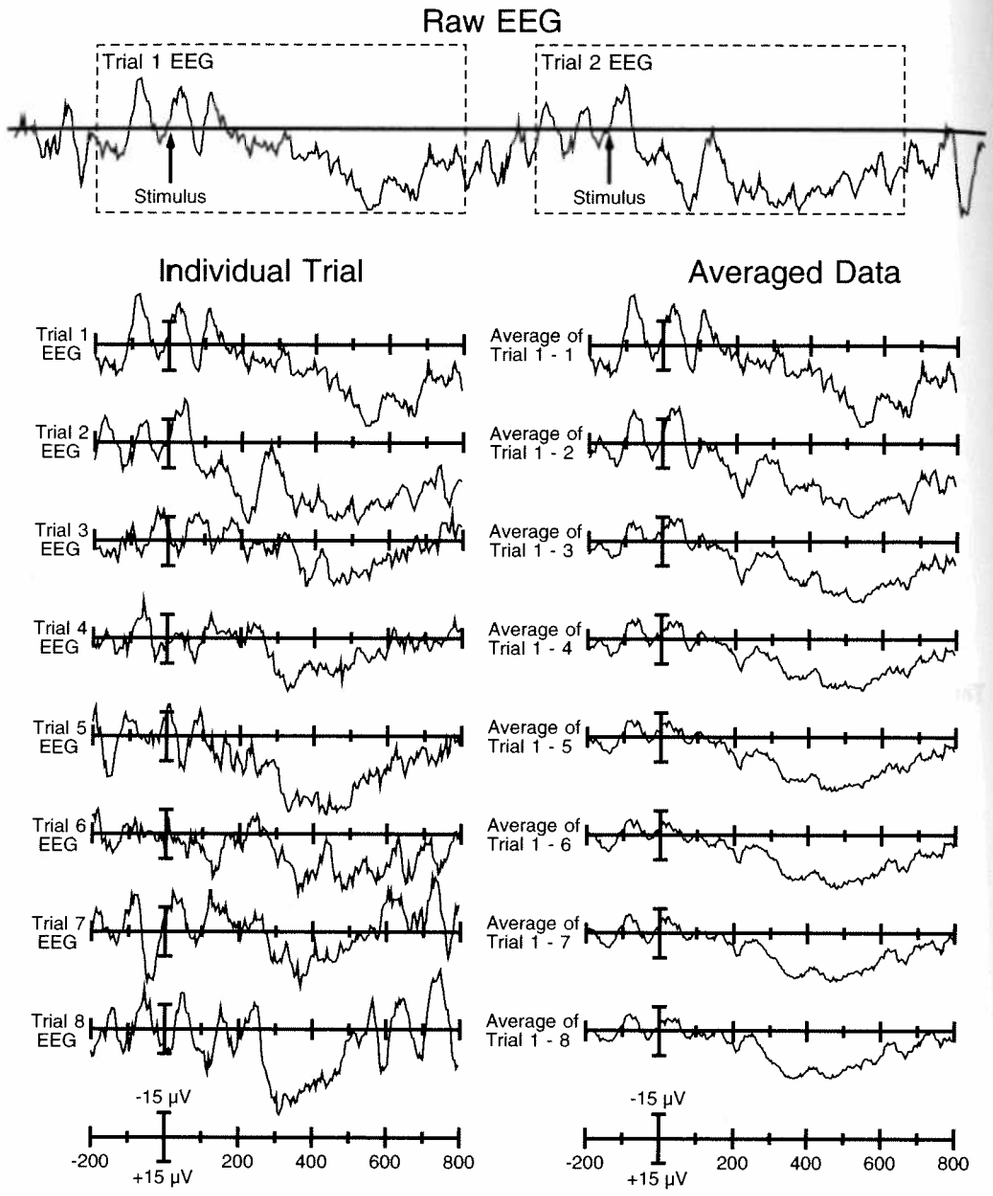
4 Averaging, Artifact Rejection, and Artifact Correction

Because ERPs are embedded in a larger EEG signal, almost all ERP studies rely on some sort of averaging procedure to minimize the EEG noise, and the averaging procedure is typically accompanied by a process that eliminates trials containing artifacts or followed by some procedure to correct for artifacts. These procedures appear to be relatively simple, but there are many important and complex issues lurking below the surface that one must understand before applying them. This chapter will discuss the underlying issues and provide several practical suggestions for averaging and for dealing with artifacts.

The Averaging Process

Basics of Signal Averaging

Figure 4.1 illustrates the traditional approach to signal averaging. First, EEG epochs following a given type of event (usually a stimulus) are extracted from the ongoing EEG. These epochs are aligned with respect to the time-locking event and then simply averaged together in a point-by-point manner. The logic behind this procedure is as follows. The EEG data collected on a single trial is assumed to consist of an ERP waveform plus random noise. The ERP waveform is assumed to be identical on each trial, whereas the noise is assumed to be completely unrelated to the time-locking event. If you could somehow extract just the ERP waveform from the single-trial EEG data, it would look exactly the same on every trial, and averaging together several trials would yield the



same waveform that was present on the individual trials. In contrast, if you could somehow extract just the noise from the EEG data, it would be random from trial to trial, and the average of a large number of trials would be a flat line at zero microvolts. Thus, when you average together many trials containing both a consistent ERP waveform and random noise, the noise is reduced but the ERP waveform remains.

As you average together more and more trials, the noise remaining in the averaged waveform gets smaller and smaller. Mathematically speaking, if R is the amount of noise on a single trial and N is the number of trials, the size of the noise in an average of the N trials is equal to $(1/\sqrt{N}) \times R$. In other words, the remaining noise in an average decreases as a function of the square root of the number of trials. Moreover, because the signal is assumed to be unaffected by the averaging process, the signal-to-noise (S/N) ratio increases as a function of the square root of the number of trials.

As an example, imagine an experiment in which you are measuring the amplitude of the P3 wave, and the actual amplitude of the P3 wave is 20 µV (if you could measure it without any EEG noise). If the actual noise in the EEG averages 50 µV on a single trial, then the S/N ratio on a single trial will be 20:50, or 0.4 (which is not very good). If you average two trials together, then the S/N ratio will increase by a factor of 1.4 (because $\sqrt{2} = 1.4$). To double the S/N ratio from .4 to .8, it is necessary to average together four trials (because $\sqrt{4} = 2$). To quadruple the S/N ratio from .4 to 1.6, it is necessary to average together sixteen trials (because $\sqrt{16} = 4$). Thus, doubling the S/N ratio requires four times as many trials and quadrupling the S/N ratio requires sixteen times as many trials. This relationship between the number of trials and the S/N

◀ **Figure 4.1** Example of the application of signal-averaging. The top waveform shows the raw EEG over a period of about 2 seconds, during which time two stimuli were presented. The left column shows segments of EEG for each of several trials, time-locked to stimulus onset. The right column shows the effects of averaging one, two, three, four, five, six, seven, or eight of these EEG segments. Negative is plotted upward.

ratio is rather sobering, because it means that achieving a substantial increase in S/N ratio requires a very large increase in the number of trials. This leads to a very important principle: *It is usually much easier to improve the quality of your data by decreasing sources of noise than by increasing the number of trials.*

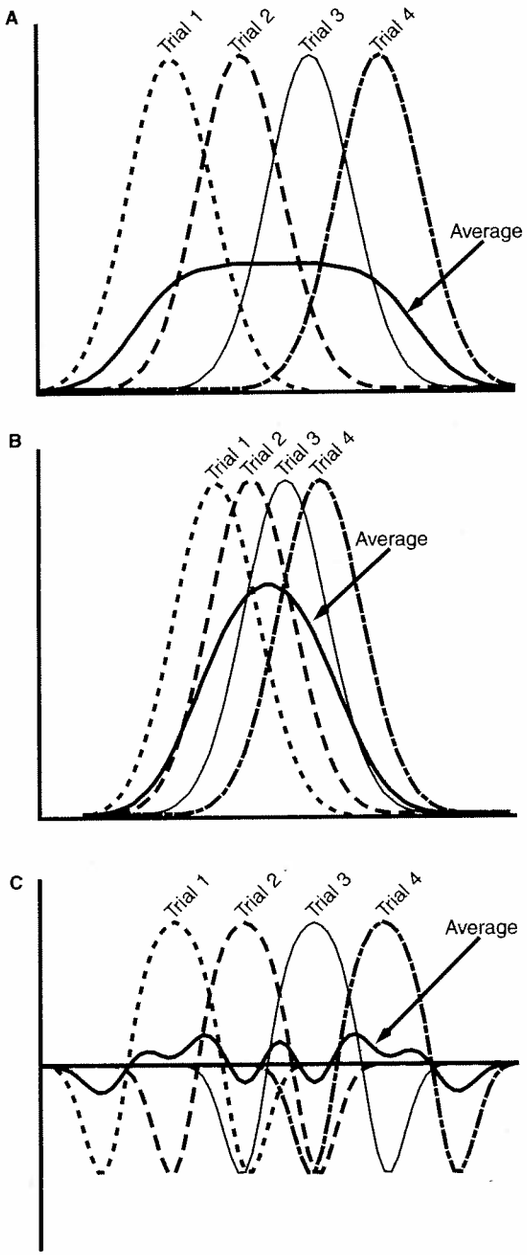
To exemplify these principles, figure 4.1 shows the application of signal averaging to a P3 oddball experiment. The top portion of the figure shows the continuous EEG signal, from which the single-trial EEG segments are taken. The left column shows the EEG segments for eight different trials in which an infrequent target was presented. The P3 wave for this subject was quite large, and it can be seen in every trial as a broad positivity in the 300–700 ms latency range. However, there is also quite a bit of variability in the exact shape of the P3 wave, and this is at least partly due to random EEG fluctuations (the P3 itself may also vary from trial to trial). The right column in figure 4.1 shows how averaging together more and more trials minimizes the effects of the random EEG fluctuations. The difference between trial 1 alone and the average of trials 1 and 2 is quite substantial, whereas the difference between the average of trials 1–7 and the average of trials 1–8 is small (even though trial 8 is quite different from the other trials). Note also that the S/N ratio in the average of trials 1–8 is 2.8 times greater than the S/N ratio on the individual trials (because $\sqrt{8} = 2.8$).

The signal-averaging approach is based on several assumptions, the most obvious of which are that (a) the neural activity related to the time-locking event is the same on every trial, and (b) only the EEG noise varies from trial to trial. These assumptions are clearly unrealistic, but violations are not problematic in most cases. For example, if the amplitude of the P2 wave varies from trial to trial, then the P2 wave in the averaged ERP waveform will simply reflect the average amplitude of the P2 wave. Similarly, one could imagine that a P1 wave is present on some trials, a P2 wave is present on other trials, and that both components are never present together on a given single trial. The averaged ERP waveform, however, would contain both a P1 wave and a P2 wave, which would incorrectly imply that the two components are part of the same waveform. However, the conclusions of most ERP experiments do not depend on the assumption that the different parts of the averaged waveform are actually present together on individual trials, so this sort of variability is not usually problematic as long as you always remember that the average is only one possible measure of central tendency.

The Problem of Latency Variability

Although trial-to-trial variability in ERP amplitude is not usually problematic, trial-to-trial variability in latency is sometimes a significant problem. Figure 4.2A illustrates this, showing four individual trials in which a P3-like ERP component occurs at different latencies. The peak amplitude of the average of these trials is much smaller than the peak amplitude on the individual trials. This is particularly problematic when the amount of latency variability differs across experimental conditions. As figure 4.2B shows, a reduction in latency variability causes the peak amplitude of the average to be larger. 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, and this could lead an investigator to incorrectly conclude that there was a difference in amplitude. Worse yet, latency variability can sometimes make it completely impossible to see a given neural response in an averaged waveform (see figure 4.2C). For example, imagine a sinusoidal oscillation that is triggered by a stimulus but varies randomly in phase from trial to trial (which is not just a hypothetical problem—see Gray et al., 1989). Such a response will average to zero and will be essentially invisible in an averaged response.

Figure 4.3 shows a real-world example of latency jitter (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, and we expected that in this condition the subjects would search one item at a time until they found the target. In this condition, therefore, we expected reaction time to increase as the number of items in the array (the set size) was increased, whereas we expected no effect of set size in the parallel search condition. This was the pattern of results that we obtained.

We also predicted that both P3 latency and P3 amplitude would increase at the larger set sizes in the serial search condition, but not in the parallel search condition. However, because the order in which the subjects search the arrays is essentially random, we also predicted that there would be more trial-to-trial variability in P3 latency at the larger set sizes in the serial search condition, which made it difficult to measure P3 amplitude and latency. 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 first item and the twelfth item. Consequently, we predicted that P3 latency would be more variable at the larger set sizes in the serial search condition (but not in the parallel search condition, in which the target was detected immediately regardless of the set size).

Figure 4.3B shows the averaged ERP waveforms from this experiment. In the parallel search condition, the P3 wave was relatively large in amplitude and short in duration, and it did not vary much

◀ **Figure 4.2** Example of the problem of latency variation. Each panel shows four single-trial waveforms, along with the average waveform. 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 for the average waveform in panel A. 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.

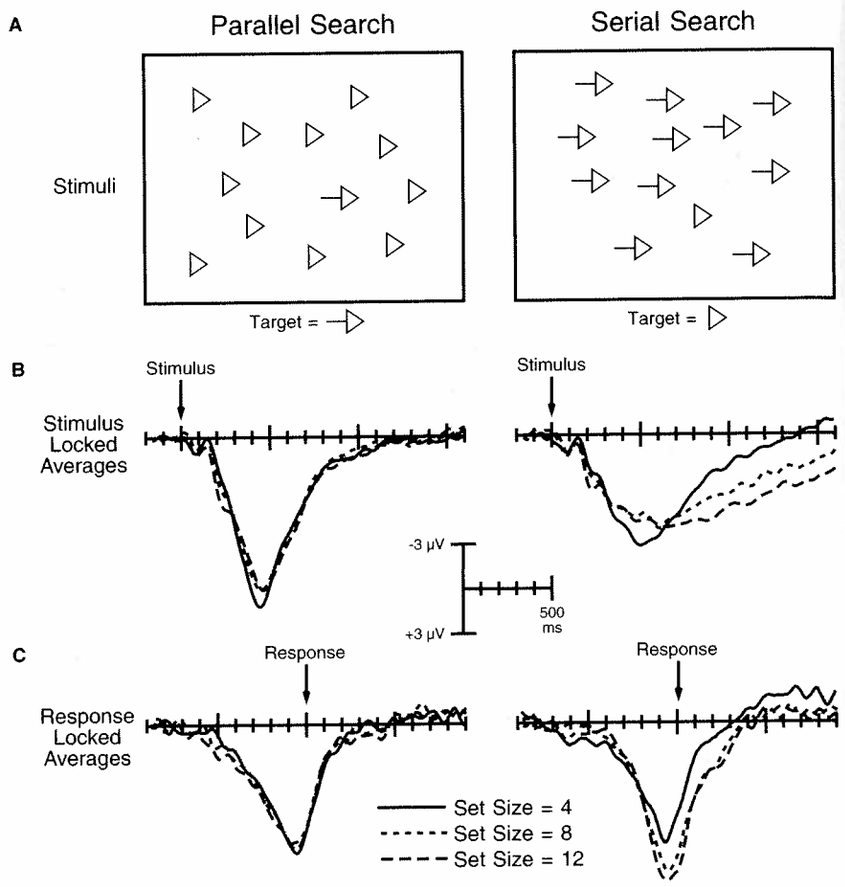


Figure 4.3 Example of an experiment in which significant latency variability was expected for the P3 wave (Luck & Hillyard, 1990). (A) Sample stimuli from the two conditions of the experiment. (B) Stimulus-locked averages. (C) Response-locked averages. Negative is plotted upward.

as a function of set size. In the serial search condition, the P3 had a smaller peak amplitude but was very broad. If you were to measure the amplitude and latency at the peak of the P3 wave in the serial search condition, you might conclude that set size didn't have much of an effect on the P3 in either the serial or parallel search conditions. However, both the amplitude and the latency of the P3 wave were significantly influenced by set size in the serial search condition, although these effects were masked by the latency variability.

I will now describe some techniques that you can use to minimize the effects of latency variability.

Area Measures In most cases, you can mitigate the reduction in amplitude caused by latency variability simply by using an area amplitude measure rather than a peak amplitude measure. The area under the curve in an average of several trials is always equal to the average of the areas under the curves in each of the individual trials, and an area measure will therefore be completely unaffected by latency variability. In the experiment illustrated in figure 4.3, for example, the area amplitude was significantly greater at larger set sizes, even though peak amplitude was slightly smaller at the larger set sizes. It is also possible to use an area-based measure of latency rather than a peak-based measure. To do this, you simply measure the area under the curve and find the time point that divides this area into equal halves (this is called a *50 percent area latency* measure). When this measure was applied to the data shown in figure 4.3B, the effects of set size on P3 latency were found to be almost identical to the effects on reaction time. Chapter 6 will discuss these area-based measures in greater detail.

Area-based measures are almost always superior to peak-based measures, and insensitivity to latency variability is just one of several reasons why I prefer area-based measures. There are two caveats, however. First, the equivalence between the area in the individual trials and the area in the average is true only when the

latency range used to measure the area spans the entire latency range of the component. When there are multiple overlapping ERP components that vary across conditions, it is sometimes necessary to use a relatively restricted measurement window, in which case the area measure is no longer completely insensitive to latency variability (although it's still usually better than a peak amplitude measure). The second caveat is that area measures can misrepresent components that are multiphasic (i.e., components with both positive and negative portions). As figure 4.2C shows, the negative and positive portions of a multiphasic waveform may cancel each other when latency variability is present, and once the data have been averaged there is no way to recover the information that is lost due to this cancellation. Thus, area-based measures are useful for mitigating the effects of latency variability under most conditions, but they are not adequate when there is variability in a multiphasic waveform or when overlapping components preclude the use of a wide measurement window.

Response-Locked Averages In some cases, variations in the latency of an ERP component are correlated with changes in reaction time, 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. Consider, for example, the visual search experiment illustrated in figure 4.3. In the serial search condition, the P3 wave was "smeared out" by latency variability in the stimulus-locked averages, leading to a low peak amplitude and a broad waveform. When response-locked averages were computed, however, the P3 wave in this condition was larger and much more narrowly peaked (see figure 4.3C). In addition, the response-locked averages show that the P3 wave was actually larger at set size 12 than at set size 4, even though the peak amplitude was larger for set size 4 in the stimulus-locked averages. Many studies have used response-locked averages in this manner.

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 S/N ratio 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.[1] 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 the component of interest (see 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.

However, even when one is examining a large component such as the P3 wave, 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 4.3, but it didn't work very well. The P3 wave in this experiment could peak anywhere between 400 and 1,400 milliseconds 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.

One should note 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 could be specified in addition to the component's waveshape, which would make it possible to reject spurious EEG deflections that may have the correct waveshape but have an incorrect scalp distribution (see Brandeis et al., 1992).

Time-Locked Spectral Averaging The final technique considered here is a means of extracting oscillatory responses that have random phase (onset time) with respect to the time-locking event. As discussed at the beginning of this section, oscillations that vary in phase will be lost in a conventional average, but one can see these oscillations using techniques that measure the amplitudes of the oscillations on single trials and then average these amplitude measures across trials. The techniques for measuring single-trial

oscillation amplitudes rely on variants of a mathematical procedure called the *Fourier transform*. As I will discuss more fully in chapter 5, 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-second 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. In this manner, you could compute the amplitude at each frequency for a single trial, and you could then average these amplitude measures across trials. Moreover, because the amplitude is measured independently of the phase, it wouldn't matter if the phase (i.e., the latency) of the oscillations varied across trials.

Although this is a useful approach, it completely discards the temporal information of the ERP technique, because the amplitude measured on a given trial is the amplitude for the entire time period. Temporal information can be retained, however, by using *moving window* techniques, such as those developed by Makeig (1993) and by Tallon-Baudry and colleagues (1996). These techniques extract a brief window of EEG from the beginning of the trial (e.g., the first 100 ms of EEG). The Fourier transform is then applied to this window to provide a quantification of the amplitude at each frequency during that relatively brief time range. The window is then moved over slightly (e.g., by 10 ms), and another Fourier transform is applied. In this manner, it is possible to compute Fourier transforms for every point in the EEG, although the values at a given time point actually represent the frequencies over a period of time (e.g., a 100-ms period). The Fourier transforms at a given time point are then averaged across trials just as the EEG amplitude would be averaged across trials in a conventional average. This is called *time-locked spectral averaging* because time-locked averages are computed for spectral (i.e., frequency) information.

Figure 4.4A[2] shows an example of this technique, presenting time-locked spectral averages from the study of Tallon-Baudry et al. (1996), who were interested in the 40-Hz oscillations elicited

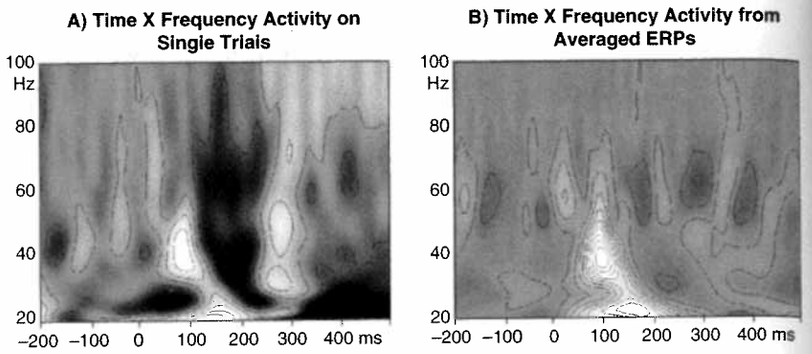


Figure 4.4 Example of time-locked spectral averaging. In panel A, the frequency transformation was applied to the individual trials and the transformed data were then averaged. This plot therefore includes activity that was not phase-locked to stimulus onset as well as phase-locked activity. In panel 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. © 1996 Society for Neuroscience.)

by visual stimuli. The X-axis in this plot is time, just as in a traditional ERP average. The Y-axis, however, is frequency, and the gray-scale level indicates the power that was present at each frequency at each time point. A band of activity between 40 and 50 Hz can be seen at approximately 100 ms poststimulus, and a somewhat weaker band of activity between 30 and 60 Hz can be seen at approximately 300 ms poststimulus. Activity can also be seen in the 20-Hz range from about 100 to 200 ms poststimulus.

The crucial aspect of this approach is that these bands of activity can be seen whether or not the oscillations vary in phase from trial to trial, whereas random-phase activity is completely lost in a traditional average. Time-locked spectral averaging thus provides a very useful technique for examining random-phase oscillations. However, it is very easy to draw an incorrect conclusion from data such as those shown in figure 4.4A, namely that the activity really consists of oscillations. As I will discuss fully in chapter 5, a brief

monophasic ERP deflection contains activity at a variety of frequencies, and the presence of activity in a given frequency band does not entail the existence of a true oscillation (i.e., an oscillation with multiple positive and negative deflections). For example, figure 4.4B shows the time × frequency transformation of the traditional ERP averages from the study of Tallon-Baudry et al. (1996), and the 40–50 Hz activity at 100 ms poststimulus can be seen in this plot just as in the single-trial data. In other words, this activity was a part of the traditional ERP and was not a random-phase oscillation. Thus, to draw conclusions about random-phase oscillations, it is necessary to apply the time × frequency transformation to the averaged ERPs as well as to the single-trial EEG data.

Overlap from Preceding and Subsequent Stimuli

Overlapping ERPs from previous and subsequent stimuli will distort averaged ERP waveforms in ways that are sometimes subtle and sometimes obvious, and it is important to understand how this arises and when it might lead you to misinterpret your data (for a detailed treatment of this issue, see Woldorff, 1988). Overlap arises when the response to the previous stimulus has not ended before the baseline period prior to the current stimulus or when the subsequent stimulus is presented before the ERP response to the current stimulus has terminated. This problem is particularly acute when stimuli are presented rapidly (e.g., 1 second or less between stimulus onsets). However, ERP waveforms can last for several seconds, and overlap can significantly distort your data even at long interstimulus intervals.

Figure 4.5 illustrates the overlap problem for a thought experiment in which a stimulus is presented every 300–500 ms. Panel A shows the actual waveform elicited by a given stimulus without any overlap. Note that the prestimulus period is flat and that the waveform falls to zero at 1000 ms poststimulus. Panel B shows the waveforms that the same stimulus would have produced if

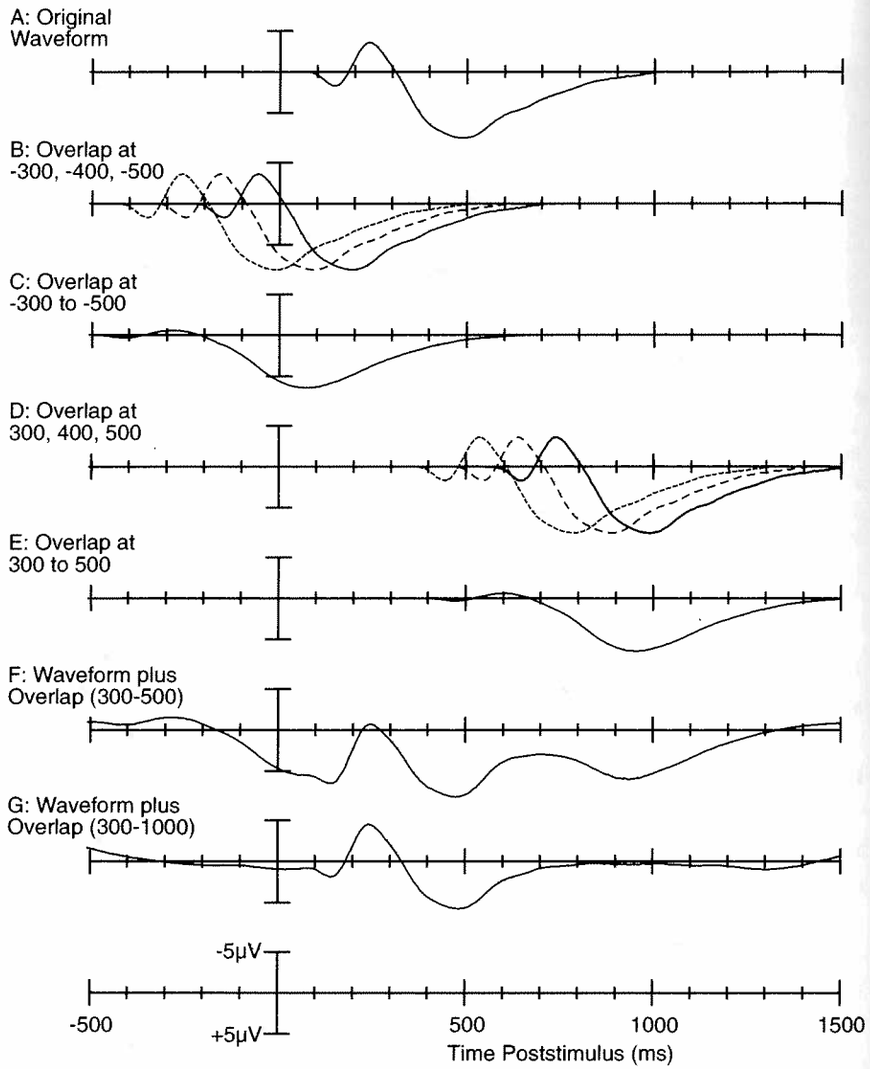


Figure 4.5 Example of the problem of overlapping waveforms. (A) An example ERP waveform. (B) Waveforms produced by the previous stimulus when it appears 300, 400, or 500 ms prior to the current stimulus. (C) Average waveform produced by the previous stimulus when it appears at random times between 300 and 500 ms prior to the current stimulus. (D) Waveforms produced by the subsequent stimulus when it appears

it appeared 300, 400, or 500 ms prior to the current stimulus; these are just the original waveform shifted to the left by various amounts. Panel C shows the average of a large number of previous waveforms, elicited by stimuli happening randomly and equiprobably between 300 and 500 ms prior to the current stimulus. This is the average overlap from the preceding stimuli. Panel D shows the responses elicited by the subsequent stimulus at 300, 400, or 500 ms, and panel E shows the overlap that would occur with stimuli occurring randomly 300–500 ms after the current stimulus.

Note that the jittered timing in this thought experiment leads to a "smearing" of the averaged waveform for the overlapping stimuli. That is, the relatively sharp positive and negative peaks at the beginning of the original waveform are mostly (but not entirely) eliminated in the overlapping waveform. The effect of temporal jitter between stimuli is equivalent to filtering out the high frequencies from the original waveform. As the range of time delays between the stimuli becomes wider and wider, the jitter reduces lower and lower frequencies from the overlap. However, even with a broad jitter, some low-frequency overlap will still occur (chapter 5 will describe a set of mathematical formalizations that you can use to understand in detail the filtering properties of temporal jitter).

Panel F shows the sum of the original waveform and the overlapping waveforms. This sum is exactly what you would obtain if you simply averaged together all of the stimuli in this thought experiment. The distortion due to overlap is quite severe. First, the prestimulus baseline is completely distorted, and this leads the initial positive component to seem much larger than it really is (compare the first positive component in panel G to the original waveform

◀ **Figure 4.5** (continued)
300, 400, or 500 ms after the current stimulus. (E) Average waveform produced by the subsequent stimulus when it appears at random times between 300 and 500 ms after the current stimulus. (F) Sum of the original waveform and the overlapping waveforms from (C) and (E). (G) Sum of the original waveform and the overlapping waveforms that would be produced if the interval between stimuli was increased to 300–1000 ms. Negative is plotted upward.

in panel A). Second, the first positive component appears to start before time zero (which is always a good indication that something is amiss). Third, there is a late positive peak that is completely artifactual.

Overlap is particularly problematic when it differs between experimental conditions. Imagine, for example, that the same target stimulus is presented in condition A and in condition B, and the preceding stimulus elicits a large P3 wave in condition A and a small P3 wave in condition B. This difference in the preceding stimuli may influence the prestimulus baseline period, and this in turn will influence the apparent amplitude of the P3 elicited by the target stimuli on the current trial. Note that this can happen even if the ERP response to the preceding stimulus has returned to baseline before the P3 wave on the current trial: if the baseline is affected, then the whole waveform will be affected. This can also have a big impact on attempts to localize the generator of an ERP component, because the scalp distribution of the overlapping activity in the prestimulus period will be subtracted away from the scalp distribution of the component that you are trying to localize, distorting the apparent scalp distribution of the component.

There are some steps that you can take to minimize the effects of overlap. The first and most important step is to think carefully about exactly what pattern the overlap will take and how it might differ across conditions (for an example, see Woldorff & Hillyard, 1991). The key is to think about how the task instructions may change the response to the preceding stimulus across conditions, even if it is the same physical stimulus. You may even want to simulate the overlap, which isn't conceptually difficult (for details, see Woldorff, 1988). It's computationally trivial to do this in a programming environment such as MATLAB, and it can even be done fairly easily in Excel.

A second approach is to use the broadest possible range of time delays between stimuli. Panel G of figure 4.5 shows what happens if the jitter in our thought experiment is expanded to 300–1000 ms. There is still some overlap, but it is much smaller.

A third approach is to use high-pass filters to filter out any remaining overlap. As mentioned earlier in this section, jittering the interstimulus interval is equivalent to filtering out the high frequencies from the overlap, and the remaining low frequencies can be filtered offline with a high-pass filter. High-pass filtering can distort your waveforms in other ways, however, so you must do this cautiously (see chapter 5 for details).

A fourth approach is to design the experiment in a way that allows you to directly measure and subtract away the overlap. In our thought experiment, for example, we could include occasional trials on which no stimulus was presented and create averaged waveforms time-locked to the time at which the stimulus would ordinarily have occurred. These waveforms will contain only the overlap, and these overlap waveforms can be subtracted from the averaged waveforms that contained the response to an actual stimulus along with the overlap. This requires some careful thought, however, because the subject might notice the omission of a stimulus triggering an ERP (see, e.g., Picton, Hillyard, & Galambos, 1974). I have frequently used this approach in my own research (see, in particular, Luck, 1998b; Vogel, Luck, & Shapiro, 1998).

A fifth approach is to estimate the overlap from the data you have collected and subtract the estimated overlap from the averaged ERP waveforms. Woldorff (1988) has developed a technique called the *ADJAR* (adjacent response) filter that has been used for this purpose in a number of experiments (e.g., Hopfinger & Mangun, 1998; Luck et al., 1994; Woldorff & Hillyard, 1991).

Transient and Steady-State ERPs

If stimuli are presented at a constant rate rather than a variable rate, the overlap from the preceding and subsequent stimuli is fully present in the averaged ERP waveforms. In fact, the preceding and subsequent stimuli are perfectly time-locked to the current stimulus, so it makes sense that they would appear in the averaged ERP waveforms. Most cognitive ERP experiments therefore use some

jitter in the interstimulus interval unless the interstimulus interval is fairly long.

It is sometimes possible to make overlap into a virtue. Specifically, if a series of identical stimuli are presented at a fast, regular rate (e.g., eight stimuli per second), the system will stop producing complex *transient* responses and enter into a *steady state*, in which the system resonates at the stimulus rate (and multiples thereof). Typically, steady state responses will look like two summed sine waves, one at the stimulation frequency and one at twice the stimulation frequency.

Figure 4.6 shows an example of a steady state response. The upper left portion of the figure shows the transient response obtained when the on-off cycle of a visual stimulus repeats twice per second. When the stimulation rate is increased to 6 cycles per second, it is still possible to see some distinct peaks, but the overall waveform now appears to repeat continuously, with no clear beginning or end. As the stimulation rate is increased to 12 and then 20

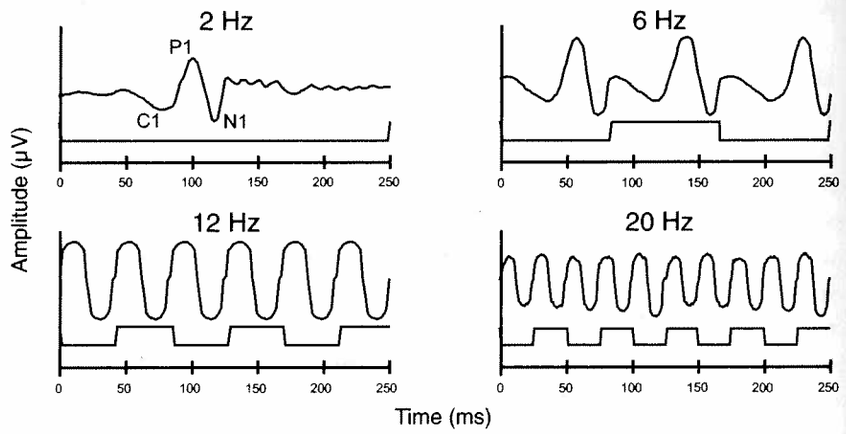


Figure 4.6 Transient response to a stimulus presented at a rate of two on-off cycles per second and steady-state response to a stimulus presented at 6, 12, or 20 Hz. Positive is plotted upward. (Adapted with permission from Di Russo, Teder-Sälejärvi, & Hillyard, 2003. © 2003 Academic Press.) Thanks to Francesco Di Russo for providing an electronic version of this figure.

cycles per second, the response is predominantly a sine wave at the stimulation frequency (with a small, hard-to-see component at twice the stimulation frequency).

This steady-state response can be summarized by four numbers, the amplitude (size) and phase (temporal shift) of each of the two sine waves. This is a lot simpler than a complex transient response with a separate amplitude value at each point in time. As a result, steady-state ERPs are widely used in the study of sensory systems and in the diagnosis of sensory disorders.

Steady-state ERPs have a significant shortcoming, however, which is that they do not provide very precise temporal information. For example, if stimuli are presented every 150 ms, the voltage measured at 130 ms after the onset of one stimulus consists of the sum of the response to the current stimulus at 130 ms, the response to the previous stimulus at 280 ms, the response to the stimulus before that at 430 ms, and so on. Because steady-state ERPs lack the high temporal resolution of transient ERPs, they are used only rarely in cognitive studies (for a review of some recent cognitive steady-state studies, see Hopfinger, Luck, & Hillyard, 2004).

Artifact Rejection and Correction

Now that we have considered the averaging process, we will move on to the artifact rejection procedures that typically accompany it.

There are several types of artifacts that can contaminate EEG recordings, including blinks, eye movements, muscle activity, and skin potentials. These artifacts can be problematic in two ways. First, they are typically very large compared to the ERP signals and may greatly decrease the S/N ratio of the averaged ERP waveform. Second, some types of artifacts may be systematic rather than random, occurring in some conditions more than others and being at least loosely time-locked to the stimulus so that the averaging process does not eliminate them. Such artifacts may lead to erroneous conclusions about the effects of an experimental manipulation.

For example, some stimuli may be more likely to elicit blinks than others, which could lead to differences in amplitude in the averaged ERP waveforms.

There are two main classes of techniques for eliminating the deleterious effects of artifacts. First, it is possible to detect large artifacts in the single-trial EEG epochs and simply exclude contaminated trials from the averaged ERP waveforms (this is called *artifact rejection*). Alternatively, it is sometimes possible to estimate the influence of the artifacts on the ERPs and use correction procedures to subtract away the estimated contribution of the artifacts (this is called *artifact correction*). In this section, I will discuss both approaches. However, I would first like to make a point that should be obvious but is often overlooked. Specifically, it is always better to minimize the occurrence of artifacts rather than to rely heavily on rejection or correction procedures. This is really just a special case of Hansen's Axiom: there is no substitute for good data. In other words, time spent eliminating artifacts at the source will be well rewarded. This section will therefore also include hints for reducing the occurrence of artifacts.

The General Artifact Rejection Process

Before I get into the details of how to detect specific types of artifacts, I would like to provide a general framework for conceptualizing the artifact rejection process.[3] Detecting artifacts is, in essence, a signal detection problem, in which the artifact is treated as the to-be-detected signal. As an example, imagine that you have lost a valuable ring on a beach, and you have rented a metal detector to help you find it. The metal detector has a continuously variable output that tells you the extent to which there is evidence of nearby metal, but this output is quite variable due to random fluctuations in the mineral content of the sand. If you started digging in the sand any time there was a hint of nearby metal, you would make very slow progress because you would start digging every few feet. However, if you only started digging when the metal

detector's output was very high, you might miss the ring altogether because it is small and doesn't create a large change in the detector's output. Thus, if you dig only when the detector's output reaches a very high level, you will probably pass right over the top of the ring, but if you dig whenever the output exceeds some small level, you will frequently be digging in vain.

The key aspects of this example are as follows. You are trying to detect something that is either there or not (the ring) based on a noisy, continuously variable signal (the metal detector's output). You select a threshold value, and if the signal exceeds that value, you make a response (digging). In this context, we can define four outcomes for each patch of sand: (1) a hit occurs when the sought-after object is present, the signal exceeds the threshold, and you respond (i.e., the metal detector's output exceeds a certain value, you dig, and you find the ring); (2) a miss occurs when the object is present, the signal fails to exceed the threshold, and you don't respond; (3) a false alarm occurs when the object is absent, the signal exceeds the threshold due to random variation, and you respond; (4) a correct rejection occurs when the object is absent, the signal doesn't exceed the threshold, and you don't respond. Hits and correct rejections are both correct responses, and misses and false alarms are both errors. Importantly, you can increase the number of hits by choosing a lower threshold (i.e., digging when the metal detector's output is fairly low), but this will also lead to an increase in the number of false alarms. The only way to increase the hit rate without increasing the false alarm rate is to get a better metal detector with an output that better differentiates between the presence or absence of small metal objects.

Now imagine that you are trying to detect blinks in a noisy EEG signal. When a subject blinks, the movement of the eyelids across the eyeball creates a voltage deflection, and it is possible to assess the presence or absence of a blink by measuring the size of the largest voltage deflection within a given segment of EEG (just like assessing the presence or absence of the ring by examining the output of the metal detector). If the voltage deflection exceeds a

certain threshold level, you conclude that the subject blinked and you discard that trial; if the threshold is not exceeded, you conclude that the subject did not blink and you include that trial in the averaged ERP waveform. If you set a low threshold and reject any trials that have even a small voltage deflection, you will eliminate all of the trials with blinks, but you will also discard many blink-free trials, reducing the signal-to-noise ratio of the averaged ERP waveform. If you set a high threshold and reject only trials with very large voltage deflections, you will have more trials in your averages, but some of those trials may contain blinks that failed to exceed your threshold. Thus, simply changing the threshold cannot increase the rejection of true artifacts without also increasing the rejection of artifact-free trials. However, just as you can do a better job of finding a ring in the sand by using a better metal detector, you can do a better job of rejecting artifacts by using a better procedure for measuring artifacts.

Choosing an Artifact Measure

Many software systems assess blinks by measuring the maximal voltage in the EOG signal on a given trial and rejecting trials in which this maximal voltage exceeds a threshold, such as ±75 μV. However, the peak amplitude in the EOG channel is a very poor measure of blink artifacts, because variations in the baseline voltage can bring the EOG signal far enough away from zero that small noise deflections will sometimes cause the voltage to exceed the threshold voltage, causing false alarms. Variations in baseline voltage can also bring the EOG signal *away* from the threshold voltage so that a true blink no longer exceeds the threshold, leading to misses.

Figure 4.7 illustrates these problems. Panel A of the figure shows an EOG recording with a blink that exceeds the 75 μV threshold and would be correctly rejected. Panel B shows an epoch in which a blink is clearly present, but because the baseline has shifted, the 75 μV threshold is not exceeded (a miss). Panel C

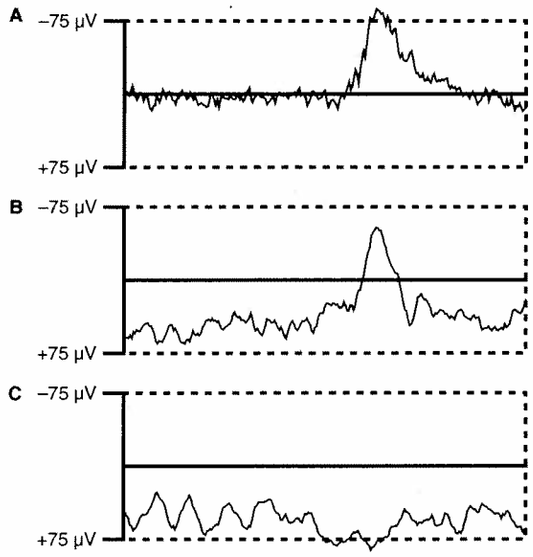


Figure 4.7 Example of the use of an absolute voltage threshold for artifact rejection. Each waveform shows single-trial activity recorded from an EOG electrode under the left eye, with a right-mastoid reference. In panel A, a blink is present and exceeds the threshold. In panel B, a blink is again present, but because the voltage had drifted downward, the threshold is not exceeded. In panel C, no blink is present, but because of a downward drift in the signal, the threshold is exceeded. Negative is plotted upward.

shows an epoch in which no blink was present, but a shift in baseline causes the 75 μV threshold to be exceeded by simple noise (a false alarm).

An alternative approach is to measure the difference between the minimum and maximum voltages within an EOG epoch and to compare this peak-to-peak voltage with some threshold voltage. This peak-to-peak measure is less distorted by slow changes in baseline voltage, reducing the impact of this possible source of misses and false alarms, and increasing the sensitivity of the artifact rejection process. Thus, it is possible to increase the rejection of trials with artifacts without increasing the rejection of

artifact-free trials by choosing a measure that can accurately distinguish between trials with and without artifacts. Later in this chapter I'll offer some suggestions for good measures.

The choice of the most sensitive measure will depend on what type of artifact you are measuring. For example, although peak-to-peak amplitude is a sensitive measure of blinks, it is not a very sensitive measure of the presence of alpha waves, which are frequently no larger than the background EEG activity. A more sensitive measure of alpha waves would be, for example, the amount of power in the 10-Hz frequency range, which would be high on trials contaminated by alpha waves and low for uncontaminated trials. Thus, a good artifact rejection system should allow the use of different measures for different types of artifacts.

It is important to note that a measure such as peak-to-peak amplitude is not really a measure of blinking, but is simply a numeric value that you can calculate from the data and use to differentiate probabilistically between trials with and without blinks. The artifact rejection process can thus be conceptualized in the general case as a two-step process, in which a "function" is applied to the data to compute a specific value and then this value is compared to the threshold. You can use different functions and different criteria in different cases, depending on the nature of the artifact that you are trying to detect.

Some investigators visually inspect the EEG on each trial to determine which trials contain artifacts, but this process is conceptually identical to the procedure that I just outlined. The only difference is that it uses the experimenter's visual system instead of a computer algorithm to determine the extent to which an artifact appears to be present and uses an informal, internal threshold to determine which trials to reject. The advantage of this approach is that the human visual system can be trained to do an excellent job of differentiating between real artifacts and normal EEG noise. However, a well-designed computer algorithm may be just as sensitive, if not more so. And computer algorithms have the advantages of being fast and not being prone to bias. Thus, it is usually best to use a good automated artifact rejection system rather than spending hours trying to identify artifacts by eye.

Choosing a Rejection Threshold

Once you have chosen an appropriate measure of an artifact, you must choose the threshold that will be used to determine whether to reject an individual trial. One possibility is to pick a threshold on the basis of experience and use this value for all subjects. For example, you may decide to reject all trials with a peak-to-peak EOG amplitude of 50 μV or higher. However, there is often significant variability across subjects in the size and shape of the voltage deflections a given type of artifact produces and in the characteristics of the EEG in which these voltage deflections are embedded, so a one-size-fits-all approach is therefore not optimal. Instead, it is usually best to tailor the artifact rejection process for each individual subject.

There is at least one exception to this rule, however: experiments that use different subject groups and in which the artifact rejection process could lead to some sort of bias. For example, it might not be appropriate to use different artifact rejection criteria for different subjects in a study that compared schizophrenia patients with normal controls, because any differences in the resulting averaged ERP waveforms could reflect a difference in artifact rejection rather than a real difference in the ERPs. However, using the same criteria for all subjects could also be problematic in such a study if, for example, one group had smaller blink amplitudes than another, resulting in more contamination from artifacts that escaped rejection. The best compromise in between-subject studies is probably to set the criteria individually for each subject, but to be blind to the subject's condition when the criteria are set. In addition, it would be worthwhile to determine whether the results are any different when using the same threshold for all subjects compared to when tailoring the threshold for each subject—if the results are the same, then the threshold is not causing any bias in the results.

If the threshold is set individually for each subject, the settings are usually based on visual inspection of a portion of the raw EEG. This can be accomplished by the following sequence of steps. First, select an initial threshold for a given subject as a starting point (usually on the basis of experience with prior subjects). Then apply this threshold to a set of individual trials and visually assess whether trials with real artifacts are not being rejected or if trials without real artifacts are being rejected. Of course, this requires that you are able to determine the presence or absence of artifacts by visual inspection. In most cases, this is fairly straightforward, and the next section provides some hints. After the initial threshold has been tested on some data, it can be adjusted and retested until it rejects all of the trials that clearly have artifacts without rejecting too many artifact-free trials (as assessed visually). Some types of artifacts also leave a distinctive "signature" in the averaged waveforms, so it is also possible to evaluate whether the threshold adequately rejected trials with artifacts after you have averaged the data.

It can also be useful to ask the subject to make some blinks and eye movements at the beginning of the session so that you can easily see what that subject's artifacts look like.

Detecting and Rejecting Specific Types of Artifacts

In this section, I will discuss several common types of artifacts and provide suggestions for reducing their occurrence and for detecting and rejecting them when they do occur.

Blinks Within each eye, there is an electrical gradient with positive at the front of the eye and negative at the back of the eye, and the voltage deflections recorded near the eye are primarily caused by the movement of the eyelids across the eyes, which modulates the conduction of the electrical potentials of the eyes to the surrounding regions. Figure 4.8 shows the typical waveshape of the eyeblink response at a location below the eyes (labeled VEOG) and

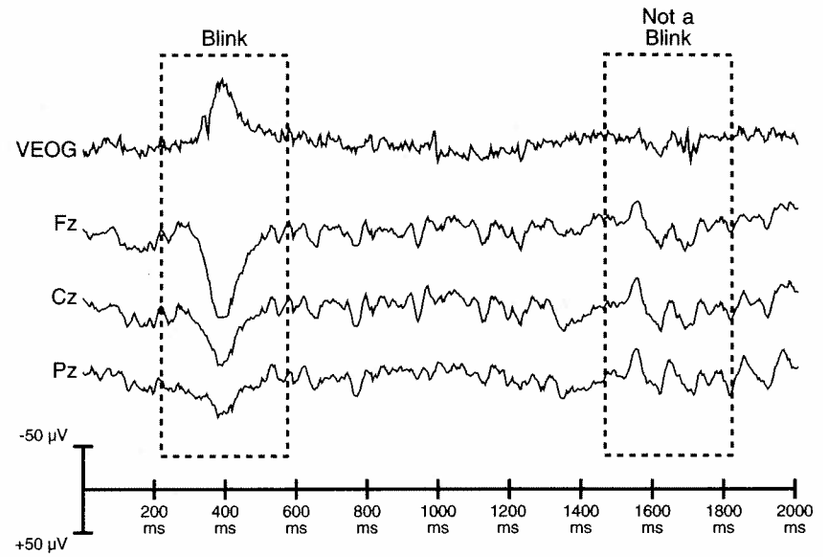


Figure 4.8 Recordings from a vertical EOG (VEOG) electrode located under the left eye and EEG electrodes located at Fz, Cz, and Pz, with a right mastoid reference for all recordings. A blink can be seen at approximately 400 ms, and it appears as a negative deflection at the VEOG electrode and as a positive deflection at the scalp electrodes. Note that the deflection is quite large at Fz and then becomes smaller at Cz and even smaller at Pz. The area labeled "Not a Blink" contains moderately large voltage deflections in all of these channels, but these deflections do not reflect a blink because the polarity is not inverted at the VEOG electrode relative to the scalp electrodes. Negative is plotted upward.

at several locations on the scalp (all are referenced to a mastoid electrode). The eyeblink response consists primarily of a monophasic deflection of 50–100 μV with a typical duration of 200–400 ms. Perhaps the most important characteristic of the eyeblink response, however, is that it is opposite in polarity for sites above versus below the eye (compare, for example, the VEOG and Fz recordings in figure 4.8). This makes it possible to distinguish between a blink, which would produce opposite-polarity voltage shifts above versus below the eye, and a true EEG deflection, which would typically produce same-polarity voltage shifts above and below the eye. The

right side of figure 4.8 shows an example of a true EEG deflection, where similar-polarity deflections appear at the VEOG and Fz sites.

Because of the polarity reversal exhibited by blinks, you should always be suspicious of an experimental effect that is opposite in polarity at electrode sites above versus below the eyes. Although such a pattern is possible for a true ERP effect, it should serve as a warning signal indicating that the averaged ERP waveforms may be more contaminated by blinks in one condition than in the other.

Reducing the occurrence of an artifact is always better than rejecting trials with artifacts, and there are several ways to reduce the number of blinks. The first is to ask subjects who normally wear contact lenses—which cause a great deal of blinking—to wear their glasses instead of their contact lenses. These individuals tend to blink more than average even when wearing glasses, and it is therefore useful to keep a supply of eyedrops handy (although you should offer them only to individuals who normally use eyedrops, and you should use single-use bottles to avoid infection risks). Another method for reducing the occurrence of blinks is to use short trial blocks of 1–2 minutes, thus providing the subjects with frequent rest breaks for blinking (this also helps to keep the subjects more alert and focused on the task). The use of such short trial blocks tends to slow down the progress of a recording session, but I have found that this slowing can be mitigated by using trial blocks of 5–6 minutes that are subdivided into "miniblocks" of 1–2 minutes, with automatic breaks of 20–30 seconds interposed between the miniblocks and somewhat longer, experimenter-controlled breaks between the full blocks.

If you see a lot of blinks (or eye movements), it is important to let the subject know. Don't be shy about telling subjects that they need to do a better job of controlling these artifacts. My students tell me that it took them a long time to become comfortable doing this, but you really need to do it, even if it makes you uncomfortable at first.

Blinks are relatively easy to detect on single trials, and a peak-to-peak amplitude measure is usually an adequate artifact rejection function (a simple voltage threshold, however, is clearly inadequate, because a threshold that is sufficiently low to reject all blinks often leads to a large number of false alarms). The peak-to-peak amplitude function can sometimes be "fooled" by slow voltage shifts that cause one end of the epoch to be substantially different in voltage from the other end, and high-frequency noise (e.g., muscle activity) can exacerbate this. Both of these problems can be minimized by a measure that I call a "step" function, which basically looks for step-like changes in voltage. This function is similar to performing a cross-correlation between the EEG/EOG epoch and a function that looks like a step (i.e., a flat low interval followed by a flat high interval). One first defines the width of the step, with a typical value of 100 ms. For each point in the epoch, the mean value of the preceding 100 ms is then subtracted from the mean value of the subsequent 100 ms (or whatever the desired step width is). After this has been computed for each point in the epoch, the largest value is compared with the threshold to determine whether the trial should be rejected. This computationally simple procedure is effective for two reasons. First, averaging together the voltage of a 100-ms interval essentially filters out any high-frequency activity. Second, computing the difference between successive 100-ms intervals minimizes the effects of any gradual changes in voltage, which corresponds with the fact that a blink produces a relatively sudden voltage change.

Whenever possible, one should obtain recordings from an electrode below the eye and an electrode above the eye, with both electrodes referenced to a common, distant site (e.g., an EOG electrode located below one of the two eyes and a frontal EEG site, both referenced to a mastoid electrode). This makes it possible to take advantage of the inversion of polarity exhibited by blinks for sites above and below the eyes, which is especially useful when inspecting the single-trial data during the assessment of the adequacy of the artifact rejection function and the threshold. As discussed above, this also makes it possible to determine whether the ERP averages are contaminated by blinks, which leads to an inversion in polarity for sites above versus below the eyes. In addition,

this montage makes it possible to implement an even more powerful artifact rejection function, which I call the "differential step" function. This function is just like the step function, except that one computes values at each time point for a location below and a location above the eyes and then subtracts these two values from each other. The maximum value of this difference is then compared with a threshold to determine whether a given trial should be rejected. The subtraction process ensures that a large value is obtained only when the voltage is changing in opposite directions for the electrodes above and below the eyes. A simpler but nearly equivalent alternative is to apply the standard step function to a bipolar recording in which the electrode beneath the eye is the active site and the electrode above the eye is the reference.

Eye Movements Like blinks, eye movements are a result of the intrinsic voltage gradient of the eye, which can be thought of as a dipole with its positive end pointing toward the front of the eye. When the eyes are stationary, this dipole creates a constant DC voltage gradient across the scalp, which the high-pass filter of the amplifier eliminates. When the eyes move, the voltage gradient across the scalp changes, becoming more positive at sites that the eyes have moved toward. For example, a leftward eye movement causes a positive-going voltage deflection on the left side of the scalp and a negative-going voltage on the right side of the scalp. It is easiest to observe these deflections with bipolar recordings, in which an electrode lateral to one eye is the active site and an electrode lateral to the other eye is the reference site.

Hillyard and Galambos (1970) and Lins et al. (1993a) have systematically measured the average size of the saccade-produced deflection. These studies yielded the following findings: (a) the voltage deflection at a given electrode site is a linear function of the size of the eye movement, at least over a 15-degree range of eye movements; (b) a bipolar recording of the voltage between electrodes at locations immediately adjacent to the two eyes will yield a deflection of approximately 16 μV for each degree of eye movement; and (c) the voltage falls off in a predictable manner as the distance between the electrode site and the eyes increases (see tables V and VI of Lins et al., 1993, for a list of the propagation factors for a variety of standard electrode sites).

Note also that eye movements cause the visual input to shift across the retina, which creates a visual ERP response (saccadic suppression mechanisms make us unaware of this motion, but it does create substantial activity within the visual system). This saccade-induced ERP depends on the nature of the stimuli that are visible when the eyes move, just as the ERP elicited by a moving stimulus varies as a function of the nature of the stimulus. Procedures that attempt to correct for the EOG voltages produced by eye movements—discussed at the end of this chapter—cannot correct for these saccade-induced ERP responses.

Unless the subject is viewing moving objects or exhibiting gradual head movements, the vast majority of eye movements will be *saccades*, sudden ballistic shifts in eye position. The top three waveforms of figure 4.9 show examples of eye movement recordings. In the absence of noise, a saccade would consist of a sudden transition from the zero voltage level to a nonzero voltage level, followed by a gradual return toward zero caused by the amplifier's high-pass filter (unless using DC recordings). In most cases, subjects make a saccade in one direction and then another to return to the fixation point, which would lead to a boxcar-shaped function in a DC recording and a sloped boxcar-shaped function when using a high-pass filter. This characteristic shape can be used to distinguish small eye movements from normal EEG deflections when visually inspecting the individual trials.

Because of the approximately linear relationship between the size of an eye movement and the magnitude of the corresponding EOG deflection, large eye movements are relatively easy to detect on single trials, but small eye movements are difficult to detect. If one uses a simple voltage threshold to detect and reject eye movement artifacts, with a typical threshold of 100 μV, eye movements as large as 10 degrees can escape detection (e.g., if the voltage

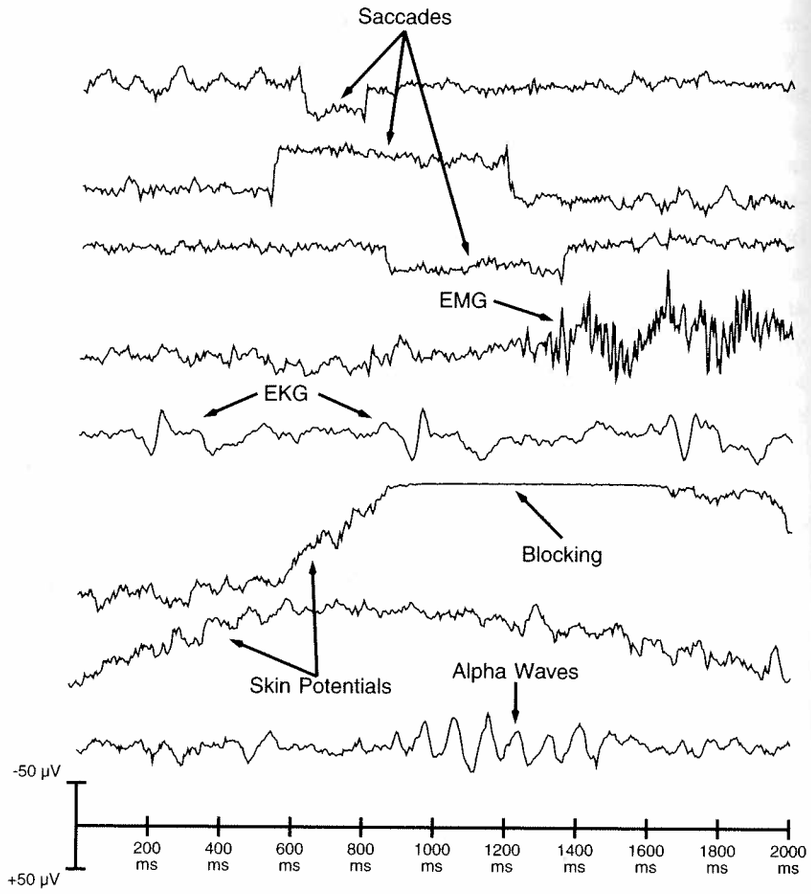


Figure 4.9 EOG and EEG recordings showing several types of artifacts. The saccades were recorded from a horizontal EOG configuration, with the active electrode adjacent to the right eye and the reference electrode adjacent to the left eye. The EMG, blocking, and skin potential artifacts were recorded at Cz with a right mastoid reference. The EKG artifacts were recorded at the left mastoid with a right mastoid reference. The alpha waves were recorded at O2 with a right mastoid reference. Negative is plotted upward.

starts at −80 μV, a 10-degree eye movement in the appropriate direction will cause a transition to +80 μV, which would be entirely within the allowable window of ±100 μV). Of course, a 10-degree eye movement greatly changes the position of the stimulus on the retina, which can be an important confound, and the resulting voltage deflection is quite large relative to the size of a typical ERP component, even at scalp sites fairly far from the eyes. However, using a lower threshold will lead to a large number of false alarms, and a simple threshold function is therefore an inadequate means of rejecting trials with eye movements. Peak-to-peak amplitude is somewhat superior to a threshold, but can be fooled by slow shifts in voltage. The step function described above for detecting blinks is better yet, because it is sensitive to temporally circumscribed shifts in voltage. Perhaps the best approach, however, would be to adapt the algorithms developed by vision researchers in models of edge detection, which is a conceptually similar problem. To my knowledge, however, no one has yet applied these algorithms to the problem of saccade detection.

Using a step function, it is possible to detect eye movements as small as 1 to 2 degrees on individual trials, but the S/N ratio of the EOG signal makes it impossible to detect smaller eye movements without an unacceptably large number of false alarms. However, it is sometimes possible to use averaged EOG waveforms to demonstrate that a given set of ERPs are uncontaminated by very small systematic eye movements. Specifically, if different trial types would be expected to elicit eye movements in different directions, you can obtain virtually unlimited resolution by averaging together multiple trials on which the eye movements would be expected to be similar. For example, if an experiment contains some targets in the LVF and other targets in the RVF, one can compute separate averaged EOG waveforms for the LVF and RVF targets and compare these waveforms. Any consistent differential eye movements will lead to differences in the averaged EOG waveforms, and even very small eye movements can be observed due to the improvement in S/N ratio produced by the averaging process.

This procedure will not allow individual trials to be rejected, nor will it be useful for detecting eye movements that are infrequent or in the same direction for both LVF and RVF targets. However, it can be useful when combined with the rejection of individual trials with large eye movements in a two-tiered procedure. The first tier consists of the rejection of individual trials with large saccades (> 1 degree) by means of the step function. You can then examine residual EOG activity in the averaged EOG waveforms, and exclude any subjects with differential EOG activity exceeding some criterion (e.g., 1.6 μV, corresponding to 0.1 degree) from the final data set.

Note that the techniques described above are useful for detecting saccades, but are not usually appropriate for detecting slow shifts in eye position or for assessing absolute eye position. To assess these, it is usually necessary to record the EOG using a DC amplifier, although you can use high-pass filtered EOG recordings for this purpose under some circumstances (Joyce et al., 2002).

Slow Voltage Shifts Slow voltage shifts are usually caused by a change in the impedance of the skin or the impedance of the electrodes (see *skin potentials* in figure 4.9). There is a small voltage between the superficial and deep layers of skin, and this voltage changes as the impedance changes in accordance with Ohm's Law, which states that voltage is proportional to the product of current and resistance (E = IR; see the appendix). If you increase the resistance of an electrical current without changing the current flow, the voltage must necessarily increase; thus, increasing the resistance actually increases the voltage. Impedance is simply the AC analog of resistance, so increases in impedance also lead to increases in resistance. When subjects sweat (even slightly), this causes a decrease in impedance, and the resulting slow voltage shifts are called skin potentials. The best way to reduce skin potentials is to reduce the impedance of the skin before applying the electrodes. Because electricity preferentially follows the path of least resistance, a change in impedance at one spot on the skin won't influence the overall impedance much if there is also a nearby spot with very low impedance. In general, the greater the initial impedance, the greater will be the changes in impedance due to sweating (see Picton & Hillyard, 1972). It is also helpful to maintain the recording chamber at a cool temperature and low humidity level. On hot and humid summer days, my lab runs a fan in the recording chamber between trial blocks.

Voltage shifts can also be caused by slight changes in electrode position, which are usually the result of movements by the subject. A change in electrode position will often lead to a change in impedance, thus causing a sustained shift in voltage. You can reduce this type of artifact by making sure that the subject is comfortable and does not move very much (a chin rest is helpful for this). If electrodes are placed at occipital sites, the subject should not place the back of his or her head against the back of the chair. You can also greatly reduce slow voltage shifts by using a high-pass filter during data acquisition, which will cause the voltage to return gradually toward 0 μV whenever a shift in voltage occurs (see chapter 5 for details).

It is not usually necessary to reject trials with slow voltage shifts, as long as they are rare. If the voltage shifts are slow and random, they shouldn't distort the averaged ERPs very much. However, a movement in the electrodes will sometimes cause the voltage to change suddenly to a new level, and you can detect this by means of a peak-to-peak amplitude function or a step function applied to each of the EEG channels (you'll want to keep the threshold fairly high to avoid rejecting trials with large ERP deflections, however). In some cases (e.g., when using very long epochs), you may wish to reject trials with gradual shifts in voltage, and you can accomplish this by computing the slope of the EEG across the trial and rejecting trials with slopes above some threshold.

Amplifier Saturation Slow voltage shifts may sometimes cause the amplifier or ADC to saturate, which causes the EEG to be flat for some period of time (this is also called *blocking*). If this happens

frequently, you should simply use a lower gain on the amplifier; if it never happens, you may wish to use a higher gain. As figure 4.9 illustrates, amplifier blocking is relatively easy to spot visually, because the EEG literally becomes a flat line. You could reject trials with amplifier saturation by finding trials in which the voltage exceeds some value that is just below the amplifier's saturation point, but in practice this would be difficult, because the saturation point may vary from channel to channel and may even vary over time. Another possibility would be to determine if there are a large number of points with identical voltages within each trial, but this isn't quite optimal because the voltages might not be exactly the same from moment to moment. A better procedure is to use a function that I call the *X-within-Y-of-peak* function, which Jon Hansen developed at UCSD. This function first finds the maximum EEG value within a trial (the peak), and then counts the number of points that are at or near that maximum. X is the number of points, and Y defines how close a value must be to the peak to be counted. For example, you might want to reject any trial in which thirty or more points are within 0.1 μV of the peak (i.e., $X = 30$ and $Y = 0.1$ μV). Of course, you must apply the same function to both the positive peak voltage and the negative peak voltage, and you should apply it to every channel.

Alpha Waves Alpha waves are oscillatory EEG deflections around 10 Hz that are largest at posterior electrode sites and occur most frequently when subjects are tired (see figure 4.9). The best way to reduce alpha waves is to use well-rested subjects, but some individuals have substantial alpha waves even when they are fully alert. Alpha waves can be particularly problematic when using a constant stimulus rate, because the alpha rhythm can become entrained to the stimulation rate such that the alpha waves are not reduced by the averaging process. Thus, it is useful to include a jitter of at least ±50 ms in the intertrial interval.

It is not usually worthwhile to reject trials with alpha waves: because ERPs can contain voltage deflections in the 10-Hz range, it is possible that trials with large ERPs will be rejected along with trials containing alpha artifacts. If it is necessary to reject trials with alpha, the best procedure is to compute the amplitude at 10 Hz on each trial and reject trials on which the amplitude exceeds some threshold value.

Muscle and Heart Activity The voltages created during the contraction of a muscle are called the electromyogram or EMG (see figure 4.9). These voltages are very high in frequency, and much of the EMG is usually eliminated by the amplifier's low-pass filter. You can also minimize the EMG by asking the subject to relax the muscles of the neck, jaw, and forehead and by providing a chinrest or some other device that reduces the load on the neck muscles.[4] As discussed above, it is also possible for the subject to recline in a comfortable chair, but this can cause movements of the posterior electrodes, resulting in large artifactual voltage shifts. Relaxation of the muscles below the neck is usually not important, because the EMG tends not to propagate very far.

It is not usually necessary to reject trials with EMG, assuming that you have taken appropriate precautions to minimize the EMG. However, if it is necessary to reject trials with EMG activity, you can detect EMG in several ways. The best method is to perform a Fourier transform on each trial and calculate the amount of high-frequency power (e.g., power above 100 Hz). A simpler method is to calculate the difference in voltage between every consecutive pair of points in a given trial and reject the trial if the largest of these differences exceeds a particular value.

Note that some stimuli will elicit reflexive muscle twitches. These are particularly problematic because they are time-locked to the stimulus and are therefore not attenuated by the averaging process. These also tend to be sudden, high-frequency voltage changes, but they are usually limited to a very short time period and are therefore difficult to detect by examining the high-frequency power across the entire trial. To reject these artifacts, it is best to look for sudden shifts in voltage during the brief time period during which they are likely to occur (usually within 100 ms of stimulus onset).

The beating of the heart (the EKG) can also be observed in EEG recordings in some subjects; figure 4.9 shows its distinctive shape. The EKG is usually picked up by mastoid electrodes, and if a mastoid is used as a reference, the EKG is seen in inverted form in all of the electrode sites. The EKG can sometimes be reduced by slightly shifting the position of the mastoid electrode, but usually there is nothing that can be done about it. In addition, this artifact usually occurs approximately once per second during the entire recording session, so rejecting trials with EKG deflections will usually lead to the rejection of an unacceptably large proportion of trials. Fortunately, this artifact is almost never systematic, and it will simply decrease the overall S/N ratio. In other words, there isn't much you can do, and it's not usually a significant problem, so don't worry about it.

The previous paragraph raises an important point. If you see an artifact or some type of noise equally in all of your EEG channels, it is probably being picked up by the reference electrode. Most artifacts and noise sources will be more prominent at some electrodes than at others, but any signals picked up by the reference electrode will appear in inverted form in all electrodes that use that reference. However, if you are using bipolar recordings for some of your channels (e.g., for EOG recordings), these recordings will not have artifacts or noise arising from the main reference electrode. This can help you identify and eliminate the sources of noise and artifacts.

Artifact Correction

Artifact rejection is a relatively crude process, because it completely eliminates a subset of trials from the ERP averages. As Gratton, Coles, and Donchin (1983) discussed, there are three potential problems associated with rejecting trials with ocular artifacts. First, in some cases, discarding trials with eye blinks and eye movements might lead to an unrepresentative sample of trials. Second, there are some groups of subjects (e.g., children and psychiatric patients) who cannot easily control their blinking and eye movements, making it difficult to obtain a sufficient number of artifact-free trials. Third, there are some experimental paradigms in which blinks and eye movements are integral to the tasks, and rejecting trials with these artifacts would be counterproductive. Under these conditions, it would be useful to be able to subtract away the voltages due to eye blinks and eye movements rather than rejecting trials with these artifacts.

Researchers have developed several artifact correction procedures for this purpose (e.g., Berg & Scherg, 1991a, 1994; Gratton, Coles, & Donchin, 1983; Lins et al., 1993b; Verleger, Gasser, & Moecks, 1982). When the eyes blink or move, voltages are created around the eyes that propagate to the scalp electrodes, and the voltage recorded at a given site will be equal to the value at the eyes multiplied by a propagation factor, plus any EEG activity present at that site. The simplest way to correct for eye artifacts is to calculate the propagation factor between the eyes and each of the scalp electrodes and subtract a corresponding proportion of the recorded EOG activity from the ERP waveform at each scalp site. For example, Lins and colleagues (1993a) found that 47 percent of the voltage present in an EOG recording propagated to the Fpz electrode, 18 percent to the Fz electrode, and 8 percent to the Cz electrode. To subtract away the EOG contribution to the averaged ERP waveforms at these electrode sites, it would be possible to subtract 47 percent of the EOG waveform from the Fpz electrode, 18 percent from the Fz electrode, and 8 percent from the Cz electrode.

There is a very significant problem with this approach, however. Specifically, the EOG recording contains brain activity in addition to true ocular activity and, as a result, the subtraction procedure ends up subtracting away part of the brain's response as well as the ocular artifacts. There are additional problems with this simple-minded subtraction, such as the assumption that the propagation factors will be the same for eye blinks and eye movements (a problem first addressed by Gratton, Coles, & Donchin, 1983). More sophisticated versions of this approach address these additional problems, and can work fairly effectively. For example, one can

use dipole modeling procedures to isolate the ocular activity (Berg & Scherg, 1991b), which works fairly well because the approximate locations of the dipoles are known in advance.

Although these artifact correction techniques can be useful or even indispensable for certain tasks and certain types of subjects, they have some significant drawbacks. First, some of these techniques can significantly distort the ERP waveforms and scalp distributions, making the data difficult to interpret. On the basis of a detailed comparison of several techniques, Lins et al. (1993b) concluded that source analysis procedures provided the least distortion, and other techniques (such as those of Gratton, Coles, & Donchin, 1983; Verleger, Gasser, & Moecks, 1982) can yield significant distortion. However, even the source analysis procedures may yield some distortion, especially when non-optimal parameters are used.

A newer and promising approach is to use independent components analysis (ICA). This approach is well justified mathematically, and recent studies demonstrated that this technique works very well at removing blinks, eye movements, and even electrical noise (Jung, Makeig, Humphries et al., 2000; Jung, Makeig, Westerfield et al., 2000) (see also the similar technique developed by Joyce, Gorodnitsky, & Kutas, 2004). However, these studies were conducted by the group who originally developed ICA, so they may not have been motivated to find conditions under which ICA performs poorly. In particular, this approach assumes that the time course of the artifacts is independent of the time course of the ERP activity, which may not always (or even usually) be a correct assumption. For example, if detecting a target leads to both a P3 wave and a blink, the blink and the P3 wave will have correlated time courses, and this could lead to inaccurate artifact correction. Until an independent laboratory rigorously tests this technique, it will be difficult to know whether this sort of situation leads to significant distortions.

A second problem with artifact correct techniques is that these techniques may require significant additional effort. For example,

Lins et al. (1993b) recommended that recordings should be obtained from at least seven electrodes near the eyes. In addition, one must conduct a set of calibration runs for each subject and carry out extensive signal processing on the data. Thus, it is important to weigh the time saved by using artifact correction procedures against the time required to satisfactorily implement these procedures.

A third problem with these techniques is that they cannot account for the changes in sensory input caused by blinks and eye movements. For example, if a subject blinks at the time of a visual stimulus, then this stimulus may not be seen properly, and this obviously cannot be accounted for by artifact correction techniques. In addition, as the eyes move, the visual world slides across the retina, generating a sensory ERP response. Similarly, eye blinks and eye movements are accompanied by motor ERPs. Artifact correction procedures do not typically address these factors, which are especially problematic when task-relevant stimuli trigger the blinks or eye movements.

Because of these limitations, I would recommend against using artifact correction procedures unless the nature of the experiment or subjects makes artifact rejection impossible. When artifact correction is necessary, I would recommend using one of the newer and less error-prone techniques, such as ICA or the source localization techniques Lins et al. (1993b) discuss. Moreover, I would strongly recommend against using the simpler techniques that are often available in commercial ERP analysis packages (such as the procedure of Gratton et al., 1983). When you use these techniques, it is difficult to know the extent to which the artifact correction procedures distort the results.

Suggestions for Further Reading

The following is a list of journal articles and book chapters that provide useful information about averaging, artifact rejection, and artifact correction.

Berg, P., & Scherg, M. (1994). A multiple source approach to the correction of eye artifacts. *Electroencephalography & Clinical Neurophysiology*, *90*(3), 229–241.

Gratton, G., Coles, M. G. H., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, *55*, 468–484.

Hillyard, S. A., & Galambos, R. (1970). Eye movement artifact in the CNV. *Electroencephalography and Clinical Neurophysiology*, *28*, 173–182.

Jung, T. P., Makeig, S., Humphries, C., Lee, T. W., McKeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, *37*, 163–178.

Jung, T. P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2000). Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects. *Clinical Neurophysiology*, *111*, 1745–1758.

Lins, O. G., Picton, T. W., Berg, P., & Scherg, M. (1993). Ocular artifacts in EEG and event-related potentials I: Scalp topography. *Brain Topography*, *6*, 51–63.

Lins, O. G., Picton, T. W., Berg, P., & Scherg, M. (1993). Ocular artifacts in recording EEGs and event-related potentials. II: Source dipoles and source components. *Brain Topography*, *6*, 65–78.

Picton, T. W., Linden, R. D., Hamel, G., & Maru, J. T. (1983). Aspects of averaging. *Seminars in Hearing*, *4*, 327–341.

Verleger, R., Gasser, T., & Moecks, J. (1982). Correction of EOG artifacts in event-related potentials of the EEG: Aspects of reliability and validity. *Psychophysiology*, *19*(4), 472–480.

Woldorff, M. (1988). Adjacent response overlap during the ERP averaging process and a technique (Adjar) for its estimation and removal. *Psychophysiology*, *25*, 490.

5 Filtering

This chapter discusses the application of filters to the EEG during data acquisition and to ERP waveforms before and after the averaging process. It is absolutely necessary to use filters during data acquisition, and it is very useful to apply filters offline as well, but filtering can severely distort ERPs in ways that ERP researchers frequently do not appreciate. For example, filters may change the onset and duration of an ERP component, may make monophasic waveforms appear multiphasic, may induce artificial oscillations, and may interfere with the localization of generator sources. This chapter will explain how these distortions arise and how they can be prevented. To avoid complex mathematics, I will simplify the treatment of filtering somewhat in this chapter, but there are several books on filtering that the mathematically inclined reader may wish to read (e.g., Glaser & Ruchkin, 1976). Note also that the term *filter* can refer to any of a large number of data manipulations, but this chapter will be limited to discussing the class of filters that ERP researchers typically use to attenuate specific ranges of frequencies, which are known as *finite impulse response filters*.

ERP waveforms are generally conceptualized and plotted as *time-domain* waveforms, with time on the X axis and amplitude on the Y axis. In contrast, filters are typically described in the *frequency-domain*, with frequency on the X axis and amplitude or power on the Y axis.[1] Because ERP researchers are typically more interested in temporal information rather than frequency information and because temporal information may be seriously distorted by filtering, it is important to understand filtering as a time-domain operation as well as a frequency-domain operation.[2] This chapter therefore describes how filters operate in both the time and