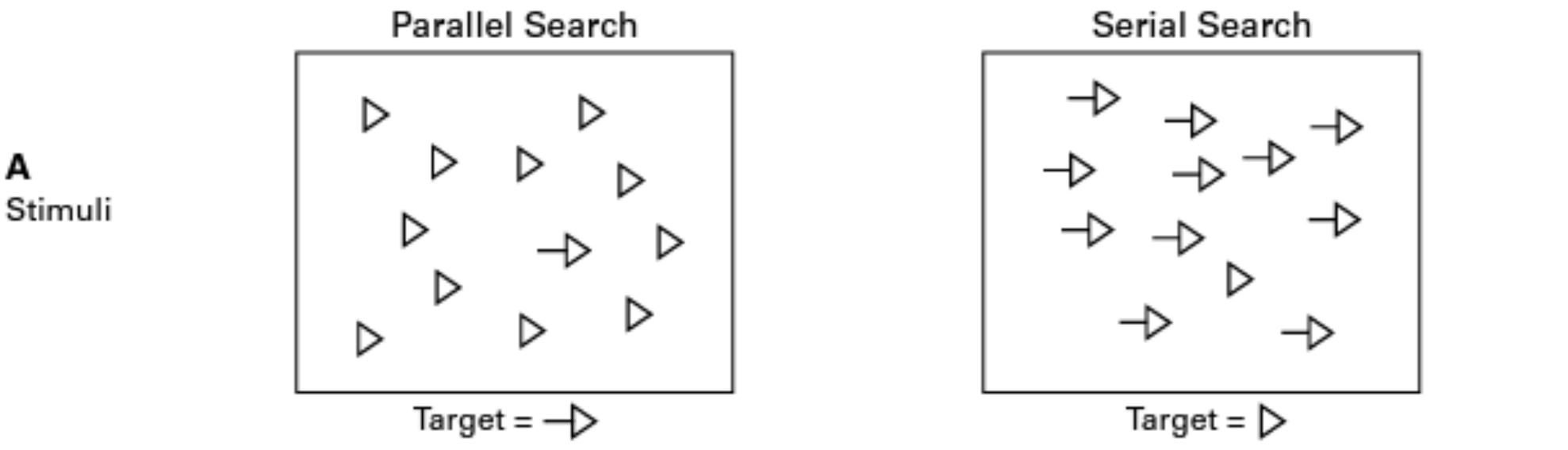

Figure 8.8
Example of an experiment in which significant latency variability was expected for the P3 wave (Luck & Hillyard, 1990). (A) Sample stimuli from the two conditions of the experiment. (B) Stimulus-locked averages from the Pz electrode site. (C) Response-locked averages from the Pz electrode site.

The averaged ERP waveforms from this experiment are shown in figure 8.8B, time-locked to the onset of the stimulus array. In the parallel search condition, the P3 wave was relatively large in amplitude and short in duration, and it did not vary much as a function of set size. In the serial search condition, the P3 had a smaller peak amplitude than in the parallel search condition but was very broad. In addition, the P3 peak in the serial search condition was greater at set size 4 than at set sizes 8 and 12. If you did not take into account the effects of latency variability, you might conclude that the brain generally produced a smaller P3 response when it performed the serial task than when it performed the parallel task, and that within the serial task the brain produced a smaller P3 at set sizes 8 and 12 compared to set size 4. However, given that the smaller peak amplitudes were observed in the conditions that were expected to have greater trial-by-trial latency variability, and that greater variability artificially reduces peak amplitudes in averaged waveforms (see figure 8.7), the peak amplitudes could be misleading. If we just look at the stimulus-locked P3 peaks, we can't tell whether the single-trial P3 amplitudes differed across conditions or if the peak differences are an artifact of differences in P3 latency variability.

How, then, can we factor out the effects of latency jitter to determine what really happened to P3 amplitude on single trials in each condition of this experiment? The following sections will describe some simple but effective methods for dealing with this type of latency jitter. Online chapter 11 will provide a more detailed analysis of latency jitter, introducing the mathematical concept of *convolution*, which will give you a deeper understanding of exactly how latency jitter influences the averaged ERP waveform.

Dealing with Latency Jitter

Area Measures In many cases, you can eliminate the effects of latency jitter simply by measuring mean or area amplitude rather than peak amplitude. These measures are particularly effective for *monophasic* ERP components (components that are entirely positive or entirely negative at a given electrode site). When a component is monophasic, the area under the curve in an average of several trials is equal to finding the average of the area under the curve on the individual trials and then taking the average. Consequently, latency variability does not impact the area under the curve for monophasic components. For example, the area under the curve for the averaged waveform shown in figure 8.7A is the same as the area under the curve for the averaged waveform shown in figure 8.7B, even though the latency variability was greater in the former than in the latter. Mean amplitude is also unaffected by latency variability; the mean amplitude is equivalent for the averaged waveforms in figure 8.7A and B despite the difference in latency variability and peak amplitude.

We can therefore apply mean or area amplitude measures to the data from the experiment shown in figure 8.8 to determine whether the differences in P3 peak amplitude in the averaged waveforms reflect differences in the single-trial P3 amplitudes. When I measured mean amplitude from the waveforms shown in figure 8.8B, I found that P3 amplitude was actually larger

in the serial search task than in the parallel search task (whereas the opposite was true for peak amplitude) and P3 amplitude increased as the set size increased in the serial search condition (whereas peak amplitude decreased). This implies that the single-trial P3 amplitudes were actually larger for the serial search task than for the parallel search task and that they increased as the set size increased in the serial search task. The next section will provide converging evidence for this conclusion using response-locked averages. Note that I would have erroneously reached exactly the opposite conclusion if I had only looked at the peak amplitudes.

In addition to being insensitive to latency jitter, area and mean amplitude measures have many additional advantages over peak amplitude measures, as will be discussed in detail in chapter 9 (which will also describe how to measure latency in a situation like this). However, these measures are ineffective if the jittered component is multiphasic (i.e., when it consists of both positive and negative periods). This is illustrated in figure 8.7C, which shows that the positive and negative deflections cancel each other when the onset time is variable. In this situation, a time–frequency analysis can be helpful, as will be described near the end of the chapter. In many cases, a single late component (e.g., the P3 wave) has substantial latency jitter, but this component is superimposed on the other positive and negative peaks in the waveform (e.g., P1, N1, P2, and N2). The overall waveform is multiphasic in this situation, but only a single monophasic component varies much in latency from trial to trial (P3). Area and mean amplitude measures will typically work quite well in this situation. The key is whether the jittered activity—not the overall waveform—is multiphasic. As long as the jittered brain activity is monophasic, area and mean amplitude measures are appropriate.

Response-Locked Averages In some cases, the latency of an ERP component is tightly coupled with RT, and in these cases latency variability can be corrected by using response-locked averages rather than stimulus-locked averages. In a response-locked average, the response rather than the stimulus is used to align the single-trial EEG segments during the averaging process. As an example, figure 8.8C shows the response-locked waveforms from our visual search experiment, in which we expected the P3 to peak at approximately the same time as the behavioral response on each trial. In the serial search condition, the P3 was narrower and taller in the response-locked averages than in the stimulus-locked averages, which is exactly what would be expected if the P3 was more tightly time-locked to the response than to the stimulus. In addition, the response-locked P3 was larger in the serial search condition than in the parallel search condition and increased as set size increased in the serial search condition. Recall that this is exactly what we found when we measured mean amplitude in the stimulus-locked averages but is the opposite of the pattern found for peak amplitude in the stimulus-locked averages. Thus, when brain activity is likely to be more closely time-locked to the response than to the stimulus, response-locked averages can be very useful for figuring out how single-trial amplitudes change across conditions or groups (for an example in the context of group differences, see Luck et al., 2009).

The Woody Filter Technique A third technique for mitigating the effects of latency variability is the *Woody filter* technique (Woody, 1967). The basic approach of this technique is to estimate

the latency of the component of interest on individual trials and to use this latency as the time-locking point for averaging. The component is identified on single trials by finding the portion of the single-trial waveform that most closely matches a template of the ERP component. Of course, the success of this technique depends on how well the component of interest can be identified on individual trials, which in turn depends on the SNR of the individual trials and the similarity between the waveshape of the component and the waveshape of the noise.

The Woody filter technique begins with a best-guess template of the component of interest (such as a half cycle of a sine wave) and uses cross-correlations to find the segment of the EEG waveform on each trial that most closely matches the waveshape of the template.[3] The EEG epochs are then aligned with respect to the estimated peak of the component and averaged together. The resulting averaged ERP can then be used as the template for a second iteration of the technique, and additional iterations are performed until little change is observed from one iteration to the next.

The shortcoming of this technique is that the part of the waveform that most closely matches the template on a given trial may not always be the actual component of interest, resulting in an averaged waveform that does not accurately reflect the amplitude and latency of this component (Wastell, 1977). Moreover, this does not simply add random noise to the averages; instead, it tends to make the averages from each different experimental condition more similar to the template and therefore more similar to each other (this is basically just regression toward the mean). Thus, this technique is useful only when the component of interest is relatively large and dissimilar to the EEG noise. For example, the P1 wave is small and is similar in shape to spontaneous alpha waves in the EEG, and the template would be more closely matched by the noise than by the actual single-trial P1 wave on many trials. The P3 component, in contrast, is relatively large and differs in waveshape from common EEG patterns, and the template-matching procedure is therefore more likely to find the actual P3 wave on single trials.

Even when a large component such as the P3 wave is being examined, Woody filtering works best when the latency variability is only moderate; when the variability is great, a very wide window must be searched on the individual trials, leading to more opportunities for a noise deflection to match the template better than the component of interest. For example, I tried to apply the Woody filter technique to the visual search experiment shown in figure 8.8, but it didn't work very well. The P3 wave in this experiment could peak anywhere between 400 and 1400 ms poststimulus, and given this broad search window, the algorithm frequently located a portion of the waveform that matched the search template fairly well but did not correspond to the actual P3 peak. As a result, the averages looked very much like the search template and were highly similar across conditions.

It should be noted that the major difficulty with the Woody filter technique lies in identifying the component of interest on single trials, and any factors that improve this process will lead to a more accurate adjustment of the averages. For example, the scalp distribution of the component can be specified in addition to the component's waveshape, which makes it possible to reject spurious EEG deflections that may have the correct waveshape but have an incorrect scalp distribution (see Brandeis, Naylor, Halliday, Callaway, & Yano, 1992).

Basics of Time–Frequency Analysis

As shown in figure 8.7C, latency jitter can cause single-trial activity to disappear from the averaged ERPs if the single-trial activity consists of an oscillation (because an uppie on one trial will cancel a downie at the same time on another trial). The past decade has seen an explosion in research on these EEG oscillations, mainly because *time–frequency analysis* techniques have been developed to aggregate across trials in a manner that avoids the cancellation that occurs with conventional signal averaging. I will discuss these techniques in greater detail in online chapter 12, but I will also provide a brief overview here because these techniques are really just a fancy version of signal averaging. If you haven't already read the section on Fourier analysis in chapter 7, you should go back and read that section before proceeding. I would also recommend reading the excellent review paper by Roach and Mathalon (2008).

Time–Frequency Analysis via Moving Window Fourier Analysis

Time–frequency analysis is based on variants of the *Fourier transform*. As was discussed in chapter 7, the Fourier transform converts a waveform into a set of sine waves of different frequencies, phases, and amplitudes. For example, if you were to apply the Fourier transform to a 1-s EEG epoch, you would be able to determine the amount of activity at 10 Hz, at 15 Hz, at 20 Hz, or almost any frequency. Fourier analysis measures the amplitude of an oscillation independently of its phase, and this allows us to avoid the problems associated with trial-by-trial variations in latency (which is closely related to phase).

However, we can't use the standard Fourier transform here because it doesn't give us any information about the time course of an oscillation. That is, it would provide a single value for each frequency, representing the power (amplitude squared) of that frequency for the entire epoch. Instead, we want a method that gives us the power of a given time frequency at each time point in the waveform. Power isn't actually defined for a single time point, but we can provide an approximation by looking at the power of a given frequency over a short time window (e.g., 200 ms) and using that power to represent the value at the middle latency of the time window. There are two basic approaches to accomplishing this, a *moving window* version of Fourier analysis and a *wavelet* analysis.

In the moving window Fourier analysis, a Fourier transform is performed for each of several consecutive time windows (Makeig, 1993). This is illustrated in figure 8.9, which shows four single trials in which a 10-Hz oscillation occurred, but starting at different time points on each trial. The gray rectangle represents the time window that is being used in the analysis. A Fourier analysis is done in this time window, and the power at each frequency is then assigned to the midpoint of that window. The window is then slid over to the right by one sample period (e.g., 4 ms), and a new Fourier analysis is done in the new time window. Three windows are shown in the figure to demonstrate the concept of a window, but an actual analysis would have many overlapping windows, one centered on each sample point in the waveform. This gives us an estimate of the power at each frequency at each sample point in the waveform. Note, however,

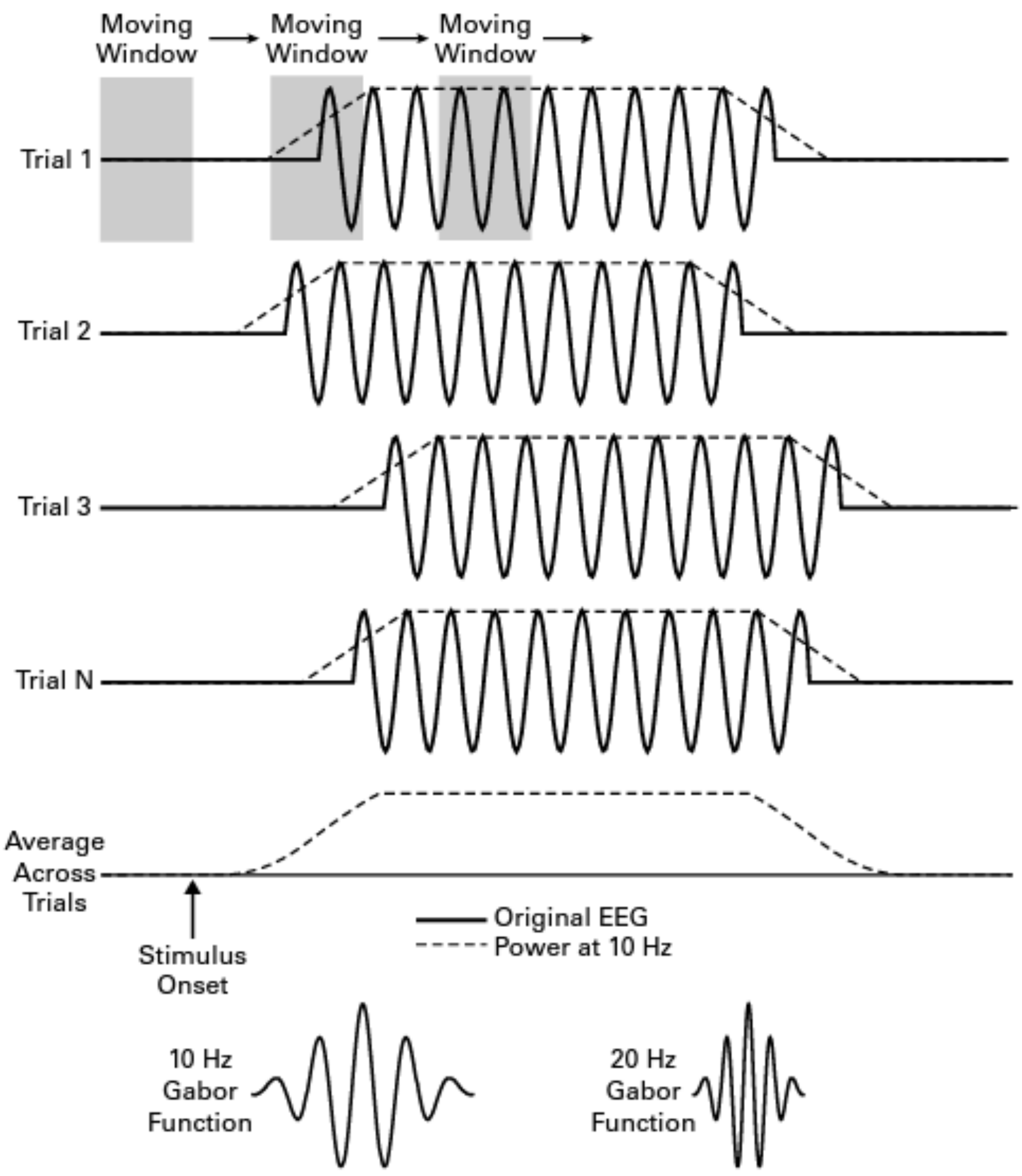

Figure 8.9
Basics of time–frequency analysis. On each trial of this example, a 10-Hz EEG oscillation is elicited at a variable time after stimulus onset (solid lines). If we were to compute a conventional average across trials, the oscillations would largely cancel, yielding a flat line in the average. To perform a time–frequency analysis, a window (200 ms wide in this example) is moved across the EEG data from each trial, and the power at a given frequency is estimated for each window position (dashed lines). Three different windows are shown here, but this is actually done for a very large number of overlapping windows, one centered on every sample point in the waveform. The power is measured at a given frequency for each sample point in the single-trial EEG waveforms, creating a single-trial power waveform for each trial. These single-trial power waveforms are then averaged together (shown in the bottom dashed waveform here). An alternative approach is to convolve the EEG with a set of Gabor functions, like the two shown at the bottom of the figure.

that we have lost some temporal resolution because the power at a given time point really reflects the entire time window centered at that time point.

Figure 8.9 illustrates how this works for a single frequency (10 Hz), using a 200-ms moving window. The EEG on each trial is shown with the solid lines, and the power at 10 Hz at the middle of each 200-ms window is shown with the dashed lines. The 10-Hz power is zero at the beginning of each EEG epoch because each epoch begins with a flat line. Once the window begins to reach the oscillation, the power at 10 Hz ramps up. The 10-Hz power levels off once the entire window is filled with the 10-Hz oscillation, and then it falls back to zero when the oscillation ends.

Once this has been done for every trial, the waveforms representing power at a given frequency on each trial can be averaged together, just as you would ordinarily average the original voltage waveforms across trials. This is shown near the bottom of figure 8.9. Whereas the oscillations will cancel out in a conventional average, leading to a flat line, the 10-Hz power in the average does a good job of representing the 10-Hz power on the single trials.

Time–Frequency Analysis via Wavelets

Although the moving window Fourier analysis method shown in figure 8.9 has some advantages, it has two disadvantages. First, it treats the power within the window as if it was the power at the center of the window, even though the entire window contributes equally to the power measurement.[4] Second, the same size window is used to calculate the power at each frequency, even though this yields lower precision for low frequencies than for high frequencies. Most people now use a different method based on *wavelets* that addresses both of these problems.

An example of a wavelet is shown at the bottom left of figure 8.9. There are many kinds of wavelets, and this particular wavelet is a Gabor function. It was created by taking a 10-Hz sine wave and multiplying it by a Gaussian (bell curve) function. Whereas the original sine wave was infinite in duration, the multiplication by the Gaussian function causes the oscillations to taper down over time. This solves the first of the two problems that arise from moving window Fourier transforms: rather than treating every point within a time period equally, a wavelet gives the greatest weight to the center of the time period. We have still lost some temporal resolution, because the power at a given time point is influenced by a range of surrounding time points, but this problem has been reduced somewhat because more distant time points receive lower weight.

The second problem—different precision for different frequencies—is solved by using Gabor functions with different widths for different frequencies. For example, the 20-Hz wavelet shown at the bottom right of figure 8.9 has the same number of cycles as the 10-Hz wavelet, thus yielding the same precision, but its duration is half as great. When multiple wavelets are created that are all identical but are squeezed or expanded horizontally to represent different frequencies, these wavelets are called a *wavelet family*. When each wavelet is a Gabor function, the family is called a *Morlet wavelet family*.

You may be wondering how a wavelet family is used to calculate the power at a given frequency and time point. The answer is a mathematical operation called convolution. The general

idea of convolution is described in online chapter 11, and online chapter 12 explains exactly how time–frequency analysis is achieved with convolution, using simple math. In my view, you shouldn't use time–frequency analysis without understanding the basics of how it works, and you should definitely read online chapters 11 and 12 if you are planning to do time–frequency analysis in the near future. In addition, Mike Cohen has recently published a book that provides both the mathematical details and practical advice for performing time–frequency analyses (Cohen, 2014), and I strongly encourage you to read that book if you want to use this technique.

Using Time–Frequency Analysis to See Random-Phase Activity

Figure 8.9 shows what time–frequency analysis looks like for a single frequency, but people usually do it for many frequencies at the same time. The results are usually plotted with a *heat map*, which uses different colors (or different shades of gray) to indicate the power at a given time and frequency. An example is shown in figure 8.10A, which shows data from the study of Tallon-Baudry, Bertrand, Delpuech, and Pernier (1996). The *X* axis in this plot is time, just as in a traditional ERP average. The *Y* axis, however, is frequency (with frequencies from 20 to 100 Hz shown here). The gray-scale level indicates the power that was present at each frequency at each time point, with lighter shades indicating greater power. A band of activity centered at 40 Hz can be seen at approximately 90 ms poststimulus, and a somewhat weaker band of activity between 30 and 60 Hz can be seen at approximately 300 ms poststimulus.

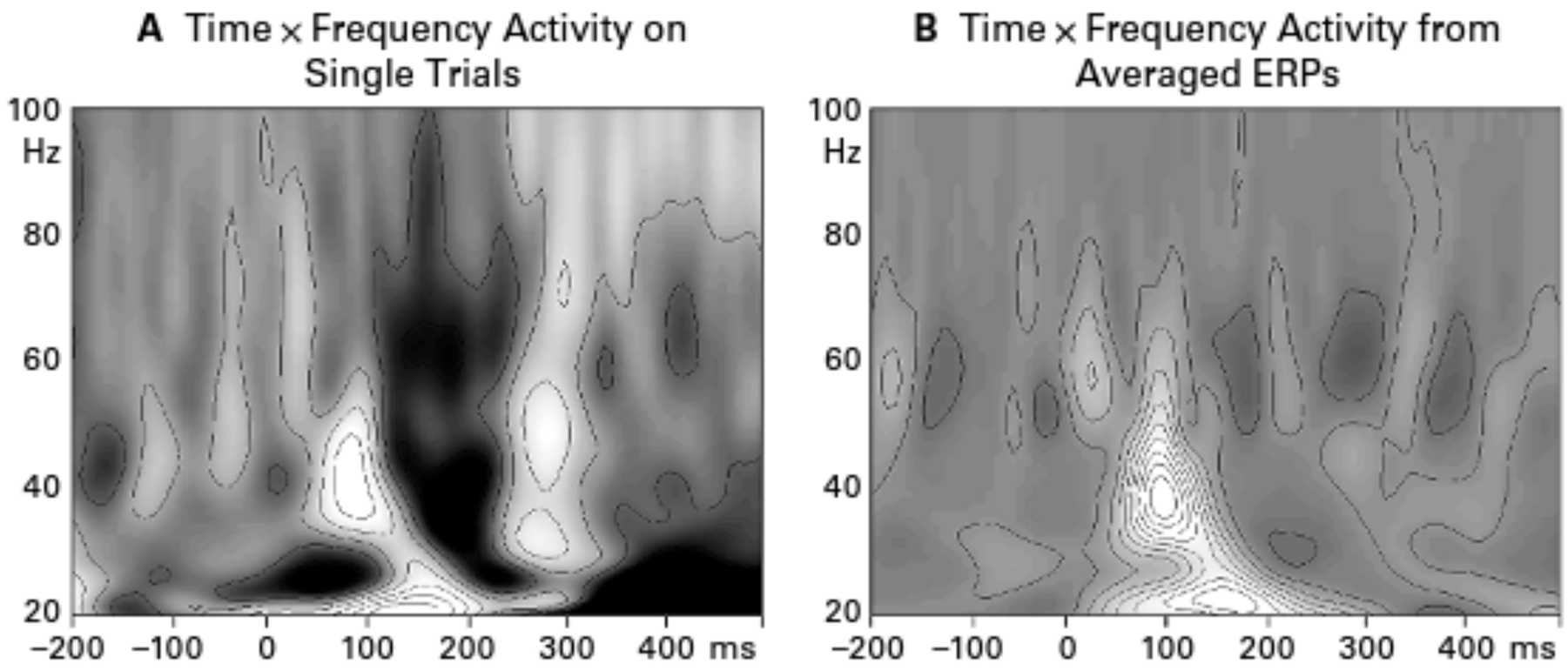

Figure 8.10
Example of time–frequency data from the study of Tallon-Baudry et al. (1996). The *X* axis is time; the *Y* axis is frequency; the intensity of the gray indicates the power at a particular time and frequency, with lighter shading for greater power. In panel A, the time–frequency transformation was applied to the individual trials, and the transformed data were then averaged. This plot therefore includes activity that was not phase-locked to stimulus onset as well as phase-locked activity. In panel B, the transformation was applied after the waveforms had been averaged together. This plot therefore includes only activity that was phase-locked to the stimulus, because random-phase activity is eliminated by the ERP averaging process. Adapted with permission from Tallon-Baudry et al. (1996). Copyright 1996 Society for Neuroscience.

The crucial aspect of this approach is that these bands of activity can be seen whether or not the activity varies in phase from trial to trial, whereas random-phase activity is completely lost in a traditional average. However, researchers often assume that they are seeing *only* random-phase oscillations in time–frequency analyses, but this is not usually a valid assumption. For example, figure 8.10B shows the results obtained by Tallon-Baudry et al. (1996) when the time–frequency transformation was applied to the averaged ERP waveform rather than to the single-trial EEG. The averaged ERP waveform should contain only activity that has a consistent phase from trial to trial, so any activity in the time–frequency transformation of the averaged ERP cannot be a result of random-phase oscillations (except under the conditions described in box 8.2). You can see that the 40-Hz activity at 90 ms is visible in the time–frequency transformation of the averaged ERP, so this activity was a part of the traditional averaged ERP waveform and was not a random-phase oscillation. However, the activity from 30 to 60 Hz at 300 ms in figure 8.10A is not visible in figure 8.10B, so this must have been random-phase activity.

Box 8.2
How Random-Phase Oscillations Can Survive in Conventional Averages

Although random-phase oscillations normally cancel out in conventional averages, Ali Mazaheri and Ole Jensen have shown that this is not always the case (Mazaheri & Jensen, 2008). Given the biophysics of neural circuits, oscillations may not always be symmetric around the baseline voltage. As shown in the illustration that follows, the voltage may instead rise up away from the baseline and then fall back to baseline on each cycle (or the opposite, depending on the orientation of the dipole). Because the voltage never dips below baseline, the downies are near zero rather than being negative, so they don't cancel the uppies (see, e.g., the time points indicated by the dashed lines in the illustration). As a result, the average across many trials with different phases may be a broad positivity (or a broad negativity).

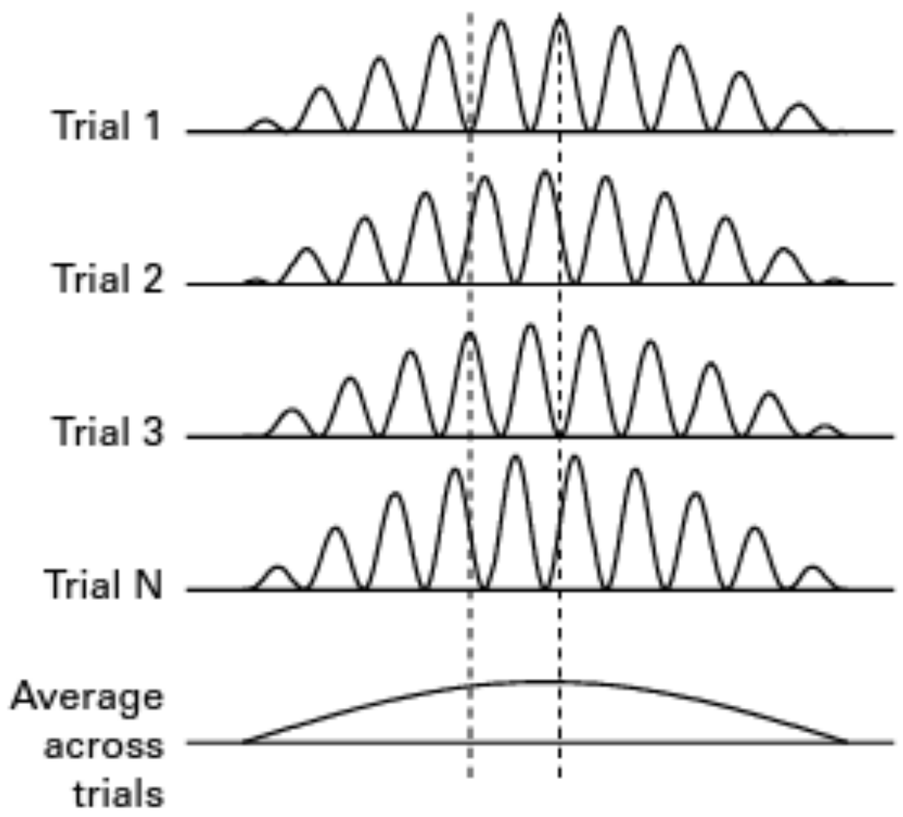

Figure 8.11 provides a closer look at how constant-phase and random-phase activity are influenced by averaging. Four trials of EEG are shown on the left of the figure, along with the conventional signal-averaged ERP. The EEG contains a 10-Hz burst at the beginning of each trial that has the same phase on every trial. It also contains a later 10-Hz burst that varies in phase from trial to trial. When the voltage waveforms are averaged together, the initial phase-locked 10-Hz burst is present in the average, but the later random-phase burst is absent. The right column contains the time–frequency transformations of the single trials, along with the average across the single-trial time–frequency transformations. You can see the two 10-Hz bursts on each single trial and also in the average. Thus, whereas only the first 10-Hz burst was present in the conventional average, both 10-Hz bursts can be seen when the averaging across trials is performed on the time–frequency transformations. Box 8.2 describes a special situation in which random-phase oscillations can be seen in conventional averages.

Although the groundbreaking study of Tallon-Baudry et al. (1996) showed the time–frequency transformation of the averaged ERP waveforms as well as the results of performing the time–frequency transformation before averaging, more recent studies often fail to show the results of applying the transformation to the averaged ERP waveforms. This makes it impossible to know whether the results reflect random-phase activity or simply reflect the traditional ERP waveform. When you read time–frequency studies, you should ask yourself whether the authors have provided the information needed to determine whether random-phase oscillations were actually present.

The Baseline in Time–Frequency Analyses

As discussed at the beginning of the chapter, the main purpose of baseline correction in conventional ERP studies is to remove large voltage offsets and slow drifts. These offsets and drifts occur at very low frequencies (mainly <1 Hz). Because time–frequency analyses provide estimates of power at specific frequencies that are usually much higher than 1 Hz, offsets and drifts have little or no impact on time–frequency results. Consequently, baseline correction is not always necessary in time–frequency analyses.

The main value of baseline correction in time–frequency analyses is to isolate stimulus-related brain activity from activity that was present prior to stimulus onset. If an experiment involves comparing activity elicited by stimuli that were presented in random order, the baseline activity should be the same for all trial types, and baseline correction is not strictly necessary (although it might be useful for minimizing random variations and therefore increasing statistical power).

However, some form of baseline correction will be necessary if the experiment involves comparing activity recorded in different trial blocks, from nonrandom stimulus orders, or from different groups of subjects. In these cases, any poststimulus differences in activity could simply be the continuation of prestimulus differences. If, for example, one condition is more boring than another, the more boring condition might yield greater alpha activity in both the prestimulus and poststimulus periods. In the absence of baseline correction, a finding of greater alpha from 200 to 300 ms poststimulus would not mean that there was greater stimulus-related alpha from

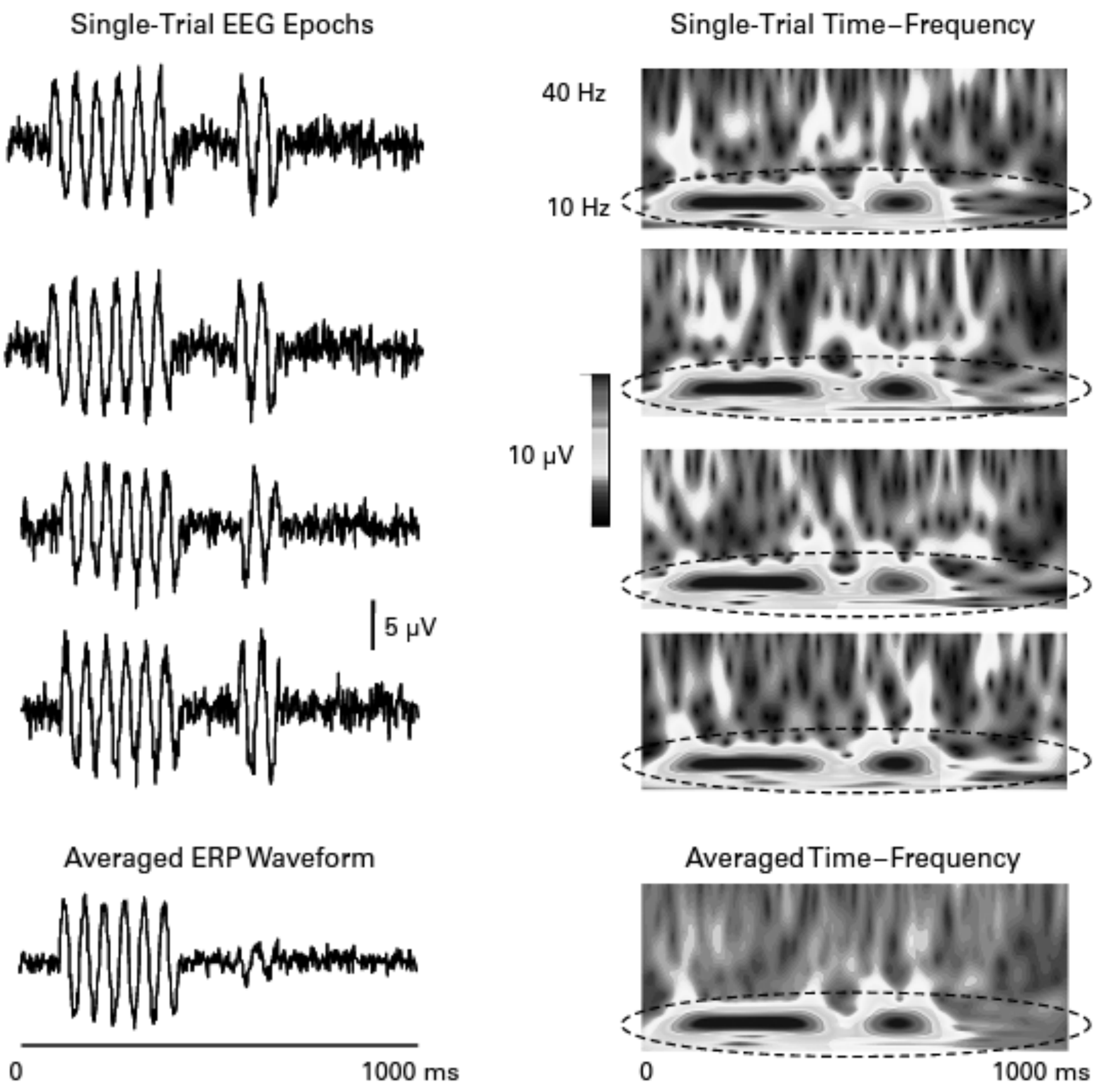

Figure 8.11
In-depth look at time–frequency analysis. The left column shows voltage waveforms for four single trials and the average. The single trials contain an initial 10-Hz oscillation that has the same phase on each trial, followed by a later 10-Hz oscillation that varies in phase from trial to trial. The constant-phase oscillation appears in the average, but the random-phase oscillation cancels out and is largely absent from the average. The right column shows the time–frequency transformation of the single trials, along with the average of the time–frequency transformations (note that this is not the time–frequency transformation of the averaged voltage waveform). You can see both the early and late 10-Hz oscillations on each trial and in the average (highlighted with the dashed ellipses). If the time–frequency transformation had been applied to the averaged ERP waveform, only the first 10-Hz oscillation would have been present. Adapted with permission from Bastiaansen et al. (2012). Copyright 2012 Oxford University Press.

200 to 300 ms, because it could just be a stimulus-independent difference in alpha power. The same is true of across-group comparisons: Without baseline correction, you cannot determine whether a difference between groups in the poststimulus period reflects a general difference that is unrelated to the stimuli or task and is also present prior to the stimulus.

However, when baseline correction is performed, you still need to be cautious about how you interpret the data. Recall from chapter 6 that alpha activity is typically suppressed shortly after the onset of a stimulus (see, e.g., the data from subject 1 in figure 6.6). Consequently, if a patient group has more alpha during the prestimulus period than a control group, but they both have the same amount of alpha from 300 to 400 ms poststimulus in the absence of baseline correction, then the patient group will appear to have less alpha than the control group after baseline correction is performed.

In my reading of the literature, studies using time–frequency analysis are quite variable in whether and how they use baseline correction. Some perform baseline correction and others don't. When baseline correction is performed, some studies subtract the average baseline power at a given frequency from the power at each poststimulus time point for that frequency. Other studies use division rather than subtraction (i.e., for each frequency, the power at each poststimulus time point is divided by the average prestimulus power). In other studies, the change between the prestimulus baseline and the poststimulus period is represented on a log scale (decibels) to take into account the fact that power typically falls off as the frequency increases. This variability across studies doesn't mean that the researchers are doing anything wrong, because different baseline correction procedures may be appropriate in different situations. However, when you read these studies, you need to look carefully at how the baseline is treated so that you understand exactly what information the time–frequency analyses are showing.

Is It Really an Oscillation?

Time–frequency analysis provides a very useful technique for making random-phase oscillations visible. However, it is very easy to draw an incorrect conclusion from time–frequency data; namely, that the brain activity actually *consists* of oscillations (i.e., a series of upward and downward deflections). As I discussed previously in chapter 7, a brief monophasic ERP deflection is mathematically equivalent to the sum of many sine waves at many different frequencies, just as a dollar bill has the same value as 100 pennies. However, the ERP deflection doesn't consist of the sine waves any more than a dollar bill consists of 100 pennies. Consequently, the presence of activity in a given frequency band does not entail the existence of a true oscillation.

As mentioned in chapter 7, a true oscillation will usually consist of a relatively narrow band of activity. In contrast, a non-oscillating transient brain response will typically lead to activity spread across a wide band of frequencies, beginning at the very lowest frequencies. For example, the alpha oscillations in figure 8.11 form a narrow band of activity at 10 Hz, with very little activity at lower frequencies. In contrast, figure 8.10 does not show any frequencies below 20 Hz, making it difficult to know if the observed effects are confined to a narrow frequency band

or if they extend all the way down to 0 Hz. If a study does not show the data from relatively low frequencies, you should not accept any conclusions they draw about oscillations per se. This is discussed in more detail in online chapter 12.

Suggestions for Further Reading

Bastiaansen, M., Mazaheri, A., & Jensen, O. (2012). Beyond ERPs: Oscillatory neuronal dynamics. In S. J. Luck & E. S. Kappenman (Eds.), T*he Oxford Handbook of ERP Components* (pp. 31–49). New York: Oxford University Press.

Cohen, M. X. (2014). *Analyzing Neural Time Series Data: Theory and Practice*. Cambridge, MA: MIT Press.

Mazaheri, A., & Jensen, O. (2008). Asymmetric amplitude modulations of brain oscillations generate slow evoked responses. *Journal of Neuroscience, 28*, 7781–7787.

Roach, B. J., & Mathalon, D. H. (2008). Event-related EEG time-frequency analysis: An overview of measures and an analysis of early gamma band phase locking in schizophrenia. *Schizophrenia Bulletin, 34*, 907–926.

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.